Spring 5-13-2016

Machine Learning Approaches for Assessing Moderate-To-Severe Diarrhea in Children < 5 Years of Age, Rural Western Kenya 2008-2012

Tracy L. Ayers

Machine learning approaches for assessing moderate-to-severe diarrhea in children < 5 years of age, rural western Kenya 2008-2012

by

Tracy Leigh Ayers

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Epidemiology

in the

School of Public Health

of the

Georgia State University

Dissertation Committee:

Dr. Christine Stauber, SPH, GSU

Dr. Ruiyan Luo, SPH, GSU

Dr. Robert M. Hoekstra, DFWED, CDC

Dr. Ciara O'Reilly, DFWED, CDC

Spring 2016

Contents

Abstract

Worldwide diarrheal disease is a leading cause of morbidity and mortality in children less than five years of age. Incidence and disease severity remain the highest in sub-Saharan Africa. Kenya has an estimated 400,000 severe diarrhea episodes and 9,500 diarrhea-related deaths per year in children. Current statistical methods for estimating etiological and exposure risk factors for moderate-to-severe diarrhea (MSD) in children are constrained by the inability to assess a large number of parameters without the limitations of sample size, complex relationships, correlated predictors, and model assumptions of linearity. This dissertation examines machine learning statistical methods to address weaknesses associated with using traditional logistic regression models. The studies presented here investigate data from a 4-year, prospective, matched case-control study of MSD among children less than five years of age in rural Kenya from the Global Enteric Multicenter Study (GEMS). The three approaches include: Least Absolute Shrinkage and Selection Operator (LASSO), use of classification trees, and random forest.

A principal finding in all three studies was that machine learning methodological approaches are useful and feasible to implement in epidemiological studies. All provided additional information and understanding of the data beyond using only logistic regression models. The results from all three machine learning approaches were supported by comparable logistic regression results indicating their usefulness as epidemiological tools. This dissertation offers an exploration of methodological alternatives that should be used more frequently in diarrheal disease epidemiology, and in public health in general.

List of Tables

List of Figures

Chapter 1. Introduction

Background and Statement of Problem

Worldwide diarrheal disease is the second leading cause of mortality in children under-five years old and is responsible for nearly a million deaths each year (Levine 2013, WHO 2013, Liu 2012)). While there have been modest decreases in the incidence of diarrhea over the last several decades, it remains one the most frequent causes of hospital admissions in children worldwide, with the least amount of improvement seen in sub-Saharan Africa(UNICEF 2-13, Walker 2013). Kenya remains among the 15 countries with the highest burden for diarrhea associated mortality in children (Das 2014). Kenya has an estimated 400,000 severe diarrhea episodes and 9,500 diarrhea-related deaths per year in children (Walker 2013).

Efforts aimed at reducing diarrhea related morbidity and mortality are targeted through appropriate clinical case management (e.g.- prompt rehydration, administration of zinc, and use of antimicrobials where indicated), promotion of early recognition of severe illness in the community and availability, knowledge and use of oral rehydration solution at home, promotion of increased fluids and continued feeding during diarrheal episodes, implementation of water, sanitation and hygiene (WASH) interventions and vaccines. For these efforts to have the greatest impact, information on the relative contribution of diarrheal etiologies is essential to guide empiric clinical treatment, implementation of effective WASH interventions, and highlight potential vaccine development needs. In addition, assessment of environmental exposures is needed to prioritize community based interventions to reduce transmission.

The Global Enteric Multicenter Study (GEMS), a case-control study of moderate-to-severe diarrhea (MSD) in seven countries in South Asia and sub-Saharan Africa, was undertaken to assess the etiologic burden of MSD in its study sites (Levine 2012). Since vaccines have been identified to have the greatest impact in reducing the burden of pediatric diarrheal disease, it is important to identify the enteric pathogens present in children with MSD (Levine 2012). Identifying etiologies and their relative contribution to the burden of pediatric diarrheal disease using current data sources is paramount to targeting vaccine interventions.

For many pathogens, vaccine development will be a lengthy process that will work better for some than others. Currently, the only enteric vaccine becoming widely-available is for rotavirus. Efforts to reduce diarrhea related mortality include appropriate clinical case management at health facilities in the form of appropriate use of rehydration and antimicrobials Creating clinical profiles to distinguish between viral and bacterial causes will aid in the judicious use of antibiotics and focus attention on rehydration needs for viral infections. It is estimated that nearly 50% of antibiotic use in health facilities for the treatment of diarrhea is unnecessary from a survey conducted in Kenya and Ghana (Spreng 2014). Misuse may be high due to the absence of diagnostic panels for enteric pathogens are not readily available to distinguish infections.

Many intervention programs aimed at reducing transmission of diarrheal diseases in developing countries include structural and behavioral changes to improve water supply and quality, sanitation, and hygiene. Figure 1.1 presents a model of diarrheal disease transmission and the complex nature of water, sanitation and hygiene exposures. As highlighted in the figure, it is a complex and difficult process to assess and prioritize targeted interventions. Identifying the relative importance of each transmission pathway is challenging due to the complex inter-relationships.

Figure 1.1. Transmission by exposures
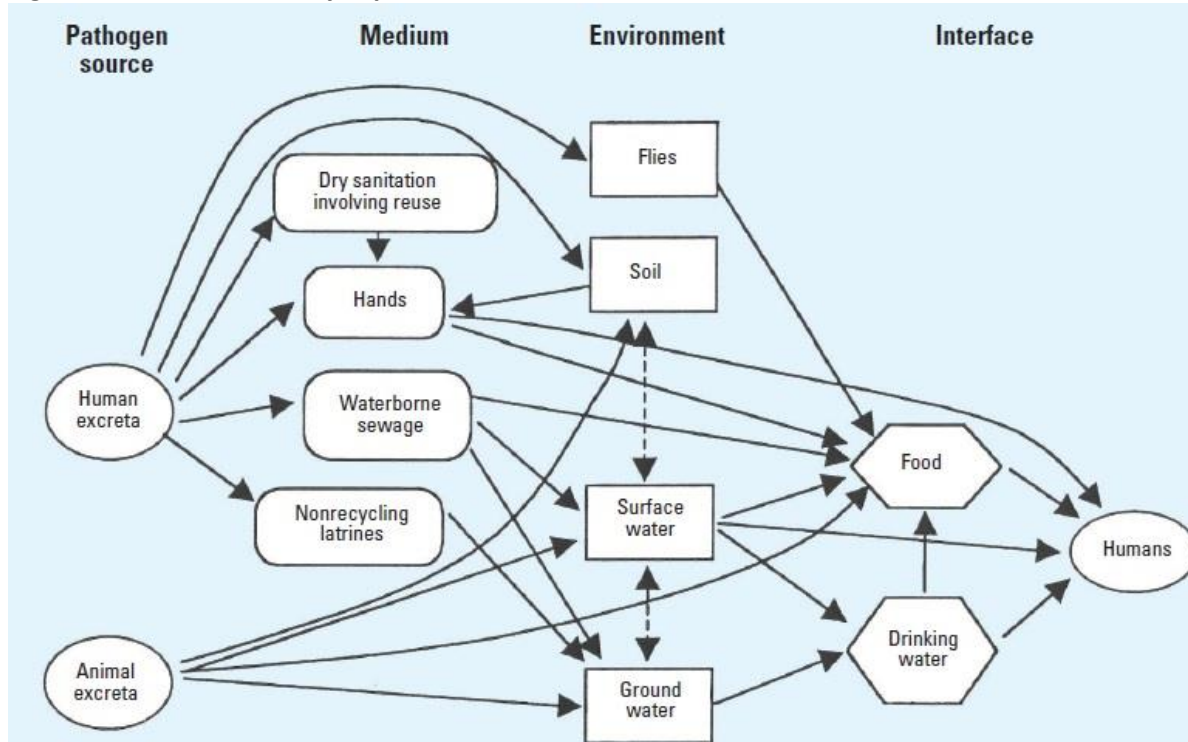
Prüss, Kay, Fewtrell, and Bartram. Estimating the Burden of Disease from Water, Sanitation, and Hygiene at a Global Level. Environmental Health Perspectives . 2002 : 110 , 5

Description of the Global Enterics Multicenter Study (GEMS)

GEMS study design and background

The purpose of GEMS was to estimate the burden, etiology, risk factors, and complications of MSD in children less than 5 years old. GEMS was funded by the Bill and Melinda Gates Foundation and was coordinated by the University of Maryland, Center for Vaccine Development. The seven study sites of GEMS were Basse , The Gambia; Siaya County (formerly Nyanza Province), Kenya; Bamako, Mali; Manhiça, Mozambique; Mirzapur, Bangladesh; Kolkata, India; and Karachi, Pakistan. In each site, GEMS targeted three age strata: infants (0‑11 months), toddlers (12‑23 months), and children (24‑59 months). The Demographic Surveillance System was used to enroll both cases and controls. (Kotloff 2012). For case enrollment, sites selected sentinel health facilities (SHFs) where DSS children sought care for diarrheal illnesses.

A matched case-control study was conducted in all seven country sites during 2007-2011. A case of MSD was defined as a child with a diarrheal illness <7 days duration comprising ≥3 loose stools in 24 hrs and ≥1 of the following: sunken eyes, skin tenting, dysentery, required IV rehydration, or hospitalization. Controls were selected using the DSS database to identify community matched controls. Controls were enrolled within 14 days of the case, and were required to be without diarrhea in the 7 days prior to enrollment and able to provide a stool specimen (Kotloff 2012). One to three controls were selected per case, depending on age stratum, and matched on age, gender, same or nearby village.

Study area and population

In Kenya, between January 31, 2008 and January 29, 2011, 3,359 children (1,476 cases and 1,883 controls) were enrolled into GEMS-1. Subsequent to this timeframe, the study was funded for an additional 11 months in Kenya. Between October 31, 2011 and September 30, 2012, 868 children (302 cases and 566 controls) were enrolled; this time period is known as GEMS-1a. The GEMS-1 Kenya study site was located in rural western Kenya in the districts of Gem and Asembo in Siaya County (formerly Nyanza Province). During GEMS-1a the study site was located in the districts of Asembo and Karemo in Siaya County as the Kenya Medical Research Institute (KEMRI)/CDC Kenya DSS moved to a new area during this time period (Figure 1.2). All papers in this dissertation utilize data from both GEMS-1 and GEMS1a.

Figure 1.2. KEMRI/CDC HDSS study area (Asembo, Gem and Karemo) where GEMS Kenya Study was conducted
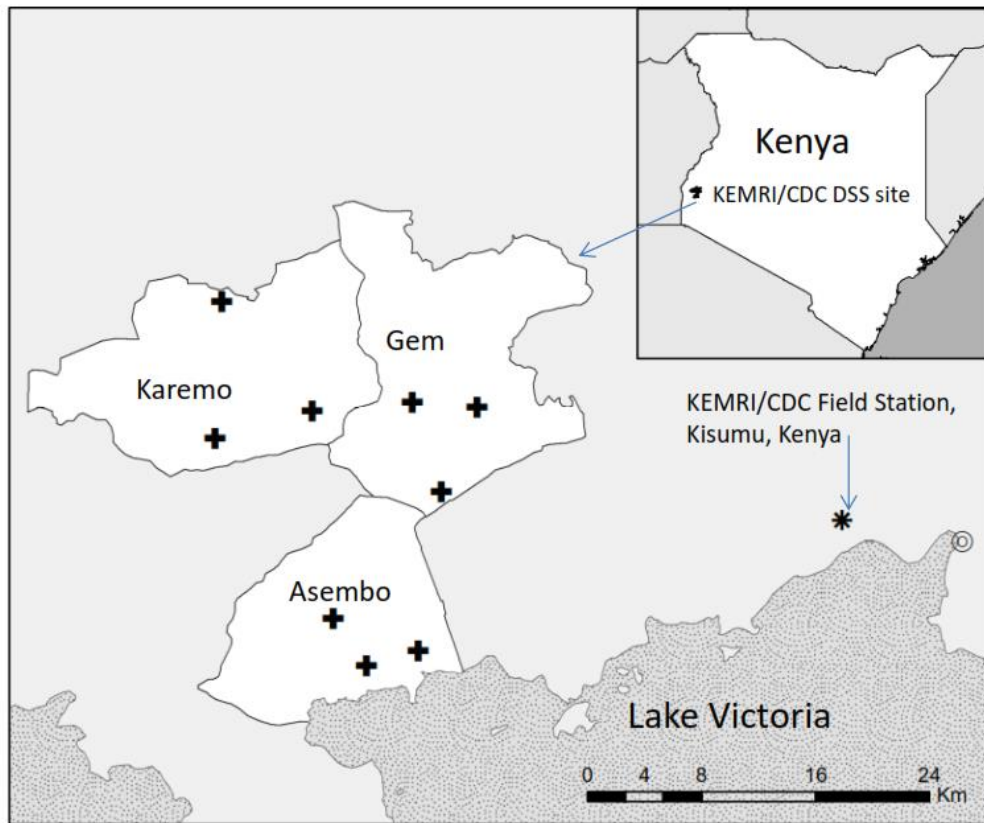
+ GEMS Sentinel Clinics where GEMS cases were enrolled in Western Kenya

Data collection

At enrollment, clinical assessments, anthropometric measurements, and stool specimens were collected from both MSD cases and their matched controls. This enabled laboratory testing for a full spectrum of 23 bacterial, viral and parasitic enteric pathogens, and subsequent characterization. Data were collected via survey questionnaires and scanned into a database which was reviewed for quality control by the centralized Data Coordinating Center (DCC) (Biswas 2012). Data was provided in SAS and Stata formats to sites.

Statement of purpose

Current statistical methods for estimating etiological and exposure risk factors for moderate-to-severe diarrhea (MSD) in children are constrained by the inability to assess a large number of parameters without the limitations of sample size, complex relationships, correlated predictors, model linearity assumptions and

instability, and as well as biased estimates. This dissertation examines machine learning statistical methods to assess etiologies, clinical profiles, and exposures associated with MSD in children less than five years of age in rural western Kenya.   All three studies presented investigate data from a 4-year, prospective, matched case-control study of MSD among children less than five years of age in rural Kenya from the Global Enteric Multicenter Study (GEMS). This dissertation examines using new machine learning statistical methods to address weaknesses associated with using traditional logistic regression for the purposes of assessing etiologies and exposures associated with MSD in children less than five years of age in rural western Kenya.

The first paper addresses the impact of model selection strategies on the description and variability of diarrheal etiologies associated with MSD. Newer statistical methods have been developed to handle exploring a large number of variables relative to sample size, known as 'shrinkage and selection methods'. Despite these statistical advantages over traditional model selection methods, they are not widely used in epidemiological studies. This first paper compares feature selection across five different logistic regression model selection approaches including two traditional stepwise procedures and three variants of Least Absolute Shrinkage and Selection Operator (LASSO) approaches.

While an extensive panel of laboratory diagnostics were supported during the active study period, they are not sustainable nor are they available for point-of-care treatment in all areas of Africa. The ability to differentiate enteric viruses from bacterial causes at the clinical setting, in the absence of laboratory diagnostics, is imperative for judicious use of antibiotics and in reducing antimicrobial resistance as a result of over prescribing. The second paper will apply classification tree methodologies for developing a clinically based decision tree for classifying rotavirus infections and highlighting high risk sub-populations.

The third paper will address the contributions of WASH exposures to MSD. WASH exposures are often correlated and nested within transmission routes, random forest methods provide a novel approach for summarizing groups of exposures in a potentially more efficient and appropriate manner. Risk factor analyses of diarrheal diseases are complicated because of numerous potential exposures that are often correlated. This

study investigates the use of machine learning approaches, specifically random forest (RF), to identify WASH

factors associated with diarrheal disease in children less than 5 years old. The newer approach of RF is compared

to the traditional analytic approach of using logistic regression models for evaluating WASH risk factors for MSD.

Estimates from the GEMS study are some of the only available burden of illness estimates for enteric

pathogens and WASH in developing countries, it is important to assess and explore multiple approaches to such

computations. All three studies will investigate the use of machine learning approaches to describe and

characterize etiologies, clinical profiles, and WASH factors associated with MSD. This dissertation will offer an

exploration of methodological alternatives to address common pitfalls of logistic regression analyses that exist

not only in diarrheal disease studies, but in epidemiology in general.

Chapter 2. Comparing novel shrinkage model selection methods for assessing etiologies associated with moderate-to-severe diarrhea in children <5 years old, rural western Kenya 2008-2012

Ayers TL[1, 6], Luo R[6], Omore R[2,3], Ochieng B[2,3], Farag TH[4], Nasrin D[4], Panchalingam S[4], Nataro JP[4], Kotloff KL[4], Levine MM[4], Oundo J[5], Parsons MB [1], Bopp C[1], Laserson K[2], Stauber CE[6], Breiman RF[7], Mintz E[1] , O'Reilly CE[1] and Hoekstra RM[1]

Abstract

Background: Multivariable model variable selection is one of the most difficult tasks for epidemiological analyses. The goal of achieving unbiased estimates and uncovering new relationships, such as interactions, is severely limited by sample size and computational capacity. Despite the emergence of newer statistical methods, such as penalized regression, they have not been applied in epidemiological studies outside of genomics. It is important to describe how alternative model selection methods influence model composition using real world epidemiologic data.

Methods: Using data from a large matched case-control study of moderate-to-severe diarrhea (MSD) model selection methods were compared. Specimens from both cases and controls were tested for 23 major enteric pathogens. Backward Elimination (BE) stepwise selection was applied to select pathogens associated with MSD case status. Alternative subset selection using penalized regression in the form of Least Absolute Shrinkage Selector Operator (LASSO) was investigated. Analysis was stratified by three age groups: infants (0-11 months), toddlers (12-23 months), and older children (24-59 months). Selection of pathogen variables and two-way interactions across methods was examined.

Results: Pathogens with stable and large effects were selected consistently across models. Pathogen and interaction selection varied the greatest in the infant subgroup analysis and demonstrated uncertainty in pathogen estimation. By comparison, the toddler and older children analyses demonstrated greater consistency across models. Pathogens selected by BE were also selected by LASSO methods. LASSO methods permitted tuning model complexity and always selected models that still converged when applied in standard multivariable logistic regression models.

Conclusions: This study demonstrated the feasibility of implementing newer LASSO model selection methods in epidemiological studies. LASSO methods permitted the inclusion of more pathogens and did not compromise the detection of the pathogens with clear associations with MSD. In some subgroups, model selection varied

more greatly than others. It is beneficial to apply different model selection strategies and consider the agreement and disagreement in making epidemiological conclusions.

Introduction

An important component of epidemiological analyses is determining which variables should be included in multivariable models. The goal of the multivariable model is to produce unbiased estimates while controlling for confounders. Including all possible variables in the model for full control of confounding often leads to model convergence issues (Greenland 2008). It is computationally prohibitive to perform best subset model selection strategies when there is a large number of independent variables. As a result, automatic stepwise model selection methods remain the most widely used methods for variable selection in epidemiology (Guo 2015). Variable selection in epidemiological modeling is a common and well known problem (Rothman, 2009). Limitations and concerns with using stepwise procedures have been well documented (Derkesen 1992, Whittingham 2006, Weigand 2010, Mundry 2009). Despite these well-known limitations, automatic stepwise regression remain the dominant multivariable modelling approach in epidemiological research (Walter 2009).

