Published in final edited form as:

Ann Epidemiol. 2010 February ; 20(2): 151–158. doi:10.1016/j.annepidem.2009.10.004.

Comparison of Two Types of Epidemiological Surveys Aimed at Collecting Daily Clinical Symptoms in Community-Based Longitudinal Studies

GWENYTH LEE, MHS, VITALIANO CAMA, PhD, DVM, ROBERT H. GILMAN, MD, LILIA CABRERA, RN, MAYUKO SAITO, MD, WILLIAM CHECKLEY, MD, PhD, and CONTENT INVESTIGATORS

From the Department of International Health, Bloomberg School of Public Health, The Johns Hopkins University, Baltimore, MD (G.L., R.H.G., M.S., V.C., W.C.); Asociacion Benefica PRISMA, Lima, Peru (L.C., M.S.); Universidad Peruana Cayetano Heredia (R.H.G, M.S.), and Division of Pulmonary and Critical Care, School of Medicine, The Johns Hopkins University, Baltimore, MD (W.C.)

Abstract

BACKGROUND—Investigators use prospective community-based studies to collect longitudinal information on childhood diarrhea. The interval in which data are collected may affect the accuracy and interpretation of results. Our objective was to compare data of reported daily clinical symptoms from surveys conducted daily versus twice-weekly surveys.

METHODS—We conducted our study in Lima, Peru, between October and December 2007. We asked 134 mothers to report daily symptoms by using a twice-weekly survey. We conducted daily surveys for the same data on 25% of participants randomly selected each day. We analyzed intersurvey variability by using Cohen's kappa and Signal Detection Theory (SDT).

RESULTS—We collected 6157 and 1181 child-days of data through the twice-weekly and daily surveys, respectively. The prevalence of diarrhea, fever, vomiting, and cough were 6.4%, 1.6%, 2.1%, and 22.7% from the twice-weekly survey and, 6.4%, 2.0%, 2.4%, and 26% from the daily survey, respectively. Despite similar prevalence, 20% of days with reported diarrhea were discrepant between the two surveys, and agreement in the report of diarrhea decreased as time between the interviews increased ($p = .03$).

CONCLUSIONS—Although twice-weekly surveys provide an adequate estimate of diarrheal prevalence compared with daily surveys, the prevalence of other symptoms based on dichotomous questions was lower under the former. Additionally, the agreement between the two surveys in the report of diarrhea decreased as the recall period increased, suggesting that data from daily interviews were of greater quality. Our analysis is a novel application of SDT to measure respondent certainty and bias, from which better inference about the quality of collected data may be drawn.

Keywords

Cohort Studies; Diarrhea–Infantile; Epidemiologic Methods; Signal Detection; Psychological

Address correspondence to: William Checkley, MD, PhD, The Johns Hopkins University, School of Medicine, Division of Pulmonary and Critical Care, 1830 Monument Street, Fifth Floor, Baltimore, MD 21205. Tel.: 443-287-8741; Fax: 410-955-0036. wcheckl1@jhmi.edu.

CONTENT investigators: Professor Dermot Kelleher (Trinity College, Ireland), Dr Henry Windle (Trinity College, Ireland), Professor Jean Crabtree (St. James' University Hospital, United Kingdom), Professor Robert Gilman (Universidad Peruana Cayetano Heredia, Peru), Dr. William Checkley (A.B. Prisma, Peru).

INTRODUCTION

Investigators frequently use prospective longitudinal studies to collect information on childhood diseases. The time interval in which the surveys are conducted commonly varies from daily to weekly or even monthly. Daily visits are more likely to minimize recall biases; however, they are used infrequently because of logistical and financial limitations. Indeed, other factors can also affect the prevalence of reported symptoms. These include what definitions are used (1); how questions are asked (i.e., open-ended or prompted) (2); who is the interviewee (3); and the sequence of questions (2). However, the time interval between interviews is a methodological issue, and few studies have determined the quality or validity of data collected at intervals of 2 days or greater in relation to the actual daily occurrence of symptoms.

