PLOS ONE

RESEARCH ARTICLE

Modeling household transmission dynamics: Application to waterborne diarrheal disease in Central Africa

Casper Woroszyło[1☯], Boseung Choi[2☯], Jessica Healy Profitós[3], Jiyoung Lee[3,4], Rebecca Garabed[5], Grzegorz A. Rempala[1,6*]

1 Mathematical Biosciences Institute, The Ohio State University, Columbus, 43210 Ohio, United States of America, **2** Department of National Statistics, Korea University, Sejong, 30019, Republic of Korea, **3** Division of Environmental Health Sciences, College of Public Health, The Ohio State University, Columbus, 43210 Ohio, United States of America, **4** Department of Food Science and Technology, The Ohio State University, Columbus, 43210 Ohio, United States of America, **5** Department of Veterinary Preventive Medicine, The Ohio State University, Columbus, 43210 Ohio, United States of America, **6** Division of Biostatistics, College of Public Health, The Ohio State University, Columbus, 43210 Ohio, United States of America

☯ These authors contributed equally to this work.
* rempala.3@osu.edu

Citation: Woroszyło C, Choi B, Healy Profitós J, Lee J, Garabed R, Rempala GA (2018) Modeling household transmission dynamics: Application to waterborne diarrheal disease in Central Africa. PLoS ONE 13(11): e0206418. https://doi.org/10.1371/journal.pone.0206418

Editor: Iratxe Puebla, Public Library of Science, UNITED KINGDOM

Received: June 17, 2017

Accepted: October 13, 2018

Published: November 7, 2018

Data Availability Statement: All relevant data are in the Supporting Information files.

Funding: The funders include the Mathematical Biosciences Institute at The Ohio State University through its National Science Foundation grant (NSF-DMS1440386 to GR) and the National Research Foundation of Korea grant (NRF-2017R1D1A3B03031008 to BC). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Introduction

We describe a method for analyzing the within-household network dynamics of a disease transmission. We apply it to analyze the occurrences of endemic diarrheal disease in Cameroon, Central Africa based on observational, cross-sectional data available from household health surveys.

Methods

To analyze the data, we apply formalism of the dynamic SID (susceptible-infected-diseased) process that describes the disease steady-state while adjusting for the household age-structure and environment contamination, such as water contamination. The SID transmission rates are estimated via MCMC method with the help of the so-called synthetic likelihood approach.

Results

The SID model is fitted to a dataset on diarrhea occurrence from 63 households in Cameroon. We show that the model allows for quantification of the effects of drinking water contamination on both transmission and recovery rates for household diarrheal disease occurrence as well as for estimation of the rate of silent (unobserved) infections.

Conclusions

The new estimation method appears capable of genuinely capturing the complex dynamics of disease transmission across various human, animal and environmental compartments at the household level. Our approach is quite general and can be used in other epidemiological settings where it is desirable to fit transmission rates using cross-sectional data.

Competing interests: The authors have declared that no competing interests exist.

Software sharing

The R-scripts for carrying out the computational analysis described in the paper are available at https://github.com/cbskust/SID.

Introduction

Diarrhea often occurs as a symptom of an infection in the intestinal tract caused by a bacterial, viral or parasitic organism. Such infections are typically spread through drinking water, contaminated food, or from animal-to-person and from person-to-person as a result of poor hygiene [1, 2]. Most people who die from diarrheal diseases actually die from severe dehydration and fluid loss. Children who are malnourished or have impaired immunity as well as people living with HIV are most at risk of life-threatening diarrhea. Indeed, diarrhea is one of the primary killers of the young children worldwide, with an estimated 1.7 billion annual cases of diarrhea among children under 5 resulting in over 500,000 deaths, the majority occurring in low and middle income countries [3].

Although diarrheal disease is common across all economic settings, it has the most potential to cause severe consequences when resources and medical care are limited or when co-morbidities are present. Acute episodes of disease more quickly lead to dangerous dehydration, while chronic gastrointestinal infection is now thought to be linked to environmental enteric disorder (EED), which results in a chronically damaged gut, reduced immunity, and stunted growth [4]. Loss of linear growth, particularly in a child's first years of life, can have long lasting impacts on cognitive and motor development [5]. However, consequences of disease aside, it remains unclear how differences in exposures and susceptibility play a role in the overall difference observed between children's and adults' diarrhea incidence.

In looking at household exposures to pathogens that may cause diarrhea, it appears that the interaction with animals (whether pets, livestock, or wildlife) plays an important role [3, 4, 6]. However, the potential of childhood animal exposure to modulate immunity and allergies [7–10] and of livestock ownership to improve nutrition and economic stability for families [4, 6] means that the specific role of household animals in transmission of diarrheal disease is complicated and needs to be clarified and better quantified with the help of a more mechanistic model.

The investigation of mechanisms behind household transmission of pathogens that cause diarrhea is not easy due to the complexity of the disease and its persistent endemic state in the global human population [3]. In general, the disease may be caused by both human-specific as well as zoonotic pathogens that have a variety of life cycles and the sheer number of potential culprits makes determining the specific cause of all observed cases on any sort of large scale practically impossible [3, 11]. In many developing countries (including most of Africa, see [12]) the problem is additionally compounded by the fact that most of the health surveillance programs operate with limited resources, and the data to assess transmission of diarrhea is generally limited to demographics, reported incidence of diarrhea, and possibly some outcome measures on households or individuals testing positive for a particular pathogen [11, 13]. Although such data may be used with the traditional mechanistic models to ascertain the role of different pathogens and transmission pathways on incidence of diarrhea [14, 15], the traditional models have difficulty adjusting for the presence of unrelated, endemic baseline of diarrhea occurrences [16, 17].

Table 1. Example of several data records from the dataset of $M = 63$ Maroua households. Full dataset provided in S1 Data.

