Epidemiological methods in diarrhoea studies—an update

Wolf-Peter Schmidt,1* Benjamin F Arnold,2 Sophie Boisson,1 Bernd Genser,3,4 Stephen P Luby,5 Mauricio L Barreto,3 Thomas Clasen1 and Sandy Cairncross1

1Department for Disease Control, Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, UK, 2Division of Epidemiology, University of California, Berkeley, CA, USA, 3Instituto de Saúde Coletiva, Federal University of Bahia, Salvador, Brazil, 4Mannheim Institute of Public Health, Medical Faculty Mannheim, University of Heidelberg, Heidelberg, Germany and 5International Centre for Diarrhoeal Disease Research Bangladesh, Dhaka, Bangladesh

*Corresponding author. Department for Disease Control, Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK. E-mail: wolf-peter.schmidt@lshtm.ac.uk

Accepted 14 September 2011

Background Diarrhoea remains a leading cause of morbidity and mortality but is difficult to measure in epidemiological studies. Challenges include the diagnosis based on self-reported symptoms, the logistical burden of intensive surveillance and the variability of diarrhoea in space, time and person.

Methods We review current practices in sampling procedures to measure diarrhoea, and provide guidance for diarrhoea measurement across a range of study goals. Using 14 available data sets, we estimated typical design effects for clustering at household and village/neighbourhood level, and measured the impact of adjusting for baseline variables on the precision of intervention effect estimates.

Results Incidence is the preferred outcome measure in aetiological studies, health services research and vaccine trials. Repeated prevalence measurements (longitudinal prevalence) are appropriate in high-mortality settings where malnutrition is common, although many repeat measures are rarely useful. Period prevalence is an inadequate outcome if an intervention affects illness duration. Adjusting point estimates for age or diarrhoea at baseline in randomized trials has little effect on the precision of estimates. Design effects in trials randomized at household level are usually <2 (range 1.0–3.2). Design effects for larger clusters (e.g. villages or neighbourhoods) vary greatly among different settings and study designs (range 0.1–25.8).

Conclusions Using appropriate sampling strategies and outcome measures can improve the efficiency, validity and comparability of diarrhoea studies. Allocating large clusters in cluster randomized trials is compromised by unpredictable design effects and should be carried out only if the research question requires it.

Keywords Diarrhoea, cluster randomized trial, sampling, methods
Introduction

Diarrhoeal diseases remain a leading cause of morbidity and mortality in children worldwide.1 Reliable field data from epidemiological studies are required to study diarrhoea epidemiology and the effect of interventions,2,3 but diarrhoea remains a condition difficult to measure.4 Systematic reviews of diarrhoea interventions have found a great variety of approaches to measure diarrhoea.5,6 The past decade saw a trend towards less intensive active diarrhoea surveillance,7 the use of repeated diarrhoea prevalence measures instead of incidence as outcome measure8 and a greater recognition of recent advances in the design of cluster randomized trials.9,10 In this article we review current practices in conducting epidemiological studies on diarrhoeal diseases with an emphasis on randomized controlled trials (RCTs) in low-income populations, including cluster randomized trials. We discuss crucial methodological problems to be considered in the planning stage of a trial, but several issues should also be relevant for observational studies.

Literature search methods and data sets

We searched the database MEDLINE for the years 1970–2009 without language restrictions, using the search terms [diarrhoea AND trial], [diarrhoea AND measurement], [diarrhoea AND recall] and [diarrhoea AND longitudinal prevalence]. We screened the reference lists of relevant articles and contacted authors and experts in the field for further identification of relevant articles. We further used original data sets from different field sites across the world (in part described previously21) to address issues of design effect and adjustment for baseline variables in RCTs. These data sets came from the authors of this article or were made available to us by other researchers in the field (see ‘Acknowledgements’).

Reporting and recording diarrhoea symptoms

Case definitions for diarrhoea commonly are either based on reported signs and symptoms (stool frequency, presence of blood or mucus) or based on local disease perception. For example, in Ghana identified seven different local terms for symptoms compatible with diarrhoea.12 Relying on local disease definitions requires extensive qualitative research and piloting,13 but such work can provide important insights that are useful for a study as a whole. Most studies continue to use the WHO definition of diarrhoea,14 defined as ‘the passage of 3 or more loose or liquid stools per day, or more frequently than is normal for the individual’.15,16 A stringent definition that does not depend on local disease concepts may reduce subjectivity and perhaps also the risk of bias but this has not yet been shown in practice. While not necessarily having more clinical validity, using the WHO definition facilitates comparison across sites (Box 1). Asking study participants specifically for the presence or absence of ‘3 or more loose or liquid stools per day’ may unnecessarily force a decision by the respondent that may be prone to bias. Therefore, some diarrhoea trials record stool frequency and then apply the WHO definition post-hoc.17,18

Studies have shown that the longer the recall period, the greater the imprecision (especially underestimation) of prevalence estimates.19–22 By assuming that reported prevalence in the last 24 h was 100% accurate, these studies may have overestimated recall error, since the higher diarrhoea prevalence closer to the day of the visit may indicate that people remember diarrhoea during the past 7 days as having occurred more recently than was actually the case. In a study from Peru, mothers reported the correct diarrhoea prevalence and then often were inaccurate in reporting the exact day when it occurred.23 Recall error depends on the severity and duration of symptoms.24 A decline in reported diarrhoea with time on study (independent of treatment) has been noted in diverse populations.10,25–29 Intensive surveillance including frequent home visits can lower the reported diarrhoea prevalence,30 perhaps due to ‘reporting fatigue’. Recall can be more complete in groups of higher socio-economic status, leading to bias when comparing different populations.31

Recall error may not be a big problem in studies exploring disease trends or comparing diarrhoea risk between treatment arms, if it can be assumed that recall error is non-differential among the groups compared. This assumption, however, is difficult to verify in unblinded trials. There are numerous theoretical possibilities for treatment effects to be biased. For example, allocation to the control group may lead to diarrhoea episodes being remembered more acutely out of frustration of not receiving the intervention. Alternatively, allocation to a treatment group may lead participants to not report disease episodes or field staff to not record disease episodes under the expectation that the intervention is effective. Due to such biases, even a diarrhoea reduction of 50% observed in unblinded trials may be compatible with no true effect.32

Given the complexity of validating reported diarrhoeal disease in community-based surveys, investigators should take steps to minimize measurement error whenever possible. A 7-day recall period is commonly used in diarrhoea trials,33 but a shorter recall period may reduce subjectivity of reporting and possibly bias in unblinded trials. Using a 2- or 3-day recall often leads to only a small to moderate loss of power compared with a 7-day recall period, especially if diarrhoea is common, if the number of measurements per individual exceeds 10 or 12 and in cluster randomized trials.34 Instead of asking for diarrhoea in the previous 24, 48 or 72 h, one could consider asking for...
whole calendar days only (Did you have diarrhoea today? Yesterday? The day before yesterday?). Such questions are usually easier to ask and to answer. Numerous studies have demonstrated that symptom recall beyond 7 days is unreliable, and we do not recommend it.13,19–23 In any case, the final choice of the recall period should only be done after pilot-testing different approaches in a given setting.