Automatic stepwise selection procedures reduce the quantity of model subsets considered by either adding or removing variables one at a time (forward or backward selection). The computational simplicity and ease of implementation have led to automatic stepwise selection popularity (Morzova 2015). While these selection strategies are more manageable, they are limited in identifying the best possible model since selection is always based on at least one fixed parameter in the model (James, Witten, Hastie, Tibshirani 2014). Other drawbacks for automatic selection procedures include the reliance on sufficient sample size, bias towards variables with greater frequency, coefficients may be biased upwards, and underestimation of standard errors (Morozova 2010, Rothman 2008, Derksen 1992).

Backwards elimination (BE) is one of the most commonly used automatic selection procedures that preferably starts with a fully saturated model that contains all variables and all interactions to be considered. However, this often leads to convergence issues and the model cannot be assessed. As result, two additional modifications are taken. One is to 'screen' variables based p-values from single variable models and only those

variables with a p-value below a predetermined threshold are considered for the backwards elimination model selection. While this approach reduces the number of main effects considered, the sample size may still restrict the ability to include interactions. Thus, an additional modification to the approach is to only consider main effects for backwards elimination selection and then consider interactions only among the remaining main effects chosen (Rothman 2009, Walter 2009)

Alternative model selection strategies, such as shrinkage or penalized regression, have emerged to as a method to handle high dimensional data in the late 1990's. Using a penalty, coefficients of unstable model parameters are 'shrunk' to zero (or close to zero). By imposing a constraint on the total value of coefficients, parameters with the largest variability are shrunk to zero and thus can be used for model selection (Steyerberg 2001, Hastie and Tibsihirani 2009, Walter 2009). The Least Absolute Shrinkage and Selection Operator (LASSO) technique is one of the key 'shrinkage with selection' methods developed and was first proposed by Tibshirani (Tibshirani 1996). LASSO penalization will constrain some parameters to exactly zero, thus a useful method for performing model reduction and selection without the need of multiple statistical tests to assess p-values. The ability for LASSO methods to search across a large number of variables without the constraint of sample size has sparked its wide use in high dimensional data scenarios, such as genomic data (Wei 2011, Won 2015).

We used data from a case-control study designed to identify the burden of MSD illnesses to specific enteric pathogens. Using a comprehensive panel of microbiological assays, stool specimens from both cases and controls were tested for 23 major pathogens. All pathogen variables are expected to have some association with the outcome of MSD. Shrinkage or penalized model methods have existed for nearly two decades, but have not been implemented or explored in epidemiological studies (Walter 2009). This study aims to compare backwards elimination (BE) and LASSO penalized logistic regression model selection methods using real world epidemiological data. While a definitive best method cannot be identified, this study will highlight the influence of model selection methods on conclusions and any potential uncertainty in estimates.

Materials and Methods

Dataset

Global Enteric Multi-centre Study (GEMS) case-control data from the Kenya site during 2008-2012 was used for this analysis. This data was collected as part of a 4-year, prospective, age-stratified, matched case-control study designed to investigateMSD in children aged 0-59 months. Case children were recruited from sentinel health facilities and matching controls were selected using the Demographic Surveillance System (DSS). Community controls were matched on age, sex, and geographic proximity. At enrollment, fecal samples were collected from both cases and controls to identify enteric pathogens. The panel and methods used to identify enteric pathogens, which includes 9 bacterial, 7 viral, and 3 parasitic etiologies are described in detail (Pachalingham 2012, Kotloff 2012). Pathogen results were recorded for each pathogen tested as simply present or absent and are considered independent variables. Both cases and controls could have more than one pathogen present.

Statistical analysis

Statistical analyses were performed using both SAS 9.3 (Cary, NC) and R Statistical Software (Foundation for Statistical Computing, Vienna, Austria). In keeping with the GEMS study design, analysis was stratified by age groups in which infants (0-11 month), toddlers (12-23 month) and older children (24-59 months) were modeled for the outcome of MSD separately. *Giardia lamblia* was excluded from all models because its role as a causal pathogen of illness is undetermined, particularly in non-industrialized settings (Muhsen 2012, Bilenko 2004, Cotton 2015). Among the 22 pathogen variables considered, only pathogens that were present in at least 10 observations were considered for multivariable modeling. Conditional logistic regression models were performed to preserve the matched case-control sets. However, because conditional penalized models are limited in their capacity to detect interactions and preserve model hierarchy, unconditional logistic regression models were computed for comparison. Model selection was performed on four model types to account for the

penalized vs non-penalized approach and the conditional vs unconditional model structure for each age strata. Two-way pathogen interactions were assessed in all models.

Backwards elimination model selection

For the traditional stepwise approach, we utilized backward elimination (BE) strategies. All pathogen variables were first screened using univariable models. Pathogens with $p \leq .20$ in simple logistic regression were considered in the multivariable model. All pathogens considered in the multivariable model were then included in the model at the start and removed one at a time based on the largest p-value, until all pathogens remained significant at $p \leq .05$. Subsequently, among the pathogens that remained in the multivariable model, all possible two-interactions were considered and removed one at a time until all parameters in model remained significant at $p \leq .05$. This backward elimination approach was performed using both conditional and unconditional logistic regression models.

Penalized and shrinkage model selection

Penalized logistic regression models were computed using the *clogitL1* R package for conditional models and *hierNet* R package for unconditional models. We used the method for computing a penalized logistic regression for matched case-controls studies developed by Reid and Tibshirani (2014). For the conditional LASSO approach, we first only considered the main effects of pathogens and selected the Lambda (penalty parameter) based on 10-fold cross-validation with the lowest error. For all conditional LASSO models, we selected the largest within one standard error from the minimum error to produce the most parsimonious model (referred to as λ max 1SE). A larger value of lambda leads to a sparser model with less predictors. One limitation of the conditional LASSO, is its inability to maintain model hierarchy when evaluating interactions. That is, the algorithm may select an interaction term in the model but omit the main effects involved in the interaction. For this reason we used the conditional LASSO to select pathogen main effects, but interactions were assessed using traditional model approaches. To illustrate the process of variable coefficients and selection in relation to Lambda values we provide trace plots for the main effects considered in the conditional LASSO model.

In addition, we performed model selection using a hierarchical group LASSO algorithm (HG LASSO) designed to address the ability of searching for two-way interactions while maintaining model hierarchy (Bien and Tibshirani 2013). This method permits searching for all two-way interactions while simultaneously considering all main effects to produce a final model that preserves variable hierarchy. Similar to the conditional LASSO approach, we selected the largest Lambda value within one standard error from the minimum error (referred to as λ max 1SE). Since this was the only model selection approach in which it was computationally feasible to fully evaluate interactions, we also ran models with the smallest Lambda value whose error was equal to the minimum cross-validation error. A smaller lambda leads to a more complex model with more predictors, where interactions are more likely to be included. Since the hierarchical group LASSO considers all main effects and two-way interactions simultaneously, a very large number of parameters were considered, and we did not provide variable coefficient and selection plots by Lambda for these models. We explored all LASSO methods in this paper as feature selection methods, but report Odds Ratios (ORs) and 95% Confidence Intervals (CIs) based on standard logistic regression coefficients for ease of interpretation and comparison.

Results

Among the pathogens considered, four pathogens were excluded for insufficient frequency (Enterohaemorrhagic *E. coli* (EHEC), *Aeromonas, Vibrio* spp*.,* and *Salmonella typhi*). Overall frequencies of cases and controls are presented in Supplemental table 1 (S1). For conditional LASSO models, the relationship between pathogen variable selection and the tuning parameter Lambda is illustrated in Figure 2.1. Each line corresponds to a pathogen variable coefficient, demonstrating that as Lambda increases (from left to right) some of the coefficients are shrunk to zero. In the infant model plot, rotavirus is the last parameter to have coefficient shrunk to zero thus demonstrating the stability of the rotavirus relationship with MSD. Lines below zero on the y-axis demonstrate a protective association with MSD. When a pathogen remains in the models, its coefficients remain stable and only change when shrunk to zero demonstrating that all the pathogens have relatively stable coefficients.

In the infant models, 7 of the 18 pathogens were selected in all five models (Figure 2.2 and Supplemental Table 2.1). Among the two conditional models, the LASSO model selected one additional main effect and one additional interaction term. Both the conditional BE and LASSO models selected the same pathogens, only the LASSO model also included *Campylobacter jejuni.* The primary difference between the two model selection strategies was the selection of interaction terms. Conditional BE models detected a statistically significant interaction between enteropathogenic *E. coli* (EPEC) and *Shigella* at exactly p=.05.   Among the three unconditional models, as expected, BE selected the fewest parameters (5 main effects and 2 interactions) while the hierarchical group LASSO with the smallest Lambda value selected the largest number of parameters (13 main effects and 2 interactions). Among the 18 pathogens selected to be included in any of the unconditional logistic regression models, as either main effects or as part of pathogen interactions, 5 pathogens were consistently chosen. Unconditional BE model selection identified a statistically significant interaction (p =.04) between rotavirus and norovirus-GII. However, the hierarchical group LASSO model with the larger Lambda value included three additional viral pathogens as main effects and did not detect any interactions (Figure 2.2 and Supplemental Table 2.2).

In either toddler or older children analysis, no interactions were selected across any of five model selection methods. For the toddler specific analysis, conditional LASSO and BE models both selected 7 pathogens, but differed on the inclusion of the protective effect of atypical EPEC and omission of *Entamoeba histolytica* in the LASSO model. In unconditional models, BE and LASSO models selected the same 6 pathogens across all models. Only the LASSO model with the minimized Lambda value selected additional 7 pathogens (Figure 2.33 and Supplemental Table 2.3). For the older children specific analysis, 4 pathogen main effects which included enterotoxigenic *E. coli* any ST, *Shigella*, rotavirus, and *Cryptosporidium*, were selected across all model selection methods. The conditional LASSO model selected 3 additional pathogens, all with protective effects, compared to the conditional BE model. Using unconditional analysis, hierarchical group LASSO selected two additional pathogens than BE methods. Adenovirus 40/41 and *Entamoeba histolytica* demonstrated large,

but insignificant, effects and were included in both hierarchical group LASSO approaches. Thus, the change in

shrinkage parameter had no effect on parameter selection in the unconditional older children specific analysis

(Figure 2.4 and Supplemental Table 2.4).

Discussion

This study demonstrates the variability and impact of methods used for model selection. While there

was consistency across the major pathogens such as rotavirus, *Shigella, and Cryptosporidium,* there was

variability in identification of interactions and covariates. The consistent detection of the major pathogens

across all methods supports the use of alternative model selection strategies as they are stable in selecting

important pathogens, but these approaches may add to our understanding and estimation by depicting

uncertainty in estimates. In the infant model, traditional conditional logistic regression selected an interaction

between *Shigella* and EPEC that was not selected by any of the other models. It is likely that because these two

pathogens were the more frequently identified pathogens and that this selection is based on sample size power.

It is important to note that all five model selection approaches selected pathogen interaction effects within the

infants. However, none of the models selected the same interaction effects. It is an important finding that there

were inconsistent conclusions across the infant models and points to an area of uncertainty about the impact

and interpretation of co-infections in infants. More longitudinal cohort designed studies are needed to better

access which infections are occurring when and the role of co-infections in infants, such as the Mal-ED project

(Platts-Mills 2015).

There are several limitations to our comparison of models. We did not utilize the coefficients from the

penalized LASSO models, therefore the estimates do not reflect the shrinkage estimated by LASSO and estimates

still reflect the upward bias from non-penalized estimates. Ability to easily implement a method for computing

standard errors of LASSO derived coefficients is currently lacking (Steyerberg 2000, Morzova 2015). Since this

study used real world epidemiological data, there is no way to conclude how the model selection methods

compared to a 'true' or 'correct' model. Thus, in applied epidemiological settings none of the model selection

methods obviate the need for external epidemiological guidance. Finally, this study represents a case study and is limited in its generalizability to other epidemiology studies.

LASSO models are known to increase model stability over stepwise methods and should be considered in more epidemiological studies for a more coherent comparison of effects across studies (Morozova 2015). In addition, LASSO models have demonstrated superior performance to automatic stepwise procedures especially when data sets are small (Steyerberg 2000, Steyerberg 2001). These newer 'shrinkage and selection' methods should be explored more frequently in epidemiological studies. While LASSO methods are not yet widely available in all statistical software, they are currently easy to implement using R software. Epidemiologist should use multiple model selection methods, including penalized regression, to explore stability in conclusions especially in situations where sample size is small relative to the number of variable of interest, when the frequency of predictors are unbalanced, or when identification of effect modifiers is important. This study highlights the feasibility of applying new LASSO techniques for epidemiological studies. The conditional LASSO and hierarchical group LASSO algorithms specifically address methods needed for epidemiological studies, such as case-control designs. These LASSO methods are promising, especially in situations in which there is a large number of variables and interactions to consider relative to the sample size and should be used more frequently in epidemiological studies

Figure 2.1. Trace plots of pathogen variable coefficients by Lambda for Infants, toddlers, and older children
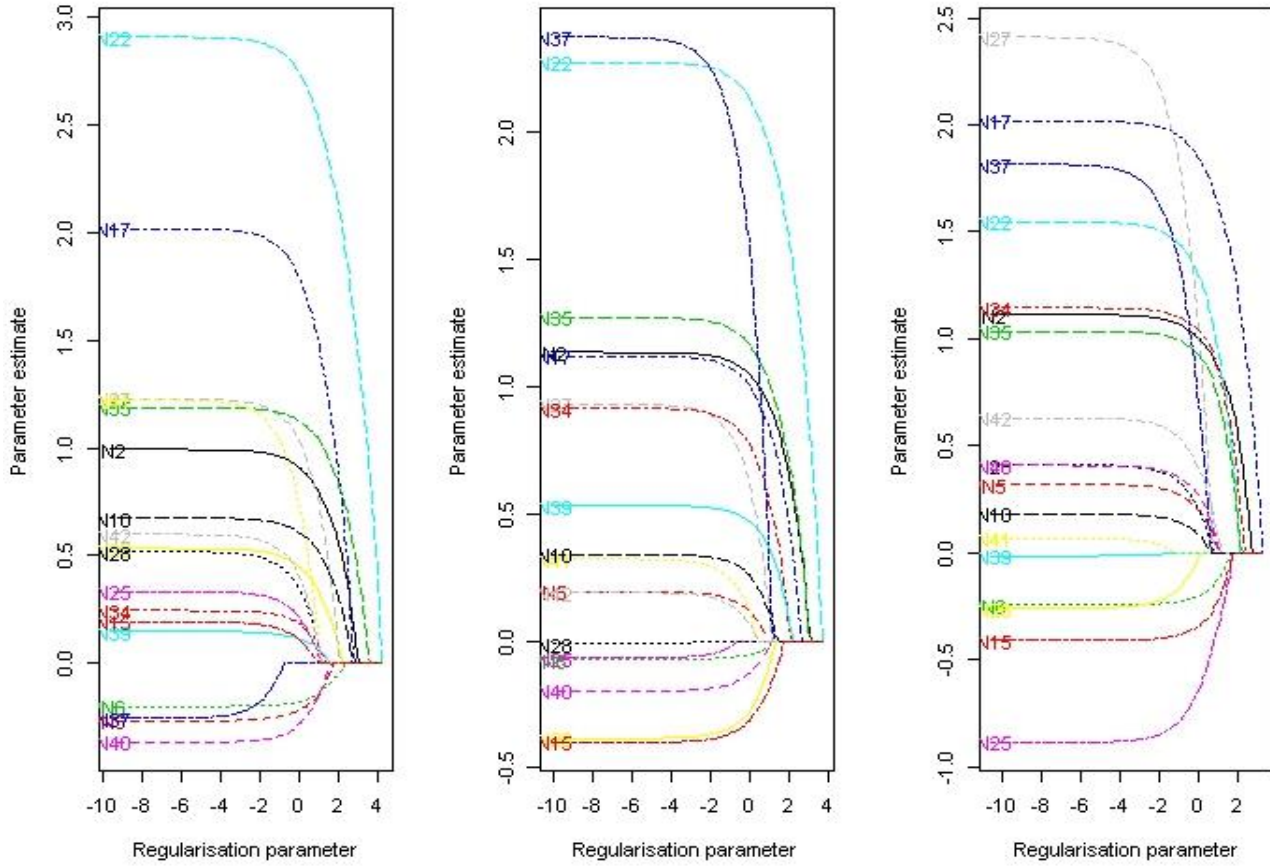
Figure 2.2. Pathogen variables and interactions selected by model selection method for infants (0-11 months)*
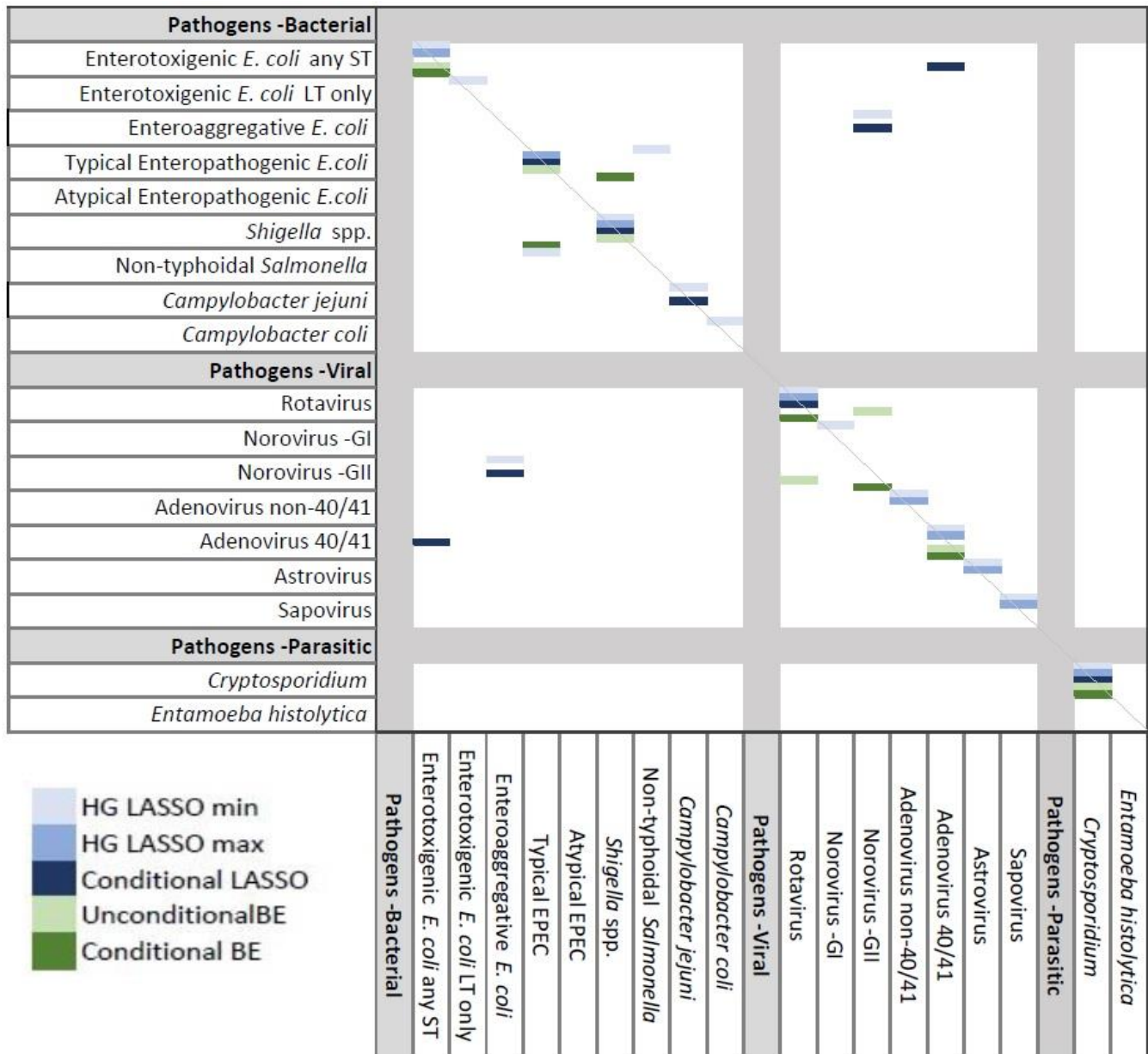
*HG LASSO – Hierarchical Group LASSO, BE –Backwards Elimination