Household Id	Water Contamination	Adults Sympt/Total	Juveniles Sympt/Total
1	No	0/2	0/0
4	Yes	0/6	0/4
24	Yes	1/4	0/2
51	No	2/2	0/0

https://doi.org/10.1371/journal.pone.0206418.t001

Our contribution

In this paper we propose a method of modeling transmission pathways of diarrhea using symptoms occurrence data from individual households consisting of family members with different susceptibility (for instance, children and adults) in the environment subject to waterborne pathogen contamination and possibly also other risks effecting baseline incidence rates. Our approach is quite general and allows to adjust not only for these different causes of diarrhea, even with data of poor resolution, but also for a variety of confounders typically encountered in similar observational studies. The particularly relevant confounders in our setting are the cases of non-symptomatic infectives and uninfected symptomatics. An example of a dataset of interest, as obtained from a cross-sectional study of Cameroon households, is presented in Table 1. The dataset is especially interesting as it matches the results of household drinking water testing for pathogens with detailed household health survey and demographic data. For the type of observational data in Table 1 we propose to first fit the household-level occurrence model and then to apply parametric resampling technique akin to *the synthetic likelihood* (see, e.g., [18, 19]) to obtain approximate distribution of the mean occurrence across households. Due to some general approximation results for a wide class of counting processes (see [20] chapter 11) we may assume here that the mean of the diarrhea occurrence is well-approximated by the stationary state of a certain system of ordinary differential equations (ODEs) with additive normal noise. This main idea behind our proposed approach is summarized in Fig 1.

Note that without the intermediate resampling step it is in general not possible to obtain the estimates of the transmission dynamics simply from cross-sectional occurrence. However, under the assumption of a constant risk (which is typically tacitly made in similar studies) we may consider the observed cases of diarrhea as a statistical sample from a stationary disease process. In this case, the ODEs parameters may be identified using the Bayesian inference techniques with the help of an MCMC algorithm (see, e.g., [21]). Using this approach we are able to obtain all relevant posterior estimates, including transmission rates and the expected count of latent infections (infection present but no diarrhea symptoms) as well as disease-unrelated occurrences (diarrhea symptoms present but no infection). Details of the analysis method are provided in the next section. To our knowledge, ours is the first application of the synthetic likelihood/resampling method to observational data on household diarrhea occurrence. We hope that similar approaches can be applied to larger datasets and consequently help improve current guidelines for treatment and intervention for diarrhea [2].

Materials and methods

Occurrence data and observed likelihood

To perform our analysis we use data from the observational study investigating the relationship between household drinking water quality and diarrhea occurrence in Maroua,

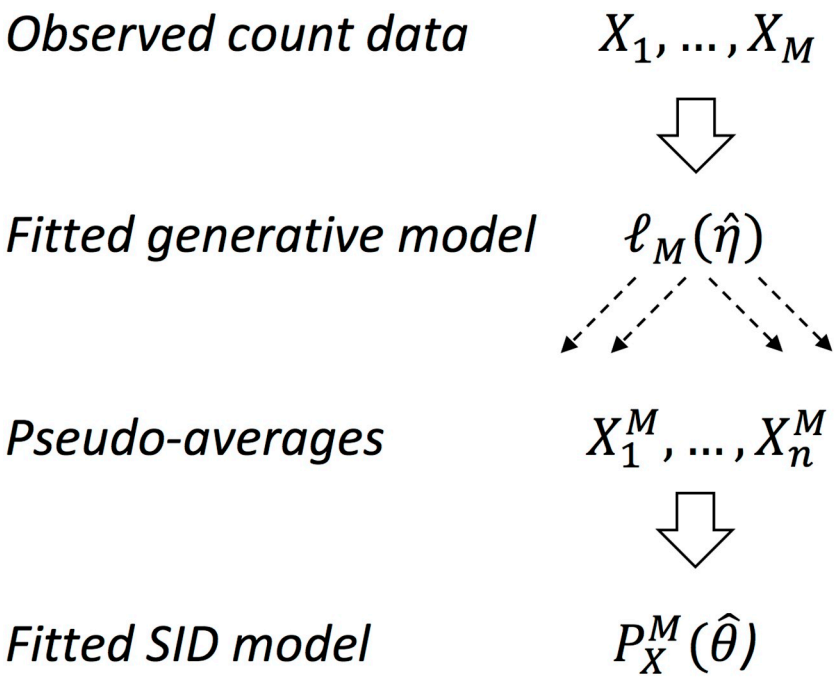

Fig 1. Synthetic inference based on some data X_{obs} and the SID likelihood $P_X^M(\hat{\theta})$. The count data $X_1, \ldots X_M$ represents household level diarrhea cases among adults and juveniles under and contaminated ($V = 1$) and clean ($V = 0$) environments and is used to fit the generative model (observed likelihood) $\ell_M(\hat{\eta})$ based on (1). The generated pseudo-data $X_1^M, \ldots X_n^M$ are then used to fit the SID model $P_X^M(\hat{\theta})$ based on (2) and (3).

https://doi.org/10.1371/journal.pone.0206418.g001

Cameroon. The data was described in [22] and more recently in [12]. Briefly, the study examined the relation between the occurrences of diarrhea and the presence of gastrointestinal pathogens within home drinking water sources in four urban neighborhoods in Maroua, the regional capital of northern Cameroon. For the purpose of the study diarrhea was defined as three or more loose bowel movements ("selles molles" in French) per day.

Heads of household assented to participation in the study with the use of a verbal consent script. In addition, other members of the household present for the survey assented to the survey and water sampling. Assent was recorded through use of a verbal consent script by the technicians collecting samples and administering the survey. The protocol was approved by the Ohio State University Institutional Review Board/ Human Research Protection Program (Federal-wide Assurance #00006378 from the Office for Human Research Protections in the Department of Health and Human Services: protocol 2010B0004). Within this ethical review for the survey the protocol was approved for a waiver of signed consent forms due to the low literacy of the population and cultural inappropriateness of obtaining signatures to record consent.

Diarrhea occurrence data and water samples from home water storage containers were collected from $M = 63$ households. Pathogen contamination was assessed using qPCR method, targeting several potential zoonotic pathogens including *Campylobacter* spp., Shiga toxin producing *Escherichia* coli (*stx*1 and *stx*2), and *Salmonella* spp. Microbial source tracking (MST) targeted three different host-specific markers: HF183 (human), Rum2Bac (ruminant) and GFD (poultry) to identify fecal contamination sources. For the purpose of our analysis below the pathogen/MST levels in each household were encoded as binary outcomes (water contamination present/absent) and combined with collected demographic information on the number

of household members, their age and the history of diarrhea symptoms within last 14 days. Two neighborhoods tested positive for most pathogens/MST while the others only tested positive for one or two. As *E.coli* was found in all water samples, it was excluded from our contamination criterion. Spatial variation of pathogens/MST existed between sources, storage containers, and neighborhoods but was not included in the set of covariates for current analysis due to small sample sizes of different spatial patterns. Differing population density and ethno-economic characteristics could potentially explain and correct for the variation but for the sake of simplicity we have not performed such analysis here. For illustration, several data points from the Cameroon dataset are listed in Table 1 where the diarrhea occurrences are recorded separately for adult and juvenile (under 15 years old) household members. The total number of adults and juveniles in the water contaminated (resp. uncontaminated) households was $N_J(1) = 103$ and $N_A(1) = 111$ (resp. $N_J(0) = 99$ and $N_A(0) = 155$).

