WSUP
Water & Sanitation for the Urban Poor

DISCUSSION PAPER

DP#001 • OCT 2011

Evaluating the health impact of urban WASH programmes: an affordable approach for enhancing effectiveness

This paper argues for more widespread evaluation of the health impacts of WASH interventions: not with the aim of demonstrating that WASH can improve health (we know it can), but rather with the aim of assessing the impact of particular interventions. We suggest that more frequent evaluation could contribute to improved effectiveness, by encouraging investors and implementers to focus on impacts rather than outputs (such as number of toilets constructed).

More widespread health impact evaluation would also enable more objective comparative assessment of the value-for-money of different types of urban WASH intervention. Further, we argue that health impact evaluation need not be as costly as is widely thought. We discuss available methods, and suggest that the most appropriate approach in urban WASH evaluation contexts will often be the before-after concurrent control (BAC) design.

shareSanitation and Hygiene Applied Research for Equity

This Discussion Paper is co-published by Water & Sanitation for the Urban Poor (WSUP) and the Sanitation and Hygiene Applied Research for Equity (SHARE) Research Consortium. It is presented at this stage as a basis for sector debate, and should not be considered a definitive statement of the views of these organisations.

Contents

A. Summary

A key justification for investment in water, sanitation and hygiene (WASH) is to reduce the burden of diseases transmitted by the faecal-oral route, most notably child diarrhoea. In dense low-income urban communities with poor water and sanitation services, there can be little doubt that genuine improvements to WASH can have a substantial positive impact on health: in other words, there can be little doubt about the general relationship between WASH and health. However, it seems likely that many *specific* WASH interventions do not achieve a significant health impact, because they do not sufficiently influence disease transmission pathways. Currently, health impact evaluations are rarely carried out for WASH interventions, reflecting a widespread perception that they are too expensive.

This paper argues for more widespread evaluation of the health impacts of specific interventions: not with the aim of demonstrating that WASH *can* improve health (we know it can) but rather with the aim of assessing the impact of particular interventions in specific settings. Affordable methodologies suited to this type of aim are available but are under-utilised. We suggest that more frequent evaluation of the health impacts of specific interventions could contribute to improved effectiveness, by encouraging investors and implementers to focus on impacts rather than outputs (such as number of toilets constructed). More widespread health impact evaluation could also enable more objective assessment of the value-for-money of different types of urban WASH intervention.

This paper does not suggest that health impact evaluations will be appropriate for all urban WASH programmes; but we do argue that they are more affordable and more feasible than is widely assumed. We discuss the various available methods, and suggest that the most appropriate design in urban WASH programme evaluation contexts will often be the before-after concurrent control (BAC) design. Randomised controlled trials (RCTs), while clearly the 'gold standard' design in some contexts, will often not be appropriate or feasible for urban WASH programme evaluation. We offer detailed practical guidance on the BAC design, with particular reference to achieving acceptable costs and integrating with the logistics of programme planning and implementation. Finally, we suggest practical ways in which investors and implementers might cooperate to enable more frequent evaluation of health impacts.

1. Introduction

A key justification for investing in urban water, sanitation and hygiene (WASH) is to reduce the burden of diseases transmitted by the faecal-oral route. In particular, WASH improvements are expected to reduce the high child mortality associated with diarrhoea, prevent outbreaks of cholera and lessen the impact of debilitating parasitic infections. As discussed in Section 9 of this paper, reduced disease burden is not the only benefit of urban WASH interventions; but it is certainly one of the most important benefits.

Nevertheless, urban WASH interventions rarely include health impact evaluation (HIE) within their evaluation procedures. In the authors' experience, this reflects a widespread view that HIEs are too expensive. An influential 1980s report by Briscoe et al. (1986) stated that rigorous HIEs using "*the standard concurrent-control quasi-experimental design*" are extremely expensive; the authors cite a World Bank (1976) report which concluded that the benefits of HIEs using designs of this type do not justify the very high costs.

However, several more recent publications (notably PREM 2006 and Clasen et al. 2010) have argued for more frequent evaluation of the health impacts of WASH interventions. Clasen et al. (2010) point to the value of rigorous observational studies in programme evaluation contexts.

This Discussion Paper argues, like PREM (2006), that sufficiently rigorous HIE need not be as prohibitively expensive as is widely assumed. We present practical guidelines on how to achieve sufficiently valid health impact estimates at reasonable cost, focusing on one particular study design (before-after with concurrent control). We suggest that randomised controlled trials (RCTs) will often not be an appropriate choice for the evaluation needs discussed here, particularly in urban settings. Further, we argue that there is an additional reason for carrying out HIEs that has not been sufficiently considered to date: that more frequent evaluation of the health impacts of specific interventions will tend to encourage more relevant intervention designs, and tend to favour increased intervention effectiveness.

2. Scope of this paper

- This paper focuses specifically on urban WASH interventions, particularly in settings with *high population density*.[1]
- Within the dense urban context, this paper will focus on community/neighbourhood-level interventions, as opposed to household- or individual-level interventions.[2]
- The primary focus of this paper is on *HIE methods that minimise cost while maintaining adequate rigour*: this reflects a primary concern not with "generalisability" (i.e. with demonstrating in a general sense that sanitation interventions can improve health), but rather with assessing the extent to which a particular intervention has achieved a health impact.[3]

Sections 5 and 6 of this paper discuss study designs, and in particular the before-after concurrent control (BAC) design, in some technical depth. Readers without a particular interest in these more technical aspects may prefer to leave these two sections aside: the central arguments of this paper are covered in the remaining sections.

[1] However, many of the conclusions of this paper may be equally applicable to rural settings and quasi-rural peri-urban settings.

[2] This has important implications for health impact evaluation. In a neighbourhood-level intervention to reduce open defecation (OD) or construct a sewerage network, we can expect a given person's health to be affected by district-level impacts of OD reduction or sewering (i.e. by what everyone else does), over and above individual-level impacts due to that person stopping OD or connecting to the sewer. By comparison, household water treatment (for example) can be expected to have household-level impacts only.

[3] However, as further discussed in Section 7, BAC studies as outlined here may contribute to cumulative bodies of evidence from which generalisable inferences can be drawn.

3. Impacts of WASH on health

This paper is not a review of previous evaluations of the impact of WASH interventions on health: for this, the reader is directed to existing systematic reviews including Esrey (1985), Esrey (1991), Fewtrell (2005), Clasen et al. (2009), Waddington (2009), Clasen et al. (2010) and Norman et al. (2010). Several of the more recent of these studies (Fewtrell 2005; Waddington 2009; Norman et al. 2010) have included meta-analyses, and in all cases the conclusion has been that WASH interventions of different types tend to substantially reduce disease burden, notably child diarrhoea. However, all systematic reviews to date have highlighted the methodological difficulties of demonstrating the health impacts of WASH interventions, the paucity of rigorous evidence from the field, and the lack of comparability of results from different studies. Indeed, Clasen et al. (2010) state that *"Differences in study populations and settings, in baseline sanitation levels, water, and hygiene practices, in types of interventions, study methodologies, compliance and coverage levels, and in case definitions and outcome surveillance limit the comparability of results."*

4. Why health impact evaluation?

Debates about how to assess, and indeed whether to assess, the health impacts of WASH interventions have been ongoing since the 1970s (see especially Blum & Feachem 1983; Briscoe, Feachem & Rahaman 1986; Esrey 1986). Briscoe, Feachem & Rahaman (1986; pages 12/13) suggest that HIEs may be done a) to assess generalisable causal relations i.e. aiming for external validity, or b) to assess location-specific causal relations as a basis for ongoing investment decisions in the same location. PREM (2006) propose a longer categorisation of possible reasons.[4]

In this paper, we suggest that the search for a generalisable causal relation between WASH improvements and health risks is of limited usefulness. In dense low-income urban communities with poor baseline water and sanitation services, there can be little doubt that WASH interventions can – *if properly implemented* – have a substantial positive impact on health: we can reasonably assume this on the basis of the biological plausibility of faecal-oral disease transmission, the imperfect but rich evidence of health impact evaluations to date, and arguably from historical experience. However, to expect a consistent magnitude of effect (a 10% reduction in diarrhoea incidence? 30%? 75%?) is clearly unrealistic: the health impacts of WASH interventions can be expected to vary greatly depending on very diverse factors including local disease ecology, baseline water and sanitation quality, baseline hygiene practices, and the precise nature of the intervention.

