Optimal Recall Period for Caregiver-reported Illness in Risk Factor and Intervention Studies: A Multicountry Study

Benjamin F. Arnold*, Sebastian Galiani, Pavani K. Ram, Alan E. Hubbard, Bertha Briceño, Paul J. Gertler, and John M. Colford, Jr

* Correspondence to Dr. Benjamin F. Arnold, Division of Epidemiology, School of Public Health, University of California, Berkeley, 101 Haviland Hall, MC7538, Berkeley, CA 94720 (e-mail: benarnold@berkeley.edu).

Initially submitted February 28, 2012; accepted for publication May 25, 2012.

Many community-based studies of acute child illness rely on cases reported by caregivers. In prior investigations, researchers noted a reporting bias when longer illness recall periods were used. The use of recall periods longer than 2–3 days has been discouraged to minimize this reporting bias. In the present study, we sought to determine the optimal recall period for illness measurement when accounting for both bias and variance. Using data from 12,191 children less than 24 months of age collected in 2008–2009 from Himachal Pradesh in India, Madhya Pradesh in India, Indonesia, Peru, and Senegal, we calculated bias, variance, and mean squared error for estimates of the prevalence ratio between groups defined by anemia, stunting, and underweight status to identify optimal recall periods for caregiver-reported diarrhea, cough, and fever. There was little bias in the prevalence ratio when a 7-day recall period was used (<10% in 35 of 45 scenarios), and the mean squared error was usually minimized with recall periods of 6 or more days. Shortening the recall period from 7 days to 2 days required sample-size increases of 52%–92% for diarrhea, 47%–61% for cough, and 102%–206% for fever. In contrast to the current practice of using 2-day recall periods, this work suggests that studies should measure caregiver-reported illness with a 7-day recall period.

bias; diarrhea; cough; fever; outcomes assessment; outcome measurement errors; survey methodology

Abbreviations: MSE, mean squared error; PR, prevalence ratio.

Community-based epidemiologic studies often use caregiver-reported symptoms to measure acute cases of diarrhea, respiratory infections, or other child illnesses because it is often impractical to use more objective measures. For example, a case of diarrhea is often defined as 3 or more loose or watery stools in 24 hours or any loose stool with blood (1), and a case of acute lower respiratory infection is defined using World Health Organization guidelines that include cough or difficulty breathing with a raised respiratory rate (2). Symptom-based definitions have reasonably good sensitivity and specificity compared with medical diagnosis over recall periods less 14 days but are recognized to be imperfect (see Katz et al. (3) for a summary of articles and Kalter et al. (4) and Kalter (5) for validation methods).

One problem with symptom-based case definitions is that they are subject to reporting errors. Reporting errors can result from failure to remember illness, “forward telescoping” illness episodes that occurred before the recall period into the recall period, confusion that results from poor-quality questionnaires, and poor interviewer-participant communication (6–8). Most studies in which interviewers asked caregivers about their child’s symptoms retrospectively showed a consistent pattern: Symptom reporting began to decline after 2–3 days (9–16). Reporting decline can be lower for more severe episodes of diarrhea (10, 16) and when the questions apply to younger children (16).

Although studies cannot identify whether the reporting decline seen with longer recall periods is due to overreporting of symptoms close to the interview or underreporting of symptoms further from the interview, investigators generally assume that respondents forget about illness in the past; measures over the previous 2–3 days, therefore, are believed to
be unbiased measures of disease (10, 13, 15, 16). Studies that have provided guidance on optimal recall periods for illness symptoms have focused on recall periods that minimize measurement bias and have discouraged the use of recall periods longer than 2–3 days (13, 15, 16).

