# DRAFT V3: Burden of diarrheal disease from contaminated water: analysis proposal using Multiple Indicator Cluster Surveys’ water quality data to estimate population attributable risks of poor water quality.

Andrew Mertens1, Benjamin F. Arnold2, and John M. Colford Jr1

1 Division of Epidemiology & Biostatistics, University of California, Berkeley, 2121 Berkeley Way Rm 5302 Berkeley, CA 94720-7360

2 Francis I. Proctor Foundation, University of California, San Francisco, 95 Kirkham St, San Francisco, CA 94122

## Background

Diarrheal disease causes over a million deaths each year, and is a leading cause of child mortality. Inadequate drinking water, sanitation and hygiene (WASH) are important risk factors for diarrheal disease, particularly in low- or middle- income countries (LMIC).3 Diarrheal disease from inadequate WASH conditions was estimated to cause 829,000 deaths in 2016, including 5.3% of deaths (297,000) in children under 5 years old.4 In particular, inadequate water was estimated to cause 484,741 deaths from diarrheal disease.4 In addition to increasing the risk of diarrheal disease, poor WASH conditions are associated with acute respiratory infections, poor child growth, and increased mortality.4

The WASH-attributable burden of disease estimates of the WHO for the year 2016 used access to point-of-use water treatment and safe water storage as the low-risk drinking-water reference level, rather than using access to continuous, high quality piped water. 4 This choice was due to a lack of high-quality studies estimating the decrease in risk of improving water quality through better operation of piped water systems in LMICs.5 However, studies from Puerto Rico and Egypt that compare uncontaminated or continuous piped water with contaminated or intermittent piped water supply find that having a high quality piped water supply was associated with a significantly decreased risk of child diarrhea,6,7 though a study from India only found an effect of continuous piped water on child diarrhea in low-income households.8

Where possible, risk factor-attributable burden of disease estimates are based on randomized or matched control trials or prospective cohorts,3 which are currently not available to estimate the effect of continuous, high quality piped water across a range of contexts, and are not necessarily representative samples. Historical experience in the WASH field,9 combined with null results of recent trials intervening to provide more basic levels of services, indicates that providing higher service levels is a public health policy imperative, and more robust evidence would help support the case for investment. Due to these limitations in the WASH epidemiologic literature regarding transitions to higher levels of service we propose alternative estimation methods that use new observational, nationally-representative cross-sectional surveys that include water quality data. The WHO/UNICEF Joint Monitoring Programme for Water Supply, Sanitation and Hygiene (JMP) draws upon household surveys, censuses and administrative data on WASH to produce SDG reports. Since 2012, the JMP team has worked with the UNICEF team that manages the Multiple Indicator Cluster Surveys (MICS) programme to develop a module for testing water quality in household surveys, and produce data on *E. coli* concentrations in drinking water. The module has now been implemented in over 30 surveys from countries with a range of WASH contexts, and these survey data can be used to identify both what types of water supplies are used (e.g. piped water, other improved sources such as protected groundwater, unimproved sources, and surface water) and the degree to which these water supplies are contaminated with faecal indicator bacteria.10 We propose to use these MICS datasets to estimate the change in risk of diarrheal disease and other child health outcomes as populations move from lower levels of drinking water to safely managed water, as well as estimate the change in risk in moving to improved sanitation, improved hygiene, and the combination of water, sanitation, and hygiene. Following recommendations from prior WASH burden of disease research, we will estimate the population attributable risks and fractions of poor WASH conditions.11

## Objectives

1. Estimate the relative risk of diarrheal disease within populations of MICS survey countries between levels of water, sanitation, and hygiene service ladders, including between contaminated and uncontaminated water.12
2. Estimate the population risks and fractions of diarrheal disease attributable to different tiers of the WASH service ladders, the number and proportion of cases of diarrhea that could have been prevented if all households in a population had improved to higher levels of WASH conditions.
3. Estimate the combined effects of improving water, sanitation, and hygiene conditions on reducing diarrheal disease.
4. Examine unmeasured confounding in relative risk estimates from MICS data by comparing them to the relative risks from existing meta-analyses of randomized trials.
5. Repeat the above objectives for secondary outcomes of respiratory infections, and child anthropometry measurements.