In fact, there is already an extensive evidence base for the causal relationship between WASH and diarrhoea, which can be used (and is being used) to convince decision-makers and the charitable public of the likely health impacts of improving WASH in poor communities: we suggest that, rather than striving to "demonstrate" these impacts, it may be more useful to focus on identifying the most efficient means for achieving these impacts across different contexts.[5]

[4] *"First, demonstrating that a particular WSS [Water Supply and Sanitation] program yields health, socioeconomic, and poverty reduction benefits can be used to build support for program expansion or modification. Second, even though specific WSS programs show great promise, they might not work under all field conditions. Program outcomes can be highly variable, with some interventions and programs in some settings showing little impact. Good evaluations can identify why this might happen and what adjustments can be made to correct it. Third, if small-scale WSS projects are to make an important contribution to government policy, they need to be expanded or "scaled up". It is important to know what aspects of these projects lead to greater or lesser success. Finally, disseminating results of WSS outcomes will contribute to the economic development community's broader understanding of water and sanitation service delivery tools".* (PREM 2006, page 2).

[5] That said, we certainly think that local impact evaluations – in the words of Briscoe, Feachem & Rahaman (1986) cited above, "to assess location-specific causal relations as a basis for ongoing investment decisions in the same location" – are of value not just for objective decision-making, but also for political advocacy.

Secondly, we suggest that there is a very important reason for carrying out HIEs that has been under-appreciated to date: *as a metric and thus driver of programme effectiveness*. If a financing or implementing organisation knows that its performance will be judged on the basis of health impact, we suggest that this will tend to improve performance in this regard, at both the design and implementation phases.

We suggest that a requirement to demonstrate health impact may help focus investment and effort on intervention locations and intervention types that can genuinely be expected to achieve strong health outcomes. In contrast, implementing agencies are currently often obliged by governments and donors to express their targets in terms of indicators like "increase in number of people with improved sanitation", and this may sometimes lead to selection of intervention districts on the basis of how easy it is to meet this type of target, rather than in terms of health priority (Norman & Pedley 2011). In the case of sanitation in particular, we suggest that more frequent health impact evaluation would encourage more inclusive district-wide "total sanitation" strategies: so for example, providing improved toilets for 50% of a district's population may be insufficient to achieve substantial health impact if the waste of the remaining 50% continues to contaminate the environment. Likewise, providing improved latrines for 100% of the district's population may be of little value if the district still floods twice a year, leading to widespread latrine overflow.

This approach ties in to ideas of results-based aid (RBA) (see Pearson 2011 and Trémolet 2011), in which finance is disbursed only after verification that the desired outputs, outcomes or impacts[6] have been achieved. Examples of RBA modalities include output-based aid (OBA), cash-on-delivery aid (COD) and conditional cash transfers (CCTs) (Pearson 2011); see also the related concept of Progress-Linked Finance (WSUP/ODI 2011).

In a typical OBA application in the WASH sector, the implementing agency receives payment only after verification of satisfactory completion of all outputs. It is conceivable that the disbursement could be at least partially tied to verified health impacts (as opposed to more immediate outputs or outcomes); though certainly the implementing agency might not be prepared to assume this high level of uncontrolled risk.[7] A variant to overcome the uncontrolled risk problem might be to use direct output or outcome measures as criteria for disbursement, but for bonus payments to be made if health impact is demonstrated. This could potentially generate incentives for effective operation and maintenance over time.

[6] The terms *output*, *outcome* and *impact* refer to different stages in achievement of the project/programme goals: for example, the immediate output of a sanitation intervention might be construction of latrines; the desired outcomes might include hygienic use of the new latrines; while the desired eventual impacts would include improved health.

[7] If disbursement is on the basis of the construction output, the implementing agency (e.g. a water utility) has strong control of this output, and therefore can more readily accept the risk of not achieving the output and thus the disbursement. But disbursement on the basis of health impact requires the utility to assume a risk that it may judge to be insufficiently under its control; so the utility might not be prepared to enter into an OBA agreement.

> There are many economic and sociological factors related to both WASH service levels and disease burdens

5. Methods of HIE in WASH interventions: an overview

[As noted, Sections 5 and 6 consider technical aspects: the reader without a strong interest in these aspects may prefer to go straight to Section 7.]

The following section discusses the following candidate study designs for health impact evaluation:

1) Ecological analyses
2) Case-control and cohort studies
3) Randomised controlled trials
4) Before-after studies
5) Before-after studies with concurrent control

It is often suggested that decisions as to the merit of different WASH interventions should be based on the "best available evidence". This makes sense, but only if the often significant limitations of the best available evidence are adequately recognised. For example, the association between sanitation and childhood mortality is often studied using country or state-level data, because data from trials and observational studies is scarce or unavailable. It is often argued that in this case one should conduct an ecological analysis, e.g. attempting a linear regression analysis with area-level sanitation coverage as exposure and area-level child mortality as outcome, and then treating the result as the "best available evidence". However, simple ecological analyses of this type have so many analytical problems that the result cannot be trusted (see overleaf).

Public health decisions should not be based on evidence that must be assumed *a priori* to be fundamentally flawed. If no reasonably unbiased evidence is available, the merits of different interventions are better judged on the basis of biological plausibility, non-health benefits, risks of adverse effects, and aspects of scalability and logistics (Ross et al. 2006; Schmidt & Cairncross 2009). The following description of different HIE methods for WASH interventions therefore does not aim to identify "the best available method", but rather tries to identify methods that can be expected *a priori* to provide reasonably unbiased estimates.

5.1 Ecological analyses

Conventional ecological analyses typically use cross-sectional data collected at the level of the population rather than the individual. For example, one could plot sanitation or water-access coverage against child mortality or proportion of children with malnourishment. There are two major problems with this, ecological fallacy and confounding. Ecological fallacy means for example that sanitation is not related to child mortality when analysing aggregate data at country or state level, despite an association between the two found among individual households in each state or country. The classic example comes from HIV research: the US has both a higher circumcision rate and a higher HIV prevalence than European countries, which might be taken to suggest that circumcision increases the risk of HIV; but when we consider individuals (as opposed to countries), circumcision proves to be protective against HIV.

In the WASH field, however, the main concern with ecological analysis is confounding. There are many economic, developmental and sociological factors that are related to both WASH service levels and disease burdens, producing a strong association between the two that is not entirely attributable to WASH effects. It has been shown theoretically that there are severe difficulties in accounting for this type of spurious association by multivariate analysis (Kaufman et al. 1997). The great potential for residual confounding due to unknown confounders and imprecise confounder measurement virtually precludes using conventional ecological analysis for the evaluation of WASH interventions.

> “Adoption of WASH improvements is highly related to socio-economic status”

5.2 Case-control and cohort studies

Case-control and cohort studies are the classic observational study designs. Case-control studies in principle can be treated as a special type of cohort study, if one regards the control group as a random sample of the whole cohort. In each study type, the study population is divided into [exposed] and [non-exposed], and into [experiencing the outcome] and [not experiencing the outcome]. In conventional case-control and cohort studies, exposure status (e.g. having a latrine) does not change over the course of the study. To the extent that selection of controls is not random, case-control studies are at risk of selection bias. Case-control studies and cohort studies have in common that they are highly susceptible to confounding, i.e. failure to account for factors that are related to both exposure (e.g. sanitation) and outcome (e.g. diarrhoea), producing a spurious association. Case-control and cohort studies are a necessary evil in evaluating some types of exposure that cannot be randomised, such as radiation or smoking, and have been found to be useful a) if the association under study is very strong (say a four-fold or higher difference in risk) and b) if the potential for confounding is not very large.