**Incidence or longitudinal prevalence as outcome measure**

Diarrhoea can be measured as incidence (new episodes per person-time) or prevalence (disease presence at time t, Box 1). Incidence does not account for episode duration,8 an important risk factor for adverse outcomes.35,36 In settings where diarrhoea is common, it can be difficult to distinguish one episode from the next. Two to three days have been suggested as the most appropriate period to separate distinct episodes,5,31 an approach widely used in today. If diarrhoea is quite rare, it makes sense to use a longer gap (e.g. 6 days27,38) to separate distinct episodes, since episodes are unlikely to occur close together by chance. Such definitions are, to some extent, arbitrary and will inevitably cause some misclassification.14,39

Methods have been developed to allow the comparison of studies using different definitions.14 Measuring incidence especially in high-risk settings can require close disease surveillance (e.g. one to three times/week) to establish beginning and end of episodes.18 However, a rough incidence estimate can be obtained by repeated period prevalence measurements assuming that diarrhoea preceded by a period without diarrhoea represents a new episode.46

Incidence is an appropriate measure if the duration of illness is not of particular interest. A new episode can be interpreted as a case of pathogen transmission to a new host, which for disease control or vaccine research can be more important than episode duration. This applies to disease surveillance in middle- and high-income settings with low risk of malnutrition and diarrhoea-related mortality. For example, a study in the UK compared the incidence of diarrhoea in the community with cases reported to surveillance agencies.41 The duration of episodes was of little importance. Health service and vaccine researchers are often more interested in incident episodes than prevalence, focusing on the incidence of episodes with pre-defined characteristics, e.g. episodes of long duration or with blood/mucus, or watery diarrhoea for the surveillance of cholera.36,42

Such studies often use passive case finding instead of intensive active surveillance, e.g. by measuring the incidence of hospital admissions. This approach allows obtaining detailed clinical data and causative agents assessed by health professionals, often at a higher standard compared with field data. Measuring the incidence of hospital admission biases the data towards severe episodes, which are often the episodes of highest public health interest. Since only a fraction of diarrhoea episodes are seen at hospitals, the study population receiving the intervention will have to be large. On the other hand, there is no need for repeated surveillance visits. Passive recording of the incidence of hospital admissions may be less prone to observer and responder bias than diarrhoea incidence recorded through active surveillance because, although bias cannot be excluded, study participants are less likely to decide on health-care use based on treatment allocation. If the aim of a study is to obtain detailed clinical data on all, not just severe, episodes, close active surveillance (e.g. contacting participants at least once a week) is usually required, especially if stool samples are collected.41

Outside clinical studies, prevalence rather than incidence is often the outcome measure of choice, especially if prevalence can be measured repeatedly in the same individual. Repeated measurements provide an estimate of an individual’s proportion of time ill, also termed ‘longitudinal prevalence’ (LP).8 The ideal settings for using LP as outcome are low-income, high-risk populations where preventing adverse outcomes such as death and malnutrition is important. LP is a better predictor of such complications than incidence.8,43 Table 1 shows the results from two large RCTs conducted in Guatemala44 (Household water treatment intervention) and Brazil17 (Vitamin A supplementation). In the Guatemala trial, the interventions reduced the incidence of diarrhoea by 24%, whereas the mean LP (days with diarrhoea/days observed) was reduced by only 14%. This was because the intervention mostly prevented short episodes.44 In contrast, the Brazil study achieved only a small reduction in the incidence, which, however, masks the impact of the intervention on the duration of illness, leading to an LP reduction of 12% (note the differences in the P-values for LP vs incidence). In both cases it can be argued that longitudinal prevalence is the more appropriate way to measure public health impact.

Individuals tend to differ more in the number of disease days than in the number of episodes they experience, since the variation in the duration of episodes increases the standard deviation (SD) of LP compared with incidence.11 LP studies may require a larger sample size than incidence studies, if the

### Table 1: Incidence vs longitudinal prevalence of diarrhoea: impact on study results and interpretation in two randomized trials

<table>
<thead>
<tr>
<th></th>
<th>Incidence reduction (%)</th>
<th>P</th>
<th>Mean LP reduction (%)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Guatemala</strong> (n = 2982)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Water treatment</td>
<td>−24</td>
<td>0.001</td>
<td>−14</td>
<td>0.185</td>
</tr>
<tr>
<td><strong>Brazil</strong> (n = 1180)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vitamin A</td>
<td>−7</td>
<td>0.18</td>
<td>−12</td>
<td>0.06</td>
</tr>
</tbody>
</table>
exposure variable has no effect on episode duration. If, however, the exposure variable is associated with shorter episodes (as in the Brazil Vitamin A study), using LP increases power because the effect size should be larger (Table 1). If an intervention reduces predominantly short episodes as in Guatemala (Table 1), incidence may be more powerful, but may also be less informative for public health purposes. Table 2 summarizes advantages and disadvantages of using incidence vs prevalence measurements.

### At what intervals should diarrhoea prevalence be measured?

Diarrhoea prevalence can be measured at long and irregular intervals because a prevalence measurement requires no information of when an episode started. Incidence may also be estimated by infrequent or irregular sampling, e.g. by assuming that any diarrhoea occurring within the recall period is a new episode if no diarrhoea was present early in the recall period. This, however, is when recall error may be greatest, making such incidence estimates potentially unreliable.

Infrequent sampling can reduce costs and may increase validity, since frequent measurements may compromise the willingness of participants to report illness. Frequent measurements may lead to a better compliance with the intervention and a lower reported prevalence of diarrhoea, at least if each visit includes procedures that are clearly related to the intervention (e.g. water testing in a household water chlorination intervention).