The recall period, the time interval, the occurrence of symptoms, and when respondents are asked to report information may have a significant effect on the accuracy of study results. The authors of several studies (4,5) have shown a decrease in the reporting of diarrhea as the recall period increases, presumably because respondents forget and therefore do not report some episodes. Therefore, twice-weekly or weekly data collection has been recommended as an acceptable compromise when conducting community-based studies (4).

Even when the overall prevalence does not change, disagreement between data from initial and validation surveys may signal important loss of information. For example, Byass and Hanlon (4) paired 4-day with 8-day results and found a significant amount of disagreement between them. Ross and colleagues (2) repeated a set of interviews within 48 hours and found a significant degree of discrepancy.

A simple analysis of agreement does not attempt to quantify why such concurrence may be good or poor. However, a methodology frequently used in the field of psychology, signal detection theory (SDT), may be applied for this purpose. SDT is commonly used in studies of human memory (6,7) and has been used in public health research to study bias effects (8,9). It is an application of the receiver operating characteristic (ROC) curve methodology to human sensory discrimination (6), first described by Green and Swets in 1966 (6), who adapted methodology originally developed by radar engineers during the second World War (10). With the use of SDT, a measure of “memory” can be estimated, as can the strength of subject “bias” (a tendency to report or not report a symptom given uncertainty).

SDT uses the terms detectability (d') and criterion (C), respectively, to describe these measures (6,7). Detectability is defined as the probability of correct signal identification minus the probability of an incorrect identification. Criterion is the extent to which one response is more probable than another. For example, when asked whether a child was irritable during a given previous period, a subject might have poor memory of the event (low detectability) but a high tendency to say “yes” in the face of that uncertainty (a low criterion). In terms of the ROC curve, the detectability is a measure of the distance between the line of no discrimination and the curve, and criterion indicates the position along the response curve. Overall agreement, as measured either by Cohen’s kappa, or ROC through sensitivity, specificity and the area under the curve (AUC), might be driven by either or both of the effects of recognition and bias. In this study, we compared results from data collected through daily and twice-weekly surveys by using kappa statistics and ROC curves and also discuss the potential strengths of the application of SDT in data derived from subject interviews.

MATERIALS AND METHODS

Study Population

The study site was the urban-marginal community of Pampas de San Juan de Miraflores, a peri-urban community located approximately 25 kilometers from the city center of Lima, Peru. It currently has approximately 57,000 residents, most of whom are rural migrants. The estimated annual median income is USD 2100 (11). Diarrheal prevalence in this community has been estimated between three and four episodes per child-year in children younger than 5 years of age (12,13).

Study Design

We enrolled 134 participants from an existing birth cohort designed to investigate the effects of *Helicobacter pylori* infection on risk of childhood diarrhea (14), in which children younger than 3 months of age had been previously identified via a community census and invited to participate. Healthy children with birth weights equal or greater than 1500 g, in families with no plans to move outside the community within 6 months, were eligible for inclusion. Study children were under twice-weekly surveillance to record daily clinical symptoms, and all were invited to participate in this new study. The study and analysis of data was approved by the Institutional Review Boards of A.B. PRISMA, Universidad Peruana Cayetano Heredia, and the Johns Hopkins Bloomberg School of Public Health. All participants enrolled in our study provided written informed consents.

We calculated the sample size a priori for our study to detect 5% or more disagreement in the report of symptoms between the two types of interviews, with 90% power and 95% confidence. Thus, we were required to collect data from 134 families during an 8-week period. Frequent daily surveys might prompt respondents to report differently at the following twice-weekly survey than they would have otherwise. We tried to minimize this bias by only visiting a small, random set of participants each day. Therefore, a total of 34 different families (25% of the total) were randomly selected each day to receive a daily interview in addition to the twice-weekly interview. The daily surveys were conducted every day except Sundays. When a daily visit occurred on a Monday, data was collected for the 2 days previous.

The surveys collected data on frequency and consistency of bowel movements and specific clinical symptoms such as irritability, vomiting, fever, and coughing in each of the 3–4 days previously. Daily interviews were almost identical to the twice-weekly interviews, except in that they only collected information from the preceding day. To minimize potential biases related to the order or form of questioning (2), data were transcribed into almost-identical paper forms. Additionally, two field workers conducted the daily surveys and worked independently of the study personnel conducting the twice-weekly surveys. Each family received an approximately equal number of visits from each daily field worker. On days when daily and twice-weekly interviews coincided, they occurred within 4 hours and either was equally likely to occur first.