An exclusive focus on bias in caregiver-reported illness does not account for the variance of parameters of interest in risk factor and intervention studies. The choice of recall period used for outcomes to estimate a quantity like the prevalence ratio (PR) should depend on both the bias and variance of estimates generated with different recall periods. In many applications, a slightly biased but more precise estimate is preferable to an unbiased but imprecise estimate because on average, it will be more accurate. This is the bias-variance tradeoff often used to select between estimators (17, 18). (See Clarke et al. (19) for an example in the context of recall errors.) The use of longer symptom recall periods to measure illness increases recall error, but it also increases the number of cases detected and, with period prevalence measures, reduces outcome variability over the measurement period (20). Higher outcome prevalence and lower variability can reduce the required study size and related costs. In the context of symptom recall, Schmidt et al. (20) noted the increased statistical power and bias due to measurement error with longer recall periods but did not provide empirically-based guidance for appropriate recall periods given the implied bias-variance tradeoff.

The objective of the present study was to determine the optimal recall period for caregiver-reported illness measurement while accounting for both bias and variance. We estimated the PR between groups using recall periods of 1–14 days and used a function that combines bias and variance—the mean squared error (MSE)—to identify recall periods with the best combination of bias and variance. We based our calculations on data from 5 cohorts that included a total of 12,191 children less than 24 months of age. The cohorts and symptoms we considered (diarrhea, persistent cough, and fever) captured a wide range of prevalence, duration, and reporting conditions. Although this article focuses on empirical estimates and power calculations based on the PR, we extend these results to the risk difference in Web Appendices 1–3 (available at http://aje.oxfordjournals.org/).

We provide specific guidance for investigators on appropriate recall periods with which to measure caregiver-reported morbidity when the outcomes are used to compare groups, as is commonly done in an intervention study. We expect that the results will apply not only to studies conducted in the low-income countries from which these examples are drawn but also broadly in the many situations in which caregiver-reported illness is the outcome.

MATERIALS AND METHODS

Study population and outcome measurement

The cohorts were enrolled as part of a multicountry study of the World Bank’s Water and Sanitation Program Global Scaling Up Initiatives in rural sanitation and hand washing promotion. Households were randomly sampled within rural villages from a list of all households with at least 1 child less than 24 months of age. The cohorts were distributed over wide geographic areas and included a large number of rural sampling clusters (usually villages): 80 in Himachal Pradesh, India, 80 in Madhya Pradesh, India, 160 in Indonesia, 211 in Peru, and 110 in Senegal. The study baseline reports include additional details of the study design and sample characteristics (21–24).

In the present analysis, we used data from baseline surveys conducted in 2008 and 2009 in which interviewers asked caregivers about their child’s health in a household interview using a daily calendar that included the 14 days before the interview (25) (Web Appendix 4). The calendar included symptoms of gastrointestinal illness (stomach cramps, nausea, vomiting, 3 or more stools in 24 hours, loose or watery stools, blood or mucus in the stool), respiratory illness (constant cough, congestion/coryza), nonspecific symptoms (fever, refusal to eat), and other symptoms unrelated to the interventions (scrapes/abrasions, itchy skin or scalp). We classified a child as ill with diarrhea if the child had 3 or more loose or watery stools in 24 hours or 1 or more stools with blood or mucus (1). Interviewers asked caregivers about constant cough and fever using common terms in the local language. Interviewers asked caregivers to estimate the start date of each symptom (25). Field teams used identical symptom calendars in all 5 cohorts. All research was reviewed and approved by the Western Institutional Review Board in Olympia, Washington, and all caregivers provided informed consent.

From the symptom calendars we abstracted 2 common measures of disease. The first was daily prevalence, which was defined as the number of days ill divided by the total days of observation (each survey contributes multiple measurements per child, 1 for each day of recall). The second was period prevalence, in which a child is classified as ill if he/she had the symptom at anytime during the recall period (1 measurement per survey, whatever the recall period).