## Data

We will analyze data from household surveys with integrated water quality testing conducted as part of the UNICEF Multiple Indicator Cluster Surveys (MICS) program. We will use both nationally representative MICS6 surveys as well as sub-national surveys. The household water quality data includes *E. coli* concentrations, and the household surveys include questionnaire responses used to categorize the level of WASH service based on the JMP WASH ladder definitions. In addition, household surveys contain information on number of children, parental education, religion/ethnic group, household size and construction quality, and asset and animal ownership. Separate questionnaires in the surveys collect health Information for children under 5 years old including measures of diarrhea, respiratory infections, fever, anthropometry, breastfeeding, child mortality, birthweight, early child development index, and nutritional status.

## Approach

**Outcomes:** The primary outcome of interest is 14-day caregiver recall of diarrhea disease in children under 5 years. Secondary outcomes of interest are symptoms of acute respiratory infections in children <5, child anthropometry status (length for age Z-scores and weight-for-length Z-scores, as well as indicators for child stunting and wasting), and under-5 child mortality.

**Exposures:** The primary exposure contrast of interest is between households with drinking water free of contamination at the point of use compared to households with contaminated water, as measured by *E. coli* concentration. We will also examine water contamination at the drinking water point of collection and point of use, and we will contrast all levels of the JMP WASH service ladder, using more improved categories as the reference:

Level of water quality (point of use):

Primary stratification:

1. Contaminated household drinking water supply (*E. coli* concentrations ≥1 per 100ml)
2. Uncontaminated household drinking water supply (*E. coli* concentrations <1 per 100ml)

Secondary stratification:

1. Very high risk (>100 E coli/100ml)
2. High risk (11-100 E coli/100ml)
3. Moderate risk (1-10 E coli/100ml)
4. Low risk (<1 E coli/100ml)

Level of water quality (point of collection):

Primary stratification:

1. Contaminated source drinking water supply (*E. coli* concentrations ≥1 per 100ml)
2. Uncontaminated source drinking water supply (*E. coli* concentrations <1 per 100ml)

Secondary stratification:

1. Very high risk (>100 E coli/100ml)
2. High risk (11-100 E coli/100ml)
3. Moderate risk (1-10 E coli/100ml)
4. Low risk (<1 E coli/100ml)

Level of water supply:

1. Surface water (no service)
2. Unimproved - drinking water from an unprotected dug well or unprotected spring
3. Limited - drinking water from an improved source for which collection time exceeds 30 minutes for a roundtrip including queuing.
4. Basic - drinking water from an improved source (piped water, boreholes or tubewells, protected dug wells, protected springs, rainwater, and packaged or delivered water), provided collection time is not more than 30 minutes for a roundtrip including queuing,
5. Continuous drinking water on premises from an improved water source (highest measured level of service).

Level of sanitation service

1. Open defecation (no service)
2. Unimproved - use of pit latrines without a slab or platform, hanging latrines or bucket latrines
3. Limited - use of improved facilities (flush/pour flush to piped sewer system, septic tanks or pit latrines; ventilated improved pit latrines, composting toilets or pit latrines with slabs) shared between two or more households
4. Basic - use of improved facilities which are not shared with other households
5. Basic with high community coverage (>75% of the population in the community use basic sanitation services, highest measured level of service)

Level of hygiene service

1. None (no facility)
2. Limited - availability of a handwashing facility on premises without soap or water
3. Basic - availability of a handwashing facility on premises with soap and water (highest measured level of service measured)

Combined levels of WASH service of interest

1. Safely managed water:
   1. Primary stratification: improved source located on the premises with water available and free of *E. coli* contamination at the point of use compared to households without safely managed water.
   2. Secondary stratifications:
      1. Improved source not available when needed, with *E. coli* contamination.
      2. Improved source not available when needed, without *E. coli* contamination.
      3. Improved source located on the premises and available when needed, with *E. coli* contamination.
      4. Improved source located on the premises with water available when needed, without *E. coli* contamination (highest measured level of service).
2. Improved combined WASH:
   1. Primary stratification: Highest level of sanitation and hygiene and a safely managed water supply compared to households without any of the highest level of sanitation, hygiene, or water safely managed water.
   2. Secondary stratification: Highest level of sanitation, hygiene, and water supply (without including status of water contamination) compared to households without any of the highest level of sanitation, hygiene, or water supply.

**Statistical approach:**

The previous analysis of risk of piped water contamination from Egyptian Demographic and Health Survey data used propensity score matching (PSM) to control for confounding by wealth and location,6 but prior research on causal inference from cross-sectional data found no difference in model bias or efficiency between classical multivariate regression and PSM or more complicated techniques like targeted maximum likelihood estimation.13 Therefore, for analytic simplicity across 30 surveys, and to avoid failing to find adequate matching, we will use regression models, adjusting for covariates listed below. To estimate relative risks between levels of the JMP service ladders rather than odds ratios, we will use modified Poisson regressions for binary outcomes, and linear regressions continuous outcomes.14 We will include random effects for households nested within sampling clusters to account for the survey design.

