Published by Oxford University Press on behalf of the International Epidemiological Association
International Journal of Epidemiology 2012;41:504–513
doi:10.1093/ije/dyr194

METHODOLOGY

Where science meets policy: comparing longitudinal and cross-sectional designs to address diarrhoeal disease burden in the developing world

Amanda R Markovitz,[1,6] Jason E Goldstick,[1] Karen Levy,[2] William Cevallos,[3] Bhramar Mukherjee,[4] James A Trostle[5] and Joseph N S Eisenberg[1]*

[1]Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI, USA, [2]Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, GA, USA, [3]Department of Microbiology, Universidad San Francisco de Quito, Quito, Ecuador, [4]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI, USA and [5]Anthropology Department, Trinity College, Hartford, CT, USA

[6]Present address: Department of Clinical Epidemiology and Biostatistics, Blue Cross Blue Shield of Michigan, Detroit, MI, USA

*Corresponding author. Department of Epidemiology, School of Public Health, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48104, USA. E-mail: jnse@umich.edu

Accepted 1 November 2011

Background Longitudinal studies are considered preferable to cross-sectional studies for informing public health policy. However, when resources are limited, the trade-off between an accurate cross-section of the population and an understanding of the temporal variation should be optimized. When risk factors vary more across space at a fixed moment in time than at a fixed location across time, cross-sectional studies will tend to give more precise estimates of risk factor effects and thus may be a better source of data for policy judgments.

Methods We conducted a diarrhoeal disease surveillance of 5616 individuals within 19 Ecuadorian villages. This data set was used to mimic cross-sectional and longitudinal studies by restricting focus to a single week and a single village, respectively. We compared the variability in risk factor effect estimates produced from each type of study.

Results For household risk factors, the effect estimates produced by the longitudinal studies were more variable than their cross-sectional counterparts, which can be explained by greater spatial than temporal variability in the risk factor distribution. For example, the effect estimate of improved sanitation was almost twice as variable in longitudinal studies.

Conclusions In our study, cross-sectional designs yielded more consistent evaluations of diarrhoea disease risk factors when those factors varied more between villages than over time. Cross-sectional studies can provide information that is representative across large geographic regions and therefore can provide insight for local, regional and national policy decisions. The value of the cross-sectional study should be reconsidered in the public health community.

Keywords Diarrhoea, cross-sectional studies, longitudinal studies

Introduction

Diarrhoea continues to be one of the leading causes of death and loss of disability adjusted life years in the developing world.[1] Estimates from the past four decades indicate that, although there has been a decrease in mortality rates, diarrhoeal disease-related morbidity has remained high.[2] Interventions to reduce diarrhoea incidence generally focus on water supply, water quality, sanitation and hygiene for all age groups, as well as breastfeeding, adequate nutrition and immunizations specifically for children < 5 years of age. Although there has been some attempt to utilize large-scale and widely available cross-sectional studies such as the Demographic and Health Surveys (DHS),[3,4] most guidelines and policies relevant to diarrhoea in developing countries, such as those developed by the WHO with reference to household storage or disinfection of water, micronutrient supplementation or use of oral rehydration solution, have been informed by longitudinal intervention trials that are conducted with smaller cohorts followed over time.[5] In this analysis we aim to revisit and assess this preference.

Longitudinal studies are generally considered to have fewer limitations to making aetiological inferences[6] and are essential to use when tracking changes among individuals. However, when researchers are interested in how risk factors affect populations, and know enough about a disease to assess the temporal direction of cause and effect, whether to choose a longitudinal over a cross-sectional design is less obvious. Longitudinal studies are logistically more challenging than cross-sectional studies, leading to smaller sample sizes given similar costs.[7] Longitudinal studies also suffer from losses to follow up and non-compliance. We reviewed 13 diarrhoea intervention trials spanning $\geqslant 6$ months, and drop-out rates ranged from 1%[8] to 33%.[9] In addition, misclassification rates are documented to increase during follow-up within a longitudinal study due to factors such as reporting fatigue.[10] The act of sampling the participants may also change their later behaviours, which can lead to biased estimates of the effect of behaviours that are repeatedly surveyed.[11]

The purpose of this work is to show that cross-sectional studies can provide more accurate risk factor effect estimates than longitudinal studies, given the same sample size and sources of bias. In order to take repeated measurements on individuals over time, longitudinal studies sacrifice on the total number of individuals sampled. This can lead to an incomplete sample of the risk factor distribution, which can cause unstable estimates of risk factor effects and loss of external validity. In particular, when the risk factor distribution is characterized more completely by observation across spatial locations at a single point in time rather than observation across time in a more limited area, a cross-sectional study is likely to perform better in terms of risk factor effect estimation. Examples of such risk factors include aspects of the built environment, such as improved sanitation facilities, which tend to vary greatly by region but stay more consistent over time.