Figure 2.3. Pathogen variables selected by model selection method for toddlers (12-23 months)
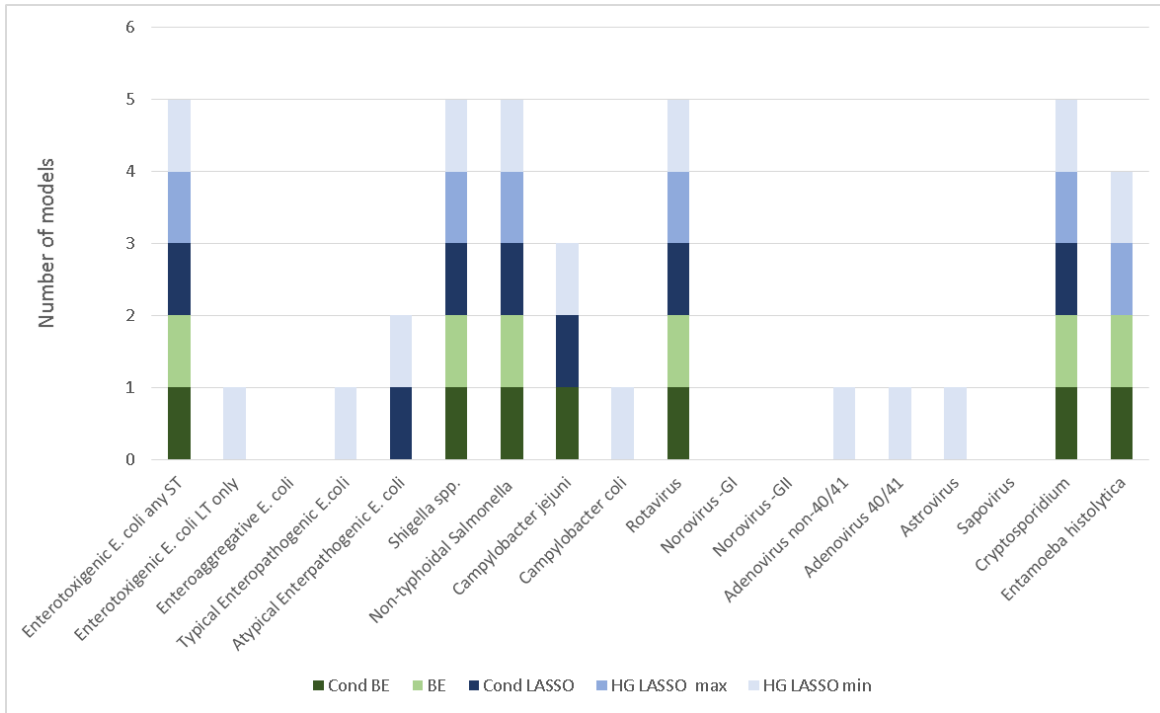
Figure 2.4. Pathogen variables selected by model selection method for older children (24-59 months)
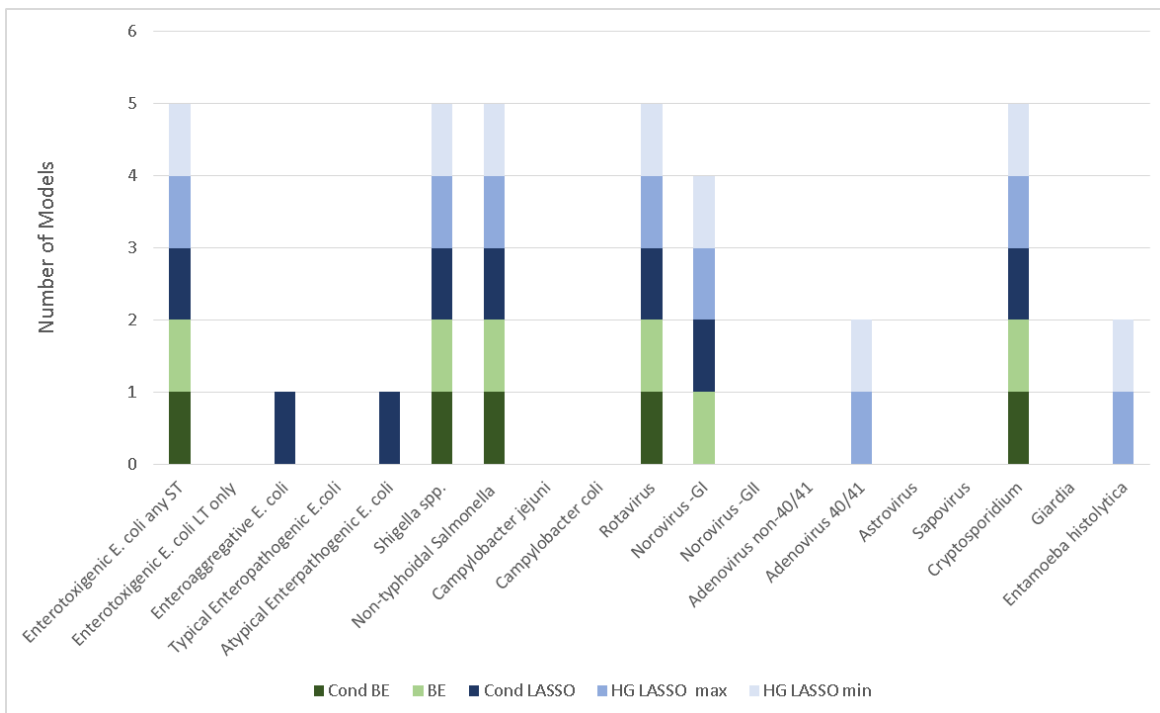
References

Avalos, M., Pouyes, H., Grandvalet, Y., Orriols, L., & Lagarde, E. (2015). Sparse conditional logistic regression for analyzing large-scale matched data from epidemiological studies: a simple algorithm. BMC Bioinformatics, 16 Suppl 6, S1. doi:10.1186/1471-2105-16-s6-s1

Bien, J., Taylor, J., & Tibshirani, R. (2013). A LASSO FOR HIERARCHICAL INTERACTIONS. Ann Stat, 41(3), 1111-1141. doi:10.1214/13-aos1096

Bilenko, N., Levy, A., Dagan, R., Deckelbaum, R. J., El-On, Y., & Fraser, D. (2004). Does co-infection with Giardia lamblia modulate the clinical characteristics of enteric infections in young children? Eur J Epidemiol, 19(9), 877-883.

Cotton, J. A., Amat, C. B., & Buret, A. G. (2015). Disruptions of Host Immunity and Inflammation by Giardia Duodenalis: Potential Consequences for Co-Infections in the Gastro-Intestinal Tract. Pathogens, 4(4), 764-792. doi:10.3390/pathogens4040764

Derksen, S. a. K. H. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. British Journal of Mathematical and Statistical Psychology, 45(2), 265-282. doi:DOI: 10.1111/j.2044-8317.1992.tb00992.x

Flack, V. F., & Chang, P. C. (1987). Frequency of Selecting Noise Variables in Subset Regression Analysis: A Simulation Study. The American Statistician, 41(1), 84-86. doi:10.2307/2684336

Greenland, S. (1989). Modeling and variable selection in epidemiologic analysis. Am J Public Health, 79(3), 340-349.

Greenland, S. (2008). Invited commentary: variable selection versus shrinkage in the control of multiple confounders. Am J Epidemiol, 167(5), 523-529; discussion 530-521. doi:10.1093/aje/kwm355

Guo, P., Zeng, F., Hu, X., Zhang, D., Zhu, S., Deng, Y., & Hao, Y. (2015). Improved Variable Selection Algorithm Using a LASSO-Type Penalty, with an Application to Assessing Hepatitis B Infection Relevant Factors in Community Residents. PLoS One, 10(7), e0134151. doi:10.1371/journal.pone.0134151

Lim, M., & Hastie, T. (2015). Learning interactions via hierarchical group-lasso regularization. J Comput Graph Stat, 24(3), 627-654. doi:10.1080/10618600.2014.938812

Morozova, O., Levina, O., Uuskula, A., & Heimer, R. (2015). Comparison of subset selection methods in linear regression in the context of health-related quality of life and substance abuse in Russia. BMC Med Res Methodol, 15, 71. doi:10.1186/s12874-015-0066-2

Muhsen, K., & Levine, M. M. (2012). A systematic review and meta-analysis of the association between Giardia lamblia and endemic pediatric diarrhea in developing countries. Clin Infect Dis, 55 Suppl 4, S271-293. doi:10.1093/cid/cis762

Mundry, R., & Nunn, C. L. (2009). Stepwise model fitting and statistical inference: turning noise into signal pollution. Am Nat, 173(1), 119-123. doi:10.1086/593303

Pavlou, M., Ambler, G., Seaman, S., De Iorio, M., & Omar, R. Z. (2015). Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. Stat Med. doi:10.1002/sim.6782

Platts-Mills JA, Babji S, Bodhidatta L, et al. 2015. Pathogen-specific burdens of community diarrhoea in developing countries: a multisite birth cohort study (MALED). Lancet Glob Health. 2015;3(9):e564-75

Reid, S., & Tibshirani, R. (2014). Regularization Paths for Conditional Logistic Regression: The clogitL1 Package. J Stat Softw, 58(12).

Steyerberg, E. W., Eijkemans, M. J., Harrell, F. E., Jr., & Habbema, J. D. (2000). Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med, 19(8), 1059-1079.

Steyerberg, E. W., Eijkemans, M. J., Harrell, F. E., Jr., & Habbema, J. D. (2001). Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. Med Decis Making, 21(1), 45-56.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society, 58(Series B), 267-288.

Walter, S., & Tiemeier, H. (2009). Variable selection: current practice in epidemiological studies. Eur J Epidemiol, 24(12), 733-736. doi:10.1007/s10654-009-9411-2

Wei, F., Huang, J., & Li, H. (2011). VARIABLE SELECTION AND ESTIMATION IN HIGH-DIMENSIONAL VARYING-COEFFICIENT MODELS. Stat Sin, 21(4), 1515-1540. doi:10.5705/ss.2009.316

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? J Anim Ecol, 75(5), 1182-1189. doi:10.1111/j.1365-2656.2006.01141.x

Wiegand, R. E. (2010). Performance of using multiple stepwise algorithms for variable selection. Stat Med, 29(15), 1647-1659. doi:10.1002/sim.3943

Won, S., Choi, H., Park, S., Lee, J., Park, C., & Kwon, S. (2015). Evaluation of Penalized and Nonpenalized Methods for Disease Prediction with Large-Scale Genetic Data. Biomed Res Int, 2015, 605891. doi:10.1155/2015/605891

Chapter 3. Identifying clinical profiles for rotavirus among children < 5 years of age with moderate-to-severe diarrhea in rural western Kenya 2008 2012: a classification tree approach

Ayers TL[1, 6], Luo R[6], Omore R[2,3], Ochieng B[2,3], Farag TH[4], Nasrin D[4], Panchalingam S[4], Nataro JP[4], Kotloff KL[4], Levine MM[4], Oundo J[5], Parsons MB[1], Bopp C[1], Laserson K[2], Stauber CE[6], Breiman RF[7], Mintz E[1] , O'Reilly CE[1] and Hoekstra RM[1].

Abstract

Background: Laboratory diagnostics at the point-of-care for children with moderate-to-severe diarrhea (MSD) are lacking. The ability to differentiate enteric viruses from bacterial and parasitic causes at the clinical setting, in the absence of laboratory diagnostics, is imperative for judicious use of antibiotics and in reducing antimicrobial resistance as a result of over prescribing.

Methods: Data from a 4-year, prospective, case-control study of MSD among children less than five years of age in rural Kenya. Cases with MSD were enrolled at sentinel health facilities in Kenya and were assessed for demographic and clinical features. Classification trees using clinical profiles for identifying rotavirus infections were developed.

Results: Both the recursive partitioning classification tree and a conditional inference tree highlighted that the at risk sub-population of children less than 18 months of age with rotavirus are likely to predominantly present for care with vomiting during warm-dry months. The rotavirus classification trees presented a useful algorithm for understanding the data structure and identifying high-risk groups among correlated clinical features.

Conclusions: The classification tree methodology identified homogeneous subgroups of cases based on clinical presentation as they related to rotavirus positivity. The risk magnitude of given risk factors within the subgroup were highlighted. While a useful classification method offers visualization of clinical decision making and structure of the data, it does not eliminate the need for more detailed clinical evaluation.

Background

For children under 24 months of age, rotavirus was the most frequent cause of illness in Kenya, and the most frequent cause of moderate-to-severe (MSD) in children under 12 months of age in all 7 GEMS country sites (Kotloff 2013). Unlike most bacterial enteric pathogens, a vaccine currently exists for rotavirus and was

introduced in July 2014 after the GEMS study period vaccination campaigns in Kenya are in the process of being

launched (personal communication Omore et al. In clearance). While laboratory diagnostics were available to

confirm the presence of rotavirus and other pathogens, this intensive and exhaustive testing is not sustainable in

resource limited settings, such as rural Kenya. Since laboratory diagnostics will not continue to be available after

initial GEMS study period, a statistically supported approach to diagnosis based on demographic and clinical

features would assist with point-of-care treatment and monitoring the impact of the vaccine introduction.

As the most frequently identified pathogen, it is important to provide a tool for clinicians in the field to

quickly identify whether a child with MSD is likely to have rotavirus or not in order to begin appropriate

treatment early. Children with rotavirus who are adequately rehydrated at the health facility are likely to

survive their infection compared to children with other enteric pathogens, such as a multi-drug-resistant

bacterial infection (O'Reilly 2012). Therefore, it is imperative that children with MSD be identified and managed

expeditiously so that focus can be spent on other complex diagnoses. In addition, because diagnostics are often

not available at the point of care, it is important to provide assistance for rapid clinical decisions at clinical

presentation.

Since laboratory diagnostics are not widely available, patients are often treated with broad spectrum

antibiotics, when available. Resistance to antibiotics has developed in many bacterial enteric pathogens,

including *Salmonella* and *Shigella* spp (Gebrekidan 2015, Brooks 2006).   In a recent assessment healthcare

workers in Kenya and Ghana, only 14% correctly identified a case patients as having a viral infection among

acute gastroenteritis patients. In addition, antibiotics were prescribed at rates of nearly 50% (Spreng 2014). In

order to combat the continued increase in antibiotic resistant pathogens, a tool for improving identification of

viral infections is necessary.

Decision tree methods for clinical decision making have been predominantly used in industrialized

settings, where multiple treatment options and resources are available (Walsh 2014, Jung 2015, Mody 2015, Van

Hlst 2015, Varma 2004). The advantage for using tree-based analytic methods for clinical decisions are that they

can model more complicated and non-linear relationships (Berk 2009, Jung 2015). Recursive partitioning of the data provides a classification tree that can describe effect modification and specific risk sub-groups. In addition, recursive partitioning methods are particularly useful for their ability to generate output that is easily interpreted, even when describing higher order interactions (Auston 2012, Van Hulst, 2015).

The primary objective of this paper is to identify key demographic, including seasonality, and clinical features that are immediately observable, for classifying children < 5 years of age with rotavirus using decision tree methods. A secondary objective is to examine how well the resulting classification algorithm can be used for prediction of rotavirus, for purposes of monitoring rotavirus in the absence of diagnostics.

Methods

Data

Data was collected as part of a prospective matched case-control study. Cases with MSD were enrolled at sentinel health facilities in Kenya in the districts of Gem, Asembo, and Karremo in Siaya County (formerly Nyanza Province) from January 31, 2008 to September 30, 2012. An MSD case was defined as a child with a diarrheal illness <7 days duration comprising ≥3 loose stools in 24 hrs and ≥1 of the following: sunken eyes, skin tenting, dysentery, required IV rehydration, or hospitalization. Controls were selected using the Demographic Surveillance System database to identify community matched controls.

At enrollment, demographic, clinical, epidemiological information and stool samples were collected. Rotavirus VP6 antigen was detected in the whole stool specimen by a well-validated commercial enzyme-linked immunosorbent assay (ELISA) (ProSpecT rotavirus kit, Oxford, Basingstoke, UK). Detailed laboratory methods are described elsewhere (Pachalingham 2013)

Statistical analysis

To assess clinical profiles of MSD children positive for rotavirus, we restricted the analysis to cases in which a pathogen was identified and further reduced to only observations with a single pathogen identified.

During enrollment at healthcare facilities demographic and clinical features were assessed from caretaker self-reporting, health facility assessment, and medical assessment. All features were explored using frequencies and simple logistic regression odds ratios (ORs) and 95% confidence intervals (CIs) are reported.

Recursive partitioning methods, from the *rpart* algorithm implemented in R statistical software version 3.2.3, were used for generating Classification Trees (CT). In the process of building a tree, an iterative process is performed to select a variable and variable values to split the data into two groups so that the outcome rotavirus presence or absence is the most homogenous in both groups. The homogeneity of the subgroups identified is measured based on the Gini index, a measure of 'purity', and splits are chosen to maximize the index value. The recursive process repeats and variables were chosen with replacement such that the same predictor could be chosen again. The process stopped when no additional splits are possible or nodes no longer contained a minimum number of 20 observations. To reduce overfitting, the tree was reduced, or 'pruned', using a complexity parameter cut off (Berk 2009, Zang 2010). The complexity parameter evaluates the cost of adding another variable with the gain in accuracy. Using 10-fold cross validation, the complexity parameter was selected and used as a stopping criterion to control the size of the tree.

Since decision tree classification algorithms are biased towards classifying based on the majority group, and this data was imbalanced, we incorporated a loss matrix into the classification tree by weighting how much to penalize false negative classification in a given choice split (Japkowicz 2001).   This approach has been shown to have superior properties to using either over-sampling or under-sampling to balance groups (Wan, 2014, Drummond 2003,Batistia 2004). We evaluated a non-weighted tree and a tree weighted based on the ratio of the outcome in our data by considering model performance in terms of sensitivity, specificity, area under the curve (AUC), and interpretability of the tree.

In parallel, we explored conditional inference trees, another type of recursive partitioning model using *party* package in R. Candidate predictors applied to split nodes to minimize misclassification and was based on the permutation test to compute p-values (Berk 2009, Hothorn 2006). Predictors were chosen based on the

smallest p-value first, and performed iteratively within subgroups until no other statistically significant predictors were found. Using this algorithm, cases could be weighted to account for imbalance in outcome. The weights were included in the computation of permutation test p-values. We considered several values of weights and assessed model performance using AUC.

Results

The study enrolled, 1778 cases with MSD during 2008to 2012. Pathogens were identified in 1,436 cases of which 719 cases had only one pathogen identified. Of the 719 single pathogen cases, 90 (12.5%) were rotavirus positive. In univariable analysis, the following clinical features were associated with rotavirus positivity at alpha =.05: age, onset in a warm-dry month, vomiting $\geq$3 times in 24 hours, not having a fever, restless or irritable, presence of dry mouth, abnormal mental status, and being admitted to the hospital.