Many repeat measurements of diarrhoea prevalence often provide little additional study power compared with fewer measurements. Clustering of disease in high-risk individuals means that if an individual reports being diseased at the time of a survey, he/she is more likely to have been ill on any other day than an individual reported healthy. The more illness is clustered in individuals, the more disease absence or presence in an individual at one point in time is representative of the true disease experience. Consider as an example, a study in which weekly surveillance visits are conducted over 1 year, each time recording the daily point prevalence of diarrhoea over the past 7 days since the last visit (a 1-week recall period), an approach resulting in continuous daily diarrhoea data. It has been shown that a study in which visits are conducted every 4 weeks instead of every week (again using a 1-week recall period) only requires a 15–30% larger sample size, while reducing the number of visits by 75%. In cluster randomized trials, the sample size increase in this example would even be smaller. Many measurements in the same cluster (e.g. more than 12 per year) yield little additional power, especially if within-cluster correlation of disease or cluster size is large.

Of note, studies using diarrhoea as the ‘exposure’ variable require more precise estimates of an individual’s burden of diarrhoea than studies with diarrhoea as ‘outcome’. For example, many studies have examined the effect of diarrhoea LP (the exposure variable) on mortality, malnutrition or the risk of other infectious diseases (the outcomes). Imprecision in the measurement of diarrhoea as an ‘exposure’ variable (e.g. due to infrequent sampling) usually biases the effect estimate towards no effect (‘regression dilution bias’). Often, more than 15 to 20 visits will be required to limit bias. Also, when measuring diarrhoea as an ‘exposure’ variable, a short

<table>
<thead>
<tr>
<th>Table 2 Comparison between incidence and longitudinal prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Incidence</strong></td>
</tr>
<tr>
<td>Suitable setting</td>
</tr>
<tr>
<td>Suitable research objectives</td>
</tr>
<tr>
<td>Data interpretation</td>
</tr>
<tr>
<td>Sampling frequency</td>
</tr>
<tr>
<td>Definition to separate episodes</td>
</tr>
<tr>
<td>Study power</td>
</tr>
</tbody>
</table>

---

**EPIDEMIOLOGICAL METHODS IN DIARRHOEA STUDIES**
recall period (e.g. 3 days) may be preferable to minimize bias.\textsuperscript{46} If diarrhoea is the ‘outcome’ measure, imprecise diarrhoea estimates due to infrequent visits will only affect the precision of the effect estimate, not the effect size.\textsuperscript{35}

Temporary absence of study participants and logistical constraints can cause prevalence measurements to be taken at irregular intervals. This is not a problem if irregularity occurs at random or at least similarly between comparison groups, which often should be the case. The later analysis should be weighted by the number of measurements in an individual.\textsuperscript{36}

**Point prevalence vs period prevalence**

While some investigators choose to record point prevalence (‘On which of the last X days did you suffer from diarrhoea?’), others collect period prevalence data (‘Did you have diarrhoea at any time during the last X days?’ Box 1). Period prevalence data are often used in large demographic and health surveys. Recording period prevalence may be simpler, but can reduce the difference (if expressed as a prevalence ratio) between two study groups because an individual with an episode several days long may be recorded as having the same disease experience as a person in the other group with only 1 disease day, i.e. period prevalence data bias the prevalence ratio towards no effect, especially if the disease is common (e.g. more than five episodes per person-year).\textsuperscript{34}

Perhaps counter-intuitively, period prevalence as an imprecise outcome measure can achieve a higher study power than point prevalence even if the recall period is the same, because differences between individuals (i.e. the coefficient of variation of the mean LP) are reduced.\textsuperscript{34} However, period prevalence data are inappropriate to capture changes in illness duration. Effect sizes will be strongly biased towards no effect if an intervention primarily works by reducing episode duration,\textsuperscript{34} and will be exaggerated if an intervention primarily reduces short episodes (Table 1, Guatemala study).

To conclude, investigators need to balance the advantages of using period prevalence data (easy to collect, slightly more powerful in many situations) with the risk of bias, which depends on the effect of the factor under study on illness duration. The collection of daily point prevalence with a limited recall period provides flexibility to use either outcome measure, but investigators should specify in advance which is to serve as the main study outcome to protect against selectively choosing a measure that provides the result most aligned with the investigators’ pre-conception.

**Adjusting for baseline diarrhoea and age**

In many trials, investigators measure diarrhoea at baseline (before randomization). In general, baseline measurements in trials may serve to (i) verify randomization success, (ii) adjust the final analysis for imbalances and (iii) increase precision of the treatment effect by including the baseline measure as a covariate in an adjusted analysis. The latter two uses require that the baseline measure be strongly associated with the later outcome to be effective.\textsuperscript{37,38} Concerns over imbalances in diarrhoea prevalence at baseline have in the past prevented or severely delayed publication of trials.\textsuperscript{39} However, caution is warranted in interpreting baseline diarrhoea data, specifically when used to verify the success of randomization. Most demographic variables commonly assessed at recruitment, such as date of birth, gender, family size or socio-economic status, do not change rapidly (if at all) and may later be used to adjust for imbalances. In contrast, diarrhoea prevalence is highly variable over time.\textsuperscript{60} If an individual has diarrhoea at baseline, it indicates that they may be more prone to diarrhoea during the follow-up period, but this depends on the within-person clustering of disease in a given setting. Typically, diarrhoea trials are designed to detect a certain difference between trial arms given a pre-specified number of repeat measurements (often more than 10), assuming a chance of false positivity of, say 0.05. A ‘single’ measurement at baseline in the same number of people has a considerable chance of suggesting a relevant imbalance where there may be none. It has been suggested that multiple baseline measurements collected during a run-in period could improve the efficiency of studies with both continuous and incidence rate outcomes,\textsuperscript{61} but this may not necessarily apply to diarrhoea. For example, Figure 1 plots the village-level diarrhoea incidence in 11 control villages from a randomized trial of solar water disinfection in Bolivia. The baseline diarrhoea measurement included 6 weeks of surveillance (six measurements per individual) that were collected 6 months before the intervention. As Figure 1 illustrates, baseline incidence bears no relation to incidence during the year-long intervention period. Several factors may have contributed to this, such as the long gap between baseline measurement and the actual trial and, in particular, the high spatial and temporal variability of diarrhoea often observed in the field.\textsuperscript{60} This contrasts with strong associations between baseline village HIV prevalence and subsequent incidence,\textsuperscript{62} or between baseline height-for-age Z-scores and subsequent height measurements.\textsuperscript{53}

Table 3 shows the effect of adjusting for baseline diarrhoea (single measurement) or age on the effect estimate and standard error (SE) in studies available to us (age usually is a strong predictor of diarrhoea). In some cases, adjusting for baseline diarrhoea or age can have a relevant effect on the effect estimates (e.g. Kenya and Colombia), and often reduces the SE. However, adjusting for covariates in RCTs by using statistical models in general can lead to bias, and should be conducted with caution.\textsuperscript{64} The protocol for adjusted analyses in randomized trials to gain study power or reduce bias should be pre-specified\textsuperscript{58} and reserved for large studies, where statistical models...
may be less biased. The age adjustments shown in Table 3, however, do not suggest a great gain in study power in large studies. In cluster randomized trials the gain in power due to baseline adjustment may be even lower than in individually randomized trials, especially if the between-cluster variation is high. Based on these results and Figure 1, we infer that baseline diarrhoea would make a poor matching or stratification variable in a trial’s design.