In this study we examine the additional utility, if any, afforded by a longitudinal study when the goal is to evaluate water, sanitation, hygiene and other interventions on diarrhoeal disease prevalence. We use a large longitudinal data set collected in 19 Ecuadorian villages from 2004 to 2007 to show empirically—countering the conventional wisdom—that there are situations in which a longitudinal study will perform worse, in terms of ability to precisely estimate risk factor effects, than a cross-sectional study with a similar sample size. We demonstrate that the relative utility of the two study designs depends on whether there is greater variation in the risk factor at a fixed time across space or over time at a fixed point in space and argue for the continuing importance of cross-sectional studies in developing policy.

Methods

Study population

This study was located in the northern coastal Ecuadorian province of Esmeraldas in the canton of Eloy Alfaro. This area contains approximately 150 small villages, located on three rivers, the Cayapas, Santiago and Onzole, which all drain towards Borbón, the main population centre of the region. More details about the study population can be obtained elsewhere.[12]

A sample of 21 villages was selected using block randomization to ensure that villages throughout the study region were represented. All houses in these 21 villages were recruited for the study. A systematic evaluation of the consistency and quality of data collection indicated that data from two villages were not of sufficient quality. These villages were therefore excluded from this analysis. Consent was obtained at both the village and household level. Institutional review boards at the University of Michigan, University of California (Berkeley), Trinity College and Universidad San Francisco de Quito approved all protocols.

Study design and variables

All consenting households in the villages were visited on a weekly basis from 18 February 2004 through 4 July 2007 by 25 community health workers employed by the study. Community health workers were recruited from the study area and trained in interviewing techniques. The self-identified heads of each household were questioned about whether anyone in the house had felt sick in the previous week and, if

they said yes, were asked more detailed questions about symptoms and treatment. A diarrhoea episode was defined as at least one 24-hour period in the past week in which an individual experienced three or more loose stools. Incident cases included only episodes where the individual had not experienced diarrhoea in the preceding week.

Explanatory variables used in the analysis included individual, household and village characteristics and were collected from two sources: an ongoing census of the villages, which was conducted annually or biennially during the study period, and from a concurrent case–control study of diarrhoea, conducted in the same villages between August 2003 and October 2008. In this 5-year period the villages were visited seven times each for 15 days at a time, during which all cases of diarrhoea were identified and three controls were selected. Two of these controls were matched on the village and one was matched on household. Demographic data (including birth date, gender, education and job type) and household information (including counts of household members, type of sanitation facility and ownership of various items) were collected through the census. Households were also mapped when they were enrolled in the study and materials used to construct both the house and roof were noted at that time. Data on hygiene, water sources and water treatment were collected as part of the case–control study.

Age and gender were assessed as potential predictors of diarrhoea status at an individual level. Exposures of interest at the household level included number of people in the house, whether the house had an improved sanitation facility, and indicators of socio-economic status (SES). Household SES indicators were based on education level, job status and wealth (housing construction and assets).[13]

At the village level, overall levels of hygiene, access to improved water sources and use of water treatment were considered. These variables were collected through the case–control study and therefore were not available for each household. Because these variables were not collected from a simple random sample, the village level averages were weighted, with individual weights based on the inverse probability of the individual being chosen for inclusion in the study. This adjusts for the fact that these individuals had a greater probability of being selected. As the case control and census both occur annually or semi-annually, each week of observation had to be linked to the closest measured variable. The predictors of interest in this analysis are not likely to fluctuate greatly over short time scales, therefore their lack of dense temporal observation is not likely to have a large impact on the substantive conclusions. Definitions of household and village level variables are available in Table 1.