Overview of the analysis

Here, we outline the 3 main steps of the analysis; the sections below include the details and rationale for each step. First, we chose 3 child characteristics (anemic, stunted, underweight) to stratify the data and compare illness prevalence between children with or without each characteristic. For each risk factor, we calculated the PR using recall periods of between 1 and 14 days. Second, we calculated bias in the PR estimated with each recall period by assuming that the estimates based on a 2-day recall period were unbiased. Third, we calculated the variance and MSE of the estimators by bootstrapping the data sets and re-estimating the PR in each iteration. We repeated our analysis in each cohort. We conducted our analysis using R, version 2.13.1 (R Project, Vienna, Austria) and Stata, version 12 (StataCorp LLP, College Station, Texas). Replication files are available from the authors.

Choice of risk factors and parameter of interest

We chose 3 dichotomous characteristics to use as risk factors for illness, and we compared illness risk between...
groups: anemic (hemoglobin <110 g/L) versus not anemic, stunted (height-for-age z score < –2) versus not stunted, and underweight (weight-for-age z score < –2) versus not underweight. We chose these risk factors to demonstrate the impact of recall period on the PR because we hypothesized that children who were anemic, stunted, or underweight would be at elevated risk for diarrhea, cough, and fever (26).

We chose the PR as the parameter of interest because it is commonly estimated in observational studies and randomized controlled trials (27). For episodic diseases with short mean durations, the prevalence over a period of 1–14 days (a restricted risk period) is a reasonable estimate of risk over the period, and the PR estimates the risk ratio (28).

The PR can be defined in terms of the observed data with different recall periods. Let \( Y(t) \) be a binary indicator variable for reported daily illness (e.g., diarrhea), where \( t \) indexes the day of recall before the interview (\( t = 1, 2, \ldots, 14 \)). The outcome definition depends on the choice of daily prevalence versus period prevalence and on recall period length. Period prevalence outcomes are calculated as: \( Y^*(T) = \max[Y(1), Y(2), \ldots, Y(T)] \), which is a binary indicator that the individual was ill on any day back to day \( T \). Daily prevalence outcomes simply include multiple observations for each individual back to day \( T \): \( \hat{Y}(T) = [Y(1), Y(2), \ldots, Y(T)] \). Let \( X \) be a binary risk factor (e.g., 1 = anemic, 0 = not anemic) or a binary indicator of intervention status (1 = treated, 0 = untreated). We can then define \( \theta_T \), which estimates PR over recall period \( T \) using, for example, the period prevalence ratio:

\[
\theta_T = \frac{E[Y^*(T)|X = 1]}{E[Y^*(T)|X = 0]}. \tag{1}
\]

Common estimators for \( \theta_T \) include a nonparametric ratio of means (used in this analysis) or a log-binomial regression model (29). In Web Appendix 5, we show that controlling for possible confounders with log-binomial regression does not change our results.

**Analysis of the bias-variance tradeoff**

We investigated the tradeoff between bias and variance in the context of different recall periods by comparing PR estimates based on different recall periods (\( \theta_T \)) using the MSE. The MSE is a standard objective criterion often used to select estimators with good overall accuracy based on bias and variance (17, 18). In this analysis, we assumed that the PR estimated with a 2-day recall period was an unbiased estimate of the true PR (i.e., \( \theta = \theta_2 \)) because consistent with earlier studies (9–16), there was little evidence of symptom underreporting until after day 2 (Figure 1). The MSE for each recall period \( T \) is then defined as:

\[
\text{MSE}(\hat{\theta}_T) = E[(\hat{\theta}_T - \theta)^2]. \tag{2}
\]
The MSE can be decomposed into separate terms for bias and variance (17):

$$\text{MSE}(\hat{\theta}_T) = E[(\hat{\theta}_T - \theta)^2] + E[(\hat{\theta}_T - E[\hat{\theta}_T])^2] = \text{Bias}(\hat{\theta}_T, \theta)^2 + \text{Var}(\hat{\theta}_T). \quad (3)$$

To calculate the variance and MSE of $\hat{\theta}_T$ for each recall period, we used a clustered bootstrap method that resampled independent village clusters with replacements (following the original sampling design) and 10,000 iterations. In each bootstrap iteration, we re-estimated the PR of each risk factor over each of the 14 recall periods ($\hat{\theta}_1, \ldots, \hat{\theta}_{14}$) using equation 1. We used the standard deviation of the bootstrapped PR estimates to calculate the standard error for the PR estimated with each recall period.