One problem with using case-control or cohort designs to evaluate WASH interventions is that adoption of and compliance with water, sanitation and hygiene improvements is highly related to socio-economic status, education and “modern lifestyle” (Schmidt et al. 2009a), factors which are difficult to measure accurately but which are themselves highly related to the risk of diarrhoea. Consider for example the difficulties of defining and measuring “poverty” as a potential confounder: poverty is perhaps the most important “upstream” risk factor for diarrhoea and inherently related to WASH, but is very difficult to objectively define and accurately measure.

The potential for confounding in case-control and cohort studies studying factors related to WASH is very large, in fact so large that one must *a priori* regard such evidence as unreliable. It is little wonder that previous cohort and case-control studies evaluating WASH interventions have shown very large effect sizes, suggesting massive effects of WASH on diarrhoea, child mortality or height-for-age Z-score (HAZ) (Azurin & Alvero 1974; Young & Briscoe 1988; Hoque et al. 1996; Nanan et al. 2003): we suggest that these estimates need to be treated with great caution.

Recently, it has been suggested that by adequate matching of cohort populations, a high degree of comparability between intervention and control communities can be obtained, thus overcoming issues of confounding and bias in unmatched cohort studies. Arnold and colleagues used propensity score matching to retrospectively match several intervention clusters that had received a multifaceted WASH intervention to control clusters (Arnold et al. 2009). Propensity scores are a method to calculate the probability of receiving an intervention based on known variables such as education and socio-economic status (Rubin 1997).

Intervention and control communities were in this study matched according to this probability (propensity score), achieving a great similarity between control and intervention clusters with regard to potential confounders. Based on this analysis they found that intervention and control clusters did not differ in terms of children’s nutritional status, and concluded that the intervention had no effect on this. It has been suggested that propensity score matching may be slightly better at controlling for confounding than multivariate regression approaches, though perhaps only in small studies (Cepeda et al. 2003). As with conventional multivariate analysis the biggest problem with propensity scores is that they can only account for known and observed confounders, not for unknown confounding (Rubin 1997). As outlined above, the potential for residual confounding (due to imprecise confounder measurements and unmeasured confounders) in WASH studies is high, so that propensity score matching will be of limited practical value.

> "RCTs are widely seen as the "gold standard""

In the presence of strong socio-economic confounding, matching of cohort clusters (using propensity or other methods) does not achieve better control for confounding than conventional regression analysis. Rather, in this situation it is akin to doing a conventional regression analysis but simply omitting control clusters that are not very similar to the intervention clusters. This can facilitate the logistics of the study, since control clusters that are different would not have contributed much to the analysis (Arnold et al. 2009). In terms of control for confounding, matching adds little value especially if confounding is strong.

In conclusion, we suggest that case-control and cohort studies that cannot make use of changes in exposure status by incorporating a "before and after" element are unlikely to be useful for HIE of WASH interventions.

5.3 Randomised controlled trials (RCTs)

Randomised controlled trials (RCT) are widely seen as the "gold standard" design for health impact evaluation. However, this is correct only under certain conditions. Clearly, HIEs of large-scale WASH interventions are very different from the classic context for RCTs, i.e. clinical trials to estimate drug efficacy. RCTs control for known and unknown confounding, and work well if both participants and observers can be blinded to both treatment and outcome assessment; but blinding of participants to large-scale urban WASH interventions is generally not possible. RCTs can still provide reasonably unbiased estimates if the outcome is an objective measure such as weight gain, parasites in stools, or death; but these have rarely been used in WASH trials. In contrast, RCTs using a more subjective outcome measure, such as self-reported gastro-intestinal symptoms, may often provide biased effect estimates: study participants in the intervention arm may tend to under-report disease for fear of being seen as non-compliant or impolite; those in the control arm will often have a strong incentive to over-report disease because they want to gain access to the intervention. Also, individuals of lower socio-economic status and education (who have a particularly high risk of disease) tend to be lost to follow-up more often: drop-out can be different between the intervention and control groups, thus introducing additional bias. Bias can also come from over-enthusiastic field workers who out of commitment or for the sake of job security want to demonstrate the effectiveness of an intervention (observer bias).

For these reasons, we would argue that RCTs using self-report measures of disease (the most common measure in WASH-sector RCTs to date) cannot be considered a "gold standard". The limitations of effect estimates obtained without ruling out participant (responder) bias and observer bias are well-known (see e.g. Fewtrell et al. 2005; Clasen et al. 2007), and indeed it has been suggested that effect estimates based on diarrhoea self-reports are likely to be severely biased, to the extent of calling into question effect estimates based on this measure (Schmidt & Cairncross 2009b). Other authors have argued that, with suitable protocols, bias need not be severe: i.e. diarrhoea self-reports, though certainly not an ideal measure, are of value as a health measure (Clasen et al. 2009); see also below Section 6.2. In our opinion, the risk of bias is important, and HIEs —whether RCTs or other designs— should increasingly strive to use objective measures not based on self-report. Such measures include mortality, height-for-age and helminths in stools (Schmidt et al. 2009b). Recently, a handwashing intervention against influenza in Hong Kong was evaluated using PCR of saliva from throat swabs that, importantly, were sampled independently of symptoms (Cowling et al. 2009), thus avoiding a potentially biased referral process based on reported symptoms. In the near future, objective molecular markers for recent infection with different diarrhoea pathogens may become available for WASH impact studies.

Independently of these methodological concerns, there are concerns about the generalisability of RCT findings. RCTs tend to be conducted in "ideal" settings, chosen primarily because doing an RCT is possible; in such a situation, external

validity (i.e. generalisability of results) may be poor, because most interventions are not carried out in ideal settings. For example, the London School of Hygiene and Tropical Medicine (LSHTM) is currently conducting an RCT on the health impacts of chlorine pills for household-level water treatment, in the Indian state of Orissa. Within Orissa, poverty and poor water quality are most severe in tribal areas in the hill forests, whereas the coastal regions are generally less poor: but the study is being carried out in coastal areas, because conducting an RCT in the tribal areas was not logistically viable.

Especially in urban settings, community-level WASH interventions may be difficult to randomise. For example, randomisation of sewerage interventions is politically and technically difficult or impossible: despite the theoretical possibility of "stepped wedge" designs[8] (Brown & Lilford 2006, De Allegri et al. 2008), in practice it would be extremely difficult for a major sewerage project to select its districts of intervention not on the basis of financial factors, social concerns and engineering logic, but randomly (Norman et al. 2010). In line with this, we are not aware of any previous attempt to randomise a sewerage intervention. Similar difficulties arise with networked water supply interventions. Other types of district-level intervention (for example, stop-open-defecation campaigns) are in theory more readily randomisable; though note that the requirement for randomisation would need to be assumed early on in the project planning process (*see Section 8*). Note also that randomisation is not a very flexible design feature of a research study: once randomisation is done at baseline, a trial must be conducted according to a relatively strict time line. Delays in the implementation process and/or drop-outs of whole clusters can quickly jeopardise random allocation.

Thus for a number of reasons we consider that RCTs will often not be the best choice for routine HIE of WASH interventions; nor should they necessarily be regarded as the "gold standard" design for public health interventions.