To conclude, a single baseline measurement of diarrhoea should primarily be useful to confirm trial procedures and familiarize study participants and field

---

**Figure 1** Village-level diarrhoea incidence during a 12-month follow-up period in 11 control villages that participated in an intervention trial of solar water disinfection. Vertical lines mark bootstrapped 95% confidence intervals. The follow-up incidence is plotted against baseline incidence measured over a 6-week period (A), and against the village rank in baseline incidence over that same period (B)

---

**Table 3** Effect of adjusting for baseline diarrhoea or age on point estimate and SE

<table>
<thead>
<tr>
<th>References</th>
<th>Country</th>
<th>Age range (years)</th>
<th>N</th>
<th>Crude analysis</th>
<th>Adjusted analysis</th>
<th>SE change (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PR  SE P</td>
<td>PR  SE P</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjustment for baseline diarrhoea</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clasen et al. 65</td>
<td>Bolivia</td>
<td>0–80</td>
<td>317</td>
<td>0.55 0.16 0.042</td>
<td>0.52 0.16 0.038</td>
<td>+1</td>
</tr>
<tr>
<td>Boisson et al. 66</td>
<td>Congo</td>
<td>0–84</td>
<td>1144</td>
<td>0.85 0.15 0.336</td>
<td>0.88 0.15 0.447</td>
<td>+1</td>
</tr>
<tr>
<td>Colford et al. 67</td>
<td>USA</td>
<td>55–95</td>
<td>770</td>
<td>0.90 0.06 0.119</td>
<td>0.90 0.06 0.123</td>
<td>+0.1</td>
</tr>
<tr>
<td>Clasen et al. 68</td>
<td>Colombia</td>
<td>0–82</td>
<td>684</td>
<td>0.54 0.14 0.017</td>
<td>0.54 0.13 0.015</td>
<td>−1</td>
</tr>
<tr>
<td>Boisson et al. 69</td>
<td>Ethiopia</td>
<td>0–91</td>
<td>1516</td>
<td>0.75 0.09 0.011</td>
<td>0.74 0.08 0.007</td>
<td>−3</td>
</tr>
<tr>
<td>Trotta 59</td>
<td>Peru</td>
<td>0.5–1.5</td>
<td>483</td>
<td>0.98 0.20 0.902</td>
<td>1.03 0.18 0.850</td>
<td>−9</td>
</tr>
<tr>
<td>Tiwari et al. 70</td>
<td>Kenya</td>
<td>&lt;15</td>
<td>216</td>
<td>0.37 0.13 0.004</td>
<td>0.31 0.12 0.002</td>
<td>−9</td>
</tr>
</tbody>
</table>

Adjustment for age

| Tiwari et al. 70 | Kenya   | <15               | 216 | 0.37 0.13 0.004 | 0.37 0.13 0.005  | +1            |
| Colford et al. 67 | USA     | 55–95             | 770 | 0.90 0.06 0.119 | 0.91 0.06 0.129  | +0.3          |
| VAST12          | Ghana   | 0–5               | 1918| 0.99 0.01 0.316 | 1.01 0.01 0.289  | −1            |
| Reller et al. 44 | Guatemala| 0–80              | 2980| 0.86 0.08 0.106 | 0.86 0.08 0.112  | −2            |
| Boisson et al. 59 | Ethiopia| 0–91              | 1516| 0.75 0.09 0.011 | 0.75 0.08 0.009  | −3            |
| Boisson et al. 66 | Congo   | 0–84              | 1144| 0.85 0.15 0.336 | 0.88 0.14 0.434  | −3            |
| Clasen et al. 68 | Colombia| 0–82              | 684 | 0.54 0.14 0.017 | 0.43 0.13 0.010  | −6            |
| Clasen et al. 65 | Bolivia | 0–80              | 317 | 0.55 0.16 0.042 | 0.48 0.12 0.005  | −23           |

Age adjustment was made with age as categorical variable (<1 year, 1 to <2 years, 2 to <3 years, 3 to <5 years, 5 to <10 years, 10 to <15 years, ≥15 years), except for the US elderly population (55–64, 65–74, 75–84 and 85–95 years); PR, prevalence ratio.
staff with measurement procedures. Occasionally, it has been observed that the first or a single measurement in a trial may provide implausibly high estimates compared with follow-up visits. Participants concerned about potentially not being included in a trial may over report the disease at first visit. A baseline measurement that would not be included in a later analysis may limit the impact of this possible effect.

**Group-level clustering and design effect**

Many diarrhoea studies need to consider clustering of diarrhoea in households or villages/neighbourhoods, e.g. if an intervention is randomized at group level. The effect of clustering can be expressed as the design effect DEFF, the factor by which the sample size needs to be increased to account for clustering:

\[
DEFF = 1 + (m - 1) \times ICC
\]

where \(m\) is the number of individuals per cluster, and ICC is the intra-cluster correlation coefficient. Estimating ICC and DEFF is one of the most challenging aspects in complex diarrhoea trials. Both depend on factors such as (i) mean number of persons per cluster, (ii) mean number of measurements per person, (iii) within-person correlation of diarrhoea (which strongly depends on the age range included) and (iv) the differences in diarrhoea risk between clusters (i.e. the between-cluster variability). In areas where a substantial proportion of diarrhoea occurs as localized epidemics shifting from place to place, between-cluster variability (i.e. ICC and DEFF) will be high because some areas may be experiencing an outbreak at the time of study, whereas others are not. In addition, the DEFF increases if cluster size and number of measurements per individual vary, which is usually the case in field studies.