The comparison of MSE values between different estimates is only sensible for estimates of the same quantity, $\theta$. 

### Table 1. Cohort Characteristics for Children 0–24 Months of Age in Himachal Pradesh, Madhya Pradesh, Indonesia, Peru, and Senegal, 2008–2009

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Himachal Pradesh, India ($n = 2,058$)</th>
<th>Madhya Pradesh, India ($n = 2,153$)</th>
<th>Indonesia ($n = 2,098$)</th>
<th>Peru ($n = 3,708$)</th>
<th>Senegal ($n = 2,174$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Mean (SD)</td>
<td>% Mean (SD)</td>
<td>% Mean (SD)</td>
<td>% Mean (SD)</td>
<td>% Mean (SD)</td>
<td>% Mean (SD)</td>
</tr>
<tr>
<td>General</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age, months</td>
<td>11.9 (6.8)</td>
<td>11.7 (6.5)</td>
<td>11.9 (6.3)</td>
<td>11.9 (6.9)</td>
<td>11.9 (6.8)</td>
</tr>
<tr>
<td>Female sex</td>
<td>47.8</td>
<td>48.9</td>
<td>49.2</td>
<td>48.5</td>
<td>46.2</td>
</tr>
<tr>
<td>Currently breastfed</td>
<td>66.1</td>
<td>86.5</td>
<td>80.0</td>
<td>83.8</td>
<td>80.5</td>
</tr>
<tr>
<td>Anemic (hemoglobin &lt;110 g/L)$^a$</td>
<td>35.6</td>
<td>30.6</td>
<td>71.0</td>
<td>40.9</td>
<td>90.8</td>
</tr>
<tr>
<td>Height-for-age z score $&lt; -2^a$</td>
<td>26.1</td>
<td>44.3</td>
<td>20.8</td>
<td>27.9</td>
<td>11.3</td>
</tr>
<tr>
<td>Weight-for-age z score $&lt; -2^a$</td>
<td>19.4</td>
<td>47.9</td>
<td>13.3</td>
<td>10.1</td>
<td>9.5</td>
</tr>
<tr>
<td>Household</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Electricity</td>
<td>94.8</td>
<td>80.7</td>
<td>98.8</td>
<td>69.9</td>
<td>37.7</td>
</tr>
<tr>
<td>Soil (dirt or clay) floor</td>
<td>31.8</td>
<td>76.4</td>
<td>24.8</td>
<td>69.2</td>
<td>25.0</td>
</tr>
<tr>
<td>Cooks with biofuel</td>
<td>79.4</td>
<td>90.8</td>
<td>73.8</td>
<td>68.6</td>
<td>76.6</td>
</tr>
<tr>
<td>JMP improved water source</td>
<td>89.8</td>
<td>84.7</td>
<td>87.3</td>
<td>75.0</td>
<td>69.2</td>
</tr>
<tr>
<td>JMP improved sanitation</td>
<td>32.0</td>
<td>13.7</td>
<td>48.7</td>
<td>40.0</td>
<td>70.8</td>
</tr>
<tr>
<td>Diarrhea</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24-hour point prevalence</td>
<td>2.5</td>
<td>6.5</td>
<td>4.1</td>
<td>8.3</td>
<td>7.9</td>
</tr>
<tr>
<td>48-hour period prevalence</td>
<td>3.4</td>
<td>8.8</td>
<td>4.8</td>
<td>10.4</td>
<td>9.9</td>
</tr>
<tr>
<td>7-day period prevalence</td>
<td>6.3</td>
<td>14.4</td>
<td>8.2</td>
<td>18.3</td>
<td>14.4</td>
</tr>
<tr>
<td>Episode duration$^b$</td>
<td>2.93 (1.97)</td>
<td>3.79 (2.17)</td>
<td>2.98 (2.06)</td>
<td>3.66 (2.26)</td>
<td>3.95 (3.06)</td>
</tr>
<tr>
<td>Cough</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24-hour point prevalence</td>
<td>15.0</td>
<td>7.9</td>
<td>18.0</td>
<td>23.5</td>
<td>10.7</td>
</tr>
<tr>
<td>48-hour period prevalence</td>
<td>17.0</td>
<td>9.8</td>
<td>19.9</td>
<td>25.3</td>
<td>12.4</td>
</tr>
<tr>
<td>7-day period prevalence</td>
<td>24.4</td>
<td>15.0</td>
<td>27.0</td>
<td>34.1</td>
<td>17.3</td>
</tr>
<tr>
<td>Episode duration$^b$</td>
<td>3.90 (1.92)</td>
<td>3.55 (2.02)</td>
<td>4.59 (2.74)</td>
<td>5.04 (2.58)</td>
<td>4.70 (2.79)</td>
</tr>
<tr>
<td>Fever</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24-hour point prevalence</td>
<td>12.8</td>
<td>8.4</td>
<td>12.9</td>
<td>6.1</td>
<td>18.0</td>
</tr>
<tr>
<td>48-hour period prevalence</td>
<td>16.2</td>
<td>12.0</td>
<td>17.0</td>
<td>9.0</td>
<td>23.4</td>
</tr>
<tr>
<td>7-day period prevalence</td>
<td>29.2</td>
<td>21.8</td>
<td>31.5</td>
<td>23.5</td>
<td>39.8</td>
</tr>
<tr>
<td>Episode duration$^b$</td>
<td>3.43 (1.92)</td>
<td>3.59 (2.37)</td>
<td>3.34 (2.30)</td>
<td>2.28 (1.49)</td>
<td>4.17 (2.53)</td>
</tr>
</tbody>
</table>