5.4 Before-after studies without control

Simple before-and-after studies without a concurrent control group have been used to evaluate rural and urban sanitation projects. For example, Barreto and colleagues conducted two separate cohort studies in the same population in Salvador de Bahia (Brazil), before and after an urban sanitation project with provision of sewerage connections (Barreto et al. 2007). Note that while both studies if treated separately can be regarded as cohort studies, the main analysis (comparison of diarrhoea rates before and after the intervention) is not a cohort design: it is a before-and-after design without a concurrent control group. The authors conducted a range of secondary analyses to underpin their findings. For example, they showed that in areas that had a lot of diarrhoea in the first survey, there was a stronger reduction in diarrhoea in the second survey than in areas with less diarrhoea at baseline. In fact, in areas with little diarrhoea at baseline, diarrhoea increased: this may well be a classic regression-to-the-mean effect, i.e. there were some communities in which by chance diarrhoea was exceptionally high in the first survey, and so was likely to decrease in a second survey. Likewise, there may have been communities that by chance had little diarrhoea at baseline, with subsequent upwards regression to the mean. The authors developed conceptual frameworks of the biological plausibility of different pathways by which sanitation may protect from diarrhoea, and incorporated these into quite sophisticated multivariate regression models. Nevertheless, none of these methods can address the main problem of before-after studies without a concurrent control group, i.e. that the overall 22% reduction in diarrhoea observed over a 6-year period (1998-2004) in the city of Salvador de Bahia may be explained at least partially by wider trends ("secular trends") acting independent of the intervention. Indeed this is to be expected since Brazil experienced very significant economic development during the study period.

To conclude, before-after studies without a concurrent control group are a weak design for HIE of WASH interventions.

[8] In a stepped wedge design, say 8 similar city districts are selected for the intervention (non-randomly, on the basis of need, or logistics, or some other consideration). The intervention is then planned to take place in two tranches: say 4 city districts in Year 1, and the remaining 4 districts in Year 3. Allocation of districts to tranches is then done randomly, so that the districts in the second tranche act as true experimental controls for the intervention districts in the first tranche.

5.5 Before-after studies with concurrent control (BAC studies)

Designs of this type have been referred to in the literature as concurrent-control quasi-experimental designs (e.g. Briscoe et al. 1986) and concurrent-control field trials (e.g. Kolahi et al. 2009). Baseline disease data are collected for an intervention and control population before project implementation, and disease is then again measured after the intervention. The critical point is that *separate* analyses are conducted for the intervention and control groups: first one looks at the before-after change in disease in the intervention arm, and then one looks at the before-after change in the control arm (with the aim of detecting secular trends).

This allows a "difference in difference" (DID) analysis: for example, if disease reduction was 30% in the intervention group and 10% in the control group, then the DID can be calculated as 30% - 10% = 20%. However, a meaningful statistical analysis of this difference providing valid confidence intervals around the DID can only be done if the number of geographically independent clusters in each arm is more than 4 or 5 (Hayes & Moulton 2009); most studies of this type have fewer clusters, though in practice statistical comparisons between intervention and control arm are often performed with small numbers of clusters by simply ignoring clustering (see e.g. Chavasse et al. 1999, Emerson et al. 1999);[9] such analyses are invalid (Hayes & Moulton 2009). Even if intervention characteristics are such that a somewhat greater number of clusters can be allocated to intervention and control (say 6 or 8 per arm), the power of any between-arm comparison may be very low, especially if the between-cluster variation in disease is high. This is why in a recent cluster-randomised handwashing trial the authors (including present author WS) specified the before-after analysis as the main outcome, although there were 5 clusters in each arm (Biran et al. 2009). The statistical between-arm comparison was done only as a secondary analysis because the authors had (correctly) anticipated that the power of this analysis would be very low.

Thus, the number of geographically independent clusters can be lower in a BAC study than in a cluster-randomised RCT. Often, in urban sanitation programmes the number of clusters is restricted because interventions extend over large geographic areas. For example, consider a situation in which one half of a city receives sewerage while the other half does not (or not yet): in this case there would only be two clusters, one allocated to the intervention and one to the control, and it is unlikely that this allocation could be done randomly.

In a BAC study, the number of independent clusters is from the purely statistical point of view not as critical as in an RCT. Each intervention and control cluster is compared with itself at different points in time, i.e. the unit of analysis is the sampling unit in each cluster (e.g. individuals or households). Thus, given a reasonably large sample size in the intervention and control arms (in terms of statistically independent units of analysis), the confidence interval of the before-after change in disease separately for intervention and control can be quite small (for further explanation see next section).

What about unknown confounders? BAC studies are much less prone to this problem than observational designs like case-control and cohorts, *because confounding is addressed by incorporating disease risk before the intervention*: most potential differences between the intervention and control group that are associated both with the probability of receiving the intervention and with disease risk at follow-up should be reflected in differences in disease at baseline (subject of course to random error). Nonetheless, BAC studies – unlike RCTs with a sufficient number of clusters – are subject to the possibility that secular trends in disease in the absence of an intervention could be different between intervention and control arm. This could arise due to fundamental differences in the epidemiology of WASH-related diseases in the intervention and control arms: for

[9] In fact these studies were randomised controlled trials: the comment here refers to the authors' comparison of mean before-after reductions between the intervention and control group.

> With only a small number of clusters, a BAC study can be better than an RCT

example as a consequence of being at different stages of socio-economic development, or because of major differences in risk factors and transmission routes for WASH-related infections, even if the overall disease burden is similar at baseline. If it is possible to randomly allocate clusters within a BAC design, this problem is limited. Careful selection of intervention and control clusters to make them as comparable as possible should also help to avoid including study clusters with very different infectious disease epidemiology. Again, the key difference with respect to conventional matched-cohort or case-control studies is that the baseline measurement of disease in BAC studies should incorporate a substantial part of *known and unknown* differences between intervention and control, which is in our view a critical advantage given the large potential for unknown and residual confounding in cohort and case-control studies without baseline measurement of disease.

Concerns with reporter bias in RCTs (see Section 5.3) are of course also applicable to BAC studies that use self-report measures of health: there is a particular risk of participant and observer bias if the disease survey is obviously related to the WASH intervention. But the common requirement in RCTs for informed consent to random allocation could well introduce more severe participant biases than in a BAC study. Most notably, if you are randomly allocated to a non-intervention group (and possibly have to sign a document stating this explicitly), there is a powerful incentive to exaggerate your family's ill-health, to convince "the authorities" that your district should be next in line. In a BAC study there is often a less strong requirement to link the study so explicitly to the intervention, making it possible to "sell" the study as a simple longitudinal health survey: so it may be easier in BAC studies than in an RCT to hide the purpose of the survey from participants and perhaps even field staff. In other words, with careful design of data collection protocols it should be possible to reduce reporter bias in BAC studies.

The lack of direct statistical comparison between intervention and control arms is a clear weakness of BAC studies. An RCT with sufficient power allows us to conclude that "the difference in post-intervention disease rate in the intervention group was X% lower than in the control group, with a 95% confidence interval of Y% to Z%". In a BAC study, all one can say is that "disease went down by X% in the intervention group, with 95% confidence interval Y% to Z%, versus only A% in the control group, with 95% confidence interval B% to C%". A BAC study is at risk of providing results that are difficult to interpret if a) the number of clusters is very small, i.e. one or two clusters per arm, and/or b) there are strong secular trends in disease risk in the absence of an intervention (see Section 6.5). Nonetheless, for the reasons outlined in this section and Section 5.3, RCTs may not be possible or not have sufficient power in these situations. This can be expressed in another way: if intervention characteristics mean that we have only a small number of intervention clusters, an RCT is unlikely to have sufficient power, and will provide results with a large confidence interval, which severely restrict interpretability and causal inference. Thus, if there is only a small number of clusters, a BAC study can be a better choice than an RCT.

The lack of direct statistical comparison between intervention and control arm is a drawback of BAC studies as compared to RCTs when attempting a meta-analysis, i.e. the pooling of effect estimates following a systematic review of all available trials of a given intervention. This is relatively straightforward for RCTs because they usually provide a single primary effect estimate with confidence intervals. We are unaware of established statistical methods that allow a formal meta-analysis for BAC studies. On the other hand, WASH interventions are commonly highly context specific, and trials evaluating their effects are almost inevitably heterogeneous, regardless of whether an RCT or BAC design is used. Strong heterogeneity, whether in the statistical or contextual sense, limits the applicability of formal meta-analysis in the WASH field even when pooling estimates from RCTs (Cochrane Collaboration 2011).