Assuming that the data in Table 1 constitutes a sample from the cross-sections of a stationary distribution, each datapoint may be represented as a pair of occurrences of diarrhea (D_J, D_A) observed, respectively, in adult and juvenile compartments of random size (N_J, N_A). Because the mean and variance for the juvenile and adult compartments are approximately the same, the independent Poisson distributions are assumed for their respective sizes. Given the compartment sizes and the status of water contamination, the respective numbers of occurrences within compartments are assumed to follow binomial distributions with probabilities $p_J(V)$ and $p_A(V)$ where $V \in \{0, 1\}$ denotes the presence or absence of the water contamination. Although we do not model it explicitly due to small sample sizes, we tacitly assume the functional relationship between $p_J(V)$ and $p_A(V)$. In summary, for the compartments of sizes N_J, N_A, and the number of symptomatic (diseased) individuals denoted by D_J, D_A, and the household contamination status V, we assume the following *generative model* for the data

$$\begin{aligned} N_J &\sim \text{Poisson}(\lambda_J); D_J \sim \text{Binomial}(N_J, p_J(V)) \\ N_A &\sim \text{Poisson}(\lambda_A); D_A \sim \text{Binomial}(N_A, p_A(V)). \end{aligned} \tag{1}$$

Under the above model the likelihood-based inference may be now performed to estimate the compartment- and contamination-specific vector of parameters $\eta_V = (p_J(V), p_A(V), \lambda_J, \lambda_A)$ for $V \in \{0, 1\}$. For ease of notation, in what follows we suppress the subscript V when describing the parameters. Further details are provided in S1 Appendix. The numerical values of the estimated parameters are given in the next section.

SID model and synthetic likelihood

The data in Table 1 is cross-sectional and cannot be immediately used to analyze the within-household transmission pattern. Nevertheless, the generative representation via (1) allows for valid statistical inference indirectly, using the idea of synthetic likelihood akin to that proposed in [18]. Note that if we consider the sample from (1) as a set of independent realizations of some stationary counting process, then, by a version of the central limit theorem, we could expect its mean to approximately follow the normal distribution centered at a stationary solution of a certain ODE system (see [23] chapter 5). For the particular problem in hand, it is natural to take the ODE system to be one describing a compartmental SID (susceptible-infected-diseased) model defined below. Accordingly, the fitted generative model (1) may be used to generate n independent batches of M pseudo-data (denoted X_{obs}, see below) with corresponding n averages (denoted X^M_{obs}, see below) following a normal distribution with mean determined by the stationary SID system of ODEs.

In order to describe the SID model and introduce the required notation, denote the household-observed number of non-symptomatic and non-infected, adults (resp. juveniles) by S_A

Table 2. The reaction network description of the SID model with two compartments ($i, j \in \{A, J\}$). The graphical representation of the network is provided in Fig 2 and the corresponding ODE system in (A.2) in S1 Appendix.

Rate Parameter	Transition	Rate Parameter	Transition
β_{ij}	$(S_i, I_j) \to (I_i, I_j)$	$V\phi_i$	$S_i \to I_i$
α_i	$S_i \to D_i$	δ_i	$D_i \to S_i$
ν_i	$I_i \to D_i$	$\gamma_i - \nu_i$	$I_i \to S_i$

https://doi.org/10.1371/journal.pone.0206418.t002

(resp. S_J), the non-symptomatic but already pathogen infected adults (resp. juveniles) by I_A (resp. I_J), and the symptomatic, or diseased, either infected or non-infected, adults (resp. juveniles) by D_A (resp. D_J). The complete data for a given household with environment $V \in \{0, 1\}$ comprises the vector $X = (S_J, I_J, D_J, S_A, I_A, D_A, V)$ although in practice (due to lack of symptoms among I's) only the vector X_{obs} with the aggregated counts of the non-symptomatic $\tilde{D}_A = S_A + I_A$ and $\tilde{D}_J = S_J + I_J$ as well as D_A, D_J and V is observable. Under these assumptions, the Maroua data (cf. Table 1) may be considered as the set of $M = 63$ independent observations of the random vector X_{obs}. We denote the empirical mean of X_{obs} based on M observations by X^M_{obs} and assume that it follows the normal distribution with mean given by the stationary compartmental SID model, as summarized in Table 2 and Fig 2. Since in the actual dataset only a single vector X^M_{obs} is available, we generate additional means vectors from the *pseudo-data* using (1) as described above. As seen in Table 2, depending on the status of contamination ($V \in \{0, 1\}$), our SID model is parametrized by the vector $\boldsymbol{\theta}_V$ of 12 ($V = 0$) or 14 ($V = 1$) parameters. As before, we suppress the subscript V in what follows and write

$$\boldsymbol{\theta} = (\beta_{JJ}, \beta_{JA}, V\phi_J, \gamma_J, \beta_{AA}, \beta_{AJ}, V\phi_A, \gamma_A, \alpha_J, \nu_J, \delta_J, \alpha_A, \nu_A, \delta_A)$$

to denote the appropriate rates of transmission and recovery/infection between different model compartments and types.

As summarized in Table 2, for $i, j \in \{A, J\}$, β_{ij} denotes the rate at which S^M_i, through interaction with I^M_j, converts into I^M_i; $V\phi_i$ denotes the rate at which infected environment ($V = 1$) converts S^M_i into I^M_i ($V\phi_i = 0$ for $V = 0$); α_i denotes the rate at which S^M_i converts into D^M_i and δ_i is the rate of the reverse conversion. Finally, ν_i denotes the rate at which I^M_i progresses to D^M_i and $\gamma_i - \nu_i$ denotes the rate at which I^M_i returns back to S^M_i. The graphical diagram of all the transitions and interactions in Table 2 is presented in Fig 2.