Ascertainment of Variables

We defined the recall period as the number of days between the interview and the day of the occurrence of symptoms. We used two age-based definitions of diarrhea: for children younger than 3 months of age, we defined a day of diarrhea when the mother reported the occurrence of diarrhea and the child passed six or more liquid or semiliquid bowel movements during a 24-hour period. For children 3 months of age and older, we defined day of diarrhea when a child passed three or more liquid or semiliquid bowel movements during a 24-h period.

Biostatistical Methods

From the twice-weekly data, we calculated the distribution of the number of loose stools, the prevalence of diarrhea, and other symptoms for each day during the recall period. We date-matched the data collected through twice-weekly and daily surveys. For analyses, data points were dropped whenever there were different respondents on the matched entries (e.g., if the mother completed the daily interview, and the father the twice-weekly, these paired interviews were dropped from the analysis), or when the interval between the reported symptom and interview dates was 4 days or greater.

We used the daily interview data as the gold standard for all analyses. We measured agreement between the daily and twice-weekly surveys by using kappa statistics (15). We stratified results by age less than or greater than 3 months. We also compared the AUCs (16) when the interviews were conducted on the same day versus when the twice-weekly interview was conducted 1 to 3 days after the daily interview by using a χ^2 test. This allowed us to measure the reliability of the household interview independently of recall-related effects. We used SDT calculate the detectability and criterion indices of selected clinical symptoms. We provide a detailed description of SDT calculations in the Appendix. We performed out analyses in STATA 9.0 (Statacorp, College Station, TX).

RESULTS

We collected 6157 child-days of data using twice-weekly surveys. A total of 71% of interviews were completed on time, and 86.7% were completed within 2 days of the planned date. Because twice-weekly interviews were not always completed on the day they were planned, there were 406 days of child-data in which the recall period from the twice-weekly interview was 5 days. We excluded these data from pairing with daily interviews because a recall period greater than 5 days is not representative of a twice-weekly survey.

A total of 1393 daily interviews were obtained for a completion rate of 85%; 212 daily interviews could not be paired to a twice-weekly interview, and therefore we used 1181 paired twice-weekly and daily interviews in our analyses. The median age of the 134 participating children was 106 days at entry and 157 days at the end of the study.

Differences in Diarrheal Prevalence Between Surveys

The prevalence of diarrhea was similar between the biweekly and daily interviews. However, about 20% of days ill with diarrhea were discrepant between the two interviews (Table 1). There was a slight increase in the mean number of loose stools as the recall period increased. By using data from the twice-weekly interviews only, we found that the overall prevalence of diarrhea in children ≥ 3 months of age varied from 4.7% with 1-day recall up to 5.1% with 5-day recall. The paired twice-weekly-daily data is reported in Table 2. The prevalence of diarrhea was similar between the biweekly and daily interviews.

Differences in Prevalence of Other Symptoms Between Surveys

All other symptoms reported in this paper were in a “yes/no” format. The prevalence for these symptoms when we used data for the twice-weekly interviews ranged from 0.4% (trouble swallowing) to 24.3% (coughing). The prevalence of most of these symptoms decreased as the recall period increased (Table 3). The exceptions were cough and phlegm, the two most common symptoms. The twice-weekly prevalence of fever and vomiting were 38% and 35% lower, respectively, than when the recall period was only 1 day (Table 2).

Agreement Between Surveys

When we stratified by age, we found that the agreement between data from the daily and twice-weekly interviews was lower for younger children (<3 months of age) than for older ones (≥ 3 months of age). The overall kappa statistic for fever was 0.29 and for vomiting was 0.13 when participants were younger than 3 months of age. Kappa values increased to 0.39 for fever and to 0.39 for vomiting when the two interviews occurred on the same day.

The number of loose stools varied significantly between the twice-weekly and daily interviews (kappa = 0.57). However, only some of this disagreement resulted in a different determination of diarrhea by the definition of 3 or more loose stools per day. Thus, the agreement between twice-weekly and daily interviews for diarrhea was greater (kappa = 0.78).