An example of the BAC design has recently been provided by Kolahi and colleagues who studied an urban sanitation project in Teheran (Kolahi et al. 2009). Two municipal districts received sewerage, which was accompanied by a reduction in diarrhoea prevalence of 46%. However, in the control group (neighbouring areas of each district with similar socio-economic conditions) in the absence of sanitation improvements, diarrhoea decreased by 37%, roughly the same order of magnitude as in the intervention districts (Figure 1). This example clearly demonstrates the weakness of studies without a concurrent control group. It also shows that BAC studies allocating as few as two clusters per arm can provide results that are unlikely to be severely biased or confounded. An RCT in this situation in Teheran would have been nearly impossible. Observational studies without a baseline measure (e.g. a cross-sectional survey post-intervention) would probably have been subject to strong confounding effects.

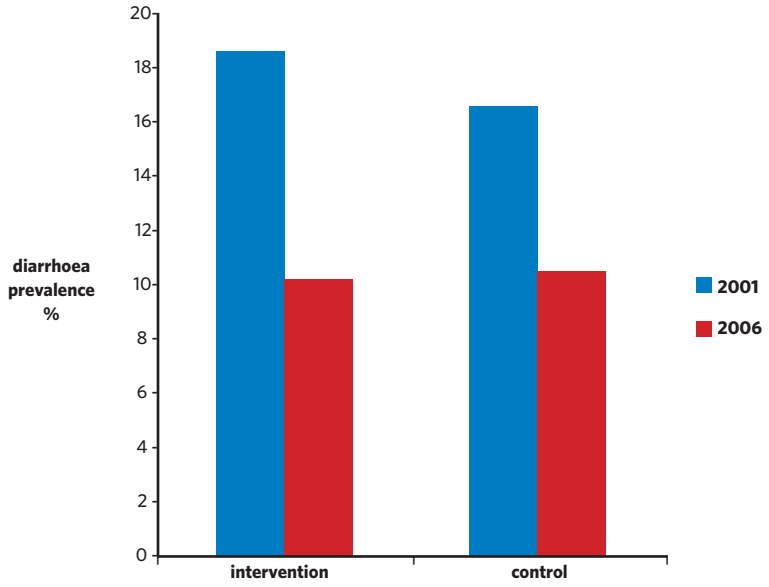

Figure 1. *Before-after study with concurrent control evaluating an urban sanitation project in Teheran (Kolahi, Rastegarpour & Sohrabi 2009).*

To conclude, the BAC design is not perfect, but we consider that it may often be the best option for assessing health impacts of urban WASH programmes in a cost-effective way that still delivers sufficient rigour. RCTs and cohort or case-control studies will, for different reasons, often not be appropriate for evaluation of the health impacts of urban WASH programmes. The BAC design is much better able to control for confounding than cohort or case-control studies. The risk of reporter bias when using self-report health measures, which is a serious limitation of RCTs to assess WASH impacts, may be relatively low (though cannot be excluded). In Section 6, we describe the BAC design in more detail, and highlight practical and methodological challenges with this design.

> By increasing the number of clusters we can increase plausibility

6. The design of BAC studies

This section discusses aspects of the design of BAC studies, and offers recommendations for their use in urban WASH contexts.

As detailed in the preceding section, the main criteria making BAC studies an appropriate choice in urban sanitation programmes are a) difficulty of randomisation and b) a small number of intervention clusters. If less than about 8-10 clusters per arm are available, then it is quite likely that a direct comparison of post-intervention disease rates will either not be possible (for less than 4 clusters per arm) or will have very low power. If the number of clusters is small, then the choice of these clusters will be critical for the validity of the study.

6.1 Choice of intervention and control clusters

The intervention group in WASH interventions is often chosen based on programme logistics or engineering considerations. BAC studies can be done with one intervention and one control cluster only, but the results will be much less convincing than if the intervention is done in more than one geographically independent cluster. Thus by increasing the number of clusters we can increase plausibility. For example in the Teheran study reported by Kolahi et al. (2009), the two intervention clusters showed similar results, and the two matched control clusters showed similar results: this adds confidence to the results. In other words, in a BAC study, one independent cluster per arm is better than none, two are better than one, and three are better than two. In general however, we do not recommend studies with only one intervention and control group (see Shadish, Cook & Campbell 2001), because the probability of obtaining a result that is difficult to interpret is high.

It may be tempting to divide a large cluster into several seemingly independent clusters, but these should not be treated as such. Selecting several smaller areas within each allocated cluster "to increase the number of clusters" for the purpose of HIE does not in fact increase the number of clusters: these smaller clusters are not geographically independent, since they belong to the same unit of allocation. The chance of being selected for intervention or control would be identical for these newly created "clusters": they will all be allocated to either intervention or control. Further, being subunits of the same large cluster (the true unit of allocation), they are likely to share similarities in terms of socio-demographic features and disease risk. The outcomes are therefore likely to be correlated and cannot be treated as statistically independent.

The most important aim in choosing intervention and control clusters is to balance baseline disease prevalence and the general epidemiological conditions influencing disease risk and disease trends as closely as possible. A BAC study will always be more convincing if the intervention and control started with roughly similar baseline disease risk (Shadish, Cook & Campbell 2001): a large difference in disease at baseline suggests that the intervention and control clusters differ in socio-economic or other aspects, and may be subject to different secular trends in disease risk (a major threat to the interpretability of BAC studies). For example, a high-risk area could experience a greater reduction in diarrhoea as a result of general economic progress, independently of an intervention, by comparison with a low-risk area where "cheap gains" in disease-risk reduction have already have been achieved as a result of earlier developments. Or it may be that at the time of study disease is dropping faster in a wealthier area due to unrelated factors.

> “Random allocation may be difficult in studies of district-level interventions”

To ensure similarity between intervention and control, allocation to intervention and control clusters is ideally done at random, similar to a classic RCT. For example, in the previously cited evaluation of a handwashing campaign in India (Biran et al. 2009), 5 villages were randomly allocated to intervention and 5 to control, and the data then analysed as a BAC study. As already noted in reference to RCTs, random allocation may be difficult in studies of district-level urban WASH interventions, or indeed practically impossible for some types of intervention; and note again that this paper centres on evaluations of already-planned interventions, not on interventions designed to optimise exploration of disease causality. When randomisation is not possible, control clusters will be often be selected after the intervention area has been selected. In this case, the most straightforward way of choosing control clusters is pair-matching, a method which has some advantages over unmatched designs (Imai et al. 2009). Indeed, pair-matching is often applied within RCT designs when the number of clusters available is small. For example, in a randomised trial on HIV prevention in Tanzania (Grosskurth et al. 1995), clusters were randomly assigned to intervention or control, and intervention and control clusters were then pair-matched. This not only increased the power of the study but also improved the plausibility of the findings, because lower HIV was observed in all matched pairs: the findings would have been less convincing (less plausible) if a reduction had only been observed in some pairs, even if the overall effect size and confidence interval had been the same.

In BAC studies, pair-matching is particularly attractive, not primarily to achieve statistical probability estimates (there is no valid “significance” comparison between pairs if the number of clusters is below 4 or 5), but to make the findings easier to interpret in terms of plausibility, as was done in the Teheran study (Kolahi, Rastegarpour & Sohrabi 2009). Such like-with-like comparison at the level of a matched pair facilitates the interpretation of secular disease trends in the absence of an intervention (a critical factor in BAC studies) because given adequate matching, they should be more similar within pairs than between pairs. For example, 3 out of 4 pairs suggesting an intervention effect would provide evidence that the intervention was indeed effective. If an effect can only be seen in half of the pairs, one would be more cautious, even if the overall before-after comparison suggests a greater disease decline in the intervention than in the control group.

Matching of control and intervention clusters could be done on the basis of existing census data or other socio-economic data sources (Arnold et al. 2009). It should not be solely based on informal observation of neighbourhoods (“eyeballing”), although other types of data (e.g. habitat-type data, data obtained from environmental health survey walks) may be used to refine matches obtained on the basis of census data. If available, data on health-care use (e.g. hospital admissions due to diarrhoea) may be a better basis for matching intervention and control clusters, because they should reflect disease risk more directly than most indicators of socio-economic status. It is also possible to match intervention and control clusters according to the baseline disease prevalence prior to the intervention.