Calculating the DEFF for diarrhoea as a binary outcome based on an ICC estimate is not straightforward, perhaps best highlighted by the many different methods available. Estimating the ICC treating diarrhoea LP as a continuous outcome is problematic since follow-up time usually differs between individuals. Alternatively, the DEFF can be estimated directly from the SEs of the log prevalence ratio or log rate ratio resulting from clustered and unclustered analyses:

\[
DEFF = \frac{SE_{\text{clustered}}^2}{SE_{\text{unclustered}}^2}
\]

where \(SE_{\text{clustered}}\) is the standard error from an analysis accounting for clustering, and \(SE_{\text{unclustered}}\) is the standard error from an analysis ignoring clustering. We calculated DEFFs from the data of several randomized trials available to us using this formula (for details, see footnote of Table 4). We calculated DEFFs separately for within-person and within-cluster correlation of disease to show the effect of group-level clustering in addition to the design effect due to within-person correlation.

DEFFs for ‘household’ clustering are quite similar across studies, ranging from one to approximately three regardless of the study design (Table 4). In contrast, we found very different design effects of up to 22 if the unit of clustering was large (villages or neighbourhoods). In one case (urban Brazil), the design effect was much smaller for the analysis accounting for neighbourhood clustering compared with the analysis accounting for within-person clustering only. In this setting an individually randomized trial may require a larger sample size than a cluster randomized trial, because children in the same cluster had very different diarrhoea risks, whereas the cluster-level diarrhoea risks were similar. For six studies with continuous diarrhoea records we did the same calculation for incidence of new episodes (Table 4), DEFFs in brackets, mostly resulting in much lower within-person DEFFs and slightly higher household DEFFs compared with prevalence data. The DEFFs for incidence vs prevalence due to village/neighbourhood clustering were quite different in three of the six studies (rural Bolivia, rural Pakistan and urban Brazil).

Overall, DEFFs in trials randomizing large clusters are difficult to predict unless previous data from the same site are available. Randomization of large clusters should perhaps be ‘avoided like the plague’ unless the research question requires it.

The DEFFs due to ‘within-person’ correlation very strongly depend on the number of measurements (Table 4), showing again that many repeated measurements contribute little to study power. Continuous surveillance of daily point prevalence generally results in extremely large within-person DEFFs because of high day-to-day correlation. DEFFs are much reduced if measurements are either reduced to period prevalence, or separated by intervals between measurements. Repeat measures add to the complexity of sample size calculations for cluster randomized diarrhoea trials, since the number of measurements also affects ‘group level’ ICC and hence DEFF. Sample size calculations for diarrhoea trials may have to be pragmatic and, even more so than for diseases that do not recur, undergo an iterative process testing different sampling intervals and cluster sizes. Several approaches are available. If diarrhoea is common, it can make sense to treat diarrhoea as a continuous variable (e.g. LP or number of episodes per person) and remove one level of complexity. This requires knowledge of the mean LP or number of episodes and the SD given a specified number of measurements (examples have been published elsewhere). The sample size resulting from simple formulae for the comparison of two means can be multiplied by a group-level DEFF deemed appropriate (Table 4). Note that the presence of several levels of clustering (e.g. person, household and area) does
Table 4 Design effects in diarrhoea studies