Abbreviations: JMP, World Health Organization/UNICEF Joint Monitoring Programme for water supply and sanitation definitions (34); SD, standard deviation.

$^a$ Some children in the cohorts had missing measurements for hemoglobin, height, and weight. Sample sizes for hemoglobin, height, and weight, respectively: Himachal Pradesh = 1580, 1972, and 1934; Madhya Pradesh = 1843, 2029, and 2067; Indonesia = 1587, 2094, and 2086; Peru = 3050, 3679, and 3702; and Senegal = 1597, 2155, and 2160. z scores were calculated using the World Health Organization 2006 standards.

$^b$ Estimated from episodes that terminated in the 7 days before the interview to avoid length-biased sampling (11, 32).
In general, the PR estimated with different recall periods estimates inherently different quantities if using period prevalence; however, the PR is approximately constant over recall periods of up to 14 days under typical incidence and duration conditions observed for caregiver-reported illness (Web Appendix 1).

Relative sample size calculations

We used a sample-size equation for binary outcomes with repeated measures (30) to demonstrate the influence of recall period on the sample size required to estimate a constant relative reduction in illness (diarrhea, cough, fever) given the outcome prevalence in the cohorts (Web Tables 1–3). For daily prevalence calculations, a single survey measurement includes an observation for each day of recall, so we additionally specified the correlation between each day of recall in the power calculations (based on empirical estimates from the data) (Web Tables 4–6). We present the results in terms of relative sample size required with a 7-day recall period as the base. The relative sample size calculations are invariant to the magnitude of the PR, the number of visits per child, and outcome correlation between repeated visits.

RESULTS

Study population

The analysis included 12,191 children who were 0–24 months of age. The 5 cohorts represent a broad range of cultural, environmental, and illness conditions (Table 1). For example, the percentage of households with a soil floor ranged from 24.8% (Indonesia) to 76.4% (Madhya Pradesh), and the 7-day period prevalence of diarrhea ranged between 6.3% (Himachal Pradesh) and 18.3% (Peru). Of the 3 risk factors that we considered in this analysis, anemia was uncorrelated with stunting ($\tau_b = -0.07$) and underweight ($\tau_b = -0.07$), whereas stunting and underweight were moderately correlated ($\tau_b = 0.43$) (31).