Comparing cross-sectional and longitudinal studies

In order to compare the performance of cross-sectional and longitudinal studies, the full data set was envisioned in two ways; as a series of weekly cross-sectional studies encompassing all villages and as a series of village-level longitudinal studies. With this in mind, each week was treated as a separate cross-sectional study, producing 176 distinct data sets of approximately the same size ($n \approx 3224$). Since villages contained a variable number of residents (53–728), the number of weeks used for a comparable longitudinal study were chosen separately for each village in order to produce a similar total number of studies for the analysis. These series of longitudinal studies varied in length from 5 to 157 weeks. Each village's longitudinal measurements were divided into the maximal number of partitions so that each time period had about 3224 observations. This produced a total of 176 cross-sectional studies, corresponding to the number of weeks in the study and 164 longitudinal studies.

Statistical analysis

Due to the binary nature of the outcome and potential correlation of the repeated measurements of individuals and villages, we used mixed effects logistic regression models in the analyses. This approach was chosen over General Estimating Equations (GEE) because of the ease of handling nested clustering, unequally sized clusters (e.g. villages) and missing-at-random data. Analysis of the full data set included random intercepts for individuals and villages while analysis of the simulated longitudinal and cross-sectional data sets included only individual- and village-level random intercepts, respectively. Including an additional random effect variable for household did not change the results of the full model so we decided not to include it for parsimony and to limit computational complexity. In the full data set, the village and individual random effects were allowed to be freely correlated.

One multiple regression model was run for the full data set and included all of the variables described above (and listed in Table 1). All models comparing the two types of studies contained only one variable at a time. This univariate approach was used to isolate the comparison of interest and to preclude artificially inflated sampling variation due to issues such as co-linearity that present differently in each of the two study designs. These univariate models were run only for the individual and household level variables described above, since the effect of village level variables cannot be estimated in data sets containing only one village. All data sets generating estimates that showed clear signs of numerical instability were excluded from summaries. *Post hoc* inspection showed that this instability was typically caused by an insufficient number of cases of diarrhoea for

Table 1 Description of variables used in analysis

	Definition	Data source	Number of households for which data is available ($n = 1130$)
Household characteristics			
Education	Highest level of education achieved by any member of the house	Census	1107
Job status	Indicator for whether anyone in the house had a stable job (government/state worker, business owner or teacher)	Census	1109
Sanitation[a]	Indicator for whether the house had an improved sanitation facility (septic tank or latrine)	Census	902
Number in household	Number of people listed in the household on most recent census	Census	1121
Ownership Index	• Score created by weighting and summing the number and type of consumer goods the household possessed • Scores range from 0 to 1 with 0 corresponding to ownership of no consumer goods	Census	890
Housing construction score	• Score based on the quality of materials used to construct house • Scores range from 0 to 1 with 0 corresponding to utilization of all unimproved housing materials	Survey of households entering study	1045
Village characteristics			
Water source[a]	• Indicator for improved water source (piped, rain or well water) • Aggregated at the village level for analysis	Case–control study	1130
Water treatment[a]	• Indicator for improved water treatment method (filtering, boiling or using chlorine) • Aggregated at the village level for analysis	Case–control study	1130
Hygiene	• Hygiene score calculated based on 23 observations of the condition of household, calculated as the percentage of questions for which improved hygiene practices were observed • Aggregated at the village level for analysis	Case–control study	1130

[a]Definitions from WHO/UNICEF Joint Monitoring Programme[14].

each level of a categorical predictor, or due to the issue of complete separability. All analyses were conducted using SAS 9.2.

As a frame of reference, we treated the regression coefficients from the univariate analyses of the full data set as the true values. The spread of the cross-sectional and village-level longitudinal estimates were quantified by the averaged square distance from this presumed true value; the bias was analogously quantified using the average difference in place of the squared distance. Since the bias was effectively equal to zero in all cases, only the average squared difference, referred to hereafter as the sampling variance, is reported.

To explain the observed performance, we looked at the spatial distribution of predictors over fixed time points and the temporal distribution of predictors over fixed regions. Since a less complete characterization of the risk factor distribution is likely to be accompanied by underestimation of the variance in the risk factor, we compared the observed sampling variation in the coefficient estimates with variance in the risk factors to check for a monotonically increasing relationship. This was done using a graphical summary of the observed variance corresponding to the cross-sectional and longitudinal data sets for risk factors where there are and are not observed differences in risk factor effect estimation precision. Larger values of the variances are expected to correspond to more precise estimation. This expectation is related to the fact that, in regression models, the standard error of a predictor effect (the regression coefficient) estimate is inversely related to the variance of the predictor distribution.