The design of a BAC study can be improved considerably by including several baseline and follow-up measures, especially if the number of allocated clusters is very small (Shadish, Cook & Campbell 2001). For example, obtaining diarrhoea data over two or even three years prior to the intervention allows a much better assessment of secular trends in the intervention and control arm independent of the intervention than surveys over a single year. A second method to “make the most of” the few clusters available can be applied if the control clusters receive the intervention at a later date.

> "Helminths in stools reflect disease exposure over a long period"

By monitoring disease trends in both study arms depending on intervention status (e.g. the extent of household coverage actually achieved in a sewerage intervention), the causal inference of BAC studies can be enhanced (Shadish, Cook & Campbell 2001). We recommend applying these two design features, especially in studies where only one cluster each can be allocated to intervention and control. Finally, choosing control clusters in which the intervention will be implemented at a later date ("pipeline studies") may be a good strategy because it will tend to mean that the control clusters have broadly similar characteristics to the intervention arm.

6.2 Choice of outcome measures

WASH interventions target a number of infectious diseases. For HIE in urban settings, diarrhoea, cholera and worm infections (i.e. helminths) are often the most relevant diseases. Nutritional markers (WAZ, HAZ) and mortality are probably the most relevant *endpoints* from the public health perspective, and they have been used in BAC studies (Hasan et al. 1989), but often the power to detect changes in these parameters may be low. Worm infections assessed by stool samples from two cross-sectional surveys (before and after) are —unlike reported diarrhoea— a fairly objective outcome measure: demonstrating a reduction in worm infections in a BAC trial can be convincing because of the low risk of both participant and observer biases (Messou et al. 1997). Furthermore, helminths in stools can be considered an integrated measure in the sense that it can be expected to reflect cumulative disease exposure over a fairly long period leading up to the time-point of measurement; by contrast, diarrhoea (even if measured objectively by stools analysis) is an acute reversible condition showing population-level temporal variability (e.g. seasonal incidence peaks), so that a single measurement may be a less reliable measure of long-term disease burden. Note though that stool samples can be tedious to collect, and compliance with providing samples can be low.

If available, hospital or clinic data (for example admissions due to diarrhoea or cholera) can be a convenient disease measure. Bias due to healthcare-seeking behaviour can usually be addressed in BAC trials because these biases should be reflected by the baseline measurement. Healthcare use data may be less biased than self-reported illness. Further, such data may allow monitoring of disease trends over long periods of time, which may increase the validity of the findings (Shadish, Cook & Campbell 2001).

The most commonly used outcome measure in WASH interventions is self-reported diarrhoea (Clasen et al. 2010). Self-reported diarrhoea is prone to bias in unblinded studies, including RCTs (Schmidt & Cairncross 2009), though the risk of bias may be reduced by careful protocol design. Often, self-reported diarrhoea will be the only measure that can be assessed if healthcare use data are not available. There is a large body of literature on how best to assess diarrhoea (Blum & Feachem 1983; Alam et al. 1989; Boerma et al. 1991; Byass & Hanlon 1994; Wright et al. 2006; Schmidt et al. 2010; Zafar et al. 2010; Zwane et al. 2011). One key issue is the need for repeated surveillance visits over time, which poses logistics problems (Schmidt et al. 2010). When self-reports are used, diarrhoea can be measured as incidence (new episodes) or prevalence (Morris et al. 1996; Schmidt et al. 2007). Repeated prevalence measurements allow calculating the proportion of time ill (the "longitudinal prevalence"), which is often the relevant measure from the public health perspective. Note that the proportion of time an individual is ill is a continuous outcome (0% to 100%), which can facilitate sample size calculation (Schmidt et al. 2010).

> “Before-after evaluations will often be several years apart”

6.3 Selection of households within clusters

Ideally, within each cluster, households are selected at random based on the same selection criteria before and after the intervention, i.e. the same household may or may not be randomly selected twice. In large-scale urban WASH interventions, before-after evaluations will often be several years apart (as in Barreto et al. 2007 and Kolahi, Rastegarpour & Sohrabi 2009), so if the same sample is used there may be fewer young children at follow-up than at baseline (which is clearly a significant problem, since age is a critical determinant of diarrhoea prevalence, nutrition markers and worm infection). By renewing random selection after the intervention, it should be possible to approximately maintain the age range of the sample.

Often it is logistically easier to conduct multi-stage sampling within a cluster, e.g. by first selecting neighbourhoods within a cluster and then households within the neighbourhoods. This approach is likely to make diarrhoea estimates less stable, even if the same neighbourhoods (but different households) are chosen before and after. Diarrhoea can be expected to vary constantly over space and time (Luby et al. 2011), i.e. simple random sampling should give a more precise overall estimate of diarrhoea burden in an area comprising different neighbourhoods than multi-stage sampling. If the outcome is more stable over time (as expected with nutrition markers or worm infections), multi-stage sampling may be less of a problem, but simple random sampling is always more appropriate to ensure that the sample is representative of the whole cluster. We recommend that if smaller clusters within larger clusters are selected for logistical reasons, then the number of clusters should be reasonably large, and chosen based on a sample-size calculation that accounts for clustering.

Sometimes, a WASH intervention is allocated to a large heterogeneous cluster including poor and rich neighbourhoods. In this case it can make sense to restrict recruitment to households living in poor neighbourhoods, and select equivalent control neighbourhoods from potential control clusters. Ideally this is done by sampling households (e.g. by simple random sampling) only in a subgroup of the whole population, for example households with children under 5 years or households with a low income. This approach can increase both the public health relevance of a study and also the statistical power, by making the study population more uniform.

6.4 Sample size and number of surveillance visits

After determining the number of clusters to be enrolled, the investigator has to decide how many households to enrol from each cluster. In a BAC study, statistical power is largely determined by the number of overall participants included. Increasing the number of clusters primarily serves to make the findings more plausible and improve generalisability, not primarily to increase precision. Each cluster included in the study is compared with itself at different points in time, i.e. the between-cluster variability in disease does not affect the statistical analysis (but can affect, non-statistically, the plausibility of the results).

However, if the sampling unit is the household, and households are selected randomly both for the pre- and the post-intervention survey, then within-household clustering of disease (which may be considerable for worm infections and diarrhoea) needs to be accounted for in the sample size calculation. This is not easy, since most published formulae for group-randomised trials do not apply if the average group size is small (e.g. a family; Hayes & Moulton 2009). One of us (WS) has recently shown for diarrhoea that multiplying the sample size by a factor of 1.5 to 2 should be sufficient in most cases to account for clustering at household level (Schmidt, unpublished data).

> “Sample size can be reduced if a stronger effect is assumed”

The sample size can be reduced if a stronger effect is assumed (i.e. if we raise the magnitude of effect that has to be achieved in order for the study to detect it). In previous WASH trials many investigators have assumed diarrhoea or worm prevalence reductions of between 20 to 30% for the sample size calculations, to be detected with 80% power. A 20% reduction in diarrhoea is often regarded as an impact that would be of interest from the public health perspective (e.g. Barreto et al. 2007). In a BAC study, it is desirable not only to detect any before-after difference, but also to measure this difference with reasonable precision (in addition to having a confidence interval that does not include the null value). A narrower confidence interval will facilitate interpretation of the findings when accounting for secular trends observed in the control group. Therefore we would ideally recommend assuming a smaller magnitude of effects than one might assume in a comparable RCT, for example 15% or (if logistically feasible) even 10%. However, this clearly depends on the amount of resources available.

The sample size can also be reduced by making the samples in each cluster more uniform, e.g. by including only poor households, or only children under 5. In this case, the between-person variability of disease will be lower than if a wide range of different age groups and socio-economic strata are included. On the other hand, it is not efficient to exclude adults if diarrhoea self-report is the health measure: if a household has to be visited anyway at a certain frequency, then it makes sense to include all householders to gain the maximum of information (and study power) from a single visit. Focusing *a priori* on poor households can make sense, as these households are likely to benefit most from reducing diarrhoea burden; but this has to be weighed against the potential for extracting more generalisable conclusions if all socio-economic strata are included.