We will estimate adjusted population attributable risks (PAR) using predictions from the modified Poisson regression models. PARs are estimated from the difference between the predicted outcomes under observed exposure and covariate distributions and outcomes predicted under a counterfactual scenario where the whole population had their exposure shifted to a lower risk level: PAR = Pr(disease) – Pr(disease | no exposure).16

We will also calculate population-attributable fractions (PAF):

Which are estimates of the percent of cases in the population that could have been prevented if the whole population had their exposure shifted to a lower risk level, by dividing the PAR by the baseline risk. Pc is the proportion of cases that are exposed. We will calculate percentile-based 95% confidence intervals (CIs) for PARs and PAFs using a nonparametric, clustered bootstrap (1000 iterations), with cluster resampling with replacement.17

We will assess potential bias in estimates of relative risk by contrasting them to comparable estimates of relative risk from existing meta-analyses of randomized controlled trials.

**Adjustment covariates:**

* 1. Asset-based wealth index (excluding WASH variables)
  2. Parental education
  3. Parental age
  4. Breastfeeding history
  5. Child age
  6. Child sex
  7. Birth order
  8. Urban/rural location
  9. Household type/construction
  10. Number of household residents
  11. Number of children under 5yrs in the household
  12. Type of flooring
  13. Presence of animals in the household
  14. Cooking stove type

We will pre-screen covariates to assess whether they are associated with each outcome prior to including them in the model. We will use a bivariate likelihood ratio test to assess the association between each outcome and each covariate and will include covariates with a p-value <0.2 in the analysis. We will also exclude or collapse covariates that have little variation in the study population (prevalence <5%) and exclude covariates with >50% missingness. For unordered categorical and indicator adjustment covariates, we will include missingness as its own category. For missingness in other adjustment covariates, we will use multiple imputation.18 For binary outcomes, due to the sparsity in some outcomes, we will only include one covariate per 10 cases of the outcome in the sparsest strata of the exposure, selecting the variables by rank ordering the likelihood ratio test.19

**Meta-analysis:**

We will estimate RRs, PARs and PAFs within each individual survey, and then pool estimates across surveys using random effects meta-analysis models.19 This modeling approach assumes that surveys are randomly selected from a hypothetical population of MICS surveys that could have been conducted in the past or future, and that this hypothetical population has a normally distributed spread of true relative risks. Random effects meta-analysis models incorporate between-survey variance in estimates and therefore have more conservative confidence intervals if there is heterogeneity across study estimates than inverse-variance weighted fixed-effects estimates, the common alternative.20 We will pool estimates both overall, by WHO region, and by UN Sustainable Development Goals region.

**Subgroup analysis:**

We will repeat estimates with households stratified into urban or rural locations to inform decisions about rural and urban investment in WASH improvements.

## Supplementary analyses

## As a sensitivity analysis, we will re-estimate relative risks using targeted maximum likelihood estimation, fit using flexible ensemble machine learning models, to confirm we are not missing any complex non-linear associations or interactions among the confounders.15 We will include simple means, generalized linear models, generalized additive models, penalized regressions, and gradient boosting machines in the machine-learning ensemble.

## We will also re-estimate pooled relative risks using a one-step hierarchical model, with children nested in households nested in clusters nested in countries. We will also re-estimate the pooled parameters using fixed effects (inverse-variance weighted) models, which assume that there is a single true parameter value.

We will repeat the analyses using a complete case analysis instead of multiple imputation on adjustment covariates, and we will repeat the analyses unadjusted for potential confounders.