We constructed a training data set, which reserves some observations for testing, with 575 MSD cases with a single pathogen. Of these 73 were rotavirus positive. The initial unweighted classification tree resulted in a single node tree, after pruning, with an AUC of .50 and thus was a non-informative model. We identified a weight of misclassifying false negatives (Type II error) to misclassifying false positives (Type I) as 8:1 respectively reflected the ratio of the outcome in the data. In addition, this misclassification weight generated the tree with the most interpretable tree, with 13 nodes, and the greatest AUC. The resulting diagnostic algorithm is shown in Figure 3.1A. The first splitting feature is the child's age in months, with the cut point of 18 months of age. Each node indicates the proportion of rotavirus cases in each subgroup, on the right side, and the proportion of rotavirus negative on the left side. Following the right most branch of the tree reveals that among children less than 18 months of age and with vomiting $\geq$ 3 times in past 24 hours that 24% of observation are rotavirus positive, 76% rotavirus negative, and that the subgroup total is 33% of the original sample. The 24% prevalence of rotavirus in this subgroup is much greater than the 13% prevalence in the un-partitioned training data. Green nodes indicate subgroups likely to be rotavirus negative while blue nodes represent sub groups with increased probability of rotavirus positive. The same diagnostic algorithm is alternatively displayed in Figure 3.1B to

demonstrate the raw frequencies of rotavirus positive cases present in each node. Among the 373 children with vomiting and less than 18 months of age, 65 are rotavirus positive (of the starting 73 positive cases in the training data set). The AUC of the classification tree was .816 on the training data and .6125 on the test data (Figure 3.2).

The unweighted conditional inference tree produced a single partition of only age, but statistically chose the cut point of 13 months (Figure 3.3). The greatest probability of rotavirus positivity among MSD cases less than or equal to 13 months of age. The weighted tree model, with rotavirus positive cases weighted two times greater than rotavirus negative cases, generated 6 partitions with age remaining the primary partition (Figure 3.4).   The largest group identified was children ≤18 months of age and with vomiting present and also had the highest probability of being rotavirus positive. The second largest group was children ≤18 months of age, without vomiting, enrolled in warm-dry months. None of the MSD cases aged ≤ 18 months of age, enrolled in cool-wet months, and with a normal skin pinch test were positive for rotavirus. We considered varying weights for the rotavirus cases, and while other weights produced higher AUC values, they produced extraordinarily large trees. For example, a condition inference tree with rotavirus cases weighted 5 times more than rotavirus negative cases produced a tree with 17 nodes. Comparison of sensitivity, specificity and AUC values are illustrated in Figure 3.5.

Discussion

Using a classification tree approach, we aimed to develop a clinical decision tool for delineating viral causes of diarrhea from bacterial or parasitic causes. The demographic feature of age was the strongest predictor of rotavirus infection as was the first node identified in both tree algorithms. The increased risk of rotavirus infection in infants is well documented (Omore 2015, Kotloff 2013). The second most important clinical feature was the presentation with vomiting of 3 or more times in 24 hours, which was recently highlighted in another study (Gasparinho 2016). This study highlighted the sub-group and the risk of rotavirus within each

group.   The classification tree also highlighted subgroups within infants that were less likely to have rotavirus. For example, infants presenting for care during cold-wet months were less likely to be rotavirus positive.

Both the recursive partitioning classification tree and the conditional inference tree offered visual interpretations of the data and groups likely to have rotavirus. However, this study had several limitations. While classification tree have demonstrated easily interpreted output that are similar to how clinicians make decisions, they are sensitive to the data used for the algorithm. If any selection bias exists in the study data and the biased variable is chosen early in the algorithm, that error will be propagated throughout the tree making it not useful in other settings. Another limitation is the clinical features presented here may not be the best indicators to distinguish rotavirus infections. Aside from vomiting, other clinical features were not strong classifiers indicating either the set of features or details may be lacking. Acute gastroenteritis infections can present similar features despite the numerous etiological causes, thus it's possible that diagnosis based on symptoms, particularly among children with MSD, is severely limited.

Our tree-based models highlight the importance of vomiting and season in considering viral causes of diarrhea. This simplified message could be useful in training health care providers in settings of high rotavirus prevalence when considering treatment options. Training using classification tree diagrams that can be displayed may offer easier reminders. Reducing antibiotic use for infants that present with vomiting in warm-dry months is a simplified formula that would have a large impact on reducing the misuse of antibiotics. However, it is important to note that other factors not collected in this study might influence the clinical decision for treatment and future studies should evaluate whether using a decision tree is effective for health care provider training and practice, in particular in rural resource limited settings.

Figure 3.1 A. Rotavirus classification tree using demographic and clinical profiles
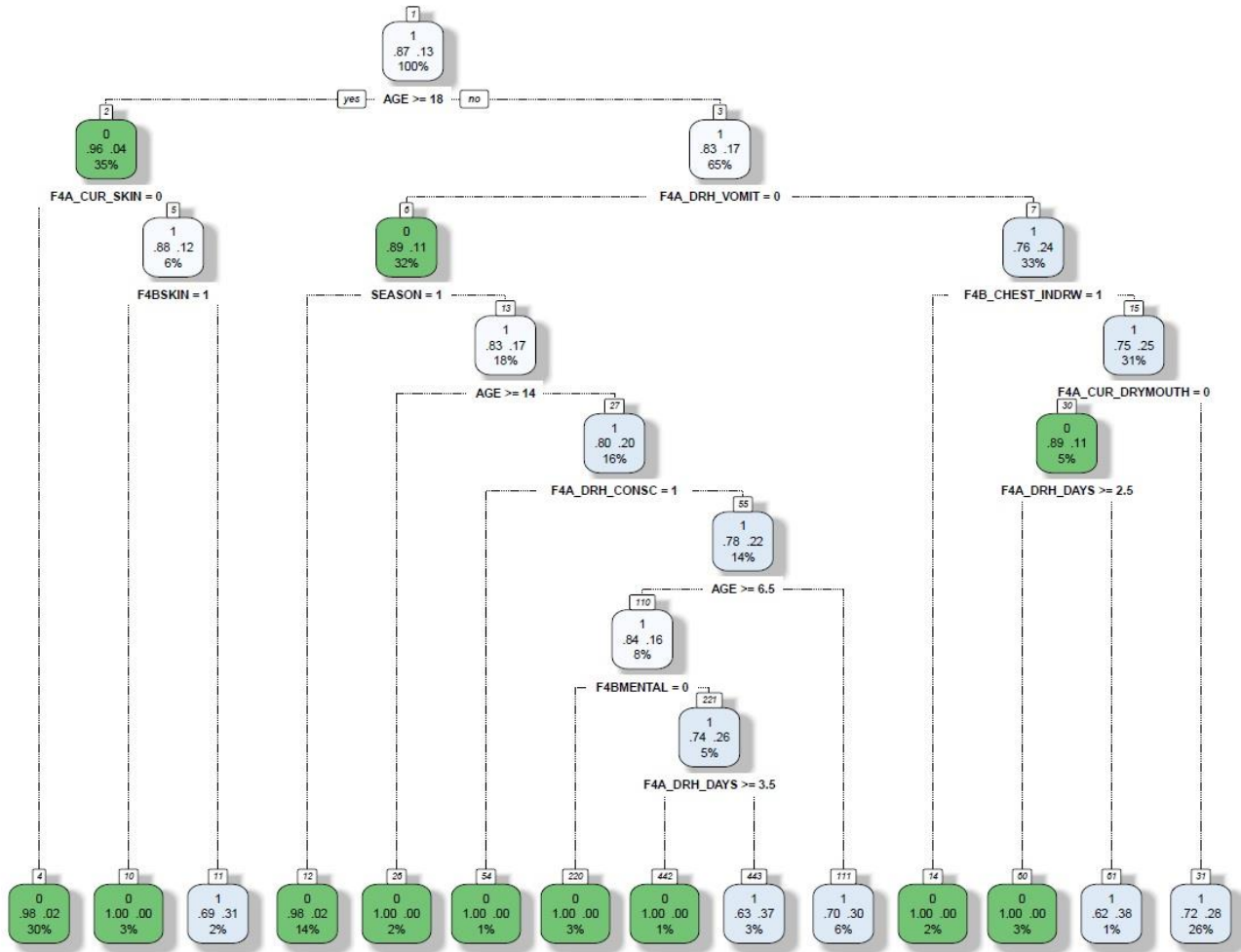
Figure 3.1B. Rotavirus classification tree using demographic and clinical profiles

Classification Tree: Rotavirus Status
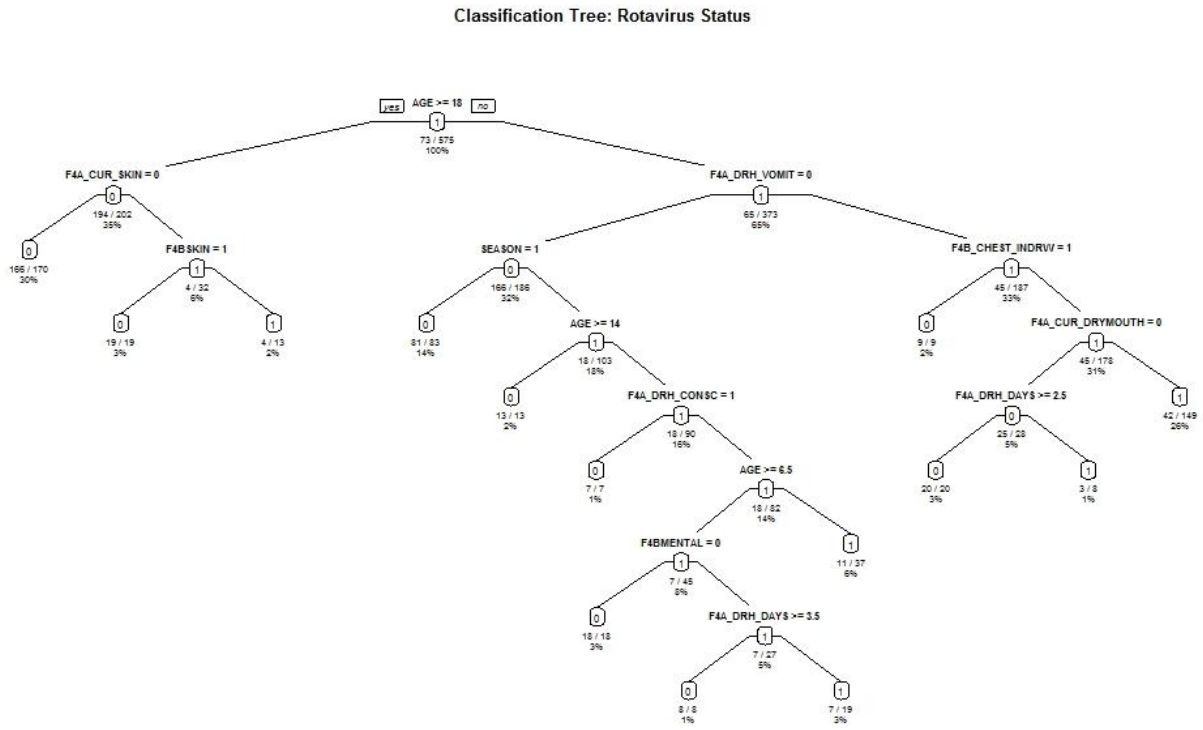
Figure 3.2. Comparison of classification tree performance on training (blue) and test (red) data.
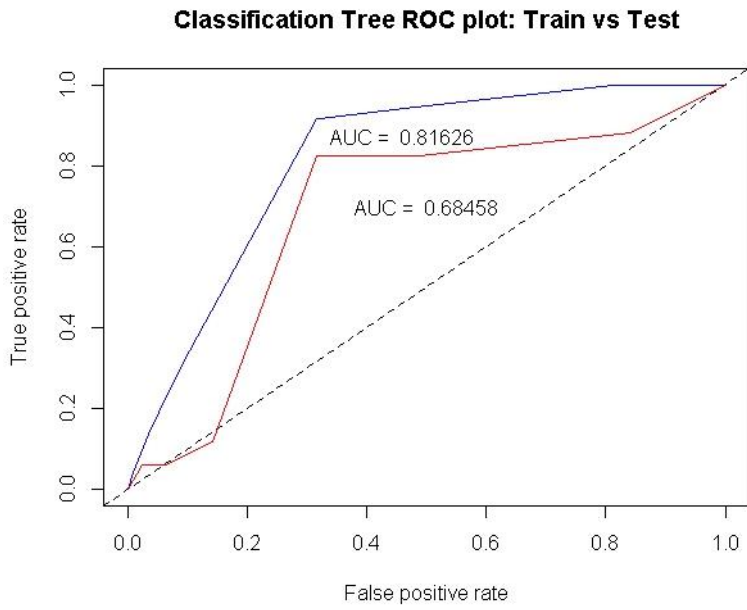
Figure 3.3. Unweighted conditional inference tree for rotavirus classification
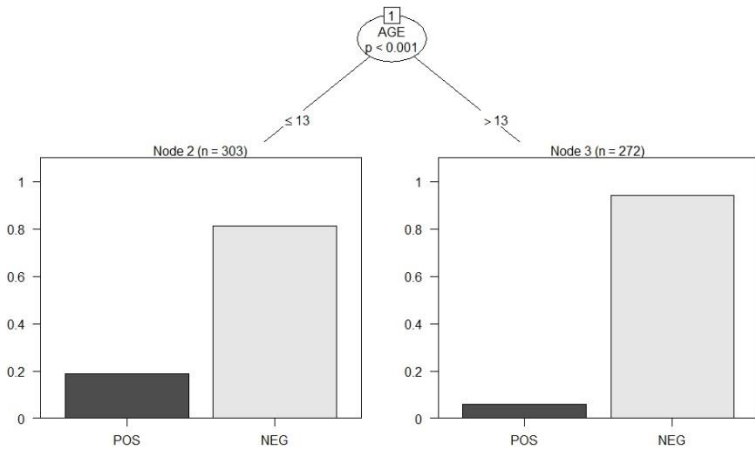
Figure 3.4. Weighted conditional inference tree for rotavirus classification
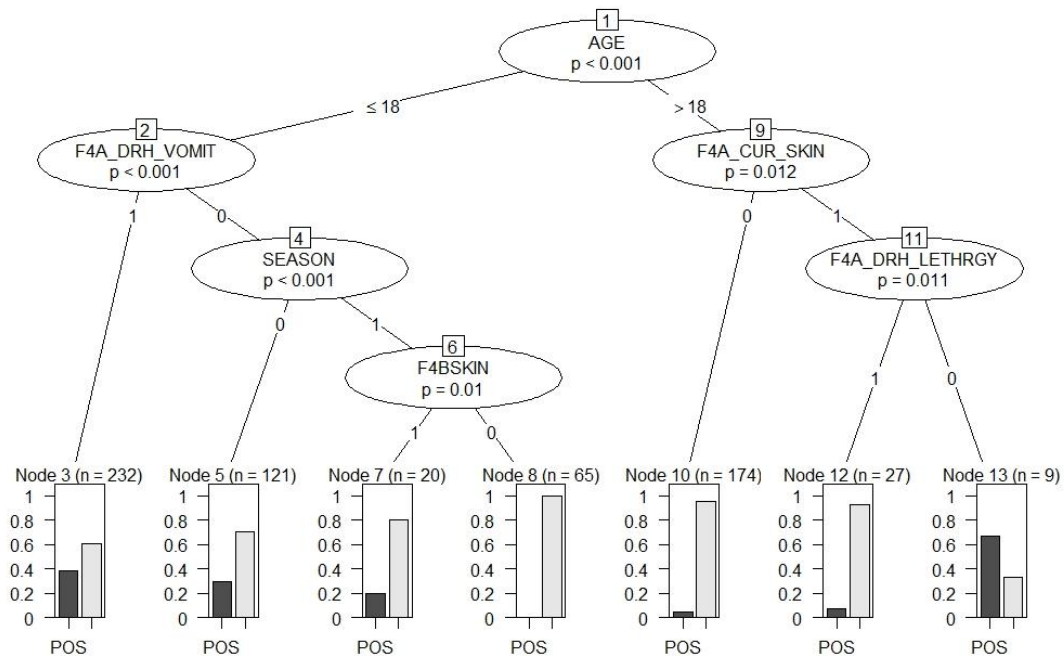
Figure 3.5. Comparison of conditional inference tree performance by weight

ROC curves comparing classification performance of ctree weighted vs unweighted
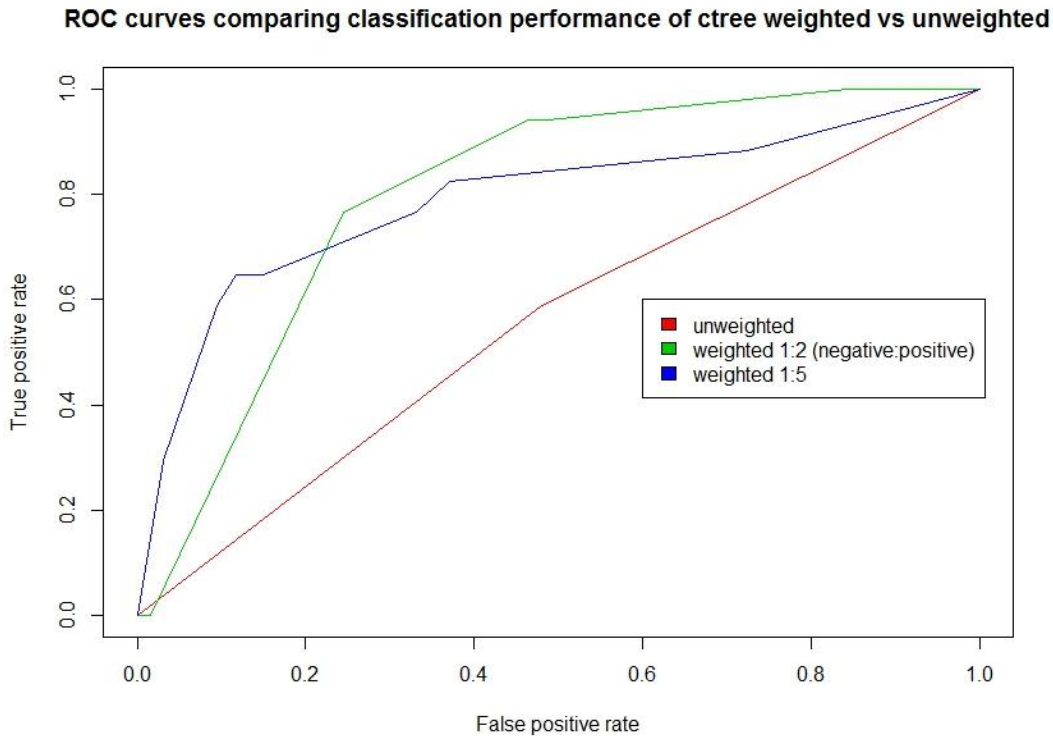
References

Alexander, K. A., & Blackburn, J. K. (2013). Overcoming barriers in evaluating outbreaks of diarrheal disease in resource poor settings: assessment of recurrent outbreaks in Chobe District, Botswana. BMC Public Health, 13, 775. doi:10.1186/1471-2458-13-775

Austin, P. C., Lee, D. S., Steyerberg, E. W., & Tu, J. V. (2012). Regression trees for predicting mortality in patients with cardiovascular disease: what improvement is achieved by using ensemble-based methods? Biom J, 54(5), 657-673. doi:10.1002/bimj.201100251

Berk. (2009). Classification and Regression Trees. Principles of Data Mining.