Where diarrhoea self-report is the health measure, we generally recommend conducting surveillance over whole years to cover all seasons. If this is not possible, one can collect data in a specific season, e.g. the wet season when diarrhoea incidence may be higher (potentially reducing the sample size): it is then of course important to measure diarrhoea after the intervention during the same season. Furthermore, due to the high seasonal and year-to-year fluctuations in diarrhoea, the interpretability of the study results may be more difficult if only one season is included.

Sample size calculations for studies using self-reported diarrhoea as the outcome are difficult because diarrhoea is typically measured using repeated measurements in the same people. A single cross-sectional survey provides a measure of diarrhoea, but often does not yield enough statistical power. For repeated measurements, the within-person correlation of diarrhoea needs to be taken into consideration. Individual measurements within a study cluster are not independent: they are correlated within individuals, i.e. people at particularly high or low diarrhoea risk disproportionally affect the outcome. To account for the added “noise” if people differ greatly in diarrhoea risk, the sample size needs to be increased. Since predicting within-person correlation is difficult, it is often easier to treat diarrhoea as a continuous outcome by using longitudinal prevalence, i.e. the proportion of time ill.

The sample size critically depends on the number of surveillance visits, but the relationship between number of repeat measurements and sample size is far from linear: in practice, conducting more than 12 surveillance rounds per time period (before vs after) will gain little additional power. The appropriate number of repeat measurements depends on many logistical factors, such as whether it is easier to recruit field workers or participants: usually 6 to 12 visits per period will provide the best balance between statistical power and costs.

Assumptions:
Minimum detectable reduction in diarrhoea = 25%
Power = 80%
Weekly diarrhoea longitudinal prevalence (LP) = 10%
(diarrhoea measured as period prevalence: disease at any time in last 7 days)
Standard deviation of the longitudinal prevalence = 14%.

Using standard formulae for the comparison of two mean LP values, the resulting sample size will be 493 people before and 493 people after the intervention in each arm (intervention and control).

This value is multiplied by 2 to account for household clustering of disease: 986 people in the intervention arm, 986 in the control arm, enrolled before and then again independently after the intervention.

This value should be multiplied by 1.2 or 1.3 to account for the fact that only 6 to 12 visits are conducted (both before and after the intervention), resulting in around 1183 people per arm (intervention/control) per study period (before/after) (Schmidt et al. 2010).

Assuming 4 people per household, we obtain 296 households to be included per arm (intervention/control) per study period (before/after).

Box 1. *Example sample size determination for a study with self-reported diarrhoea as outcome measure.*

To conclude, we recommend the following pragmatic approach for calculating the sample size for BAC studies:

- If the outcome can be measured in one cross-sectional survey, such as HAZ or worm infection, standard sample size formulae can be used. If self-reported diarrhoea is the main outcome, the sample size calculation should be based on the proportion of time ill, and its standard deviation after choosing the number of repeat measurements.
- The size of the disease reduction to be detected by the study should be as conservative as possible. The resulting sample size can then be multiplied by a factor of, say, 1.5 to 2.0 to account for within-household clustering. If the average number of people recruited per household is very small (for example if only young children are recruited) then household clustering can be ignored.

Box 1 shows an example sample size calculation for a BAC study.

We certainly consider that in many contexts it will be possible to perform a sufficiently rigorous study with around 250-500 households per arm per period, with each household visited 6 to 9 times over one year (one year before, one year after). If an objective "integrated" measure of disease is used (e.g. helminths in stools), only one visit per household will be required; as discussed further below, the use of such measures may increase total cost.

For further discussion of the issues outlined in this subsection (6.4), see Schmidt et al. (2010).

> "The interpretation of BAC studies relies on "plausibility" aspects"

6.5 Analysis and interpretation

The analysis and interpretation of BAC studies relies not only on statistical methods (as in RCTs) but to a great extent also on "plausibility" aspects that are difficult to quantify statistically. These can be crucial for interpretation and the confidence one can have in the results.

The first step in the analysis should be to display disease trends and confidence intervals of point estimates graphically, separately for intervention and control (Shadish, Cook & Campbell 2001). The difference in disease before versus after the intervention, with 95% confidence intervals, should be calculated for both arms. If there is more than one cluster per arm (hopefully there is), the results should be displayed for each individual cluster, or for each individual pair of clusters in a pair-matched design.

As a simple way of analysing the data, one can calculate the "difference in difference" (DID). For example, if the reduction in diarrhoea is 30% in the intervention arm, and 10% in the control arm, then the DID would be 20%. Importantly, it is not possible to calculate a valid confidence interval (CI) for the DID if the number of clusters per study arm is less than 4 or 5. Unless the number of clusters per arm is at least 4, CIs can only be calculated for the disease prevalence and before-after difference overall within each arm, or each cluster; a CI cannot be calculated for DID, nor does non-overlap of intervention and control CIs have a specific statistical meaning (although it can still be a noteworthy observation).

Apart from these simple descriptive statistics, the plausibility of results from BAC studies depends on a) the similarity of the intervention and control clusters, and b) the observed disease trends (Shadish, Cook & Campbell 2001). These two aspects are closely related, since clusters that are similar in many socio-economic factors are also likely to display similar disease trends.

It is further plausible that areas with similar baseline disease are more likely to show similar secular trends than areas with very different disease levels at baseline. For example, it could be that in a richer area, disease may be declining more rapidly than in a poor area. Conversely, in a richer area, a disease reduction may already have happened in the past, reaching a plateau, whereas in a poor area social development may result in rapid changes in disease. Differences in such disease trends accounting for the baseline (before) disease level are central to interpreting BAC studies.

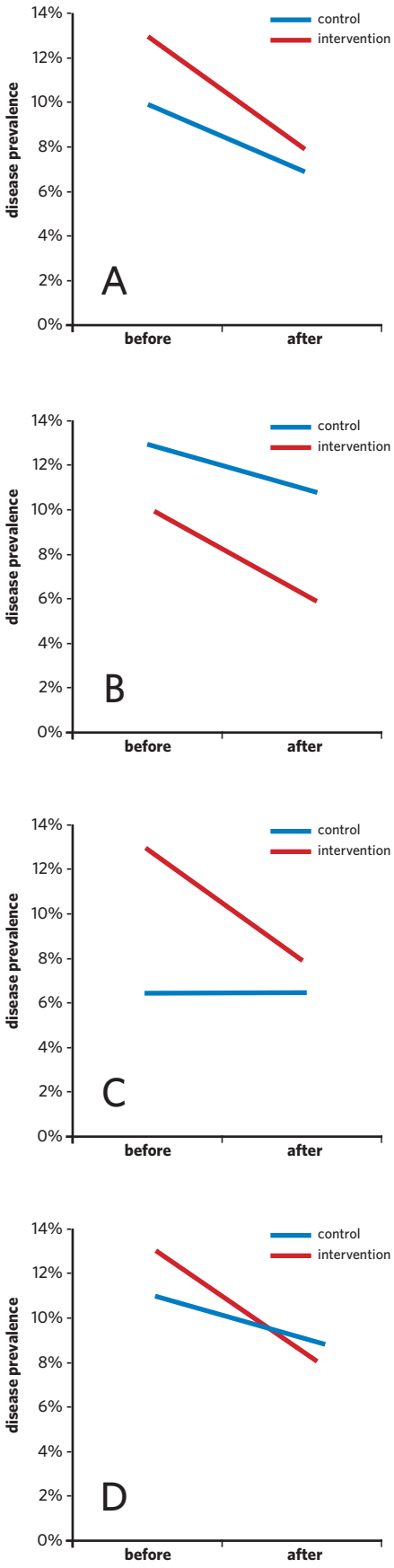

Figure 2:
Four different possible outcomes of BAC studies. All examples suggest a disease reduction in the intervention arm, but provide different levels of confidence that the reduction is causally related to the intervention.