The corresponding ODE system describing the SID dynamics is presented in (A.2) in S1 Appendix. Based on that system we may relate the generated pseudo-data to model parameters as follows. Consider the average number of household asymptomatic individuals in adult and juvenile groups and denote

$$\tilde{D}^M_A := S^M_A + I^M_A \qquad \text{and} \qquad \tilde{D}^M_J := S^M_J + I^M_J.$$

Solving the SID model ODE for its steady state we obtain, on one hand,

$$\begin{aligned} \tilde{D}^M_J &= \left(\frac{\gamma_J}{\beta_{JJ}I_J + \beta_{JA}I_A + V\phi_J} + 1\right) I_J =: f_1^{\boldsymbol{\theta}_1}(I_J, I_A) \\ \tilde{D}^M_A &= \left(\frac{\gamma_A}{\beta_{AA}I_A + \beta_{AJ}I_J + V\phi_A} + 1\right) I_A =: f_2^{\boldsymbol{\theta}_2}(I_J, I_A) \end{aligned} \tag{2}$$

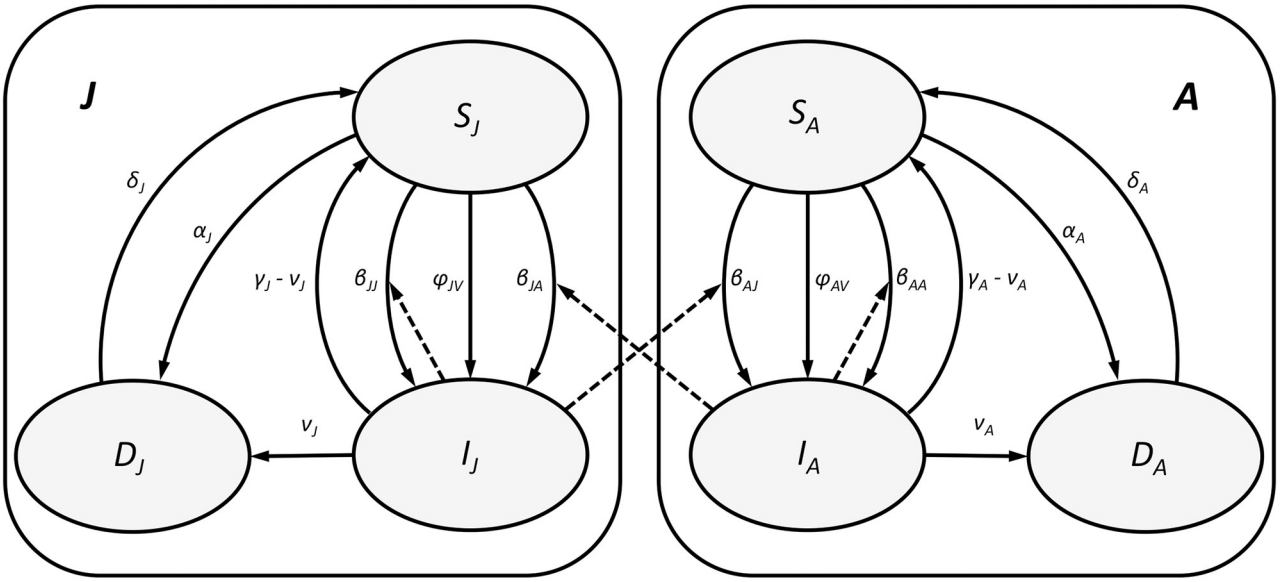

Fig 2. The graphical representation of the SID model from Table 2 with marked two compartments *J* (juveniles) and *A* (adults). Solid lines denote transitions within compartments. Dashed lines indicate transitions due to interactions (both within and across compartments) between susceptible (S) and infected (I) individuals.

https://doi.org/10.1371/journal.pone.0206418.g002

and, on the other,

$$\begin{aligned} \tilde{D}_J^M &= \frac{(\alpha_J - \nu_J)I_J + \delta_J D_J}{\alpha_J} =: f_3^{\boldsymbol{\theta}_3}(I_J) \\ \tilde{D}_A^M &= \frac{(\alpha_A - \nu_A)I_A + \delta_A D_A}{\alpha_A} =: f_4^{\boldsymbol{\theta}_4}(I_A). \end{aligned} \tag{3}$$

where the f's are defined by their left-hand sides and we denote $\boldsymbol{\theta}_1 = (\beta_{JJ}, \beta_{JA}, V\phi_J, \gamma_J)$, $\boldsymbol{\theta}_2 = (\beta_{AA}, \beta_{AJ}, V\phi_A, \gamma_A)$, $\boldsymbol{\theta}_3 = (\alpha_J, \nu_J, \delta_J)$, and $\boldsymbol{\theta}_4 = (\alpha_A, \nu_A, \delta_A)$, so that $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\theta}_4)$.

Note that the quantities $\tilde{D}_A^M$ and $\tilde{D}_J^M$ are derived from the pseudo-data X_{obs}^M obtained by sampling from the fitted model (1).

Parameter estimation

Due to a relatively small size M of the dataset, we do not attempt to evaluate the variable V dynamically but instead consider two separate SID models for contaminated and uncontaminated environments ($V = 1$ and $V = 0$). In each case, in order to estimate the vector of parameters $\boldsymbol{\theta}$ as well as two hidden states (I_A, I_J) based on the generated sample of n pseudo-averages X_{obs}^M, we employ an MCMC procedure. Its advantage is in being able to handle the latent (unobservable) variables and in providing a simple and intuitive way of validating the final model against observations in Table 1. The disadvantage is in a relatively high computational overhead due to a somewhat complicated Metropolis-within-Gibbs algorithm [24] described in Algorithm 1 below. Details on the forms of the conditional distributions are provided in S1 Appendix. To ease notation, let $\boldsymbol{\theta}_{-k}$ denote the vector $\boldsymbol{\theta}$ with its k-th component removed ($k = 1, \ldots, 4$). Recall that when $V = 0$ the ϕ parameter is 0 and hence is excluded from from $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. We estimate parameters $\boldsymbol{\theta}$ separately for $V = 0, 1$ via the following iterative procedure.

Table 3. Estimates in data generating model based on the observed likelihood (1). Estimates of λ are pooled across V values.

Water Contamination (V)	p_J	p_A	λ_J	λ_A
Yes ($V = 1$)	0.1262	0.1261	3.2063	4.2222
No ($V = 0$)	0.1414	0.0710	3.2063	4.2222

https://doi.org/10.1371/journal.pone.0206418.t003

MCMC algorithm for SID model fitting.