<table>
<thead>
<tr>
<th>Authors</th>
<th>Country</th>
<th>Age range (years)</th>
<th>Individuals (N)</th>
<th>Follow-up method</th>
<th>Weeks follow-up per person, mean (SD)</th>
<th>Outcome measure</th>
<th>Mean LP (% days or weeks ill)</th>
<th>Persons per cluster, SD</th>
<th>DEFF due to clustering by household or village</th>
<th>DEFF due to clustering by person</th>
</tr>
</thead>
<tbody>
<tr>
<td>Household clustering</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clasen et al.</td>
<td>Bolivia (rural)</td>
<td>0–80</td>
<td>317</td>
<td>Monthly visits with 7-day recall</td>
<td>4.8 (0.5)</td>
<td>Weekly period prevalence</td>
<td>3.7</td>
<td>54 (2.4)</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>Colford et al.</td>
<td>USA (rural)</td>
<td>55–95</td>
<td>770</td>
<td>Weekly visits with 7-day recall</td>
<td>45.1 (13.6)</td>
<td>Weekly period prevalence</td>
<td>7.4</td>
<td>14 (0.5)</td>
<td>2.4 (14)</td>
<td>2.3 (13)</td>
</tr>
<tr>
<td>Reller et al.</td>
<td>Guatemala (rural)</td>
<td>0–80</td>
<td>2980</td>
<td>Monthly visits with 7-day recall</td>
<td>52 (0)</td>
<td>Daily point prevalence</td>
<td>2.5</td>
<td>63 (2.4)</td>
<td>46.8 (7.1)</td>
<td>44.0 (5.9)</td>
</tr>
<tr>
<td>Arnold et al.</td>
<td>India (rural)</td>
<td>0–5</td>
<td>1184</td>
<td>Monthly visits with 7-day recall</td>
<td>11.1 (2.5)</td>
<td>Weekly period prevalence</td>
<td>1.8</td>
<td>14 (0.6)</td>
<td>1.4</td>
<td>1.3</td>
</tr>
<tr>
<td>Luby et al.</td>
<td>Bangladesh (rural)</td>
<td>0–5</td>
<td>1278</td>
<td>Monthly visits with 2-day recall</td>
<td>4.7 (1.2)</td>
<td>2-days period prevalence</td>
<td>11.0</td>
<td>13 (0.5)</td>
<td>3.1</td>
<td>2.5</td>
</tr>
<tr>
<td>Boisson et al.</td>
<td>Ethiopia (rural)</td>
<td>0–91</td>
<td>1516</td>
<td>Monthly visits with 7-day recall</td>
<td>10 (1.2)</td>
<td>Weekly period prevalence</td>
<td>3.5</td>
<td>48 (2.4)</td>
<td>1.4</td>
<td>1.2</td>
</tr>
<tr>
<td>Tiwari et al.</td>
<td>Kenya (rural)</td>
<td>0–15</td>
<td>216</td>
<td>Monthly visits with 7-day recall</td>
<td>5.3 (1.4)</td>
<td>Daily point prevalence</td>
<td>3.6</td>
<td>33 (1.4)</td>
<td>7.4</td>
<td>5.7</td>
</tr>
<tr>
<td>Mausezahl et al.</td>
<td>Bolivia (rural)</td>
<td>0–5</td>
<td>725</td>
<td>Monthly visits with 7-day recall</td>
<td>32.9 (9.9)</td>
<td>Weekly period prevalence</td>
<td>9.4</td>
<td>17 (0.7)</td>
<td>142 (6.1)</td>
<td>101 (4.0)</td>
</tr>
<tr>
<td>Boisson et al.</td>
<td>DRC (rural)</td>
<td>0–84</td>
<td>1144</td>
<td>Monthly visits with 7-day recall</td>
<td>10.6 (1.2)</td>
<td>Weekly period prevalence</td>
<td>3.5</td>
<td>48 (2.4)</td>
<td>2.5</td>
<td>1.8</td>
</tr>
<tr>
<td>Luby et al.</td>
<td>Pakistan (urban)</td>
<td>0–15</td>
<td>4691</td>
<td>Twice-weekly visit with 3–4 day recall</td>
<td>44.8 (11.0)</td>
<td>Weekly period prevalence</td>
<td>3.4</td>
<td>5.2 (2.2)</td>
<td>6.2 (4.5)</td>
<td>4.3 (2.8)</td>
</tr>
<tr>
<td>Van der Hoek et al.</td>
<td>Pakistan (urban)</td>
<td>0–80</td>
<td>1500</td>
<td>Monthly visits with 7-day recall</td>
<td>58.4 (10.0)</td>
<td>Daily point prevalence</td>
<td>1.5</td>
<td>6.7 (2.8)</td>
<td>57.5 (7.5)</td>
<td>37.5 (4.8)</td>
</tr>
<tr>
<td>Luby et al.</td>
<td>Pakistan (urban)</td>
<td>0–66</td>
<td>8949</td>
<td>Twice-weekly visit with 3–4 day recall</td>
<td>34.2 (7.6)</td>
<td>Weekly period prevalence</td>
<td>5.0</td>
<td>6.7 (2.6)</td>
<td>5.0 (3.7)</td>
<td>3.3 (2.0)</td>
</tr>
<tr>
<td>Clasen et al.</td>
<td>Colombia (rural)</td>
<td>0–82</td>
<td>684</td>
<td>Monthly visits with 7-day recall</td>
<td>3.8 (0.5)</td>
<td>Weekly period prevalence</td>
<td>7.1</td>
<td>5.0 (2.5)</td>
<td>3.2</td>
<td>1.5</td>
</tr>
<tr>
<td>Clasen et al.</td>
<td>Bolivia (rural)</td>
<td>1–86</td>
<td>278</td>
<td>Monthly visits with 7-day recall</td>
<td>3.9 (0.3)</td>
<td>Weekly period prevalence</td>
<td>16.7</td>
<td>5.6 (1.8)</td>
<td>4.6</td>
<td>1.4</td>
</tr>
<tr>
<td>Village/neighbourhood clustering</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Barretto et al.</td>
<td>Brazil (urban)</td>
<td>0–5</td>
<td>1880</td>
<td>Thrice-weekly visit with 2–3 day recall</td>
<td>37.6 (17.9)</td>
<td>Daily point prevalence</td>
<td>2.9</td>
<td>76.4 (24.5)</td>
<td>1.0</td>
<td>10.2</td>
</tr>
<tr>
<td>Reller et al.</td>
<td>Guatemala (rural)</td>
<td>0–80</td>
<td>2980</td>
<td>Monthly visits with 7-day recall</td>
<td>52 (0)</td>
<td>Daily point prevalence</td>
<td>2.5</td>
<td>248.5 (199.7)</td>
<td>50.2 (6.1)</td>
<td>44.0 (4.7)</td>
</tr>
<tr>
<td>Troia99</td>
<td>Peru (urban)</td>
<td>0–2</td>
<td>483</td>
<td>Weekly visits with 7-day recall</td>
<td>6.7 (0.6)</td>
<td>Daily point prevalence</td>
<td>9.6</td>
<td>24.2 (1.0)</td>
<td>6.6</td>
<td>5.5</td>
</tr>
<tr>
<td>Arnold et al.</td>
<td>India (rural)</td>
<td>0–5</td>
<td>1184</td>
<td>Monthly visits with 7-day recall</td>
<td>11.1 (2.5)</td>
<td>Weekly period prevalence</td>
<td>1.8</td>
<td>51.4 (17.4)</td>
<td>1.6</td>
<td>1.3</td>
</tr>
<tr>
<td>Mausezahl et al.</td>
<td>Bolivia (rural)</td>
<td>0–5</td>
<td>725</td>
<td>Monthly visits with 7-day recall</td>
<td>32.9 (9.9)</td>
<td>Weekly period prevalence</td>
<td>9.4</td>
<td>33.0 (7.6)</td>
<td>19.2 (7.4)</td>
<td>10.1 (4.0)</td>
</tr>
<tr>
<td>Van der Hoek et al.</td>
<td>Pakistan (rural)</td>
<td>0–80</td>
<td>1500</td>
<td>Weekly visits with 7-day recall</td>
<td>58.4 (10.0)</td>
<td>Daily point prevalence</td>
<td>1.5</td>
<td>146.8 (66.7)</td>
<td>115.1 (38.2)</td>
<td>37.5 (4.8)</td>
</tr>
<tr>
<td>Luby et al.</td>
<td>Pakistan (urban)</td>
<td>0–5</td>
<td>1278</td>
<td>Monthly visits with 2-day recall</td>
<td>4.7 (1.2)</td>
<td>2-days period prevalence</td>
<td>11.0</td>
<td>12 (2.7)</td>
<td>14.8</td>
<td>2.7</td>
</tr>
<tr>
<td>Luby et al.</td>
<td>Pakistan (urban)</td>
<td>0–15</td>
<td>4691</td>
<td>Twice-weekly visit with 3–4 day recall</td>
<td>44.8 (11.0)</td>
<td>Weekly period prevalence</td>
<td>3.4</td>
<td>130.3 (34.2)</td>
<td>56.7 (46.8)</td>
<td>43.2 (2.8)</td>
</tr>
<tr>
<td>Luby et al.</td>
<td>Pakistan (urban)</td>
<td>0–66</td>
<td>8949</td>
<td>Twice-weekly visit with 2–3 day recall</td>
<td>34.2 (7.6)</td>
<td>Weekly period prevalence</td>
<td>5.0</td>
<td>190 (42.7)</td>
<td>82.0 (51.5)</td>
<td>3.2 (2.0)</td>
</tr>
</tbody>
</table>