Breiman L, F. J., Olshen RA, Stone CJ. (1984). Classification and Regression Trees. Belmont, CA: Wadsworth, Inc.

Brooks JT, Ochieng JB, Kumar L, Okoth G, Shapiro RL, et al. (2006) Surveillance for bacterial diarrhea and antimicrobial resistance in rural western

Kenya, 1997–2003. Clin Infect Dis 43: 393–401.

Cieslak D, a. C. (2001). Learning Decision Trees for Unbalanced Data. In Springer (Ed.), Learning Decision Trees for Unbalanced Data (pp. pp 241-256).

Ekwochi, U., Chinawa, J. M., Obi, I., Obu, H. A., & Agwu, S. (2013). Use and/or misuse of antibiotics in management of diarrhea among children in Enugu, Southeast Nigeria. J Trop Pediatr, 59(4), 314-316. doi:10.1093/tropej/fmt016

Gasparinho, C., Mirante, M. C., Centeno-Lima, S., Istrate, C., Mayer, A. C., Tavira, L., . . . Brito, M. (2016). Etiology of Diarrhea in Children Younger Than 5 Years Attending the Bengo General Hospital in Angola. Pediatr Infect Dis J, 35(2), e28-34. doi:10.1097/inf.0000000000000957

Gebrekidan, A., Dejene, T. A., Kahsay, G., & Wasihun, A. G. (2015). Prevalence and antimicrobial susceptibility patterns of Shigella among acute diarrheal outpatients in Mekelle hospital, Northern Ethiopia. BMC Res Notes, 8, 611. doi:10.1186/s13104-015-1606-x

Green, S. T., Small, M. J., & Casman, E. A. (2009). Determinants of national diarrheal disease burden. Environ Sci Technol, 43(4), 993-999.

Hothorn. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework Journal of Computational and Graphical Statistics, 15(3), 651-674.

Japkowicz, N. (2001). The class imbalance problem: A systematic study Intelligent Data Analysis, 6(5), 429-449.

Jung, S. Y., Vitolins, M. Z., Fenton, J., Frazier-Wood, A. C., Hursting, S. D., & Chang, S. (2015). Risk profiles for weight gain among postmenopausal women: a classification and regression tree analysis approach. PLoS One, 10(3), e0121430. doi:10.1371/journal.pone.0121430

Justino, M. C., Brasil, P., Abreu, E., Miranda, Y., Mascarenhas, J. D., Guerra, S. F., & Linhares, A. C. (2016). Clinical Severity and Rotavirus Vaccination among Children Hospitalized for Acute Gastroenteritis in Belem, Northern Brazil. J Trop Pediatr. doi:10.1093/tropej/fmv098

Lewis, K. D., Dallas, M. J., Victor, J. C., Ciarlet, M., Mast, T. C., Ji, M., . . . Neuzil, K. M. (2012). Comparison of two clinical severity scoring systems in two multi-center, developing country rotavirus vaccine trials in Africa and Asia. Vaccine, 30 Suppl 1, A159-166. doi:10.1016/j.vaccine.2011.07.126

Mody, R. K., Gu, W., Griffin, P. M., Jones, T. F., Rounds, J., Shiferaw, B., . . . Hoekstra, R. M. (2015). Postdiarrheal hemolytic uremic syndrome in United States children: clinical spectrum and predictors of in-hospital death. J Pediatr, 166(4), 1022-1029. doi:10.1016/j.jpeds.2014.12.064

Spreng, C. P., Ojo, I. P., Burger, N. E., Sood, N., Peabody, J. W., & Demaria, L. M. (2014). Does stewardship make a difference in the quality of care? Evidence from clinics and pharmacies in Kenya and Ghana. Int J Qual Health Care, 26(4), 388-396. doi:10.1093/intqhc/mzu054

Tierney, N. J., Harden, F. A., Harden, M. J., & Mengersen, K. L. (2015). Using decision trees to understand structure in missing data. BMJ Open, 5(6), e007450. doi:10.1136/bmjopen-2014-007450

Van Hulst, A., Roy-Gagnon, M. H., Gauvin, L., Kestens, Y., Henderson, M., & Barnett, T. A. (2015). Identifying risk profiles for childhood obesity using recursive partitioning based on individual, familial, and neighborhood environment factors. Int J Behav Nutr Phys Act, 12, 17. doi:10.1186/s12966-015-0175-7

Varma, J. K., Katsitadze, G., Moiscrafishvili, M., Zardiashvili, T., Chokheli, M., Tarkhashvili, N., . . . Sobel, J. (2004). Signs and symptoms predictive of death in patients with foodborne botulism--Republic of Georgia, 1980-2002. Clin Infect Dis, 39(3), 357-362. doi:10.1086/422318

Walsh, P., Cunningham, P., Merchant, S., Walker, N., Heffner, J., Shanholtzer, L., & Rothenberg, S. J. (2015). Derivation of Candidate Clinical Decision Rules to Identify Infants at Risk for Central Apnea. Pediatrics, 136(5), e1228-1236. doi:10.1542/peds.2015-1825

Zang H, S. B. (2010). Recursive partitioning and applications. (2nd ed.). New York: Springer.

Zhu, Y., & Fang, J. (2016). Logistic Regression-Based Trichotomous Classification Tree and Its Application in Medical Diagnosis. Med Decis Making. doi:10.1177/0272989x15618658

Chapter 4. Identifying water, sanitation, and hygiene risk factors among children <5 years old with moderate-to-severe diarrhea in rural western Kenya, 2008-2011: using random forest methods

Ayers TL[1, 6][*], O'Reilly CE[1], Luo R[6], Omore R[2,3], Ochieng B[2,3], Farag TH[4], Nasrin D[4], Panchalingam S[4], Nataro JP[4], Kotloff KL[4], Levine MM[4], Oundo J[5], Parsons MB[1], Bopp C[1], Laserson K[2], Stauber CE[6], Breiman RF[7], Mintz E[1] , and Hoekstra RM[1]

Abstract

Objective: The use of predictive methods in the analysis of diarrheal disease research is limited. Risk factor analyses of diarrheal diseases are complicated because of numerous potential exposures that are often correlated. Newer analytic and machine learning techniques are available that are capable of handling a large numbers of variables and describing complex relationships. This study investigates the use of machine learning approaches, specifically random forest (RF) methods to identify water, sanitation, and hygiene (WASH) factors associated with diarrheal disease in children less than 5 years old. The newer approach of RF was compared to the traditional analytic approach of using logistic regression models for evaluating WASH risk factors for moderate-to-severe diarrhea (MSD).

Methods: The Global Enteric Multicenter Study (GEMS) was a prospective case-control study of children under the age of 5 with MSD from 2008 to 2012 in Kenya. Controls were matched to cases enrolled at sentinel health care facilities by age, sex, and nearby village. Cases and controls were extensively surveyed about WASH exposures and subsequently directly observed at a single follow-up visit. Both logistic regression and RF models were constructed considering over 50 WASH exposure variables collected. Models were run for the overall data and also stratified by age. RF models were then compared to logistic regression models using area receiver operating characteristic curve (AUC) statistics.

Results: There were 1,718 cases and 2,388 controls enrolled in Kenya with complete follow-up visits. Both logistic regression and RF models identified the importance of having a travel time of greater than an hour to a water source as a risk factor for MSD in all models. In addition, exposure to rodents was identified as a risk factor and having a large household size was protective for toddler and older children in both logistic and RF models. RF models identified a few different exposures, such as retaining the protective effect of household size for infant models, however model performance was similar between logistic regression and RF models.

Conclusions: In this methodological study, the RF classification approach performed similar to traditional logistic regression approaches for predicting risk factors for MSD. Future WASH research, in other settings, should consider this approach and whether it may highlight important exposure pathways previously overlooked. This study supports the use of RF analytic approaches as an alternative statistical approach for identifying WASH risk factors.

Introduction

Data mining techniques, such as random forest (RF), have predominantly been limited to the use of exploring high-throughput laboratory data (e.g.- molecular markers) and in clinical decision rules (e.g- scoring to predict outcome from surgical intervention) (Maier 2015, Kasthurirathne 2016,Walsh 2015, Forsberg 2015). The primary advantages of these newer analytic approaches is their ability to consider a large numbers of variables and ability to identify complex interactions. Despite these advantages, RF methods have yet to be considered as an epidemiological approach for identifying risk factors for diarrheal diseases.

One of the Millennium Development Goals is to reduce the mortality rate of children under five years of age by two-thirds. In order to achieve this goal improvements in water, sanitation, and hygiene (WASH) needs to occur. Identifying which interventions will have the greatest impact for diarrhea-related morbidity and mortality in children is difficult. Causal pathways for diarrhea in children <5 years of age are complex, largely in part because there is considerable number of potential environmental exposures associated with WASH and they are often inter-related.

Alternative modeling strategies for exploring exposures should be examined. Most epidemiological studies conducted to assess risk factors for diarrhea in children investigate WASH exposures using logistic regression models. Not only are these exposures often inter-related with one another, they are also nested. Several studies have attempted to disentangle the effects of water and sanitation (Gundry 2004, Eisenberg 2007, Fink 2011, and Fuller 2015). Some studies conclude that improved water has little to no impact if sanitation is not simultaneously improved while others describe that the water improvements have a substantial

impact alone but can be capitalized upon when sanitation improvements occur concurrently (Fuller 2015). It remains unclear whether these improvements are co-dependent or simply amplify the effects of each other.

Since environmental exposures are numerous, are part of transmission groupings and are often correlated, statistical methods that do not require absence of collinearity nor a linear relationship with the outcome should be explored. Identification and estimation of WASH risk factors could be considered a machine learning task. RF is one of the most attractive methods for classification problems because it can consider a combination of various classifiers to perform a classification problem jointly (Breiman 2010). RF methods provide a classification algorithm with advantages over logistic regression in that it can appropriately handle searching across a large number of binary variables that may be correlated, which is often the case for WASH variables. RF methods are a method that expands upon the classification tree method proposed by Breiman (Breiman 1998 and Breiman 2001). RF classification creates separate sub samples of the data. The results of these multiple classifiers are then assigned based on the majority vote. The data excluded from each sub-sample is used to compute the Out-Of-Bag (OOB) error rate.

The assessment of WASH risk factors have generally been constrained to logistic regression models in which a pre-selected small set of variables can be evaluated. Newer analytic techniques capable of assessing a large number of related variables may better predict MSD outcomes and facilitate prioritization of interventions. In this proof-of-concept study, a machine learning approach is evaluated for attributing illnesses to WASH risk factors.

Materials and Methods
Study design and data collection

The Global Enteric Multi-center Study (GEMS) was a prospective case-control study of moderate-to-severe diarrhea (MSD) conducted during 2008-2012. An MSD case was defined as a child with a diarrheal illness <7 days duration comprising ≥3 loose stools in 24 hrs and ≥1 of the following: sunken eyes, skin tenting, dysentery, required IV rehydration, or hospitalization. Controls were enrolled using the Demographic

Surveillance System (DSS) and were matched to cases by age, sex, and geographic proximity. Controls were

required to be free of diarrhea seven days prior to enrollment. At enrollment, both case and control caregivers

were surveyed about numerous WASH exposures including animal exposures. In addition, a single follow-up

home visit was made approximately 60 days later and additional WASH exposures were directly observed.

Study setting and population

The Kenya study site was located in rural western Kenya in the districts of Gem and Asembo in Siaya County

(formerly Nyanza Province) during January 31, 2008 and January 29, 2011. Subsequent to this timeframe, the

study was funded for an additional 11 months in the districts of Asembo and Karemo in Siaya County as the

Kenya Medical Research Institute (KEMRI)/CDC Kenya DSS moved to a new area during this time period.

Statistical analyses

Exposure variable frequencies and logistic regression models were performed using SAS 9.3

(Cary, NC). Exposures were explored using univariable logistic regression models using both conditional models,

for matched sets of cases and controls, and unconditional models (i.e. unmatched analysis). Univariable

estimates and 95% confidence intervals are reported. Since transmission routes are likely to differ by child

developmental stages (e.g. crawling vs not crawling, walking, etc.), three additional separate models were

performed for each age group. Age groups were defined as infants (0–11 months), toddlers (12–23 months),

and older children (24–59 months). Multivariable logistic regression models were developed by first selecting

variables demonstrating statistical significance in univariable logistic models and then performing backwards

elimination until all variables retained remained significant at $p \leq .05$. Gender was included in all unconditional,

multivariable logistic regression models since controls were enrolled based on this as a matching criteria. Self-

reported hygiene, especially among caregivers of severely ill children, were likely biased and were excluded from

all age stratified multivariable models (Cotzen 2015, Halder 2010, Manun'Ebo 1997 ,Curtis 1993, Stanton 1987).

An overall RF model was constructed using the 51 exposures and exposure groups to predict MSD case

or control status using R Statistical Software and the *randomForest* package. Similar to logistic models, self-

reported hygiene and sanitation variables were excluded and the reduced set of 41 exposures were considered

for all age stratified models. The number of variables randomly considered at each split (m) started with

approximately the square root of the number of independent variables under consideration (Katz, 2010).

Subsequently, other values of (m) the number of variables were considered if the OOB error was reduced by .10

or more, using the *tuneRF* function in *randomForest* package. To describe the relative importance of WASH

exposures, we computed variable of importance measures based on the mean decrease accuracy measure. The

RF models were then compared to the standard logistic regression models. Models were compared using area

under the receiver operating characteristic curve (AUC).

Results

There were 1,778 cases and 2,448 controls from the 4 year GEMS Kenya data set. Among those, 1,718

cases and 2,388 had completed follow-up surveys and were included in this analysis. Demographic and

exposure frequencies among cases and controls are presented in Supplemental Table 1. Estimates from

univariable logistic regression for both unconditional and conditional models remained similar between models,

with the exception of age, one of the matching variables. In the overall data, not stratified by age, having a large

number of household members and having finished floors were identified as having significantly protective

effect for MSD in logistic regression. Risk factors for MSD among WASH and animal exposures were identified as

the following: presence of rodents around the home, using an unimproved water source, having a travel time of

greater than an hour to a water source, and having observed feces near home or in yard in both logistic

regression and RF models. Self-reported hand washing behavior at key times demonstrated statistically

significant protective effects and were subsequently excluded for age stratified risk factor models. (S4.1).

Age stratified multivariable logistic regression models

For age stratified logistic regression models, the largest effect estimate for all three age group models

was having a travel time to water source of greater than an hour. For the infant models travel time to water

source (OR 2.69, CI 1.73 -4.18) and feces visible in house or yard (OR 1.65, CI 1.15-2.37) had the largest effect size estimates (S2). However, having an unimproved water source explained the largest number of infant cases. Toddler and children models contained the only risky animal exposure of rodents (OR 1.56 CI 1.22-1.99 for infants; OR 1.33, CI 1.03-1.72)(S4.3 and S4.4).

Overall RF results

In the overall RF model which was developed using all potential predictors including age, the variables of importance were identified and plotted (Figure 4.1). The variable with the highest importance measure was identified as the self-reported measures of not washing hands after defecation and disposing of children's feces in open. Self-reported hygiene and sanitation variable were excluded. The overall RF model generated using the reduced number of exposure variables resulted in same the same error rate. The most important variable, based on mean decrease in accuracy, for classification of MSD was having a travel time of greater than hour to water source, followed by the protective effect of having a large household size (Figure 4.2).

Age stratified RF analysis

In age stratified RF models, 800 infant, 474 toddler, and 444 older children cases were analyzed separately. In all three age specific models, the number of variables considered at each node that produced the lowest OOB error was 3. Similar to the overall model, having a travel time greater than an hour to water source followed by the protective effect of large household size, were the most important variables for predicting infant MSD cases. Important factors for classifying toddler and older children MSD cases were the child being given untreated water and access to sanitation facilities. Rodent exposure was an important predictor of MSD classification in the toddler model (Figure 4.3, 4.4, 4.5).

Comparison of RF and logistic regression models

All RF models predicted MSD cases with an AUC greater than 0.54, with the highest AUC of .61 in the toddler age group. Similarly, logistic regression models had high AUC estimates for toddler and older children specific models with .67 and .69. Only the older children specific logistic regression model was significantly outperformed the RF model (Table 4.5). In all models, logistic regression models selected similar exposures to the RF models. The RF models tended to rank 1-2 additional variables that lacked statistical significance in logistic regression in the top five. For example, the second most important exposure identified by RF in the infant model was having a large household size but lacked statistical significance in logistic regression. Having a large household size was selected in top five exposures for predicting MSD in all three age specific RF models, however was only retained in toddler and older children models.

Discussion

We found the results from RF models did not differ greatly from logistic regression models when comparing in feature selection or model performance. Both identified exposures that tended to have large effect sizes regardless of proportion of cases explained. For example, both highlighted the importance of the time it takes to reach the water source and MSD in infants in this study corroborating other studies that have also independently explored the association between travel time to water source and diarrheal disease rural western Kenya (Nygren 2016). Identifying factors in both methods supports the importance of their role in MSD. While there were many similarities between the variables highlighted in both logistic regression and RF, a primary difference was the emphasis that RF placed on the protective demographic features such household size in both toddler and older children models. In addition, having more than one child under 5 years of age was a significant protective effect in older children logistic models and was the second most important variable for RF models. In this study, RF results did not differ from logistic regression models in terms of conclusions however, in studies from other geographic regions with less homogeneity in exposures would likely results in differences between models.

This study presented a proof -of-concept analysis of the use of RF methods for identifying risk factors for MSD, but has several limitations. First, this study was limited by the exposures captured and their frequencies. Many of the exposures of interest had relatively low frequencies which reduced the statistical power in logistic regression and potentially the data space for random forest. RF and logistic regression models performed equally, but none of the models would be considered robust with AUCs below .70 for all models. RF models error rates ranged from 32-38%, thus this data might be missing important features necessary to improve classification of MSD. Secondly, many of the exposures of interest were collected based on self-reports which are known to be potentially biased (Halder 2010). In our study, cases of MSD were enrolled at sentinel health facilities when presenting for care and care givers were asked about many water and hygiene practices. Conversely controls were asked these self-reported measures during a home visit, thus reporting bias might be lower. While we attempted to address this by excluding these variables in our models, it limited our ability to identify the contribution and importance of hygiene. Finally, a limitation specific to RF, is that variable importance and inferences about variables are not as easy to interpret as estimated Odds ratios produced from logistic regression models. Machine learning techniques, such as RF, do currently suffer some interpretability issues however as their use increases in epidemiology additional methods for converting in population attributable fractions may become more widely used and understood (Gu 2015).

Despite these limitations, this study demonstrates the usefulness of applying a machine learning approach to WASH studies. RF performed as well as logistic regression for predicting MSD cases and should be considered as an alternative for variable selection is studies with a large number of exposures to explore and with potentially complex relationships including interactions. The methods employed here serve as an example that could be utilized for WASH studies in other geographic settings. In our analysis, logistic regression and RF models produced similar conclusions supporting the notion that RF methods may be an appropriate method for assessing WASH risk factors. Future analyses should evaluate the ability of these methods to highlight important exposures missed by traditional algorithms.