Results

Over the period from 18 February 2004 through 4 July 2007, data were collected from 5616 people describing whether they had three or more loose stools in any 24-h period in the past week. This amounted to a total of 567 444 person-weeks of observation, with an average of 3224 individuals asked during each week. During these weeks, 2276 incident cases of diarrhoea were observed, for an overall incidence rate of 0.21 cases per person-year [(2276 incident cases/563 112 person-weeks at risk) × 52 weeks/year] and 0.67 cases per person-year [(1121 incident cases/87 066 person-weeks at risk) × 52 weeks/year] among children <5 years of age (Table 2). A descriptive plot of these incidence rates over time is shown in Figure 1. The village-level longitudinal data sets and the weekly cross-sectional data sets averaged around 3200 observations each.

Surveyed households averaged five members and respondents included a high proportion of children <5 years of age (16%). Only 61% of households had a member who had completed primary school and 26% had a member with a stable job. Approximately one-half of the houses in the 19 villages had improved sanitation facilities. Close to one-quarter had an improved water source and a similar percentage treated their water. These characteristics varied slightly over time, but varied more widely between village-level longitudinal data sets (Table 2).

We looked at a mixed effects logistic regression model run on the full longitudinal study, which is not meant to be compared with the univariate results, to provide a frame of reference for the study population. The model revealed that age <5 years, female gender and a smaller number of household members were all significant risk factors for diarrhoea, when controlling for all other variables. Villages with a greater percentage of households who practiced improved hygiene practices, utilized improved water sources and treated their water experienced significantly fewer diarrhoea episodes (Table 3).

Table 4 provides a comparison between the cross-sectional and longitudinal designs. The first column shows the univariate (mixed effects) logistic regression of diarrhoea on each factor on the full data set to provide a frame of reference; the longitudinal and cross-sectional designs both unbiasedly estimated these coefficients (results not shown).

In terms of estimation precision, as measured by the sampling variance of the coefficient estimates, we found similar results for both individual-level variables. On the other hand, for all household-level variables, the estimation precision for the village-level longitudinal estimates exceeded that of the cross-sectional estimates (Table 4). In other words, whenever there was a clear discrepancy in the estimation precision, it was the cross-sectional design that provided more precise estimates. This is most notable when comparing the improved sanitation variable, where about twice the variation is seen in the longitudinal estimates.

Next we compared the variability of each risk factor across the 176 cross-sectional studies with the variability across the 164 longitudinal studies. We did this because with little variability in exposure the effect of the exposure is harder to estimate; i.e. the estimate will be imprecise. For improved sanitation, we observed that the variance across the longitudinal sets was less than or equal to the variance across the cross-sectional sets (Figure 2A). Notably, the variance across the cross-sectional studies was always around 0.25, whereas there were several cases where the observed variation in the longitudinal data sets was very small; these correspond to highly imprecise estimates of the effect of improved sanitation. Aggregating across these data sets resulted in an overall sampling variation (estimation precision) in the risk factor effect estimates that was nearly twice as large as the cross-sectional counterparts (Table 4).

In contrast, for the variable age <5 years, the variance from both study types clustered around 0.13, with the longitudinal study showing a roughly symmetric scatter around that average (Figure 2B). This explains the fact that the sampling variation (estimation precision) for the estimates produced by the longitudinal design was similar to those of the cross-sectional design. The construction score and stable job (as defined by any government-funded job such as a teacher) variables showed a similar but less conclusive signal, when compared with improved sanitation, that there tended to be more variability in the cross-sectional compared with the longitudinal design (Figure 2C and D).

Discussion

To efficiently estimate the association between exposure and disease requires a study design that can capture the complete distribution of the exposure. When comparing the full data set to a sample of 176 possible cross-sectional data sets occurring at different time points and 164 possible longitudinal data sets occurring throughout our study region, we found that the variance in diarrhoea risk estimates associated with household level SES and sanitation levels were consistently lower in the cross-sectional studies. For example, the variance in risks of unimproved sanitation was twice as high in the longitudinal compared with the cross-sectional data sets, indicating the longitudinal data were doing a less efficient job of capturing the effect. In our analysis of diarrhoeal disease and its associated risks, a cross-sectional study design better captured that distribution then did the longitudinal design.