1. Given the state of environment $V \in \{0, 1\}$ generate a collection $\tilde{d}_V(n)$ of n pseudo-data points $(\tilde{d}_i^A, \tilde{d}_i^J)(V)$, each of them being an average of M independent draws of the pair $(\tilde{D}_A, \tilde{D}_J)(V)$ from (1) under fitted parameters $\hat{\eta}$.
2. Initiate values of the rate vector $\boldsymbol{\theta}$ as well as $I_A(V)$, and $I_J(V)$, according to their prior distributions (see S1 Appendix).
3. Using the Metropolis-Hastings (MH) step, conditionally on $(I_A, I_J)(V)$ and $\tilde{d}_V(n)$, draw sequentially samples from the conditional distributions of $\boldsymbol{\theta}_k|\boldsymbol{\theta}_{-k}$, $k = 1, \ldots, 4$. The form of the proposal in MH step as well as the forms of the conditionals are given in (A.4)–(A.7) in S1 Appendix
4. Using the MH step, conditionally on $\boldsymbol{\theta}$ and $\tilde{d}_V(n)$, draw independently from $I_A(V)$ and $I_J(V)$ using their conditionals as given in (A.8) and (A.9) in S1 Appendix.
5. Repeat step 3 and 4 until convergence.

In our analysis, we iterated the above MCMC procedure 40,000 times retaining every 10-th iteration for $V = 0$ and 20-th iteration for $V = 1$, in order to ensure good chain mixing. We also removed the first 20,000 iterations as a burn-in set and summarized the posterior statistics based on the remaining iterations. To check for the robustness of our analysis with respect to the amount n of the generated pseudo-data, we applied the MCMC algorithm above with $n = 50$ and $n = 100$, however, since the results were virtually identical, we only report below on the case $n = 100$. Although larger values of n could be also considered, this particular value seems to strike a good balance between required MCMC precision and computational overhead.

Model validation. The final step in our model estimation procedure was validation against the observed data. This was done by comparing the posterior distributions of the model generated data samples using estimated parameters with the actually observed values from X_{obs} and looking for large departures from the posterior mode.

Software. The R-scripts for carrying out tour computational analysis described above along with the Maroua dataset adapted from [22] are available at https://github.com/cbskust/SID.

Results

The initial set of fitted parameters obtained for the generative model (1) based on the $M = 63$ Maroua households dataset is provided in Table 3. As can be seen from the entries in the table, an interesting feature of this dataset appears to be that the probability of diarrhea in the juvenile compartment is *decreased* in the households with contaminated water environment ($V = 1$). There may be several reasons for this finding which appears inconsistent with other reported observational studies [25]. First, the survey data for juveniles may be less reliable than for adults, particularly in young children who under our definition are also a part of the juvenile compartment. Second, it is known [26] that a substantial number of juvenile diarrhea

cases is, in fact, unrelated to the waterborne infections and the collected data may be simply confounded with this unrelated process. Finally, it is also possible that the contaminated environment offers some measure of immunity from diarrhea, perhaps due to non-specific activation of the immune system [26].

The numerical results of the MCMC-based fitting of $\boldsymbol{\theta}$ for SID model under both $V = 0$ and $V = 1$ are summarized in Table 4 where we list the posterior means, posterior standard deviations, and 95% credible intervals (CIs) based on the generated $n = 100$ pseudo-data points and 2000 thinned posterior samples. Complementing the table entries, the full sets of marginal densities and trace plots for the posterior distributions are provided, in S1–S4 Figs of the Supporting Information.

Although we opted not to conduct the direct comparison of the parameter values in $\boldsymbol{\theta}$ between $V = 0$ and $V = 1$, one may somewhat informally perform such a comparison based on the CI entries in Table 4. In general, if for a particular parameter in $\boldsymbol{\theta}$ its CI bounds under $V = 0$ are contained within the CI bounds under $V = 1$, or vice-versa, one would consider the corresponding posterior distributions as statistically (i.e., for given data) equal. To facilitate such analysis in Table 4 the parameters with statistical distinct posterior distributions are entered in bold. From the entries in Table 4 it therefore follows that although the posterior distributions of the transmission rates β_{JJ} and β_{AJ} are statistically different between $V = 0$ and $V = 1$, it is not so for the remaining rates β_{AA} and β_{JA}. Similarly, we find that although the average number of silent infections among juveniles under $V = 0$ and $V = 1$ (mean $I_J^M(1) = 2.2634$ vs mean $I_J^M(0) = 1.6109$) is not statistically different, this is not the case for the average number of silent infections among adults, despite the smaller absolute difference of their means. (This particular finding appears to be due to the relatively large value of the posterior standard deviation of $I_J(1)$.) Similar comparisons may be also performed between the recovery rates. Indeed, we find that while the recovery rate in the adult compartment is significantly slowed down in the contaminated environment (mean $\delta_A(1) = 0.7880$ vs mean $\delta_A(0) = 0.6314$), the rate in the juvenile compartment is not significantly changed.

Table 4. Summary of MCMC results based on $n = 100$ pseudo-households.

	Water contaminated ($V = 1$)			Water clean ($V = 0$)		
	Mean	**Std Dev**	**95% CI**	**Mean**	**Std Dev**	**95% CI**
β_{JJ}	**0.4921**	0.4212	(0.0296, 1.5966)	**0.5275**	0.4293	(0.0392, 1.6873)
β_{JA}	0.4950	0.4224	(0.0348, 1.5245)	0.4677	0.3962	(0.0355, 1.5137)
ϕ_J	0.5159	0.4227	(0.0233, 1.6213)			
γ_J	0.7700	0.5357	(0.0761, 2.1018)	0.7104	0.4933	(0.0795, 2.0047)
β_{AA}	0.4829	0.3699	(0.0404, 1.3745)	0.4563	0.3699	(0.0317, 1.4145)
β_{AJ}	**0.5562**	0.4443	(0.0298, 1.6465)	**0.4748**	0.3771	(0.0269, 1.4428)
ϕ_A	0.5318	0.4228	(0.0561, 1.7337)			
γ_A	0.7847	0.5688	(0.092, 2.3176)	0.8219	0.5571	(0.0999, 2.1469)
α_J	**0.4892**	0.4824	(0.0651, 1.7999)	**0.7266**	0.4694	(0.1424, 1.8301)
ν_J	**0.1415**	0.1409	(0.0131, 0.5821)	**0.2318**	0.1980	(0.0157, 0.7836)
δ_J	0.9223	0.4685	(0.1324, 1.9421)	0.7711	0.5515	(0.0599, 2.2047)
α_A	**0.7269**	0.5331	(0.1042, 2.1272)	**0.8243**	0.5424	(0.1194, 2.2096)
ν_A	**0.1927**	0.1728	(0.0144, 0.6356)	**0.2192**	0.1952	(0.0168, 0.7247)
δ_A	**0.7880**	0.5270	(0.0972, 1.9188)	**0.6314**	0.4654	(0.0605, 1.8727)
I_J^M	2.2634	1.0453	(0.2444, 3.7829)	1.6109	0.4300	(0.6082, 2.2321)
I_A^M	**3.3334**	0.7008	(1.4657, 4.1644)	**3.0352**	0.6349	(1.4452, 3.7943)