Figure 4.1. Variables ranked by mean decrease accuracy for all age groups

Variable Importance for top 20 predictors
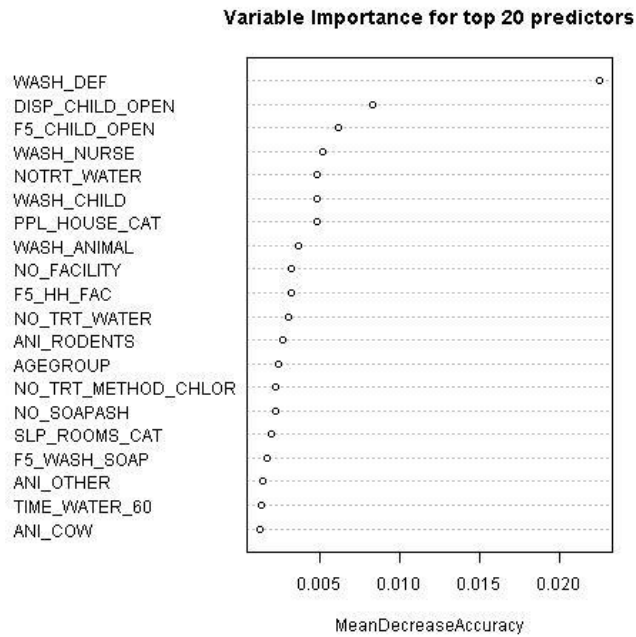
Figure 4.2. Variables ranked by mean decrease accuracy for all age groups, excluding self-reported hygiene exposures
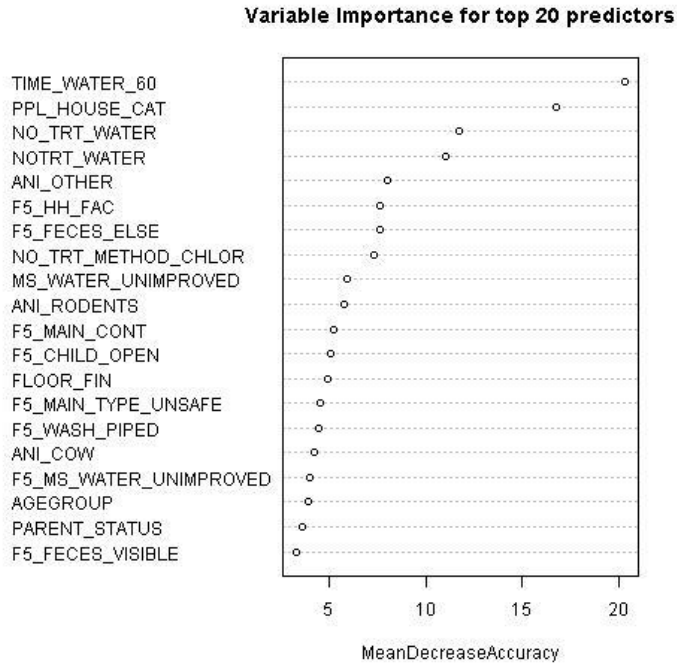
Figure 4.3 Random forest variable importance plots by age group excluding self-reported hygiene exposures

Variable Importance for Infants
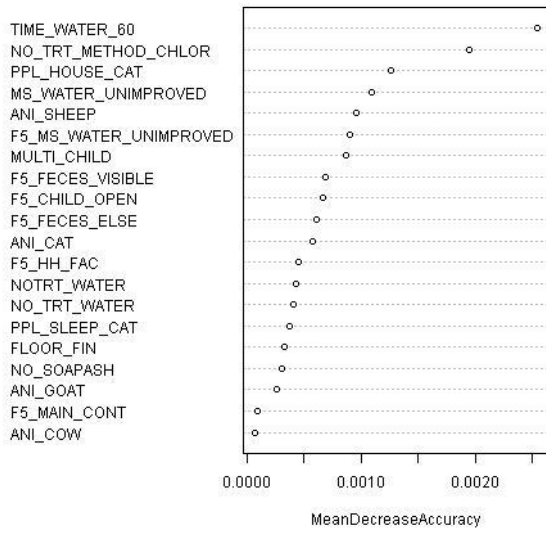
Variable Importance for Toddlers
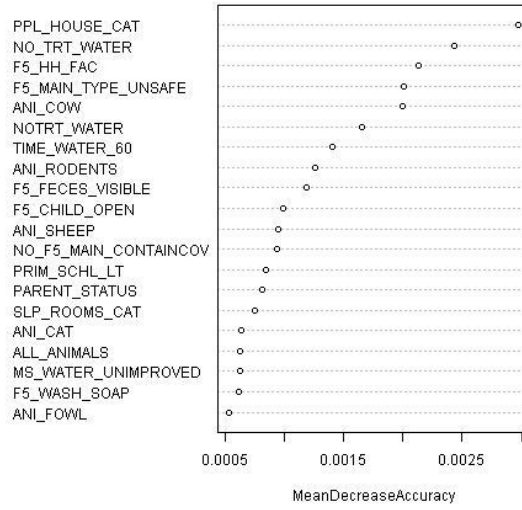
Variable Importance for Children
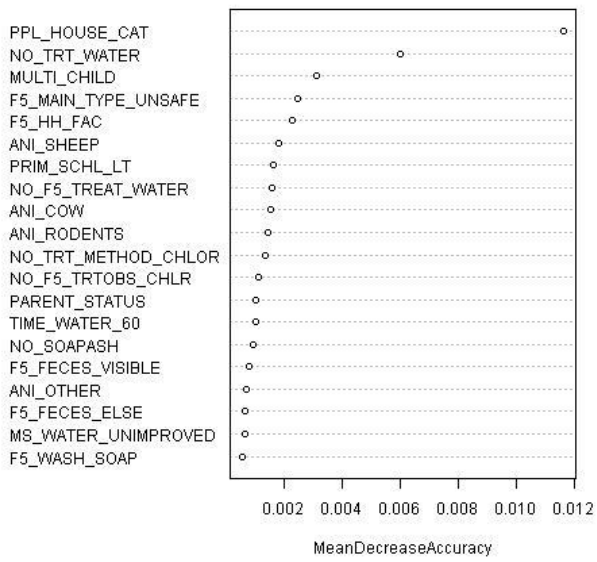
Figure 4.4 Comparison of odds ratio estimates from logistic regression and variable importance measures from random forest
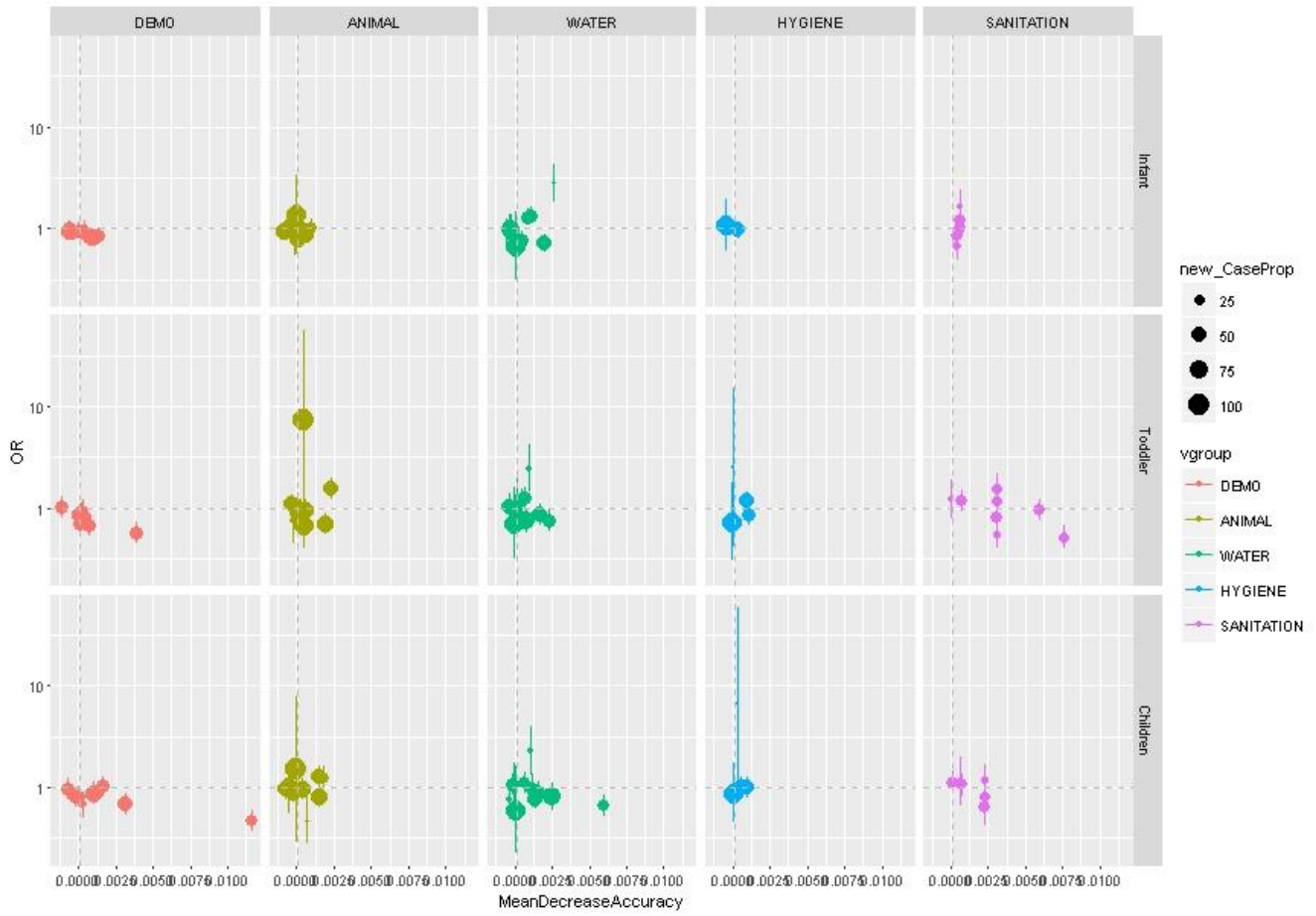
Table 4.5 Comparing model performance between logistic regression and random forest

Model	Logistic AUC	(95% CI)	RF AUC	(95% CI)
Infant models	0.59	(.57 -.62)	0.54	(.51-.57)
Toddler models	0.67	(.64 -.70)	0.61	(.58 -.64)
Older children models	0.69	(.66 - .72)	0.59	(.56 -.63)

References

Breiman L, F. J., Olshen RA, Stone CJ. (1984). Classification and Regression Trees. Belmont, CA: Wadsworth, Inc.

Contzen, N., De Pasquale, S., & Mosler, H. J. (2015). Over-Reporting in Handwashing Self-Reports: Potential Explanatory Factors and Alternative Measurements. PLoS One, 10(8), e0136445. doi:10.1371/journal.pone.0136445

Curtis VA, C. S., Mertens T, Traore E, Kanki B, Diallo I. . (1993). Structured observations of hygiene

behaviours in Burkina Faso: Validity, variability, and utility. .B World Health Organ., 71(1), 23-32.

Eisenberg, J. N., Scott, J. C., & Porco, T. (2007). Integrating disease control strategies: balancing water sanitation and hygiene interventions to reduce diarrheal disease burden. Am J Public Health, 97(5), 846-852. doi:10.2105/ajph.2006.086207

Fink, G., Gunther, I., & Hill, K. (2011). The effect of water and sanitation on child health: evidence from the demographic and health surveys 1986-2007. Int J Epidemiol, 40(5), 1196-1204. doi:10.1093/ije/dyr102

Forsberg, J. A., Potter, B. K., Wagner, M. B., Vickers, A., Dente, C. J., Kirk, A. D., & Elster, E. A. (2015). Lessons of War: Turning Data Into Decisions. EBioMedicine, 2(9), 1235-1242. doi:10.1016/j.ebiom.2015.07.022

Fuller, J. A., Westphal, J. A., Kenney, B., & Eisenberg, J. N. (2015). The joint effects of water and sanitation on diarrhoeal disease: a multicountry analysis of the Demographic and Health Surveys. Trop Med Int Health, 20(3), 284-292. doi:10.1111/tmi.12441

Gu, W., Vieira, A. R., Hoekstra, R. M., Griffin, P. M., & Cole, D. (2015). Use of random forest to estimate population attributable fractions from a case-control study of Salmonella enterica serotype Enteritidis infections. Epidemiol Infect, 143(13), 2786-2794. doi:10.1017/s095026881500014x

Gundry, S., Wright, J., & Conroy, R. (2004). A systematic review of the health outcomes related to household water quality in developing countries. J Water Health, 2(1), 1-13.

Halder, A. K., Tronchet, C., Akhter, S., Bhuiya, A., Johnston, R., & Luby, S. P. (2010). Observed hand cleanliness and other measures of handwashing behavior in rural Bangladesh. BMC Public Health, 10, 545. doi:10.1186/1471-2458-10-545

Kasthurirathne, S. N., Dixon, B. E., Gichoya, J., Xu, H., Xia, Y., Mamlin, B., & Grannis, S. J. (2016). Toward better public health reporting using existing off the shelf approaches: A comparison of alternative cancer detection approaches using plaintext medical data and non-dictionary based feature selection. J Biomed Inform. doi:10.1016/j.jbi.2016.01.008

Maier, O., Schroder, C., Forkert, N. D., Martinetz, T., & Handels, H. (2015). Classifiers for Ischemic Stroke Lesion Segmentation: A Comparison Study. PLoS One, 10(12), e0145118. doi:10.1371/journal.pone.0145118

Manun'Ebo, M., Cousens, S., Haggerty, P., Kalengaie, M., Ashworth, A., & Kirkwood, B. (1997). Measuring hygiene practices: a comparison of questionnaires with direct observations in rural Zaire. Trop Med Int Health, 2(11), 1015-1021.

Nygren, B. L., O'Reilly, C. E., Rajasingham, A., Omore, R., Ombok, M., Awuor, A. O., . . . Mintz, E. D. (2016). The Relationship Between Distance to Water Source and Moderate-to-Severe Diarrhea in the Global Enterics Multi-Center Study in Kenya, 2008-2011. Am J Trop Med Hyg. doi:10.4269/ajtmh.15-0393

O'Reilly, C. E., Jaron, P., Ochieng, B., Nyaguara, A., Tate, J. E., Parsons, M. B., . . . Mintz, E. (2012). Risk factors for death among children less than 5 years old hospitalized with diarrhea in rural western Kenya, 2005-2007: a cohort study. PLoS Med, 9(7), e1001256. doi:10.1371/journal.pmed.1001256

Papathomas, M., Molitor, J., Richardson, S., Riboli, E., & Vineis, P. (2011). Examining the joint effect of multiple risk factors using exposure risk profiles: lung cancer in nonsmokers. Environ Health Perspect, 119(1), 84-91. doi:10.1289/ehp.1002118

Stanton, B. F., Clemens, J. D., Aziz, K. M., & Rahman, M. (1987). Twenty-four-hour recall, knowledge-attitude-practice questionnaires, and direct observations of sanitary practices: a comparative study. Bull World Health Organ, 65(2), 217-222.

Walsh, P., Cunningham, P., Merchant, S., Walker, N., Heffner, J., Shanholtzer, L., & Rothenberg, S. J. (2015). Derivation of Candidate Clinical Decision Rules to Identify Infants at Risk for Central Apnea. Pediatrics, 136(5), e1228-1236. doi:10.1542/peds.2015-1825

Chapter 5. Integrated discussion

The first paper examined the impact of model selection strategies on the description and variability of diarrheal etiologies associated with MSD. The study used LASSO methods and demonstrated consistency in large signal detection across model selection methods, while identifying some variability in less frequently occurring two-way interactions. The second study addressed the feasibility of assessing etiologies in the absence of laboratory diagnostics. While an extensive panel of laboratory diagnostics were supported during the active study period, they are not sustainable nor are they available for point-of-care treatment in most areas of Africa. The ability to differentiate enteric viruses from bacterial causes at the clinical setting, in the absence of laboratory diagnostics, is imperative for judicious use of antibiotics and in reducing antimicrobial resistance as a result of over prescribing. Classification trees using clinical profiles for identifying rotavirus infections were developed. Both the recursive partitioning classification tree and a conditional inference tree highlighted that the at risk sub-population of children less than 18 months of age with rotavirus are likely to predominantly present for care with vomiting during warm-dry months. The rotavirus classification trees presented a useful algorithm for understanding the data structure and identifying high-risk groups among correlated clinical features. Finally, the third study explored methods for investigating the numerous and complex environmental exposures associated with enteric disease transmission using random forest techniques. Random forest methods offered a non-parametric alternative to logistic regression and highlighted the association of distance to water source and household member size with MSD for toddlers and older children.

A principal finding in all three studies was that machine learning methodological approaches, such as LASSO, classification trees, and random forest, are useful and feasible to implement in epidemiological studies. All three approaches used in this dissertation provided additional information and understanding of the data beyond using a singular logistic regression model. The results from all three machine learning approaches were supported by comparable logistic regression results indicating their usefulness as epidemiological tools. The limitations of feature selection and analysis of data in situations of limited sample size, large number of variables

to investigate, correlated predictors, and complex interactions are present in nearly all areas of epidemiological research, not just diarrheal diseases in developing countries. This dissertation offers an exploration of methodological alternatives that should be used more frequently in diarrheal disease epidemiology, and in public health in general.