Diarrhoeal disease prevalence can vary greatly over many time scales, with variation by year, season, week and even day. For policy decisions on allocation of resources for water, sanitation and hygiene

Table 2 Descriptive statistics of the three data sets used in the analysis

	Units	Full data set	Range for longitudinal data sets	Range for cross-sectional data sets
Sample size				
Number of observations	*N*	567 444	980–4430	1624–3804
Average number of people[a]	*N*	3224	42–718	1624–3804
Average number of observations/person	*N*	101	3–51	1
Cases of diarrhoea	*N*	2348	1–40	3–28
Incidence rate (per person-year)				
Overall	Cases/person-year	0.21	0.01–0.71	–
<5-year olds		0.67	0–2.65	–
⩾5-year olds		0.13	0–0.52	–
Individual characteristics				
Age				
<5	%	16	10–27	14–19
⩾5	%	84	73–90	81–86
Gender				
Male	%	53	42–61	52–55
Female	%	47	39–58	45–48
Household characteristics				
Highest level of education				
<6 years	%	39	12–79	34–48
>6 years	%	61	21–88	52–66
Household member has a stable job[b]	%	26	4–79	45–51
Household has improved sanitation[b]	%	51	12–100	45–57
Number of people in house	mean	5	3–7	4–5
Ownership index[b,c]	mean	0.37	0.25–0.49	0.36–0.38
Housing construction score[b,c]	mean	0.62	0.19–0.77	0.60–0.64
Village characteristics				
Average % houses with improved water source[b]	%	28	0–100	5–47
Average % houses that treat water[b]	%	29	0–94	18–42
Average % houses with improved hygiene[b]	%	57	17–88	47–67

Column 3 summarizes values for all weeks and villages, including each week as independent observations when calculating average characteristics. Columns 4 and 5 summarize at the study level a series of 164 longitudinal data sets and 176 cross-sectional data set, respectively.
[a]Average number of people = number of observations/number of weeks of observation.
[b]See Table 1 for definitions of variables.
[c]Measured on a scale of 0–1.

interventions, however, this variation in prevalence may not be the relevant measure of interest; rather, as demonstrated in our analysis, the variation in the predictor (or risk factor) variables may be the more relevant measures of interest. For example, if only a few houses in a data set have unimproved sanitary infrastructure, it will be difficult to estimate the effect of sanitation on diarrhoeal disease risk, regardless of the overall sample size, the model chosen and any other methodological considerations.

The proximal risk factors of major concern are water, sanitation and hygiene. Water and sanitation are generally measured by the water source and sanitation facility used, respectively. Hygiene is measured by observation of the cleanliness of a household and/or the presence of soap. In our study, these and other similar variables measured at a household level were more stable over time than between villages and, because of this, cross-sectional studies were able to produce more precise estimates of their effect on

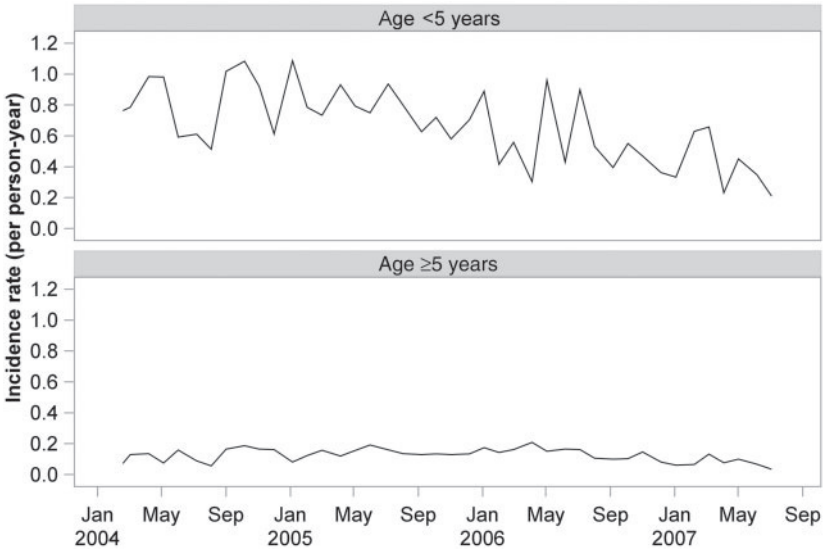

Figure 1 Average monthly incidence rates of diarrhoea for 19 villages in Ecuador, stratified by age of individuals