Design effects were calculated as $\text{DEFF} = \frac{\text{SE}_{\text{clustered}}^2}{\text{SE}_{\text{unclustered}}^2}$; diarrhoea was treated as a binary variable. Prevalence of diarrhoea (days or weeks with diarrhoea over days or weeks observed) was compared between treatment and control using log-binomial regression (family = binomial, link = log). Clustering was accounted for by using robust SE. $\text{SE}_{\text{unclustered}}$ was calculated ignoring any within-person or within-group correlation of diarrhoea. $\text{SE}_{\text{clustered}}$ to account for clustering at group level used the group ID as cluster variable (the unit of randomization). If the original data provided no treatment variable, clusters were allocated post-hoc at random to a simulated treatment and control arm. DEFF due to clustering at group level was calculated as $\frac{\text{SE}_{\text{overall}}}{\text{SE}_{\text{unclustered}}}$. $\text{DEFFs}$ show the same analyses for incidence (new episodes/days or weeks observed), calculated analogously to prevalence, using Poisson regression without $\text{SE}_{\text{unclustered}}$ and with robust SE to account for clustering ($\text{SE}_{\text{clustered}}$).
### Table 5: Examples of different epidemiological studies and suggested sampling strategy

<table>
<thead>
<tr>
<th>Study example</th>
<th>Suggested sampling strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1</strong></td>
<td></td>
</tr>
</tbody>
</table>
| Context: RCT of household level food hygiene promotion to reduce the burden of diarrhoea, delivered by community health workers to mothers of young children | Outcome measure: LP  
Surveillance duration: 1 year  
Sampling frequency: every 6–8 weeks (~6–9 contacts)  
Recall period: 3 days  
Data type: point prevalence  
Comment: Incidence is not suitable as the treatment aims to lower disease burden, for which LP is likely to be a better measure. Sampling at intervals (with a corresponding increase in the overall sample size) is chosen to decrease survey effects and bias.  
3-day recall is chosen to minimize recall error.  
The study is done over 1 year to study potential seasonal effects in food contamination.  
For the sample size within-household clustering can be ignored as the average number of young children per household is usually small (less than two). |
| Study population: children aged <5 years |                             |
| Logistics: adequate budget, trained staff and large eligible population available |                             |
| **2**         |                             |
| Context: RCT of household level food hygiene promotion to reduce the burden of diarrhoea, delivered by community health workers to mothers of young children | Outcome measure: LP  
Surveillance duration: 5 months  
Sampling frequency: every 4 weeks (~6 contacts)  
Recall period: 7 days  
Data type: period prevalence  
Comment: Incidence is not suitable as the treatment aims to lower disease burden, for which LP is likely to be a better measure. Sampling at intervals (with a corresponding increase in the overall sample size) is chosen to decrease survey effects and bias.  
7-day recall (period prevalence) is chosen to maximize power.  
The study is restricted to 5 months because of the tight budget, focussing on the hot season where food contamination may be most common.  
For the sample size within-household clustering can be ignored as the number of young children per household is small |
| Study population: children aged <5 years |                             |
| Logistics: tight budget, trained staff scarce, large eligible population available |                             |
| **3**         |                             |
| Context: RCT of household level food hygiene promotion to reduce the burden of diarrhoea, delivered by community health workers to mothers of young children | Outcome measure: LP  
Surveillance duration: 5 months  
Sampling frequency: every 2 weeks (~12 contacts)  
Recall period: 3 days  
Data type: point prevalence  
Comment: Incidence is not suitable as the treatment aims to lower disease burden, for which LP is likely to be a better measure. Frequent sampling is chosen to make the most of the small sample size. Short recall (point prevalence) is chosen to minimize recall error. Because of the short visit intervals, longer recall periods do not add much power.56  
The study is restricted to 5 months because of the tight budget, focussing on the hot season where food contamination may be most common.  
For the sample size within-household clustering can be ignored as the average number of young children per household is usually small (less than two). |
| Study population: children aged <5 years |                             |
| Logistics: tight budget, trained staff scarce, eligible population small (e.g. refugee camp) |                             |

(continued)
<table>
<thead>
<tr>
<th>Study example</th>
<th>Suggested sampling strategy</th>
</tr>
</thead>
</table>
| **Context:** RCT of a new vaccine against a pathogen causing severe diarrhoea  
Study population: children aged <5 years  
Logistics: adequate budget, trained staff and large eligible population available | **Outcome measure:** incidence  
**Surveillance duration:** 12 months  
**Sampling approach:** passive surveillance of hospital admissions  
**Recall period:** Not applicable (NA)  
**Data type:** incidence of severe episodes  
**Comment:**  
Incidence is suitable as the treatment aims to lower disease transmission of a specific pathogen.  
Passive surveillance is chosen because a vaccine can be delivered relatively easily to a large study population, focusing on episodes of particular clinical interest.  
Because hospital admissions do not allow estimating the effect of the vaccine on LP (a better marker for adverse effects on nutritional status), one could consider adding a substudy with active surveillance similar to Example 1, as was done in a Vitamin A trial in Ghana.12 |
| **Context:** cluster RCT of a large rural sanitation programme delivered at village level  
Study population: all ages  
Logistics: tight budget, trained staff scarce, large eligible population available | **Outcome measure:** LP  
**Surveillance duration:** 1 year  
**Sampling frequency:** every 6–8 weeks (~6–9 contacts)  
**Recall period:** 7 days  
**Data type:** period prevalence  
**Comment:**  
Incidence is not suitable as the treatment aims to lower disease burden, for which LP is likely to be a better measure.  
Sampling at long intervals (with a corresponding increase in the number of included villages) is chosen to limit the number of surveillance teams and transport costs. The sampling procedure aims to measure the outcome in one village per day per team. In a cluster randomized trial, more frequent surveillance rounds add relatively little power.  
7-day recall (period prevalence) is chosen to maximize power. Data on 3-day point prevalence can be obtained in addition as a secondary outcome.  
The effect of the intervention in children aged <5 years can be a secondary outcome. Because of the great uncertainties in study power due to the cluster-design, it is preferable to include all household members to maximize power. This is specifically the case if there is little reason to assume the intervention will affect young and older ages differently. |
| **Context:** observational study with recurrent infections as exposure (e.g. to study association between diarrhoea and reduction in weight-for-age Z-score)  
Study population: children aged <5 years  
Logistics: adequate budget, trained staff and large eligible population available | **Outcome measure:** LP  
**Surveillance duration:** 1 year  
**Sampling frequency:** every 2 weeks (~20–25 visits)  
**Recall period:** 3 days  
**Data type:** point prevalence  
**Comment:**  
Frequent sampling is chosen to minimize bias towards no effect. Efforts should be made to keep study participants happy and interested. Bias is not a great concern in an observational study without differential treatment of study participants.  
Short recall period is chosen to minimize bias that could exaggerate the effect size. |
Conclusion