https://doi.org/10.1371/journal.pone.0206418.t004

From the view point of waterborne disease, the most interesting are perhaps the estimates of the water contamination effects on the households diarrhea persistence in different compartments. In Table 4 our SID model quantifies the effect of water contamination ($V = 1$) in the households on average as $\phi_A = 0.5318$ and $\phi_J = 0.5159$. This indicates that despite the differences in the diarrhea prevalence patterns among juveniles and adults (see Table 3), the overall effect of waterborne pathogens is quantitatively similar. Note that the simple estimates in Table 3 which are based on the survey data (Table 1) and ignore the SID dynamics and asymptomatic infections suggest otherwise (cf. also [12]). Note also that according to the SID model the average number of infectious individuals (both pre-symptomatic and never-symptomatic) is larger in the contaminated environment, with the observed difference being significant in the adult compartment. Moreover, Eqs (2) and (3) for the specific values in Table 4 indicate that remediating contaminated water environment in the household (moving from $V = 1$ to $V = 0$) is likely to remove the symptomatic cases in the average adult compartment but not so in the juvenile one. Finally, let us also note that the observed higher prevalence of diarrhea among juveniles Table 3 in the clean environment may be explained on the basis of our SID model by the higher transmission in the juvenile compartment ($\beta_{JJ}(0)$ is significantly greater then $\beta_{JJ}(1)$) and an increase in non-pathogen/ non-household diarrhea (increased α_J).

The results of model validation are shown separately for the A and J compartments in Fig 3 where the numerical values of the means of X_{obs} (vertical lines) are plotted along with the corresponding histograms of their posterior distributions obtained from the model with estimated parameters. As seen from the plots, the observed values are within the reasonable range of the posterior mode and hence may be considered in agreement with the fitted model. This also implies that the CI bounds in Table 4 may be interpreted as plausible ranges of respective parameter values consistent with the observed data. These ranges are quite wide indicating somewhat large uncertainty, likely due to moderate sample size ($M = 63$).

Summary and discussion

In many observational disease studies we lack the ability to collect repeated measurements over time, either due to cost or practicality considerations. Consequently, disease transmission studies often have to rely on cross-sectional data containing latent variables and multiple confounders (e.g. latent infections or different disease susceptibility across population). For such data we have proposed here a statistical method for direct analysis of the transmission rates across different population compartments and different environmental risk factors. The ideas for statistical analysis came from the consideration of stationary SID model based on differential equations and synthetic likelihood with MCMC algorithm for estimating parameters. The proposed estimation method appears to be quite stable and capable of converging in a relatively large parameter space (in our example we had up to 16 parameters) even when supplied with only slightly informative prior distributions for moderate sample size.

Applying modern Bayesian approach to fit SID transmission model allowed us to better account for the uncertainty of various model components (i.e., bias or lack of accuracy) as well as the uncertainty of outcomes predictions (i.e., variance or lack of precision). It also allowed us to naturally incorporate any additional information about the model parameters. For instance, should some of the estimated compartmental diarrhea probabilities be fixed at specific values (say, based on prior studies) the fitting algorithm could easily incorporate this additional information. In such case one would expect to see both model's precision and accuracy to increase. We also note that in our example dataset the posterior marginal distributions of the parameters were all unimodal, indicating that the model parameters were identifiable, that is, their joint posterior distribution had a unique mode contained in the range of plausible