Table 3 Results from mixed effects regression model for 19 villages in Ecuador, 2004–2007

Effect	OR	95% CI
Individual characteristics		
Age		
<5	Ref (1)	
⩾5	0.20	(0.17–0.23)
Gender		
Male	Ref (1)	
Female	1.18	(1.02–1.37)
Household characteristics		
Highest level of education		
<6 years	Ref (1)	
⩾6 years	0.97	(0.83–1.13)
Whether anyone has a stable job		
No	Ref (1)	
Yes	1.14	(0.97–1.34)
Ownership index[a]	0.73	(0.42–1.27)
Housing construction[a]	1.01	(0.78–1.30)
Number of people in house[b]	0.93	(0.91–0.96)
Improved sanitation facility		
No	Ref (1)	
Yes	1.09	(0.95–1.25)
Village characteristics		
Observed hygiene[a]	0.51	(0.34–0.77)
Water source[a]	0.75	(0.64–0.89)
Water treatment[a]	0.65	(0.50–0.85)

All variables were included in one regression model with random effects for individuals and villages. OR: odds ratio; 95% CI: 95% confidence interval.
[a]OR represents an increase from the minimum to maximum.
[b]OR represents an increase of one person.

diarrhoeal disease prevalence. These results reinforce the idea that policy decisions should be informed by data that represent the spatial extent of the relevant population. In a more general setting, when there is good reason to suspect that a risk factor will show greater variation across space at a particular point in time than across time at a particular location, one may be well advised to consider a cross-sectional rather than longitudinal study. This is likely true for behavioural factors, such as hygiene, bednet use, antibiotic use, as well as socio-economic factors; all of these are associated with risks for a wide variety of environmentally mediated pathogens. A predictor's distribution over space and time will frequently be available from previous studies in the region of interest or generalizable from other regions, but in some cases may have to be estimated through pilot studies.

The cross-sectional studies performed better than the longitudinal studies in this analysis when the total sample size was kept constant; however, in reality the higher cost of longitudinal studies would necessitate a trade-off in sample size and so the difference in study performance might be even more noticeable. Other limitations of repeated sampling, including loss to follow-up, non-compliance, reporting fatigue and behavioural changes in response to being surveyed, could further improve the relative performance of cross-sectional studies.

One limitation of our study is that we did not measure risk factors on a weekly basis but assumed that semi-annual to yearly measurements would capture their variation. If these factors did change more frequently, our estimates of variation across weeks of the study could be underestimated and longitudinal studies might have performed better in comparison with cross-sectional studies. However, after working in this region for a number of years we are confident that we are not grossly underestimating the variance

Table 4 Comparison of results from mixed effects univariate regression models for 19 villages in Ecuador, 2004–2007

	Full data set		Longitudinal data sets ($n = 164$)		Cross-sectional data sets ($n = 176$)	
Effect	β estimate	95% CI	Number successfully run	Sampling variance for coefficient estimates	Number successfully run	Sampling variance for coefficient estimates
Individual characteristics						
Age						
<5	Ref (0)					
⩾5	−1.66	(−1.78, −1.54)	151	0.62	174	0.62
Gender						
Male	Ref (0)					
Female	0.22	(0.06, 0.39)	156	0.33	174	0.42
Household characteristics						
Highest level of education						
<6 years	Ref (0)					
⩾6 years	−0.12	(−0.26, 0.03)	137	0.63	163	0.58
Whether anyone has a stable job						
No	Ref (0)					
Yes	0	(−0.16, 0.16)	142	0.85	168	0.52
Ownership index	−0.75	(−1.33, −0.16)	135	8.09	176	6.58
Housing construction	−0.02	(−0.09, 0.04)	154	2.53	176	1.89
Number of people in house	−0.05	(−0.07, −0.02)	162	0.04	176	0.02
Improved sanitation facility						
No	Ref (0)					
Yes	0.07	(−0.07, 0.22)	143	0.98	171	0.5

The full data set contains all weeks and villages and contains random effects for individual and village. The village-level longitudinal data sets are of varying lengths, chosen to have a similar number of observations as the weekly cross-sectional data. Village-level longitudinal analyses contain a random effect for individual and weekly cross-sectional analyses contain a random effect for village. 95% CI: 95% confidence interval.

in the variables used for comparison in this study, including measures of household SES and owning an improved sanitation facility, by measuring them annually.