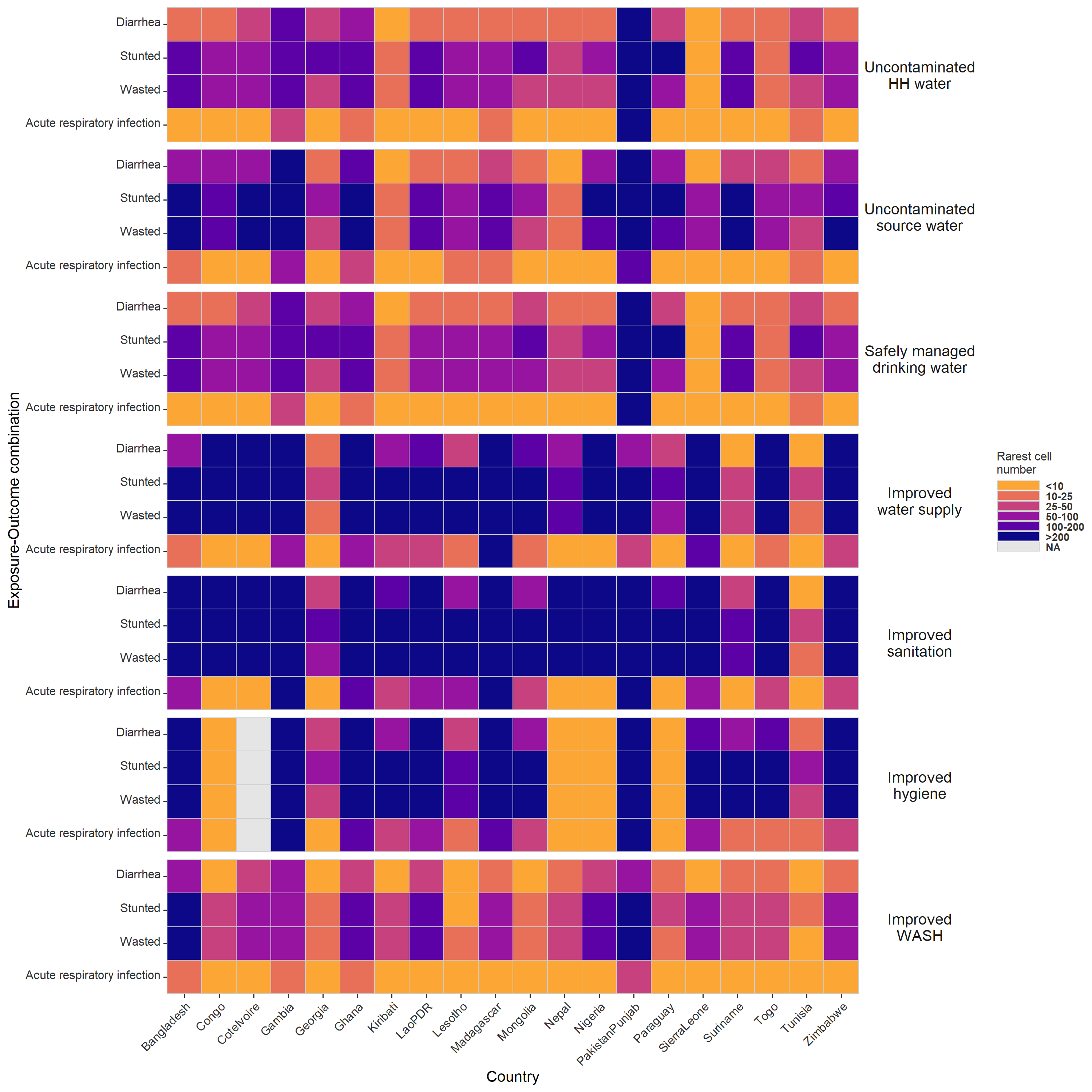
We will also estimate the effects of stochastic interventions to shift the distributions of *E. coli* concentration in drinking water as a supplementary analysis. Rather than estimating the effect of shifting the whole population to a static, low risk reference level, stochastic intervention analyses estimate the effect of a more realistic change: a downward shift in the continuous distribution of *E. coli* concentration, or a shift in the percentages of households at each category of the JMP WASH ladder.

Because *E. coli* concentrations are only measured within a subset of households, we will also impute the level of *E. coli* contamination using the pre-specified adjustment covariates and the other household WASH conditions. We will conduct multiple imputation using the *mice* package in R to impute the level of E. coli contamination, and we will compare imputation accuracy by comparing the imputed values to the observed level of contamination within the subsets of households measuring water contamination.18

## Proof of concept

We will conduct the analyses laid out above using MICS surveys from Bangladesh, Zimbabwe, and Pakistan, estimating relative risks between levels of the JMP WASH ladder present within each survey, and estimating PARs and PAFs using the highest rung of the JMP WASH ladder as the reference level. We will not estimate pooled parameters across these three studies. Surveys Bangladesh and Zimbabwe were chosen because high-quality randomized controlled trials on combined WASH Interventions (WASH Benefits and SHINE) were conducted in these countries. This will allow us to contrast estimates from MICS data to comparable estimates from randomized trials to roughly assess the amount of unmeasured confounding biasing survey-based estimates. The Pakistan survey will be analyzed because it has less data sparsity than other surveys (Figure 1). After the proof of concept analysis is finished, we will finalize the list of included MICS surveys and update the analysis plan if needed.