Empirical symptom reporting patterns and prevalence ratio estimates

Without reporting error, one would expect the daily prevalence of an outcome such as diarrhea to be stable over a 14-day period in large cohorts because it is reasonable to assume that the populations were in a steady state over the 14-day measurement period (32). Across symptoms and cohorts, the estimated daily prevalence generally declined for recall periods greater than 2–3 days, and the decline was large for recall periods greater than 7 days (Figure 1). The daily prevalence increased between days 2 and 7 in only 1 measurement (fever in Peru). By 14 days of recall, the estimated daily point prevalence was a small fraction of

Figure 2. Relative daily reporting in the 5 cohorts for diarrhea (top), cough (middle), and fever (bottom) in Himachal Pradesh (India HP), Madhya Pradesh (India MP), Indonesia, Peru, and Senegal, 2008–2009. Recall day 0 is the day of the survey. The relative daily reporting is calculated as the daily point prevalence divided by the mean point prevalence for days 1 and 2. The cohorts included children 0–24 months of age.
the 2-day average for diarrhea (16%–29%), cough (17%–30%) and fever (26%–47%). The patterns and levels of relative daily reporting were strikingly similar across cohorts and symptoms (Figure 2). There was some heaping of symptom reports on day 7 across cohorts, which was particularly evident for fever.

In the range of comparisons that we considered, we observed situations with both nondifferential and differential reporting errors by risk factor. Web Figure 1 shows examples using diarrhea. There was no consistent pattern of differential reporting by population, outcome, or risk factor, but the extent of differential reporting, if present, increased with longer recall periods; short recall periods (<3 days) were more protected from differential reporting errors. Despite large drops in the prevalence of reported daily illness with longer recall and the potential for differential reporting errors, the empirical PR was relatively stable over different recall periods in the scenarios we considered for diarrhea, cough, and fever (Web Figures 2–4).

**Optimal recall periods based on a bias-variance tradeoff**

Bias in the PR, if present, tended to increase with longer recall periods, whereas variance of the PR tended to decrease (Figure 3). The percentage of bias with a 7-day recall period was less than 10% in 30 of 45 scenarios and was broadly similar for diarrhea (range, 1%–30%), cough (range, %–13%), and fever (range, <1%–14%) (Web Figures 5–7). Bias, if present, could be toward or away from the null. Compared with a 2-day recall period, the standard error of the PR estimated with a 7-day recall period was smaller (more than 20% smaller in 35 of 45 scenarios) for diarrhea (range, −1% to −52%), cough (range, −1% to −29%), and fever (range, −21% to −57%) (Web Figures 5–7). Results based on daily prevalence were highly consistent with the period prevalence results (Web Appendix 3).

The empirical patterns in the MSE varied substantially across countries, outcomes, and comparison groups (Figure 3 and Web Figures 8–10). Despite this variation, the MSE tended to decline with longer recall periods and was minimized in most scenarios with a recall period of 6 days or longer. In 9 of 45 scenarios, the MSE increased and was larger with a 7-day recall period than with a 2-day recall period.

**The effect of recall period and outcome measure on study sample size**

Shortening the recall period from 7 days to 2 days reduced outcome prevalence and led to smaller absolute differences between groups for a constant PR; this consequently required sample size increases of 52%–92% for diarrhea, 47%–61% for cough, and 102%–206% for fever (Figure 4). The use of
daily prevalence required larger sample sizes compared with period prevalence for more than 1 day of recall. For example, the use of daily prevalence instead of period prevalence with a 7-day recall period required sample size increases of 35%–52% for diarrhea, 29–42% for cough, and 36%–47% for fever (Web Figure 11).