By using ROC curve analysis, we estimated the AUC for two instances: (i) when the twice-weekly and the daily interviews both occurred on the same day; and (ii) when the AUC when twice-weekly interview was 1 to 3 days after the daily interview. Greater AUC values imply greater agreement, and therefore more reliable information. The AUC when the interviews occurred on the same day was significantly larger than the AUC when they were further apart in time ($p = .03$).

Kappa statistics for agreement between twice-weekly and daily interviews ranged from 0.40 to 0.77. We display kappa statistics for selected symptoms in Table 3. In seven of the nine symptoms evaluated, kappa values were greater when the two surveys occurred on the same day than when they occurred 1 to 3 days apart. The two exceptions were coughing and phlegm, the two most common symptoms. Greater kappa values were associated with symptoms of greater prevalence. This is a statistical artifact because kappa penalizes for agreement as the result of chance and agreement is least likely to be the result of chance as prevalence approaches 50% (17). The ROC analysis of fever showed statistically significantly larger AUC when the two interviews occurred on the same day than when the recall period was 1 to 3 days (Table 4).

SDT Evaluation Between Surveys

Detectabilities of common symptoms from data collected through twice-weekly interviews were higher when the two interviews occurred on the same day, except for two symptoms, coughing and mother's opinion of diarrhea (Table 5). The values of criterion increased with all symptoms as the recall period increased, suggesting an increasing bias towards not reporting a symptom. When stratified by age, we saw lower detectabilities, and greater criterion for younger children than for older ones (data not presented).

Unlike kappa, detectability was not associated with prevalence in our data, thus supporting the statement that SDT provides an estimate of diagnostic accuracy that is not confounded by changing cutoff values or prevalence rates (18). The criterion, however, was sensitive to changes in prevalence. Fever had the highest detectability, suggesting that, of the "yes/no" symptoms, it was best remembered by respondents. Irritability had the lowest detectability ($d' = 1.6$), suggesting it was the most poorly remembered symptom.

DISCUSSION

We found that the overall prevalence of diarrhea estimated by the twice-weekly surveys was similar to that estimated by the daily survey. However, there were differences between the two surveys on the actual placement of days with diarrhea. Furthermore, the AUC for the ROC curve of diarrhea was significantly smaller as the recall period increased. Therefore, it appears that daily visits provide a significantly better source of information about diarrhea than do twice-weekly visits. The prevalence of other clinical symptoms, as measured by "yes/no"

questions, was underestimated by twice-weekly surveys when compared with daily surveys. For diarrhea and most other symptoms, the agreement between two interviews decreased as the recall period increased.

Given that in our study population the duration of a diarrheal episode was short, 2 to 3 days, any misplacement in illness days may significantly alter the average duration and potentially the incidence of episodes. Therefore, the misplacement of 20% of episodes between the daily and the twice-weekly interviews could be a cause for concern. However, the magnitude of this effect could not be evaluated under our study design. A previous report also showed misplaced incidence and therefore miscalculated duration of episodes (19), although that study used diaries, which may indirectly have led to that misplacement.

One factor that could have influenced our study data was that the daily interviews occurred chronologically before the twice-weekly interview. This sequence might have prompted mothers to remember and report events at the second interview more accurately. This would result in less variation between the two interviews and a bias towards the null in the results of this study. However, even when the two interviews took place on the same day, there was often a moderate amount of disagreement between them, even after attempting to minimize differences through questionnaire design. This disagreement could occur because of respondent uncertainty as to the presence of a symptom, and greater likelihood of saying that a symptom had not occurred. This is backed by the lower detectability and greater criterion for infants younger than 3 months because symptoms may be more difficult to identify in these children. If respondent uncertainty does lead to increased intraobserver disagreement, factors such as maternal age, parity, education, and stress that affect uncertainty might bear consideration in study design. This disagreement may be the result of differences in fieldworker questioning style or of a respondent's tendency while repeating an interview to answer quickly or by rote. Therefore, in some instances, they may represent a limit to the quality of information that can be reasonably gathered from any household survey.