When planning a study that measures diarrhoea, investigators must jointly consider the interdependent methodological points we have discussed in this article, which include recall periods, measures of disease occurrence (incidence vs prevalence), sampling frequencies and design effects. For example, the sampling frequency and the choice of the measure of disease occurrence can both influence the design effect. Conversely, the design effect can influence the choice of the sampling frequency or the recall period, because a strong design effect limits the study power gained from frequent sampling and long recall periods. Further, study settings differ from one another, especially in their logistics, which in turn has great implications for the study design. In some places it is difficult to recruit and train many field workers; in others it may be difficult to recruit study participants. As a consequence, it is difficult to develop universally applicable guidelines or a simple algorithm to identify the best way to measure diarrhoea in a specific study. In Table 5 we list examples of diarrhoea studies and suggest approaches to measure diarrhoea. None of our suggestions is meant to be absolute. As already suggested by Table 2, investigators must consider the

### Table 5 Continued

<table>
<thead>
<tr>
<th>Study example</th>
<th>Suggested sampling strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>7</strong></td>
<td></td>
</tr>
<tr>
<td>Context: observational study &gt;1 year, aimed at detailed exploration of clinical features of individual episodes (e.g. illness duration, severity, clinical signs and symptoms, stool testing for pathogens) Study population: children aged &lt;5 years Logistics: adequate budget, trained staff and large eligible population available</td>
<td>Outcome measure: incidence Surveillance duration: 1 year Sampling frequency: once a week (~50 contacts) Recall period: 7 days Data type: point prevalence data from which incidence can be calculated Comment: Frequent sampling is chosen to accurately establish the beginning and end of episodes, and to record clinical signs and symptoms in detail. Efforts should be made to keep study participants happy and interested. Bias is not a great concern in an observational study without treatment allocation of study participants. Continuous disease records may be needed, but depending on the budget, the surveillance period can be cut into blocks of, e.g., 6–8 weeks where surveillance is intense. This could allow capturing different seasons where different pathogens may circulate (dry cold season, wet season, hot season).</td>
</tr>
<tr>
<td><strong>8</strong></td>
<td></td>
</tr>
<tr>
<td>Context: demographic and health survey (DHS). The aim of the survey is to gain information on a range of topics, but the investigator also wishes to explore risk factors for diarrhoea (e.g. water, sanitation, socio-economic status) Study population: all ages Logistics: adequate budget, trained staff and large eligible population available</td>
<td>Outcome measure: LP Sampling frequency: one visit Recall period: 2–3 days Data type: point prevalence Comment: A short recall period is preferred to minimize recall error. A DHS usually aims to estimate prevalence as an absolute figure, not primarily to compare two groups, and therefore requires accurate data. Given the large sample size of most DHS surveys, loss of power due to a short recall period is normally not a big issue. Point prevalence data may often be easier to interpret and compare with, than period prevalence data, since diarrhoea definitions used in most DHS and epidemiological studies are based on disease experience during one day.</td>
</tr>
</tbody>
</table>
research question first, as many critical decisions depend on it. For example, incidence of diarrhoea (such as hospital admissions) could be the preferred measure in vaccine trials. Point or period prevalence measured at long intervals could be ideal for large environmental health interventions in high-risk populations where many villages and individuals need to be surveyed over a long time. A high-risk study population here means a setting where malnutrition and case fatality are a public health problem. Some studies (such as Demographic and Health Surveys) require obtaining precise absolute prevalence figures, for which collecting point prevalence data with a short recall period is most suitable.

We did not describe a number of important methodological challenges in diarrhoea trials that have been discussed elsewhere, such as the clinical definition of disease severity, or objective proxy markers for diarrhoea in trials of interventions that cannot be blinded. We also did not discuss recent advances in diagnostic tools for pathogen identification currently in use in some population-based studies.

Diarrhoea continues to be a major global health problem, and there is an ongoing debate over identifying research priorities and the development of cheap and effective interventions, given the limited funding. Whereas standard clinical trial procedures are often adequate to assess the effect of a vaccine or drug on diarrhoea in individuals, environmental interventions aiming at diarrhoea control are often much more complex, and more difficult to evaluate with randomized trials. Efficient methods to measure diarrhoea should allow more valid and generalizable results from research to be conducted with the same resources, especially in settings where resources are scarce.

<table>
<thead>
<tr>
<th>Box 1 Definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Diarrhoea day</strong></td>
</tr>
<tr>
<td><strong>Diarrhoea episode</strong></td>
</tr>
<tr>
<td><strong>Diarrhoea incidence</strong></td>
</tr>
<tr>
<td><strong>Diarrhoea point prevalence</strong></td>
</tr>
<tr>
<td><strong>Diarrhoea period prevalence</strong></td>
</tr>
<tr>
<td><strong>Recall period</strong></td>
</tr>
<tr>
<td><strong>Longitudinal prevalence</strong></td>
</tr>
</tbody>
</table>
Funding
This work was funded by the Wellcome Trust (WT082569AIA).

Acknowledgements
We thank Claudio Lanata, David Ross, Saul Morris, Adam Trotta, Wim van der Hoek, Mimi Jenkins, John Colford, Daniel Maeusezahl and Jan Hattendorf for providing data sets for the design effect calculations; Mike Kenward, Ben Armstrong and Richard Hayes for advice on calculating design effects for cluster-randomized diarrhoea trials; Zaid Chalabi and Rachel Peletz for advice on sampling strategies and comments on the draft.

Conflict of interest: None declared.

KEY MESSAGES
- The design of epidemiological studies on diarrhoea requires specifying recall periods, sampling frequencies and outcome measures that are most suitable to answer the research question in a given setting.
- Sample size calculations often need to be done based on scarce data. This article outlines how the validity and logistical efficiency of diarrhoea studies can be improved by careful consideration of these factors.

References

59 Trotta A. Statistical analysis and lessons learned from a hygiene intervention carried out 20 years ago in Lima, Peru. MSc Thesis. London School of Hygiene and Tropical Medicine, 2010.