DISCUSSION

Summary

In the present analysis of 12,191 children 0–24 months of age from 5 cohorts that include a wide range of symptom prevalence, environmental conditions, and cultural conditions, we found consistent evidence that caregiver-reported symptoms declined after 3 days of recall, and estimates typically fell below 50% of the 1–2-day average after 7 days of recall (Figures 1 and 2). If the goal of the study is to obtain an unbiased measure of disease burden, such as in demographic and health surveys, then recall should be limited to 2–3 days; the use of recall periods greater than 3 days will bias prevalence levels downward under the assumption that reporting is less biased closer to the interview. However, despite larger reporting errors over longer recall periods and the possibility of differential recall between comparison groups, we estimated the PR with little bias with up to 7 days of recall. Furthermore, a recall period of 6 days or longer minimized the MSE in most scenarios we considered (Figure 3 and Web Figures 8–10).

Our analyses suggest that studies can increase statistical power through the use of a 7-day recall period compared with a 2-day recall period at little cost in terms of bias. Calculations from the 5 cohorts demonstrated required sample size increases of 47%–206% with the use of a 2-day recall period compared with a 7-day recall period (Figure 4). Studies in which a recall period of 2–3 days has been recommended as the optimal period for symptom measurement (13, 15, 16) have failed to account for the loss of statistical efficiency with shorter recall periods, which in the present analysis outweighs the small amount of bias introduced with longer recall periods when estimating the PR between groups. Our analysis also demonstrated that studies can gain additional efficiency if they use period prevalence rather than daily prevalence: With a 7-day recall period, the use of daily prevalence required 29%–52% larger samples than did period prevalence across symptoms and cohorts (Web Figure 11).

Differential recall errors

Errors in caregiver-reported illness are a form of misclassification error. If caregivers fail to remember or report
illness, then a child is incorrectly classified as well when she or he is in fact ill. If reporting errors are mainly due to the underreporting of symptoms (forgetting) and are nondifferential by group, the PR will be unbiased with any recall period; without those 2 conditions, the bias is unpredictable (Web Appendix 2). To our knowledge, only 1 study has evaluated whether reporting errors were differential by randomized treatment, and it found errors to be nondifferential despite weekly visits and intense household water treatment promotion (16). In the present analysis, some of the PR calculations used outcomes that were reported differentially by risk factor group (Web Figure 1). Despite the presence of differential reporting errors, we observed relatively consistent estimates of the diarrhea PR over a recall period of 1–14 days for diarrhea, cough, and fever (Web Figures 2–4). We would expect that bias from differential reporting errors to be most problematic in nonrandomized, cross-sectional studies because 1) randomization will balance, on average, characteristics that could lead to differential reporting, and 2) longitudinal studies could remove the differential reporting errors by conditioning effect estimates on the caregiver (e.g., with household-level fixed-effects).

**Recommendations for practice**

We recommend that investigators collect daily illness information within 7 days of recall. This suggestion follows from 3 of our results. First, daily symptom reporting declined precipitously after 7 days of recall across symptoms and cohorts, suggesting large errors beyond 7 days of recall (Figure 1). Second, there was relatively little bias in PR estimates with 7 days of recall, and the MSE was minimized in most scenarios with close to 7 days of recall (Figure 3 and Web Figures 8–10). Finally, the gains in statistical efficiency were largest between 2 and 7 days of recall, with relatively small efficiency gains over longer recall periods (Figure 4). As a final pragmatic consideration, 7 days is a standard unit of time in many cultures (1 week) and is relatively easy to standardize across studies (compared with 6 or 8 days, for example).

With a prespecified analysis protocol, investigators can use the daily symptom data to test whether the reporting error patterns are differential by group. If errors are nondifferential, then the analysis could proceed with a 7-day recall period. If survey time is limited, we recommend that investigators measure symptoms daily for 2 days before the interview and ask an additional question to measure 7-day period prevalence. Differential reporting error can be detected with this more limited set of information using the ratio of prevalent cases to terminated cases under the assumption that episode duration is similar in the 2 groups (11). Web Appendix 4 includes example survey instruments.