For our analyses, we believe the daily interview data were the appropriate gold standard because recall period was short and the field workers conducting these interviews were assigned fewer households per day to visit than the corresponding twice-weekly field workers and therefore had more time for completing a shorter questionnaire. Kappa is the most commonly reported measure for comparing survey results between two interviewers, even when there is no clear gold standard. Within our community, field-workers spent most of their time travelling between houses. Were daily interviews to be implemented, each fieldworker would be made responsible for more interviews in a smaller geographic radius. Therefore, we estimate that the costs of a change to daily surveillance in our study might be relatively low. However, this may not be applicable to studies at other sites because there are differences in geography, local wages, weather conditions, and other logistical considerations.

The low agreement in the number of stools per day between the two interviews suggests that, in our community, respondents reported approximate rather than the exact numbers and might have slightly overestimated the number of stools as the short-term recall period increased. Although our study did not address longer recall periods, others researchers (4,5) have found that the prevalence of diarrhea does decrease with a 7-day recall. However, the specific form of questioning is important: our results suggest that asking "how many stools has your child had?" is likely to result in a diarrheal prevalence that stays roughly constant (but may lead to misplaced episodes) as recall increases, whereas asking "did your child have three or more loose or watery stools?" may lead to a prevalence that decreases with recall. Alternately, this habit of up-approximation may be a community-specific behavior, which might vary in another setting with different patterns of stool frequency in young children. This factor was reported

by Baqui et al. (20) as a potential modifier of the appropriate definition of diarrhea given a study area.

The common measurement for comparisons of survey instruments is Cohen's kappa. For comparability with other studies, we also measured agreement by this statistic. Although the ROC and kappa methodologies come from different disciplines, they can be linked mathematically (21) and are therefore related and complementary. In our study, the conclusions drawn from two approaches are mutually supportive.

We suggest that, when examining data that relies on subject memory, extending the ROC analysis to include SDT statistics is an informative way to explore agreement. In our study, detectability was less influenced by prevalence than Cohen's kappa. This makes comparisons of agreement across symptoms more meaningful. Second, the detectability and criterion indices have specific associated behavioral interpretations. They have been extensively explored in the psychological literature, although they are infrequently used in epidemiology. Through these interpretations, the detectability and criterion indices help to explain "why" agreement may be good or poor. This is of practical use in survey design and data implementation. Greater detectabilities suggest symptoms are better remembered by respondents. Criterion values indicate how strongly respondents who are "unsure" about a symptom tend to favor a "yes" or "no" response.

In cases in which overall diarrheal prevalence is the primary outcome of interest, we found that twice-weekly interviews might be sufficient to successfully capture these data. However, this recommendation is strongly tempered by the fact that a significant amount of misplacement clearly occurs as the recall period increases. This finding is evidenced by lower agreement as the recall period increases. Additionally, the displacement in days of disease creates as a risk of biased duration and incidence. Thus, daily surveillance clearly provides the best data on symptoms of diarrheal disease in the context of community-based field studies.

Acknowledgments

We thank Dr. William Pan, Dr. Margaret Kosek, Dr. Ellen Furlong, and Frank Kayanet for their helpful comments.

The CONTENT project was supported in part by the Sixth Framework Program of the European Union [INCO-CT-2006-032136]. Gwenyth Lee was supported by an International Maternal and Child Health Training Grant of the United States National Institutes of Health [T32HD046405]. Dr. William Checkley was further supported by a Clinician Scientist Award from the Johns Hopkins University.

Selected Abbreviations and Acronyms

SDT	signal detection theory
ROC	receiver operating characteristic
AUC	area under the curve

References

1. de Melo MCN, de AC, Taddei A, Diniz-Santos DR, May DS, Carneiro NB, Silva LR. Incidence of diarrhea: Poor parental recall ability. *Braz J Infect Dis* 2007;11:571–574. [PubMed: 18327469]
2. Ross D, Huttly SRA, Dollimore N, Binka FN. Measurement of the frequency and accuracy of childhood acute respiratory infections through household surveys in Northern Ghana. *Int J Epidemiol* 1994;23:608–616. [PubMed: 7960390]
3. Larson A, Mitra SN. Usage of oral rehydration solutions (ORS): A critical assessment of utilization rates. *Health Policy Plan* 1992;7:251–259.