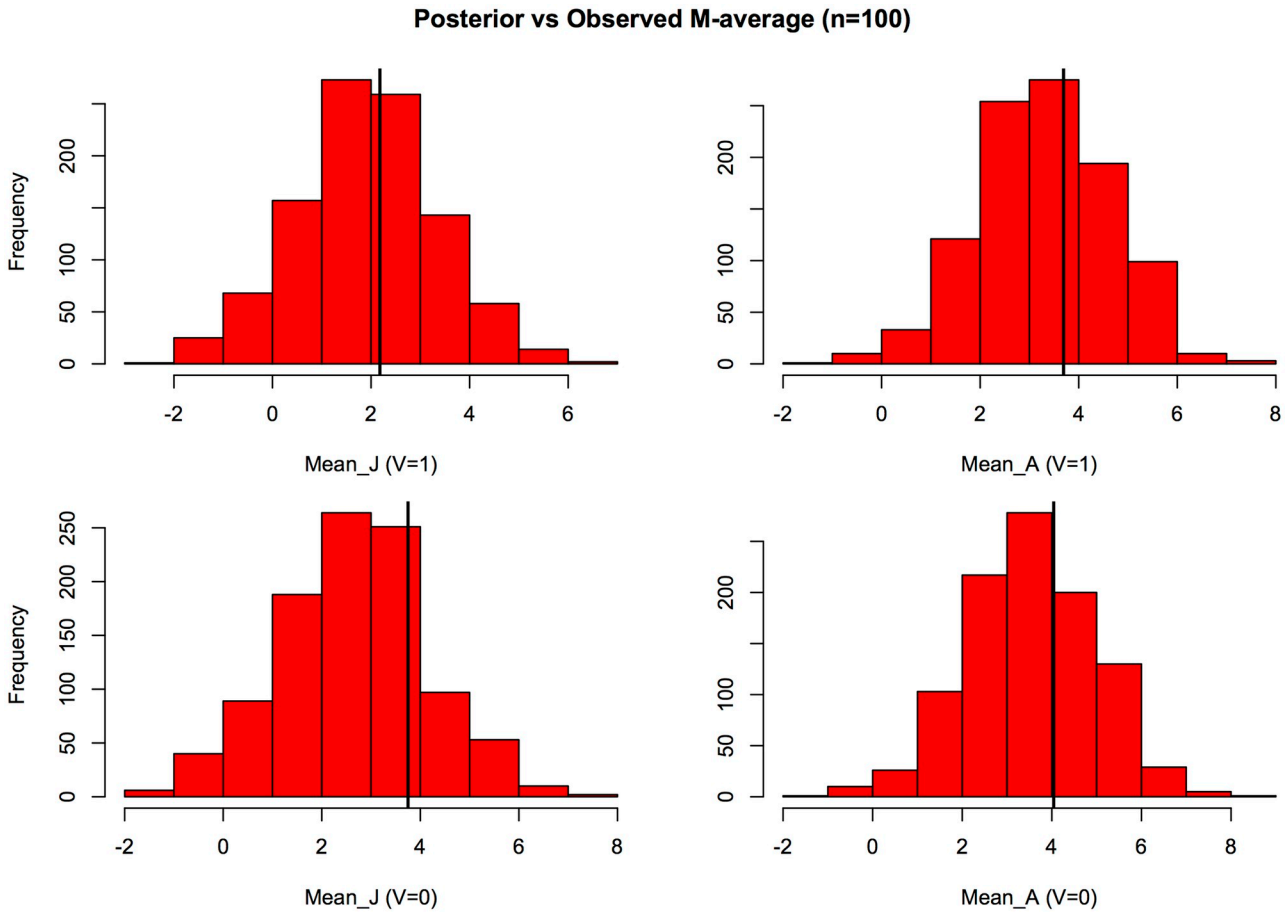

Fig 3. Model validation. The distributions of the posterior means of the counts of asymptomatic individuals in juvenile (J) and adult compartments based on the fitted SID model (2) and (3) vs the actual observed values from $M = 63$ Maroua households (cf. Table 1) marked by vertical lines.

https://doi.org/10.1371/journal.pone.0206418.g003

parameter values given the observed data. In general, our proposed statistical approach may be viewed as an alternative to a more traditional epidemiological disease risk analysis based on the odds ratios, where the Cochran-Mantel-Haenszel (CMH) stratification method is typically used to adjust for confounders.

The example dataset we have chosen was part or a larger study investigating possible links between drinking water contamination and diarrheal diseases in urban environment of Central/Sub-Saharan Africa [12, 22]. Although this particular study did not specifically examine other factors associated with gastrointestinal infections (socioeconomic status, overall sanitation, household education, storage, etc), they likely did contribute to the observed baseline (not water-related) occurrence. However, our statistical analysis indicated that in our dataset they constituted only a small minority of the observed symptomatic cases.

In order to better appreciate the possible implications of SID-type analysis for public health policies and interventions, it is helpful to compare its results (Table 4) with the results from initial, purely descriptive analysis of the Maroua dataset (Table 3) akin to that conducted previously in [22]. We note that since descriptive analysis in Table 3 is based on risks comparison (i.e., the binomial probabilities) it provides only an aggregated measure of the water contamination effect on the prevalence of diarrhea. It is not clear in particular what specific transmission pathways should be targeted for intervention in order to minimize the observed

occurrence (note that the juvenile risk is actually smaller in contaminated households). In contrast, the SID analysis in Table 4 provides (via Eqs (2) and (3)) an explicit numerical relations between transition rates and occurrence, and therefore a comprehensive picture of competing household transmission risks. Consequently, the SID analysis allows for a more detailed examination of how household occurrence risk is associated with the water environment and how it is transferred across age compartments. Such information appears essential for developing more targeted water intervention strategies beyond those that are currently recommended by WHO (see, [2] Section 11.3) for reducing diarrhea risk.

Supporting information

S1 Fig. Marginal plots for the posterior parameters of the SID model under $V = 1$ and $n = 100$ with 2,000 iterations.
(TIF)

S2 Fig. Marginal plots for the posterior parameters of the SID model under $V = 0$ and $n = 100$ with 2,000 iterations.
(TIF)

S3 Fig. Diagnostic trace plots for the posterior parameters of the SID model under $V = 1$ and $n = 100$ with 2,000 iterations.
(TIF)

S4 Fig. Diagnostic trace plots for the posterior parameters of the SID model under $V = 0$ and $n = 100$ with 2,000 iterations.
(TIF)

S1 Appendix. Appendix on statistical analysis. Contains additional formulas and derivations related to the statistical analysis.
(PDF)

S1 Data. Maroua household data. Diarrhea symptoms data from 63 households from Maroua, Cameroon. The household ID, number of juveniles (J) and adults (A) as well as the number of respective symptomatics (DJ and DA).
(CSV)

Acknowledgments

The authors thank Seungjun Lee for helping them organize and interpret the data and Will Gehring for helping with some manuscript figures. They are also indebted to the reviewers for their helpful comments on the earlier draft of the paper.

Author Contributions

Conceptualization: Casper Woroszyło, Jiyoung Lee, Rebecca Garabed, Grzegorz A. Rempala.

Data curation: Jessica Healy Profitós, Rebecca Garabed.

Formal analysis: Casper Woroszyło, Boseung Choi.

Funding acquisition: Grzegorz A. Rempala.

Investigation: Boseung Choi, Jessica Healy Profitós.

Methodology: Rebecca Garabed, Grzegorz A. Rempala.

Project administration: Grzegorz A. Rempala.

Software: Casper Woroszyło, Boseung Choi.

Supervision: Jiyoung Lee, Rebecca Garabed, Grzegorz A. Rempala.

Writing – original draft: Grzegorz A. Rempala.

Writing – review & editing: Jessica Healy Profitós, Jiyoung Lee, Rebecca Garabed, Grzegorz A. Rempala.