Certain characteristics of our analysis and our data set may have implications with respect to interpretations of the results. Examples include: (i) this analysis did not stratify by age group, although risk factors may affect these groups differently; (ii) we did not include some variables such as breastfeeding that are commonly associated with diarrhoeal disease; and (iii) our site had a low incidence of diarrhoea compared with other areas of the world. Compared with a review of publications that reported incidence rates in children <5 years of age from 20 countries, the rate found in our region (0.67 per person-year) was low, comparable with data from Indonesia, Thailand and Malaysia.[15] However, our overall conclusions are statements about estimation precision, not effect sizes. If researchers use similar variables, which vary more spatially than temporally then, all other things being equal, a cross-sectional study will likely be a more effective design for measuring effect estimates than a longitudinal study. Additionally, the fact that the direction of risk factor effect estimates observed in our full study are in general agreement with what has been published provides evidence towards the validity of our data set. The one association that is not in general agreement is that in our data set larger numbers of household members are shown to be protective for diarrhoea. Although this association is statistically significant, the small point estimate calls into question its practical significance.

There are instances where singular cross-sectional studies are not the ideal study design. When there is evidence for the presence of secular changes in environmental or social processes, e.g. when village socio-economic level changes or when improvements are made to water treatment or sanitation facilities, repeated cross-sectional studies should be considered. Regardless of the risk factor distribution over time and space, longitudinal data collection still has a

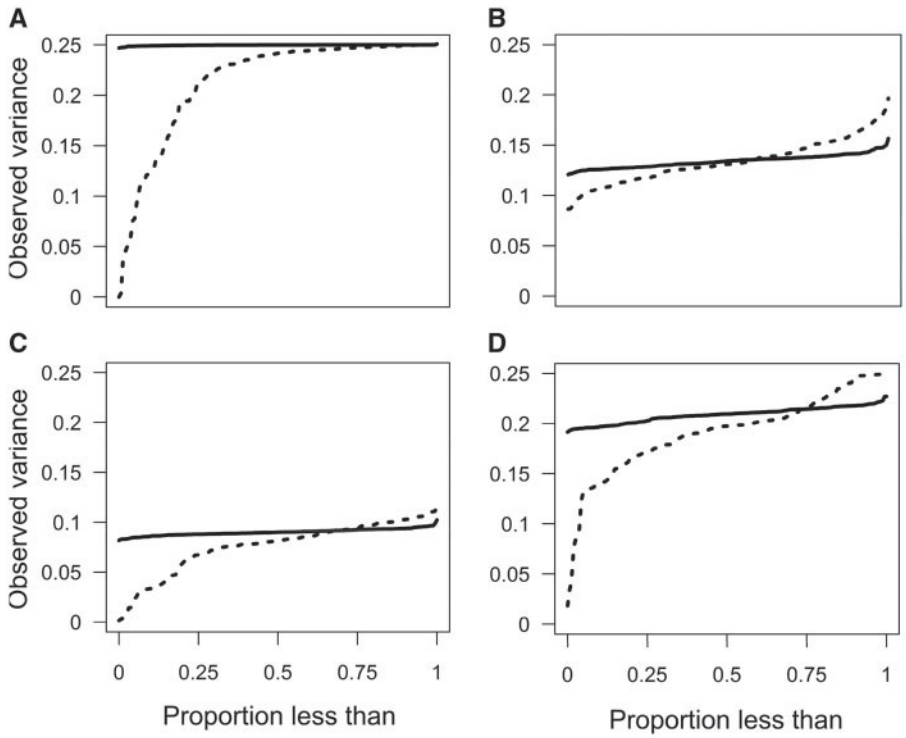

Figure 2 Quantile functions for the observed sample variances comparing the full data set to cross-sectional (solid) and longitudinal (dashed) data sets for variables (A) improved sanitation, (B) age <5 years, (C) construction score and (D) stable job

place when particular interest is paid to how individuals respond to interventions or other secular shifts.