4. Byass P, Hanlon P. Daily morbidity records: Recall and reliability. *Int J Epidemiol* 1994;23:757–763. [PubMed: 8002190]
5. Ramakrishnan R, Venkatarao T, Koya PKM, Kamaraj P. Influence of recall period on estimates of diarrhoea morbidity in infants in rural Tamil Nadu. *Indian J Public Health* 198(42):3–9.
6. Green, D.; Swets, J. *Signal Detection Theory and Psychophysics*. New York: John Wiley; 1966.
7. Swets, J.; Pickett, R. *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. New York: Academic Press; 1982.
8. Allan LG, Siegel S. A signal detection theory analysis of the placebo effect. *Eval Health Prof* 2002;25:410–421. [PubMed: 12449084]
9. Kapucu A, Rotello CM, Ready RE, Seidl KN. Response bias in “remembering” emotional stimuli: A new perspective on age differences. *J Exp Psychol Learn Mem Cogn* 2008;34:703–711. [PubMed: 18444767]
10. Marcum J. A statistical theory of target detection by pulsed radar. *IRE Trans Information Theory* 1960;6:59–267.
11. Berkman D. Effects of stunting, diarrhoeal disease, and parasitic infection during infancy on cognition in late childhood: A follow-up study. *Lancet* 2002;359:564–571. [PubMed: 11867110]
12. Checkley W, Gilman RH, Black RE, Epstein LD, Cabrera L, Sterling CR, et al. Effect of water and sanitation on childhood health in a poor Peruvian peri-urban community. *Lancet* 2004;363:112–118. [PubMed: 14726164]
13. Checkley W, Buckley G, Gilman RH, Assis AM, Guerrant RL, Morris SS, et al. Multi-country analysis of the effects of diarrhoea on childhood stunting. *Int J Epidemiol* 2008;37:816–830. [PubMed: 18567626]
14. Windle HJ, Kelleher D, Crabtree JE. Childhood *Helicobacter pylori* infection and growth impairment in developing countries: A vicious cycle? *Pediatrics* 2007;119:e754–e759. [PubMed: 17325213]
15. Cohen J. A coefficient of agreement for normal scales. *Educ Psychol Measure* 1960;37–46.
16. Zweig M, Campbell G. Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clin Chem* 1993;39:561–577. [PubMed: 8472349]
17. Vach W. The dependence of Cohen’s kappa on prevalence does not matter. *J Clin Epidemiol* 2005;58:655–661. [PubMed: 15939215]
18. McFall RM, Treat TA. Quantifying the information value of clinical assessments with signal detection theory. *Annu Rev Psychol* 1999;50:215–241. [PubMed: 15012460]
19. Haggerty P, Manunabo M, Ashworth A, Muladi K, Kirkwood B. Methodological approaches in a baseline study of diarrhoeal morbidity in weaning-age children in rural Zaire. *Int J Epidemiol* 1994;23:1940–1949.
20. Baqui AH, Black RE, Yunus M, Hoque ARA, Chowdhury HR, Sack RB. Methodological issues in diarrhoeal diseases epidemiology: Definition of diarrhoeal episodes. *Int J Epidemiol* 1991;20:1057–1063. [PubMed: 1800404]
21. Ben-David A. About the relationship between ROC curves and Cohen’s kappa. *Engin Appl Artificial Intelligence* 2008;21:874–882.

APPENDIX

SDT Example

The detectability is the difference in the probability of a correct identification of the signal from the probability of an incorrect identification of the signal converted into a Z-score. Detectability is a continuous variable that ranges from zero to infinity. (Under the assumption that the false alarm rate is less than the hit rate). A detectability score of four or greater represents nearly perfect recognition.

The criterion is the sum of the probabilities of a correct identification of a signal and of a false alarm converted to Z-scores divided by two. When the criterion is less than zero, it indicates

a bias towards “yes” responses, resulting in more hits but also more false alarms. If the criterion is greater than zero, the criterion favors “no” responses, with fewer hits and fewer false alarms.