## Figures



**Figure 1. Data availability across exposure-outcome combinations by MICS country.**

Heatmap of the number of cases of the outcome in the rarest strata of the outcome-exposure combination. Each column is a MICS country with available water-testing data. Households with uncontaminated household and source water had water quality tests with zero detectable *E. coli*, and safely managed water is uncontaminated drinking water present in the household and available on demand. Improved sanitation and improved water were coded in the harmonized MICS datasets, household were considered to have improved hygiene if they had both water and soap available at the handwashing station, and households were considered to have improved WASH if the household had water free of *E. coli* contamination, improved sanitation, and improved hygiene.

## References

1. GBD 2015 Mortality and Causes of Death Collaborators. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet Lond. Engl.* **388**, 1459–1544 (2016).

2. WHO | Disease burden and mortality estimates. *WHO* http://www.who.int/healthinfo/global\_burden\_disease/estimates/en/.

3. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet Lond. Engl.* **388**, 1659–1724 (2016).

4. Prüss-Ustün, A. *et al.* Burden of disease from inadequate water, sanitation and hygiene for selected adverse health outcomes: An updated analysis with a focus on low- and middle-income countries. *Int. J. Hyg. Environ. Health* **222**, 765–777 (2019).

5. Prüss-Ustün, A. *et al.* Burden of disease from inadequate water, sanitation and hygiene in low- and middle-income settings: a retrospective analysis of data from 145 countries. *Trop. Med. Int. Health* **19**, 894–905 (2014).

6. Roushdy, R., Sieverding, M. & Radwan, H. The impact of water supply and sanitation on child health: Evidence from Egypt. *Poverty Gend. Youth* (2012) doi:doi.org/10.31899/pgy3.1016.

7. Hunter, P. R., Ramírez Toro, G. I. & Minnigh, H. A. Impact on diarrhoeal illness of a community educational intervention to improve drinking water quality in rural communities in Puerto Rico. *BMC Public Health* **10**, 219 (2010).

8. Ercumen, A. *et al.* Upgrading a Piped Water Supply from Intermittent to Continuous Delivery and Association with Waterborne Illness: A Matched Cohort Study in Urban India. *PLOS Med.* **12**, e1001892 (2015).

9. Pickering, A. J. et al. The WASH Benefits and SHINE trials: interpretation of WASH intervention effects on linear growth and diarrhoea. Lancet Glob. Health 7, e1139–e1146 (2019).

10. Water quality monitoring | JMP. https://washdata.org/monitoring/drinking-water/water-quality-monitoring.

11. Clasen, T. *et al.* Estimating the impact of unsafe water, sanitation and hygiene on the global burden of disease: evolving and alternative methods. *Trop. Med. Int. Health TM IH* **19**, 884–893 (2014).

12. Bain, R., Johnston, R., Mitis, F., Chatterley, C. & Slaymaker, T. Establishing Sustainable Development Goal Baselines for Household Drinking Water, Sanitation and Hygiene Services. *Water* **10**, 1711 (2018).

13. Spiegelman, D. & Zhou, X. Evaluating Public Health Interventions: 8. Causal Inference for Time-Invariant Interventions. *Am. J. Public Health* **108**, 1187–1190 (2018).

14. Zou, G. A modified poisson regression approach to prospective studies with binary data. *Am. J. Epidemiol.* **159**, 702–706 (2004).

15. Gruber, S. & Laan, M. van der. Targeted Maximum Likelihood Estimation: A Gentle Introduction. *UC Berkeley Div. Biostat. Work. Pap. Ser.* (2009).

16. Greenland, S. & Drescher, K. Maximum likelihood estimation of the attributable fraction from logistic models. *Biometrics* **49**, 865–872 (1993).

17. Greenland, S. Interval estimation by simulation as an alternative to and extension of confidence intervals. *Int. J. Epidemiol.* **33**, 1389–1397 (2004).

18. Buuren, S. van & Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. J. Stat. Softw. 45, 1–67 (2011).

19. Hedges, L. V. & Vevea, J. L. Fixed- and random-effects models in meta-analysis. *Psychol. Methods* **3**, 486–504 (1998).

20. Chen, Q. *et al.* Too Many Covariates and Too Few Cases? – A Comparative Study. *Stat. Med.* **35**, 4546–4558 (2016).