Supplemental Table 2.1. All pathogens tested and frequencies among cases and controls

Pathogen variable	Cases N=1778 (%)	Controls N=2448 (%)
Pathogens -Bacterial		
Enterotoxigenic *E. coli* any ST[a]	176 (9.9)	103 (4.2)
Enterotoxigenic *E. coli* LT only[a]	103 (5.8)	142 (5.8)
Enteroaggregative *E. coli*	268 (15.1)	400 (16.3)
Typical Enteropathogenic E. *coli*	135 (7.6)	120 (4.9)
Atypical Enteropathogenic E. *coli*	97 (5.5)	157 (6.4)
Enterohaemorrhagic E. *coli* (EHEC)	0	0
Shigella spp.	130 (7.3)	55 (2.3)
Aeromonas	1 (.06)	4 (0.2)
Vibrio cholerae O1	7 (.39)	0
Vibrio O139	0	0
Salmonella Typhi	0	0
Non-typhoidal *Salmonella*	95 (5.3)	84 (3.4)
Campylobacter	253 (14.2)	327 (13.4)
Pathogens -Viral		
Rotavirus	253 (14.2)	49 (2.0)
Norovirus -GI	54 (3.0)	97 (4.0)
Norovirus -GII	88 (5.0)	100 (4.1)
Adenovirus non-40/41	45 (2.5)	58 (2.4)
Adenovirus 40/41	40 (2.3)	23 (0.9)
Astrovirus	31 (1.5)	36 (1.5)
Sapovirus	56 (3.2)	73 (3.0)
Pathogens -Parasitic		
Cryptosporidium	195 (11.0)	104 (4.3)
Giardia	329 (18.5)	578 (23.6)
Entamoeba histolytica	16 (0.9)	8 (0.3)

[a] ST (heat stable toxin) and LT (heat labile toxin)

Table 2.2 Comparison of selected pathogens by model selection method, infants (0-11 months old)

Variable	Cases n=829 (%)	Controls n= 896 (%)	Conditional BE α =.05	Conditional LASSO λ max (1SE)	Unconditional BE α =.05	Unconditional Hierarchical group LASSO λ max(1SE)	Unconditional Hierarchical group LASSO λ min(1SE)	Num of methods including variable	Num of methods variable sig.
Pathogens -Bacterial									
Enterotoxigenic *E. coli* any ST	79 (9.5)	37 (4.1)	2.75 (1.72 - 4.39)	*	2.86 (1.88 - 4.37)	2.87 (1.88 - 4.38)	2.79 (1.82 - 4.27)	5	5
and Adenovirus 40/41 Present				0.11 (0.01 - 2.55)					
and Adenovirus 40/41 Present				2.9 (1.8 - 4.69)					
Enterotoxigenic *E. coli* LT only	47 (5.7)	66 (7.4)					0.72 (0.46 - 1.11)	1	0
Enteroaggregative *E. coli*	176 (21.2)	221 (24.7)		**			*	2	1
and Norovirus -GII Present				2.41 (0.94 - 6.16)			2.23 (0.93 - 5.33)		
and Norovirus -GII Absent				0.76 (0.57 - 1.01)			0.78 (0.6 - 1.02)		
Typical EPEC	88 (10.6)	61 (6.8)	*	2.02 (1.31 - 3.11)	1.84 (1.28 - 2.65)	1.82 (1.26 - 2.61)	**	5	5
and *Shigella* spp. Present			0.24 (0.03 - 2.02)						
and *Shigella* spp. Absent			2.1 (1.36 - 3.22)						
and Non-typhoidal *Salmonella* Present							NA		
and Non-typhoidal *Salmonella* Absent							1.95 (1.34 - 2.83)		
Atypical EPEC	48 (5.8)	48 (5.4)					1.16 (0.74 - 1.81)	1	0
Shigella spp.	42 (5.1)	9 (1)	*	7.36 (3.19 - 16.98)	7.04 (3.37 - 14.71)	7.24 (3.47 - 15.13)	7.69 (3.66 - 16.14)	5	5
and typical EPEC Present			1.12 (0.17 - 7.45)						
and typical EPEC Absent			9.62 (3.7 - 25.05)						
Non-typhoidal *Salmonella*	43 (5.2)	44 (4.9)				**		1	1
and typical EPEC Present						NA			
and typical EPEC Absent						1.52 (0.95 - 2.44)			
Campylobacter jejuni	98 (11.8)	94 (10.5)		1.22 (0.85 - 1.75)			1.18 (0.85 - 1.64)	2	0
Campylobacter coli	33 (4)	45 (5)					0.78 (0.47 - 1.29)	1	0
Pathogens -Viral									
Rotavirus	159 (19.2)	19 (2.1)	16.97 (9.06 - 31.79)	17.43 (9.26 - 32.81)	*	13.78 (8.43 - 22.53)	14.48 (8.83 - 23.74)	5	5
and Norovirus -GII Present					2.34 (0.43 - 12.91)				
and Norovirus -GII Absent					15.14 (9.03 - 25.38)				
Norovirus -GI	30 (3.6)	33 (3.7)					1.38 (0.81 - 2.36)	1	0
Norovirus -GII	63 (7.6)	50 (5.6)	1.56 (1 - 2.44)	**	*		*	4	4
and rotavirus Present					0.26 (0.05 - 1.46)				
and rotavirus Absent					1.66 (1.1 - 2.51)				
and Enteroaggregative E. coli. Present				3.63 (1.6 - 8.25)			3.33 (1.55 - 7.15)		
and Enteroaggregative E. coli Absent				1.14 (0.66 - 1.96)			1.16 (0.71 - 1.9)		
Adenovirus non-40/41	18 (2.2)	14 (1.6)				1.59 (0.76 - 3.34)	1.51 (0.71 - 3.19)	2	0
Adenovirus 40/41	25 (3)	12 (1.3)	3.35 (1.53 - 7.34)	*	3.03 (1.48 - 6.19)	3.06 (1.5 - 6.25)	3.02 (1.48 - 6.19)	5	5
and Enterotoxigenic E. coli any ST Present				0.15 (0.01 - 3.24)					
and Enterotoxigenic E. coli any ST Absent				3.93 (1.74 - 8.89)					
Astrovirus	10 (1.2)	5 (0.6)				2.57 (0.85 - 7.83)	2.41 (0.78 - 7.41)	2	0
Sapovirus	19 (2.3)	16 (1.8)				1.64 (0.81 - 3.31)	1.76 (0.86 - 3.57)	2	0
Pathogens -Parasitic									
Cryptosporidium	119 (14.4)	52 (5.8)	3.28 (2.22 - 4.84)	3.25 (2.2 - 4.79)	3.52 (2.48 - 5)	3.49 (2.46 - 4.96)	3.52 (2.47 - 5.01)	5	5
Giardia	75 (9)	101 (11.3)	*excluded*	*excluded*	*excluded*	*excluded*	*excluded*		
Entamoeba histolytica	6 (0.7)	5 (0.6)						0	0
Number of main effects in model			5	6	5	8	13		
Number of interaction terms in model			1	2	1	0	2		
Number of significant variables			6	6	6	5	6		

Table 2.3 Comparison of selected pathogens by model selection method, toddlers (12-23 months old)

Variable	Cases n=491 (%)	Controls n=808 (%)	Conditional BE α =.05	Conditional LASSO λ max (1SE)	Unconditional BE α =.05	Hierarchical group LASSO λ max(1SE)	Hierarchical group LASSO λ min(1SE)	Num of methods including variable	Num of methods variable sig.
Pathogens -Bacterial									
Enterotoxigenic *E. coli* any ST	58 (11.8)	41 (5.1)	3.11 (1.94 - 5.01)	3 (1.87 - 4.82)	2.77 (1.79 - 4.26)	2.77 (1.79 - 4.26)	2.82 (1.82 - 4.38)	5	5
Enterotoxigenic *E. coli* LT only	33 (6.7)	48 (5.9)					1.35 (0.83 - 2.21)	1	0
Enteroaggregative *E. coli*	57 (11.6)	107 (13.2)						0	0
Typical EPEC	30 (6.1)	39 (4.8)					1.48 (0.89 - 2.48)	1	0
Atypical EPEC	27 (5.5)	56 (6.9)		0.68 (0.4 - 1.15)			0.73 (0.44 - 1.21)	1	0
Shigella spp.	35 (7.1)	26 (3.2)	3.04 (1.75 - 5.29)	2.91 (1.68 - 5.05)	2.98 (1.75 - 5.06)	2.98 (1.75 - 5.06)	3.1 (1.81 - 5.29)	5	5
Non-typhoidal *Salmonella*	27 (5.5)	24 (3)	2.51 (1.35 - 4.64)	2.29 (1.24 - 4.23)	2.34 (1.31 - 4.17)	2.34 (1.31 - 4.17)	2.32 (1.29 - 4.15)	5	4
Campylobacter jejuni (363)	43 (8.8)	53 (6.6)	1.73 (1.08 - 2.79)	1.72 (1.07 - 2.77)			1.4 (0.9 - 2.17)	3	2
Campylobacter coli (216)	28 (5.7)	60 (7.4)					0.75 (0.46 - 1.23)	1	0
Pathogens -Viral									
Rotavirus	75 (15.3)	21 (2.6)	9.08 (5.12 - 16.1)	9.27 (5.23 - 16.42)	7.67 (4.63 - 12.72)	7.67 (4.63 - 12.72)	7.99 (4.81 - 13.3)	5	5
Norovirus -GI	13 (2.6)	28 (3.5)						0	0
Norovirus -GII	17 (3.5)	32 (4)						1	0
Adenovirus non-40/41	16 (3.3)	30 (3.7)						0	0
Adenovirus 40/41	11 (2.2)	9 (1.1)					1.82 (0.7 - 4.74)	1	0
Astrovirus	11 (2.2)	15 (1.9)					1.69 (0.75 - 3.77)	1	0
Sapovirus	20 (4.1)	34 (4.2)						0	0
Pathogens -Parasitic									
Cryptosporidium	54 (11)	39 (4.8)	3.33 (2.01 - 5.49)	3.37 (2.04 - 5.56)	2.82 (1.81 - 4.38)	2.82 (1.81 - 4.38)	2.91 (1.86 - 4.54)	5	5
Giardia	115 (23.4)	218 (27)						0	0
Entamoeba histolytica	6 (1.2)	2 (0.2)	9.27 (1.54 - 55.96)		6.44 (1.26 - 32.9)	6.44 (1.26 - 32.9)	6.52 (1.27 - 33.42)	4	4
Number of main effects in model			7	7	6	6	13		
Number of interactions in model			0	0	0	0	0		
Number of significant variables			7	6	6	6	6		

Table 2.4 Comparison of selected pathogens by model selection method, young children (24-59 months old)

Variable	Cases	Controls	Conditional		Unconditional			Num of methods including variable	Num of methods variable sig.
			BE	LASSO	BE	Hierarchical group LASSO			
	n=458 (%)	n=744(%)	α =.05	λ max (1SE)	α =.05	λ max(1SE)	λ min(1SE)		
Pathogens -Bacterial									
Enterotoxigenic *E. coli* any ST	39 (8.5)	25 (3.4)	3.06 (1.75 - 5.33)	2.93 (1.67 - 5.16)	3.15 (1.86 - 5.34)	3.21 (1.89 - 5.44)	3.21 (1.89 - 5.44)	5	5
Enterotoxigenic *E. coli* LT only	23 (5)	28 (3.8)						0	0
Enteroaggregative *E. coli*	35 (7.6)	72 (9.7)		0.75 (0.47 - 1.22)				1	0
Typical EPEC	17 (3.7)	20 (2.7)						0	0
Atypical EPEC	22 (4.8)	53 (7.1)		0.65 (0.37 - 1.14)				1	0
Shigella spp.	53 (11.6)	20 (2.7)	6.84 (3.65 - 12.84)	6.95 (3.68 - 13.1)	5.26 (3.08 - 8.98)	5.24 (3.07 - 8.95)	5.24 (3.07 - 8.95)	5	5
Non-typhoidal *Salmonella*	25 (5.5)	16 (2.2)	3.18 (1.64 - 6.14)	3.24 (1.67 - 6.29)	3.19 (1.67 - 6.07)	3.04 (1.59 - 5.83)	3.04 (1.59 - 5.83)	5	5
Campylobacter jejuni	30 (6.6)	45 (6)						0	0
Campylobacter coli	20 (4.4)	30 (4)						0	0
Pathogens -Viral									
Rotavirus	19 (4.1)	9 (1.2)	4.48 (1.84 - 10.89)	4.45 (1.84 - 10.76)	4.28 (1.89 - 9.7)	4.36 (1.92 - 9.9)	4.36 (1.92 - 9.9)	5	5
Norovirus -GI	11 (2.4)	36 (4.8)		0.48 (0.22 - 1.04)	0.45 (0.22 - 0.92)	0.41 (0.2 - 0.86)	0.41 (0.2 - 0.86)	4	0
Norovirus -GII	8 (1.7)	18 (2.4)						0	0
Adenovirus non-40/41	11 (2.4)	14 (1.9)						0	0
Adenovirus 40/41	4 (0.9)	2 (0.3)				4.28 (0.74 - 24.93)	4.28 (0.74 - 24.93)	2	0
Astrovirus	10 (2.2)	16 (2.2)						0	0
Sapovirus	17 (3.7)	23 (3.1)						0	0
Pathogens -Parasitic									
Cryptosporidium	22 (4.8)	13 (1.7)	2.82 (1.36 - 5.87)	2.95 (1.41 - 6.17)	2.77 (1.35 - 5.65)	2.68 (1.31 - 5.5)	2.68 (1.31 - 5.5)	5	5
Giardia	139 (30.3)	259 (34.8)						0	0
Entamoeba histolytica	4 (0.9)	1 (0.1)				6 (0.56 - 64.25)	6 (0.56 - 64.25)	2	0
Number of main effects in model			5	8	6	8	8		
Number of interactions in model			0	0	0	0	0		
Number of significant variables			5	5	6	4	4		

Supplementary Table 3.1 Comparing clinical features of rotavirus positive and rotavirus negative moderate-to-severe diarrhea cases at enrollment

Characteristic	Rotavirus-positive N=163		Rotavirus -negative N=554		OR (95% CI)	P-value
	n	%	n	%		
Demographic:						
Median age in months	8	(1-46)	12	(1- 59)		**<0.001***
Female sex	80	49%	232	42%	1.34 (.94 - 1.90)	0.10
Cool-wet month	41	25%	233	42%	**0.46 (0.31 - 0.69)**	**<0.001**
Self reported symptoms at enrollment:						
Median days of diarrhea	3	(1-7)	3	(1-7)		0.44*
Stool type:						
-watery	104	64%	334	60%	*ref*	
-rice water	3	2%	8	1%	1.20 (.31 - 4.62)	0.79
-sticky mucoid	56	34%	194	35%	.93 (.64 - 1.34)	0.69
-bloody	0	0%	18	3%	NA	0.98
Dysentery	4	2%	54	10%	**0.23 (0.08 - 0.65)**	**< 0.01**
Blood in stool	6	4%	75	14%	0.24 (0.1 - 0.57)	0.00
Vomiting 3 or more times in 24 hrs	114	70%	268	48%	**2.48 (1.71 - 3.61)**	**<0.001**
Very thirsty	148	91%	472	86%	1.79 (0.99 - 3.26)	0.06
Drank less	44	27%	106	19%	**1.56 (1.04 - 2.34)**	**0.03**
Unable to drink	14	9%	17	3%	**2.97 (1.43 - 6.16)**	**< 0.01**
Belly pain	114	71%	354	65%	1.29 (0.88 - 1.9)	0.19
Fever	118	72%	404	73%	0.97 (0.66 - 1.44)	0.89
Irritable	136	83%	395	71%	**2.03 (1.29 - 3.19)**	**0.00**
Lethargy	87	53%	288	52%	1.06 (0.75 - 1.5)	0.76
Loss of conciousness	22	13%	45	8%	**1.76 (1.03 - 3.04)**	**< 0.05**
Rectal straining	18	11%	47	9%	1.35 (0.76 - 2.4)	0.30
Rectal prolapse	0	0%	10	2%		0.99
Cough	106	65%	327	59%	1.29 (0.9 - 1.86)	0.17
Difficulty breathing	35	21%	91	16%	1.39 (0.9 - 2.15)	0.14
Convulsion	2	1%	6	1%	1.13 (0.23 - 5.68)	0.88
Observed symptoms at enrollment:						
Very thirsty	137	85%	431	78%	1.51 (0.94 - 2.43)	0.09
Drinks poorly	33	20%	75	14%	**1.63 (1.04 - 2.57)**	**< 0.05**
Sunken eyes	159	98%	526	95%	2.12 (0.73 - 6.12)	0.17
Wrinkled skin	46	29%	82	15%	**2.31 (1.53 - 3.5)**	**< 0.001**
Restless/irritable	119	73%	346	62%	**1.63 (1.1 - 2.39)**	**< 0.01**
Lethargy	30	18%	50	9%	**2.27 (1.39 - 3.72)**	**< 0.01**
Dry mouth	127	78%	445	80%	0.86 (0.56 - 1.32)	0.50
Fast breathing	33	20%	77	14%	1.57 (1 - 2.47)	0.05
Medical assessment symptoms:						
Chest indrawing	5	3%	5	1%	3.47 (0.99 - 12.15)	0.05
Sunken eyes	160	98%	546	99%	0.78 (0.2 - 2.98)	0.72
Dry mouth	38	23%	120	22%	1.10 (.73 - 1.67)	0.65
Abnormal skin pinch	56	34%	145	26%	**1.48 (1.02 - 2.15)**	**< 0.05**
Mental status abnormal	127	78%	354	64%	**1.99 (1.32 - 3)**	**<0.001**
Rectal prolapse	0	0%	5	1%	NA	0.98
Bipedal edema	5	3%	6	1%	2.89 (0.87 - 9.6)	0.08
Abnormal hair	5	3%	32	6%	0.52 (0.2 - 1.35)	0.18
Wasted /very thin	13	8%	60	11%	0.71 (0.38 - 1.34)	0.29
Skin has 'flaky paint' appearance	4	2%	19	3%	0.71 (0.24 - 2.11)	0.54
Requires IV rehydration	35	21%	72	13%	**1.83 (1.17 - 2.87)**	**< 0.01**
Hospital admission	26	16%	56	10%	**1.69 (1.02 - 2.79)**	**< 0.05**

* Wilcoxon-rank sums test

Supplemental table 4.1. Case and controls exposure frequencies and univariable logistic regression estimates