CONTINGENCY TABLE

FIGURE 1 Did your child have a fever on the study date?	“Signal” “yes” at daily interview	“Noise” “no” at daily interview
Yes	16	9
“yes” at biweekly interview	(HIT)	(FALSE ALARM)
No	12	1144
“no” at biweekly interview	(MISS)	(CORRECT REJECTION)

The detectability is calculated by converting the probability of a correct identification of the signal (a “hit”) and the probability of an incorrect identification of the signal (a “false alarm”) into Z-scores, and subtracting. It is a continuous variable ranging from zero to infinity. A d' of four or greater represents nearly perfect recognition. The daily interview was used as the “gold standard.” Therefore, the detectability is the respondent’s likelihood of saying “yes” on a twice-weekly questionnaire, given that their response on a daily questionnaire was also yes. The criterion is the probability of a correct identification of a signal and the probability of a false alarm converted to Z-scores, added, and divided by two. When the criterion is less than zero, it indicates a bias towards “yes” responses, resulting in more hits but also more false alarms. If the criterion is greater than zero, the criterion favors ‘no’ responses, with fewer hits and fewer false alarms.

The false alarm rate is one minus the specificity, so the criterion and detectability of a signal are both linear combinations of the Z-scores of sensitivity and specificity.

$$d' = \Phi(\text{sens}) - \Phi(1 - \text{spec})$$

$$C = \frac{1}{2(\Phi(\text{sens}) + \Phi(1 - \text{spec}))}$$

Consider the above contingency table derived from our study data:

$$P(\text{Hit}) = \frac{16}{16+12} = 0.571$$

$$Z(\text{Hit}) = 0.180$$

$$P(\text{False Alarm}) = \frac{9}{9+1144} = 0.008$$

$$Z(\text{False Alarm}) = -2.418$$

$$d' = Z(\text{Hit}) - Z(\text{False Alarm}) = 2.6$$

$$C = -1/2(Z(\text{Hit}) - Z(\text{False Alarm})) = 1.1$$

The detectability for fever is 2.6. This is greater than detectabilities for most other symptoms in the study, implying that fever was a relatively well-recognized symptom by participants. The criterion for fever of 1.1 is average compared with that of most other symptoms, implying that there were fewer “false alarms” for fever than, for example, bloating, but more than were seen for coughing. This combination of good recognition of the symptom combined with a tendency towards providing false alarms resulted in a relatively mediocre agreement between daily and biweekly interviews for the symptom of fever overall as measured by kappa (kappa of 0.59).

TABLE 1

Contingency table showing agreement between daily and biweekly interviews for children older than 3 months of age

	Diarrhea from daily interview	No diarrhea from daily interview
Diarrhea from biweekly interview	42	11
No diarrhea from biweekly interview	11	762

TABLE 2

Prevalence of symptoms by recall times, selected variables

Question	“Yesterday” n = 1325		“Two days ago” n = 1628		“Three days ago” n = 1682		“Four days ago” n = 1116		“Five days ago” n = 406		Biweekly overall (days 1–5) n = 6157	
	Percent Prevalence {95% CI}		Percent Prevalence {95% CI}		Percent Prevalence {95% CI}		Percent Prevalence {95% CI}		Percent Prevalence {95% CI}		Percent Prevalence {95% CI}	
Vomiting	2.0% {1.2–2.7%}		1.2% {0.7–1.8%}		1.2% {0.7–1.8%}		0.8% {0.2–1.3%}		0.7% {0.0–1.6%}		1.3% {1.0–1.5%}	
Fever	3.3% {2.4–4.4%}		2.9% {2.1–3.7%}		1.6% {1.0–2.2%}		0.9% {0.3–1.4%}		0.2% {0.0–0.7%}		2.1% {1.8–2.5%}	
Cough	24.2% {21.8–26.5%}		24.4% {22.3–26.5%}		23.8% {21.7–25.8%}		24.2% {21.7–26.7%}		27.3% {23.0–31.7%}		24.3% {23.2–25.4%}	
Phlegm	19.2% {17.1–21.4%}		19.5% {17.5–21.4%}		19.2% {17.3–21.1%}		19.0% {16.7–21.3%}		20.0% {16.1–23.8%}		19.3% {18.3–20.3%}	
Diarrhea (≤3 months)	2.7% {1.1–4.4%} n = 365		2.4% {1.0–3.8%} n = 460		3.0% {1.4–4.4%} n = 484		3.3% {1.4–5.3%} n = 326		2.7% {0.0–5.8%} n = 111		2.8% {2.0–3.6%} n = 1746	
Diarrhea (>3 months)	4.7% {3.3–6.0%} n = 960		4.8% {3.6–6.0%} n = 1168		4.5% {3.3–5.7%} n = 1198		5.4% {3.9–7.0%} n = 790		5.1% {2.6–7.6%} n = 295		4.8% {4.2–5.5%} n = 4411	

CI = confidence interval.