**Caveats**

Consistent with earlier analyses of caregiver recall error (10, 15, 16), our analysis assumed that symptoms reported close to the interview were more accurate than reports further in the past. If this assumption is wrong and participants overreport symptoms close to the interview, then longer recall periods would be less biased (and more favorable) than we estimated in this analysis. Psychological research has found that event recall is less accurate over time (see references 6–8 for reviews in the context of survey measurement). However, Freij et al. (33) compared daily interviews of child morbidity with 2-week interviews in Addis Ababa and found that the 2-week interview resulted in overreporting of recent illness and underreporting of past illness relative to the daily interview measurements. This pattern could have occurred by forgetting events in the past or by misplacing days of illness closer to the interview that actually occurred further in the past. Whatever the cause, the 2 effects approximately cancelled each other out, so a prevalence measure that captured the entire period was approximately unbiased. Replicating the study by Freij et al. in other settings would be a useful future contribution because it would provide empirical evidence about whether it is reasonable to assume that symptom reporting in the 2 days before the interview is unbiased.

A second caveat is that our primary analysis estimated the PR with a nonparametric ratio of means, which did not account for possible confounding or effect modification of the relationship between the exposures of interest and outcomes. Nonetheless, adjusting for possible confounders in this data set by estimating the PR with log-binomial regression (29) in a secondary analysis did not change our results (Web Appendix 5).

Finally, we showed that the use of period prevalence is more statistically powerful than daily prevalence, which corroborates an earlier finding based on simulation (20). However, the choice between the measures should depend on whether investigators expect the intervention to act primarily through reduced episode duration (27). If an intervention reduces prevalence only through episode duration, then daily prevalence measures would be preferred because period prevalence can be biased toward the null hypothesis of no difference between groups (20).

**External validity**

We expect that our results apply to a broad range of epidemiologic field studies because our analysis included large samples drawn from diverse cultural and environmental conditions. We found highly consistent results across symptoms that cover many combinations of prevalence and episode duration. The empirical comparisons we considered included situations with both nondifferential and differential recall errors between groups. The broad consistency in reporting patterns, bias, MSE, and statistical power that we observed across the wide range of analysis scenarios suggest that our findings are likely quite general for caregiver-reported illness in young children. We repeated the calculations for the risk difference and our findings were broadly consistent with the results for the prevalence ratio (Web Appendix 3). Because the cohorts included children less than 24 months of age, our findings might not be generalizable to older populations.

**Conclusions**

The increased statistical efficiency (measured by markedly reduced sample sizes required) that resulted from the use of
ACKNOWLEDGMENTS

Author affiliations: School of Public Health, University of California, Berkeley, Berkeley, California (Benjamin F. Arnold, Alan E. Hubbard, Paul J. Gertler, John M. Colford, Jr.); Department of Economics, Washington University, Saint Louis, Missouri (Sebastian Galiani); Department of Social and Preventive Medicine, University at Buffalo, Buffalo, New York (Pavani K. Ram); The Water and Sanitation Program, The World Bank, Washington, DC (Bertha Briceño); and Haas School of Business, University of California, Berkeley, Berkeley, California (Paul J. Gertler).

This work was supported by a grant from the Bill and Melinda Gates Foundation to the Water and Sanitation Program at the World Bank.

We thank the impact evaluation teams who led the research efforts in each country. Country teams were led by Sumeeet Patil and Dr. Alicia Salvatore (India), Dr. Lisa Cameron and Dr. Manisha Shaw (Indonesia), Dr. Lucie Heizerling and Paul Wassenich (Senegal), and Alex Orsola-Vidal (with S.G.) (Peru).

The data in this article were collected through research conducted by the Water and Sanitation Program (www.wsp.org). The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the Water and Sanitation Program, the World Bank and its affiliated organizations or those of the Executive Directors of the World Bank or the governments they represent. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of interest: none declared.

REFERENCES