References

1. World Health Organization Diarrhoeal Disease Fact Sheet; May 2017 http://www.who.int/mediacentre/factsheets/fs330/en/
2. World Health Organization The Treatment of Diarrhoea. A manual for physicians and other senior health workers. 4th edition, 2005 http://www.who.int/mediacentre/factsheets/fs330/en/
3. Julian TR. Environmental transmission of diarrheal pathogens in low and middle income countries. Environ Sci Process Impacts. 2016 Aug 10; 18(8):944–55. https://doi.org/10.1039/c6em00222f PMID: 27384220
4. Headey D, Hirovnen K. Is exposure to poultry harmful to children nutrition? An observational analysis for rural Ethiopia. PLoS One. 2016 Aug 16; 11(8):e0160590. https://doi.org/10.1371/journal.pone.0160590 PMID: 27529178
5. Sudfeld CR, McCoy DC, Danaei G, Fink G, Ezzati M, Andrews KG, Fawzi WW. Linear growth and child development in low and middle-income countries: a meta-analysis. Pediatrics. 2015 May; 135(5): e1266–75. https://doi.org/10.1542/peds.2014-3111 PMID: 25847806
6. Zambrano LD, Levy K, Menezes NP, Freeman MC. Human diarrhea infections associated with domestic animal husbandry: a systematic review and meta-analysis. Trans R Soc Trop Med Hyg. 2014 Jun; 108(6):313–25. https://doi.org/10.1093/trstmh/tru056 PMID: 24812065
7. Ownby DR, Johnson CC, Peterson EL. Exposure to dogs and cats in the first year of life and risk of allergic sensitization at 6 to 7 years of age. JAMA. 2002; 288(8):963–972. https://doi.org/10.1001/jama.288.8.963 PMID: 12190366
8. Braun-Fahrländer C, Gassner M, Grize L, Neu U, Sennhauser FH, Varonier HS, Vuille JC, Wathrich B. Prevalence of hay fever and allergic sensitization in farmer's children and their peers living in the same rural community. Clin & Experimental Allergy. 1999; 29(1):28–34. https://doi.org/10.1046/j.1365-2222.1999.00479.x
9. Waser M, Von Mutius E, Riedler J, Nowak D, Maisch S, Carr D, Eder W, Tebow G, Schierl R, Schreuer M, Braun-Fahrländer C,. Exposure to pets, and the association with hay fever, asthma, and atopic sensitization in rural children. Allergy. 2005; 60(2):1398–9995. https://doi.org/10.1111/j.1398-9995.2004.00645.x
10. Simpson A, Custovic A. Pets and the development of allergic sensitization. Current Allergy & Asthma Reports. 2005; 5(3):212–220. https://doi.org/10.1007/s11882-005-0040-x
11. Schmidt WP, Arnold BF, Boisson S, Genser B, Luby SP, Barreto ML, Clasen T, Cairncross S. Epidemiological methods in diarrhoeal studies—an update. Int J Epidemiol. 2001 Dec; 40(6):1678–92. https://doi.org/10.1093/ije/dyr152
12. Healy-Profitos J, Lee S, Mouhaman A, Garabed R, Moritz M, Piperata B, Lee J Neighborhood diversity of potentially pathogenic bacteria in drinking water from the city of Maroua, Cameroon. J Water Health. 2016; 14(3):559–570. https://doi.org/10.2166/wh.2016.204 PMID: 27280618
13. Medina DC, Findley SE, Guindo B, Doumbia S. Forecasting non-stationary diarrhea, acute respiratory infection, and malaria time-series in Niono, Mali. PLoS One. 2007; 2(11):e1191. https://doi.org/10.1371/journal.pone.0001181
14. Watson CH, Edmunds WJ. A review of typhoid fever transmission dynamic models and economic evaluations of vaccination. Vaccine. 2015; 33:C42–C54. https://doi.org/10.1016/j.vaccine.2015.04.013 PMID: 25921288
15. Merler S, Ajelli M, Fumanelli L, Gomes MFC, Piontti AP, Rossi L, Chao DL, Longini IM, Halloran ME, Vespignani A. Spatiotemporal spread of the 2014 outbreak of Ebola virus disease in Liberia and the effectiveness of non-pharmaceutical interventions: a computational modelling analysis. The Lancet Infectious Disease. 2015; 15(2):204–211. https://doi.org/10.1016/S1473-3099(14)71074-6
16. Kotloff KL, Blackwelder WC, Nasrin D, Nataro JP, Farag TH, van Eijk A, Adegbola RA, Alonso PL, Breiman RF, Faruque ASG. The Global Enteric Multicenter Study (GEMS) of diarrheal disease in infants and young children in developing countries: epidemiologic and clinical methods of the case/control

study. Clinical Infectious Diseases. 2012; 55:S232–S245. https://doi.org/10.1093/cid/cis753 PMID: 23169936

17. Saidi SM, Lijima Y, Sang WK, Mwangudza AK, Oundo JO, Taga K, Aihara M, Nagayama K, Yamamoto H, Waiyaki PG. Epidemiological study on infectious diarrheal diseases in children in a coastal rural area of Kenya. Microbiology and Immunology. 1997; 41(10):773–778. https://doi.org/10.1111/j.1348-0421.1997.tb01925.x PMID: 9403500
18. Wood SN. Statistical inference for noisy nonlinear ecological dynamic systems. Nat Letters. 2010 Aug; 466:1102–07. https://doi.org/10.1038/nature09319
19. Hartig F, Calabrese JM, Reineking B, Wiegand T, Huth A. Statistical inference for stochastic simulation models and application. Ecology Letters. 2011; 14:816–27. https://doi.org/10.1111/j.1461-0248.2011.01640.x PMID: 21679289
20. Kurtz TG, Ethier S. Markov Processes: Characterization and Convergence. Wiley. 2005.
21. Brooks S, Gelman A, Jones GL, Meng X. Handbook of Markov Chain Monte Carlo. Chapman & Hall. 2011.
22. Healy-Profitos J, Mouhaman A, Lee S, Mouhaman A, Garabed R, Moritz M, Piperata B, Tien J, Bisesi M, Lee J. Muddying the Waters: A New Area of Concern for Drinking Water Contamination in Cameroon. Int. J. Environ. Res. Public Health. 2014; 11, 12454–12472. https://doi.org/10.3390/ijerph111212454
23. Andersson, H, Britton T. Stochastic Epidemic Models and their Statistical Analysis Springer. 2000.
24. Gilks WR, Best NG, Tan KKC.Adaptive Rejection Metropolis Sampling within Gibbs Sampling J R Stat Soc Series C (Appl. Stat.) 1995; 44(4): 455–472
25. Garrett V, Ogutu P, Mabonga P, Ombeki S, Mwaki A, Aluoch G, Quick RE. Diarrhoea prevention in a high-risk rural Kenyan population through point-of-use chlorination, safe water storage, sanitation, and rainwater harvesting. Epidemiol Infect. 2008; 136(11): 1463–1471 https://doi.org/10.1017/S095026880700026X PMID: 18205977
26. Brown J, Cairncross S, and JEnsink JH. Water, sanitation, hygiene and enteric infections in children. Arch Dis Child. 2013 Aug; 98(8): 629–634. https://doi.org/10.1136/archdischild-2011-301528 PMID: 23761692