Cross-sectional data sets that provide a broad representation of the target population may provide more efficient estimates of risk than longitudinal data sets that generally contain small number of individuals over a smaller geographic area. Our analysis suggests that for assessing diarrhoeal disease risks associated with water sanitation and hygiene, public health officials making policy decisions should take advantage of cross-sectional data sets such as the DHS. Epidemiologists should also re-evaluate the value of the cross-sectional design in different sites and with respect to different diseases, and consider, when appropriate, utilizing studies of a larger segment of the population at a given point in time over longitudinal studies of a smaller segment of the population in a given location.

Funding

This work was supported by the National Institute of Allergy and Infectious Diseases at the National Institutes of Health (RO1 AI050038).

Acknowledgements

We would like to thank the Ecologia, Desarrollo, Salud, y Sociedad (EcoDeSS) field team for their invaluable contributions to collecting the data, with special thanks for the members of the Asociacion de Promotores de la Salud, Borbon (APSAB) who were responsible for collecting the active surveillance data used in this study.

Conflict of interest: None declared.

KEY MESSAGE

- When choosing between cross-sectional and longitudinal studies, researcher and policy makers should take into account whether risk factors vary more across space at a fixed moment in time than at a fixed location across time. Using a large diarrhoeal disease surveillance data set collected in Ecuador we found that cross-sectional studies produced more precise estimates of risk factor effects than nearly equivalent longitudinal studies, particularly for those collected at a household level. Public health officials making policy decisions should take advantage of cross-sectional data sets such as the DHS for assessing diarrhoeal disease risks associated with water sanitation and hygiene.

References

[1] Lopez AD, Mathers CD, Ezzati M, Jamison DT, Murray CJ. Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. *Lancet* 2006;**367:**1747–57.

[2] Kosek M, Bern C, Guerrant RL. The global burden of diarrhoeal disease, as estimated from studies published between 1992 and 2000. *Bull World Health Organ* 2003; **81:**197–204.

[3] Esrey SA. Water, waste, and well-being: a multicountry study. *Am J Epidemiol* 1996;**143:**608–23.

[4] Gunther I, Fink G. Water, Sanitation and Children's Health: evidence from 172 DHS Surveys. The World Bank Development Economics Prospects Group; 2010; Policy Research Working Paper 5275.

[5] Pruss A, Kay D, Fewtrell L, Bartram J. Estimating the burden of disease from water, sanitation, and hygiene at a global level. *Environ Health Perspect* 2002;**110:**537–42.

[6] Rothman KJ, Greenland S, Lash TL. Types of epidemiologic studies. In: Rothman KJ, Greenland S, Lash TL (eds). *Modern Epidemiology*. 3rd edn. Philadelphia: Lippincott Williams and Wilkins, 2008, pp. 87–99.

[7] Ho PM, Peterson PN, Masoudi FA. Evaluating the evidence: is there a rigid hierarchy? *Circulation* 2008;**118:**1675–84.

[8] Doocy S, Burnham G. Point-of-use water treatment and diarrhoea reduction in the emergency context: an effectiveness trial in Liberia. *Trop Med Int Health* 2006; **11:**1542–52.

[9] Barros AJ, Ross DA, Fonseca WV, Williams LA, Moreira-Filho DC. Preventing acute respiratory infections and diarrhoea in child care centres. *Acta Paediatr* 1999;**88:** 1113–18.

[10] Hellard ME, Sinclair MI, Forbes AB, Fairley CK. A randomized, blinded, controlled trial investigating the gastrointestinal health effects of drinking water quality. *Environ Health Perspect* 2001;**109:**773–78.

[11] Zwane AP, Zinman J, Van Dusen E *et al*. Being surveyed can change later behavior and related parameter estimates. *Proc Natl Acad Sci USA* 2011;**108:**1821–26.

[12] Eisenberg JN, Cevallos W, Ponce K *et al*. Environmental change and infectious disease: how new roads affect the transmission of diarrheal pathogens in rural Ecuador. *Proc Natl Acad Sci USA* 2006;**103:**19460–65.

[13] Bartley M. Health inequality: an introduction to theories, concepts, and methods. Cambridge: Polity Press, 2004.

[14] UN-Water. Progress on sanitation and drinking-water: 2010 update. Geneva and New York: WHO/UNICEF Joint Monitoring Programme, 2010.

[15] Kosek M, Bern C, Guerrant RL. The global burden of diarrhoeal disease, as estimated from studies published between 1992 and 2000. *Bull World Health Organ* 2003; **81:**197–204.