	Cases (n= 1718)	Case%	Controls (n= 2388)	Control%	OR (95% CI)	mOR (95% CI)
Demographic and household exposures						
Age groups						
0-11 months (infants)	800	47%	869	36%	*ref*	*ref*
12-24 months (toddlers)	474	28%	786	33%	**0.66 (0.56 - 0.76)**	1.85 (0.15 - 23.47)
24 -59 months (children)	444	26%	733	31%	**0.66 (0.57 - 0.77)**	0.93 (0.07 - 12.2)
Female	737	43%	1037	43%	0.98 (0.86 - 1.11)	0.47 (0.06 - 3.68)
Both parents live in household	1154	67%	1653	69%	0.91 (0.8 - 1.04)	0.91 (0.79 - 1.04)
Caretaker completed primary school	769	45%	1097	46%	0.95 (0.84 - 1.08)	0.92 (0.81 - 1.05)
Household contains >5 members	709	41%	1243	52%	**0.65 (0.57 - 0.73)**	**0.64 (0.56 - 0.73)**
Household has >4 members sleeping there	799	47%	1232	52%	**0.82 (0.72 - 0.92)**	**0.81 (0.71 - 0.93)**
Household has > 1 rooms used for sleeping	774	45%	1195	50%	**0.82 (0.72 - 0.93)**	**0.82 (0.72 - 0.93)**
> 2 children <5 years old love in household	1059	62%	1590	67%	**0.81 (0.71 - 0.92)**	**0.77 (0.67 - 0.87)**
Finished floor in house*	316	18%	510	21%	**0.83 (0.71 - 0.97)**	0.87 (0.73 - 1.03)
Animal Exposures						
Goats	957	56%	1314	55%	1.03 (0.91 - 1.16)	1 (0.88 - 1.14)
Sheep	513	30%	665	28%	1.1 (0.96 - 1.26)	1.08 (0.93 - 1.24)
Dog	1091	64%	1564	65%	0.92 (0.81 - 1.04)	0.92 (0.8 - 1.05)
Cat	1147	67%	1637	69%	0.92 (0.81 - 1.05)	0.94 (0.82 - 1.08)
Cow	1121	65%	1691	71%	**0.77 (0.68 - 0.88)**	**0.76 (0.66 - 0.88)**
Rodents	930	54%	1134	47%	**1.31 (1.15 - 1.48)**	**1.27 (1.11 - 1.45)**
Fowl	1620	94%	2266	95%	0.89 (0.68 - 1.17)	0.94 (0.7 - 1.26)
Other animals	93	5%	182	8%	**0.69 (0.54 - 0.9)**	**0.67 (0.51 - 0.88)**
Any Animal contact (composite)	1707	99%	2359	99%	1.91 (0.95 - 3.83)	2.03 (1 - 4.12)
Water exposures						
Enrollment main water source is unimproved +	720	42%	859	36%	**1.28 (1.13 - 1.46)**	**1.28 (1.10 - 1.48)**
Follow-up main water source is unimproved+	755	44%	925	39%	**1.24 (1.09 - 1.41)**	**1.31 (1.13 - 1.52)**
Time to travel to water source > 60 min	138	8%	77	3%	**2.62 (1.97 - 3.49)**	**2.44 (1.8 - 3.3)**
Water not available daily	126	7%	184	8%	0.95 (0.75 - 1.2)	0.97 (0.76 - 1.24)
In last 2 weeks gave child untreated water	328	30%	373	28%	1.11 (0.93 - 1.32)	1.09 (0.87 - 1.37)
At enrollment, does not treat household water	631	37%	1055	44%	**0.73 (0.65 - 0.83)**	**0.7 (0.61 - 0.8)**
At enrollment,does not treat with chlorine	837	49%	1314	55%	**0.78 (0.69 - 0.88)**	**0.71 (0.62 - 0.81)**
At follow-up, does not treat water	706	41%	1039	44%	0.91 (0.8 - 1.03)	0.85 (0.74 - 0.97)
At follow-up, does not treat with chlorine	353	45%	482	45%	1 (0.83 - 1.2)	1.03 (0.79 - 1.35)
At follow-up, water container in home (observed)	1683	98%	2354	99%	0.67 (0.42 - 1.09)	0.77 (0.46 - 1.28)
Storage container not narrow mouthed	1328	79%	1914	81%	0.86 (0.73 - 1.01)	.83 (.70 - .98)
No cover on container	177	11%	238	11%	1.07 (0.87 - 1.31)	1.03 (0.82 - 1.28)
Hygeine						
Self reported, times at which caretaker usually wash hands:						
Before eating	1429	83%	2029	85%	0.87 (0.74 - 1.04)	0.87 (0.73 - 1.03)
Before cooking	582	34%	894	37%	**0.86 (0.75 - 0.97)**	0.87 (0.76 - 1)
Before nursing or preparing baby's food	518	30%	466	20%	**1.78 (1.54 - 2.06)**	**1.65 (1.41 - 1.92)**
After defecation	1325	77%	1292	54%	**2.86 (2.49 - 3.28)**	**2.86 (2.45 - 3.33)**
After cleaning child who defecated	446	26%	780	33%	**0.72 (0.63 - 0.83)**	**0.79 (0.68 - 0.92)**
After handling domestic animals	193	11%	517	22%	**0.46 (0.38 - 0.55)**	**0.46 (0.38 - 0.55)**
Any reported hand washing (composite)	1718	100%	2388	100%	*NA*	*NA*
Wash hands near dwelling/yard (observed at follow-up)	1717	100%	2384	100%	2.16 (0.22 - 20.74)	2.11 (0.18 - 25.27)
Station used piped water	9	1%	3	0%	**4.18 (1.13 - 15.46)**	**4.37 (1.17 - 16.39)**
Station had basin	1671	97%	2328	98%	0.92 (0.62 - 1.36)	0.86 (0.55 - 1.34)
Station had soap	887	52%	1208	51%	1.04 (0.92 - 1.18)	1.04 (0.86 - 1.25)
Station had ash	2	0%	1	0%	2.78 (0.25 - 30.61)	2.56 (0.23 - 29.12)
No soap or ash at hand wash station (composite)	831	48%	1178	49%	0.96 (0.85 - 1.09)	0.97 (0.81 - 1.16)
Sanitation						
Disposes of child's feces in open	587	34%	1108	46%	**0.6 (0.53 - 0.68)**	0.52 (0.45 - 0.6)
Disposes of child's feces in open (observed)	606	36%	786	34%	1.11 (0.97 - 1.27)	0.99 (0.85 - 1.16)
Household access to sanitation facility:						
Private household facility	216	13%	402	17%	**0.7 (0.57 - 0.86)**	**0.74 (0.59 - 0.92)**
Shares facility with 1-2 households	550	32%	778	33%	0.93 (0.79 - 1.09)	0.95 (0.8 - 1.13)
Shares facility with ≥3 households	415	24%	474	20%	1.15 (0.96 - 1.37)	1.17 (0.97 - 1.42)
No facility	465	27%	609	26%	*ref*	*ref*
Human feces visible in defecation area	588	34%	785	33%	1.06 (0.93 - 1.21)	1.03 (0.86 - 1.23)
Human feces visible in house or yard	138	8%	136	6%	**1.45 (1.13 - 1.85)**	**1.36 (1.05 - 1.75)**
Cold/wet month	700	41%	1008	42%	0.94 (0.83 - 1.07)	0.64 (0.34 - 1.18)

Supplemental table 4.2 Infant case and controls exposure frequencies and logistic regression estimates

	Cases (n=800)	Case%	Controls (n= 869)	Control%	OR (95% CI)	aOR (95% CI)
Demographic and household exposures						
Female	323	40%	351	40%	1 (0.82 - 1.22)	1.06 (.87 - 1.29)
Both parents live in household	542	68%	598	69%	0.95 (0.77 - 1.17)	
Caretaker completed primary school	362	45%	411	47%	0.92 (0.76 - 1.12)	
Household contains >5 members	366	46%	434	50%	0.85 (0.7 - 1.02)	
Household has >4 members sleeping there	392	49%	448	52%	0.9 (0.75 - 1.09)	
Household has > 1 rooms used for sleeping	371	46%	412	47%	0.96 (0.79 - 1.16)	
> 2 children <5 years old love in household	536	67%	620	71%	0.82 (0.66 - 1)	
Finished floor in house*	144	18%	166	19%	0.93 (0.73 - 1.19)	
Animal Exposures						
Goats	443	55%	474	55%	1.03 (0.85 - 1.25)	
Sheep	231	29%	247	28%	1.02 (0.83 - 1.26)	
Dog	499	62%	552	64%	0.95 (0.78 - 1.16)	
Cat	524	66%	593	68%	0.88 (0.72 - 1.08)	
Cow	522	65%	609	70%	**0.8 (0.65 - 0.98)**	**.78 (.63 - .96)**
Rodents	434	54%	443	51%	1.14 (0.94 - 1.38)	
Fowl	750	94%	812	93%	1.05 (0.71 - 1.56)	
Other animals	50	6%	67	8%	0.8 (0.55 - 1.17)	
Any Animal contact (composite)	792	99%	857	99%	1.39 (0.56 - 3.41)	
Water exposures						
Enrollment main water source is unimproved +	379	47%	348	40%	**1.35 (1.11 - 1.64)**	
Follow-up main water source is unimproved+	371	46%	346	40%	**1.3 (1.07 - 1.58)**	**1.29 (1.06 - 1.58)**
Time to travel to water source ≥ 60 min	73	9%	30	3%	**2.81 (1.81 - 4.34)**	**2.69 (1.73 - 4.18)**
Water not available daily	60	8%	73	8%	0.88 (0.62 - 1.26)	
In last 2 weeks gave child untreated water	122	24%	88	18%	**1.48 (1.09 - 2.01)**	
At enrollment, does not treat household water	290	36%	366	42%	**0.78 (0.64 - 0.95)**	
At enrollment,does not treat with chlorine	382	48%	486	56%	**0.72 (0.59 - 0.87)**	**.70 (.58 -.86)**
At follow-up, does not treat water	324	41%	382	44%	0.87 (0.71 - 1.05)	
At follow-up, does not treat with chlorine	161	44%	160	43%	1.04 (0.77 - 1.38)	
At follow-up, water container in home (observed)	785	98%	857	99%	0.67 (0.31 - 1.47)	
Storage container not narrow mouthed	630	80%	689	80%	0.99 (0.78 - 1.27)	
No cover on container	82	11%	90	11%	1 (0.73 - 1.37)	
Hygeine						
Self reported, times at which caretaker usually wash hands:						
Before eating	654	82%	711	82%	1 (0.78 - 1.28)	
Before cooking	246	31%	309	36%	**0.8 (0.66 - 0.99)**	
Before nursing or preparing baby's food	296	37%	218	25%	**1.75 (1.42 - 2.16)**	
After defecation	614	77%	449	52%	**3.09 (2.5 - 3.81)**	
After cleaning child who defecated	226	28%	304	35%	**0.73 (0.59 - 0.9)**	
After handling domestic animals	80	10%	199	23%	**0.37 (0.28 - 0.49)**	
Any reported hand washing (composite)	800	100%	869	100%		
Wash hands near dwelling/yard (observed at follow-up)	799	100%	866	100%		
Station used piped water	2	0%	0	0%		
Station had basin	777	97%	842	97%	**1.09 (0.61 - 1.94)**	
Station had soap	394	49%	421	49%	1.03 (0.85 - 1.25)	
Station had ash	0	0%	0	0%		
No soap or ash at hand wash station (composite)	406	51%	447	52%	0.97 (0.8 - 1.17)	
Sanitation						
Disposes of child's feces in open	312	39%	430	50%	0.65 (0.54 - 0.79)	
Disposes of child's feces in open (observed)	299	38%	291	34%	1.2 (0.98 - 1.47)	
Household access to sanitation facility:						
Private household facility	107	13%	147	17%	0.67 (0.49 - 0.92)	
Shares facility with 1-2 households	252	32%	272	31%	0.85 (0.66 - 1.1)	
Shares facility with ≥3 households	183	23%	192	22%	0.88 (0.67 - 1.16)	
No facility	228	29%	210	24%	ref	
Human feces visible in defecation area	264	33%	287	33%	1 (0.81 - 1.22)	
Human feces visible in house or yard	79	10%	54	6%	**1.65 (1.15 - 2.37)**	**1.61 (1.12 - 2.32)**
Cold/wet month	319	40%	354	41%	0.96 (0.79 - 1.17)	

Supplemental table 4.3. Toddler case and controls exposure frequencies and logistic regression estimates

	Cases		Controls			
	(n=474)	Case%	(n= 786)	Control%	OR (95% CI)	aOR (95% CI)
Demographic and household exposures						
Female	207	44%	338	43%	1.03 (0.82 - 1.29)	1.03 (.80 - 1.32)
Both parents live in household	322	68%	557	71%	0.87 (0.68 - 1.11)	
Caretaker completed primary school	208	44%	362	46%	0.91 (0.73 - 1.15)	
Household contains >5 members	182	38%	409	52%	**0.57 (0.46 - 0.72)**	**.67 (.52 - .86)**
Household has >4 members sleeping there	198	42%	399	51%	0.7 (0.55 - 0.88)	
Household has > 1 rooms used for sleeping	196	41%	402	51%	0.67 (0.54 - 0.85)	
> 2 children <5 years old love in household	277	58%	499	63%	0.81 (0.64 - 1.02)	
Finished floor in house*	90	19%	161	20%	0.91 (0.68 - 1.21)	
Animal Exposures						
Goats	269	57%	429	55%	1.09 (0.87 - 1.37)	
Sheep	147	31%	227	29%	1.11 (0.86 - 1.42)	
Dog	305	64%	521	66%	0.92 (0.72 - 1.17)	
Cat	325	69%	547	70%	0.95 (0.74 - 1.22)	
Cow	307	65%	570	73%	**0.7 (0.55 - 0.89)**	.63 (.49 - .83)
Rodents	253	53%	334	42%	**1.55 (1.23 - 1.95)**	1.56 (1.22 - 1.99)
Fowl	446	94%	754	96%	0.68 (0.4 - 1.14)	
Other animals	25	5%	55	7%	0.74 (0.45 - 1.2)	
Any Animal contact (composite)	473	100%	774	98%	7.33 (0.95 - 56.57)	
Water exposures						
Enrollment main water source is unimproved +	185	39%	271	34%	1.22 (0.96 - 1.54)	
Follow-up main water source is unimproved+	205	43%	295	38%	**1.27 (1.01 - 1.6)**	**1.36 (1.06 - 1.76)**
Time to travel to water source > 60 min	34	7%	24	3%	**2.46 (1.44 - 4.2)**	**2.77 (1.55 - 4.97)**
Water not available daily	37	8%	49	6%	1.27 (0.82 - 1.98)	
In last 2 weeks gave child untreated water	104	35%	139	32%	1.16 (0.85 - 1.59)	
At enrollment, does not treat household water	181	38%	354	45%	**0.75 (0.6 - 0.95)**	
At enrollment,does not treat with chlorine	240	51%	426	54%	0.87 (0.69 - 1.09)	
At follow-up, does not treat water	194	41%	328	42%	0.97 (0.77 - 1.22)	
At follow-up, does not treat with chlorine	106	48%	175	48%	1.02 (0.73 - 1.42)	
At follow-up, water container in home (observed)	462	97%	772	98%	0.7 (0.32 - 1.52)	
Storage container not narrow mouthed	349	76%	623	81%	**0.75 (0.57 - 0.99)**	
No cover on container	55	13%	86	12%	1.11 (0.77 - 1.59)	
Hygeine						
Self reported, times at which caretaker usually wash hands:						
Before eating	397	84%	682	87%	0.79 (0.57 - 1.08)	
Before cooking	175	37%	291	37%	1 (0.79 - 1.26)	
Before nursing or preparing baby's food	144	30%	153	19%	1.81 (1.39 - 2.35)	
After defecation	370	78%	427	54%	**2.99 (2.31 - 3.87)**	
After cleaning child who defecated	133	28%	261	33%	0.78 (0.61 - 1.01)	
After handling domestic animals	50	11%	174	22%	0.41 (0.3 - 0.58)	
Any reported hand washing (composite)	474	100%	786	100%		
Wash hands near dwelling/yard (observed at follow-up)	474	100%	785	100%		
Station used piped water	3	1%	2	0%	2.5 (0.42 - 14.99)	
Station had basin	465	98%	775	99%	0.73 (0.3 - 1.78)	
Station had soap	261	55%	403	51%	1.16 (0.93 - 1.46)	
Station had ash	2	0%	0	0%		
No soap or ash at hand wash station (composite)	213	45%	382	49%	0.86 (0.68 - 1.08)	
Sanitation						
Disposes of child's feces in open	130	28%	331	42%	**0.52 (0.41 - 0.67)**	**.38 (.27 - .54)**
Disposes of child's feces in open (observed)	148	32%	252	33%	0.95 (0.74 - 1.22)	
Household access to sanitation facility:						
Private household facility	57	12%	130	17%	0.81 (0.55 - 1.19)	.48 (.30 - .77)
Shares facility with 1-2 households	164	35%	256	33%	1.19 (0.88 - 1.6)	.76 (.51 - 1.12)
Shares facility with ≥3 households	121	26%	142	18%	**1.58 (1.13 - 2.2)**	.91 (.60 - 1.38)
No facility	116	24%	215	27%	ref	
Human feces visible in defecation area	166	35%	247	31%	1.17 (0.92 - 1.49)	
Human feces visible in house or yard	36	8%	49	6%	1.24 (0.79 - 1.93)	
Cold/wet month	191	0.402954	331	42%	0.93 (0.74 - 1.17)	

Supplemental table 4. 4. Children logistic case and controls exposure frequencies and logistic regression estimates

	Cases (n=444)	Case%	Controls (n= 733)	Control%	OR (95% CI)	aOR (95% CI)
Demographic and household exposures						
Female	207	47%	348	47%	0.97 (0.76 - 1.22)	.96 (.74 - 1.24)
Both parents live in household	290	65%	498	68%	0.89 (0.69 - 1.14)	
Caretaker completed primary school	199	45%	324	44%	1.03 (0.81 - 1.3)	
Household contains >5 members	161	36%	400	55%	**0.47 (0.37 - 0.6)**	**.49 (.38 - .65)**
Household has >4 members sleeping there	209	47%	385	53%	0.8 (0.63 - 1.02)	
Household has > 1 rooms used for sleeping	207	47%	381	52%	0.81 (0.64 - 1.02)	
> 2 children <5 years old love in household	246	55%	471	64%	**0.69 (0.54 - 0.88)**	
Finished floor in house*	82	18%	183	25%	**0.68 (0.51 - 0.91)**	**.62 (.45 - .86)**
Animal Exposures						
Goats	245	55%	411	56%	0.96 (0.76 - 1.22)	
Sheep	135	30%	191	26%	1.24 (0.96 - 1.61)	
Dog	287	65%	491	67%	0.9 (0.7 - 1.15)	
Cat	298	67%	497	68%	0.97 (0.75 - 1.25)	
Cow	292	66%	512	70%	0.83 (0.64 - 1.07)	
Rodents	243	55%	357	49%	**1.27 (1.01 - 1.61)**	**1.33 (1.03 - 1.72)**
Fowl	424	95%	700	95%	1 (0.57 - 1.77)	
Other animals	18	4%	60	8%	**0.47 (0.28 - 0.81)**	**.51 (.28 - .91)**
Any Animal contact (composite)	442	100%	728	99%	1.52 (0.29 - 7.86)	
Water exposures						
Enrollment main water source is unimproved +	156	35%	240	33%	1.11 (0.87 - 1.43)	
Follow-up main water source is unimproved+	179	40%	284	39%	1.07 (0.84 - 1.36)	
Time to travel to water source ≥ 60 min	31	7%	23	3%	**2.32 (1.33 - 4.03)**	**2.49 (1.33 - 4.66)**
Water not available daily	29	7%	62	8%	0.76 (0.48 - 1.2)	
In last 2 weeks gave child untreated water	102	36%	146	37%	0.96 (0.7 - 1.32)	
At enrollment, does not treat household water	160	36%	335	46%	**0.67 (0.53 - 0.85)**	**.70 (.53 - .91)**
At enrollment,does not treat with chlorine	215	48%	402	55%	0.77 (0.61 - 0.98)	
At follow-up, does not treat water	188	42%	329	45%	0.9 (0.71 - 1.14)	
At follow-up, does not treat with chlorine	86	44%	147	45%	0.96 (0.68 - 1.38)	
At follow-up, water container in home (observed)	436	98%	725	99%	0.6 (0.22 - 1.61)	
Storage container not narrow mouthed	349	80%	602	83%	0.81 (0.6 - 1.1)	
No cover on container	40	10%	62	9%	1.13 (0.75 - 1.72)	
Hygeine						
Self reported, times at which caretaker usually wash hands:						
Before eating	378	85%	636	87%	0.87 (0.62 - 1.22)	
Before cooking	161	36%	294	40%	0.85 (0.67 - 1.08)	
Before nursing or preparing baby's food	78	18%	95	13%	**1.43 (1.03 - 1.98)**	
After defecation	341	77%	416	57%	**2.52 (1.94 - 3.29)**	
After cleaning child who defecated	87	20%	215	29%	0.59 (0.44 - 0.78)	
After handling domestic animals	63	14%	144	20%	0.68 (0.49 - 0.93)	
Any reported hand washing (composite)	444	100%	733	100%		
Wash hands near dwelling/yard (observed at follow-up)	444	100%	733	100%		
Station used piped water	4	1%	1	0%	6.64 (0.74 - 59.51)	
Station had basin	429	97%	711	97%	0.88 (0.45 - 1.72)	
Station had soap	232	52%	384	52%	0.99 (0.79 - 1.26)	
Station had ash	0	0%	1	0%		
No soap or ash at hand wash station (composite)	212	48%	349	48%	1.01 (0.79 - 1.27)	
Sanitation						
Disposes of child's feces in open	145	33%	347	47%	**0.54 (0.42 - 0.69)**	**.34 (.24 - .66)**
Disposes of child's feces in open (observed)	159	37%	243	34%	1.13 (0.88 - 1.45)	
Household access to sanitation facility:						
Private household facility	52	12%	125	17%	**0.63 (0.43 - 0.94)**	**.41 (.25 - .66)**
Shares facility with 1-2 households	134	30%	250	34%	0.82 (0.6 - 1.11)	**.50 (.34 - .73)**
Shares facility with ≥3 households	111	25%	140	19%	1.21 (0.86 - 1.69)	**.65 (.43 - .98)**
No facility	121	27%	184	25%	*ref*	
Human feces visible in defecation area	158	36%	251	34%	1.06 (0.83 - 1.36)	
Human feces visible in house or yard	23	5%	33	5%	1.16 (0.67 - 2)	
Cold/wet month	190	43%	323	44%	0.95 (0.75 - 1.2)	