TABLE 3
Diarrheal prevalence and agreement between daily and twice-weekly interviews by Cohen's kappa

Variable (n = 1181)	Point prevalence from daily visits (n = 1181) {95% CI}	Point prevalence from twice-weekly visits (n = 1181) {95% CI}	Kappa (same day) {95% CI}	Kappa (1- to 3-day lag) {95% CI}	Kappa (overall) {95% CI}
Vomiting	2.0% (n = 24) {1.2–2.4%}	1.6% (n = 19) {0.9–2.3%}	0.70 {0.60–0.80}	0.37 {0.30–0.44}	0.50 {0.45–0.56}
Fever	2.4% (n = 28) {0.5–3.2%}	2.1% (n = 25) {1.3–2.9%}	0.84 {0.74–0.94}	0.46 {0.38–0.52}	0.59 {0.54–0.65}
Cough	26.0% (n = 307) {23.5–28.5}	22.7% (n = 268) {20.3–25.1%}	0.72 {0.62–0.82}	0.72 {0.62–0.82}	0.73 {0.67–0.79}
Criteria for diarrhea (<3 months)	(n = 22) 6.2% {3.7–8.7%}	(n = 12) 3.4% {1.5–5.3%}	0.53 {0.36–0.70}	0.41 {0.29,0.53}	0.45 {0.35,0.55}
Criteria for diarrhea (>3 months)	6.4% (n = 53) {4.7–8.1%}	6.4% (n = 53) {4.7–8.1%}	0.90 {0.78–1.03}	0.72 {0.64–0.80}	0.78 {0.71–0.85}
Number of loose stools (multinomial) (>3 months)	-	-	0.71 {0.63–0.78}	0.51 {0.46–0.56}	0.57 {0.53–0.61}

CI = confidence interval.

TABLE 4

Agreement between daily and biweekly interviews by the area under the receiver operating characteristic curve

	AUC (same day){95% CI}	AUC (1- to 3-day lag) {95% CI}	χ^2 statistic (1 df)	<i>p</i> value
Vomiting	0.83 {0.67–0.99}	0.66 {0.54–0.79}	2.58	.11
Fever	0.996 {0.99–1.00}	0.70 {0.56–0.81}	28.40	.00
Cough	0.85 {0.80–0.90}	0.85 {0.82–0.88}	0.01	.93
Diarrhea (<3 months)	0.71 {0.51–.91}	0.66 {0.54–0.78}	0.18	.67
Diarrhea ($\geq$ 3 months)	0.97 {0.90–1.00}	0.86 {0.78–0.93}	4.95	.03

AUC = area under the curve.

TABLE 5

Detectability and criterion values for selected symptoms

	Detectability (same day) {95% CI}	Detectability (1- to 3-day lag) {95% CI}	Criterion (same day) {95% CI}	Criterion (1- to 3- day lag) {95% CI}
Vomiting	3.0 {2.0–4.0}	2.0 {1.3–2.7}	1.1 {0.6–1.6}	1.4 {1.1–1.8}
Fever	3.9 {2.5–5.4}	2.2 {1.5–2.8}	0.4 {0.3–1.2}	1.3 {1.0–1.7}
Cough	2.3 {1.9–2.7}	2.4 {2.2–2.7}	0.5 {0.3–0.6}	0.6 {0.4–0.7}
Diarrhea (<3 months)	2.2 {1.5–3.0}	1.7 {1.2–2.1}	0.4 {0.1–0.8}	0.5 {0.3–0.8}
Diarrhea (≥3 months)	4.0 {2.8–5.0}	2.7 {2.5–3.0}	0.4 {−0.1 to 1.0}	0.8 {0.5–1.0}