Figure 2 illustrates examples of different possible outcomes of BAC studies. All examples suggest a disease reduction in the intervention arm, but offer different levels of confidence that the reduction is causally related to the intervention. In the first example (A), disease goes down in both arms, converging after a large difference in disease at baseline: there may possibly have been a reduction in disease due to the intervention, but it is difficult to exclude that the trends are due to different stages of socioeconomic development (or "maturation"; Shadish, Cook & Campbell 2001) in the two study arms, since the difference in disease prevalence at baseline is quite large.

Similar caution must be applied if the results resemble (B). Again, the trends could reflect different secular trends independent of the intervention. In both cases (A and B), additional measurements of disease at different time-points before and after the intervention would have provided a better picture of overall trends, and greater confidence in assessing the likelihood that the intervention had influenced these trends. For more details of these aspects, see Shadish, Cook & Campbell 2001.

If disease in the control arm does not change (C), a reduction in disease observed in the intervention arm can clearly be attributed to the intervention with a higher confidence than for A and B. Finally, if disease levels in the intervention arm are higher at baseline, and lower at follow-up compared to the control arm (D), i.e. the disease trends cross during the study, then it becomes difficult to ascribe this to anything other than an intervention effect (Shadish, Cook & Campbell 2001).

Finally, the interpretation of BAC studies is greatly enhanced by specifying in advance sampling procedures and an analysis plan in a detailed protocol. This will help avoid "post-hoc" reasoning and "fishing for significance" by conducting multiple analyses and then selectively publishing results most compatible with the investigators' own views. Pre-specification policies of this type, and evaluators who are independent of funder and implementer, can help to reduce the likelihood of effect exaggeration due to funder/implementer/researcher interests, as highlighted by authors like Bhandari et al. (2004) and Ionnadis (2005).

> "BAC studies can achieve strong plausibility in contexts in which RCT designs may be impossible"

7. BACs versus RCTs: "horses for courses"

This paper focuses on health impact evaluations of urban WASH interventions designed not with the aim of demonstrating in a general sense that WASH can improve health, but rather with the aim of assessing the specific impact of particular interventions (on the view that this will both improve programme efficiency and contribute to longer-term learning). In this specific context, and as detailed in Sections 5 and 6, we argue that RCTs are not necessarily the most appropriate method, and indeed in many common situations BAC studies may be more appropriate. Apart from higher financial cost, difficulties with application of RCTs in the present context are as follows:

- In some types of large-scale urban WASH intervention (for example networked water and sanitation interventions), it may be extremely difficult, or indeed practically impossible, to randomly allocate districts to intervention and non-intervention groups
- Even if randomisation is theoretically possible, in the real world it will often not be applied, so that the health impact evaluation of such interventions must necessarily use study designs without random allocation
- Where randomisation is possible, for many types of intervention the number of clusters will be too small to give sufficient power for a true RCT design: given 5 or fewer clusters in each arm, a reliable cluster RCT is not even possible, and this is very relevant because many real-world urban WASH interventions target 2 or 3 city districts at a time
- Over and above randomisation issues, the need for informed consent in RCTs can introduce an additional bias when self-report health measures are used: notably, people who have been randomly allocated to non-intervention will be strongly incentivised to exaggerate their ill-health, in order to strengthen their case that they be next in line for donor/government spending

Properly designed BAC studies offer an alternative approach for HIE of urban WASH interventions, particularly when allocation cannot be random and/or when cluster number is small. BAC studies cannot achieve the very high levels of proof achievable by large-scale RCTs in ideal contexts; however, BAC studies can achieve strong plausibility in evaluation contexts in which rigorous RCT designs may be impossible or extremely expensive, and where classic observational studies (cohort, cross sectional or case-control studies) are subject to confounding.

The table overleaf summarises our recommendations for appropriate cluster study designs in different situations:

Table 2. *Recommended study designs as a function of whether random allocation of clusters is possible, and of number of intervention clusters (i.e. number of clusters per arm).*

Intervention Clusters	Random allocation possible	Random allocation not possible
7 or more	RCT design as primary outcome. Measure and adjust for baseline disease risk to improve study power where possible.	Direct comparison (non-randomised trial design) as primary outcome. Adjust for baseline disease risk to minimise bias.
5-6	BAC as primary outcome, between arm comparison as secondary outcome (Biran et al.2009).	BAC as primary outcome, between-arm comparison as secondary outcome (Biran et al.2009).
2-4	BAC with random allocation	BAC
1	Only do if disease can be studied over two or more years before intervention or during implementation in the control arm.	Only do if disease can be studied over two or more years before intervention or during implementation in the control arm.

As noted, these guidelines refer to studies designed for programme evaluation, i.e. to assess the specific impact of already-planned interventions. However, if the aim is to explore in detail the causal pathways leading to disease (as in the current "environmental enteropathy" research drive, outlined in Humphrey 2009), then very high standards of statistical rigour are demanded, and in this situation it makes sense to purpose-design randomisable treatments that allow detailed exploration of causal pathways with very high levels of internal validity (though correspondingly reduced external validity): in such cases we agree that RCTs are likely to be the design of choice.

8. Financial costs and feasibility

So how much might it cost to run a BAC evaluation? How easy would it be to integrate such an evaluation into the time-frame and logistics of an urban WASH intervention? These are critical questions, since we are here proposing procedures for HIE of already-defined WASH interventions, as opposed to the deliberate design of WASH interventions in order to explore causality.

The costs of a BAC evaluation will of course vary depending on sampling effort and on local costs. Our experience is that, in urban areas of low-income countries, costs are often acceptably low, as a result of low labour and transport costs. Let us assume a total of 600 households (300 per arm) each visited 9 times over a one-year period (one year before, one year after), for estimation of diarrhoea incidence on the basis of self-reports; see Box 1 above. This visit rate and associated data input[10] can be readily achieved by a two-person team. In Kenya (for example), employment of an interviewer costs about $500 per month; in our example we require two interviewers working over a two-year period, which comes to approximately $24,000. In addition, someone with sufficient epidemiological knowledge will be required to design the study, develop data collection instruments, train interviewers, coordinate and quality-check data collection, carry out the data analysis, and report the findings. We suggest that, with careful cost control, it should certainly be possible to achieve a study of this type for a total amount in the order of $100,000–$200,000.

This cost range certainly lies at the bottom of the indicative range of $70,000–$1 million per study suggested by Briscoe, Feachem and Rahaman (1986).[11] But whether an expenditure of this order is judged justifiable will depend on funders' expectations and on the size of the overall budget. Clearly, it would not make sense to spend this amount on health impact evaluation within a $200,000 intervention, but within a $2 million intervention it might certainly be judged worthwhile. There is no general rule about the proportion of a development programme's budget that should be spent on evaluation, but between 5 and 10% is not untypical for public health interventions (see e.g. Global Fund 2009). This rule of thumb would suggest we might consider health impact evaluations for any project of $1 million or more, and certainly for projects/programmes of $2 million or more.

Note that this is no more than a crude ballpark assessment based on very approximate estimates of HIE cost and sector norms about how much to spend on evaluation; clearly, assessment of whether HIE is worthwhile in a specific context calls for detailed context-specific judgements, and here the reader is urged to consider not only the arguments of the present paper, but also to consult Briscoe et al. (1986), who discuss in detail the circumstances under which an HIE can be considered "useful, "sensible" and "feasible". Also note a "spin-off" advantage of extensive before-after householder surveys carried out in intervention and control districts: this will allow collection of data on diverse other relevant variables at little or no extra cost, allowing very detailed assessment of other impacts of the intervention (for example, time spent accessing and queuing for sanitation facilities; amount spent on water; etc. etc. depending on intervention characteristics). Additionally, HIEs can provide a testing ground for identifying more reliable proxy indicators of health impact (see Section 10).

[10] This assumes that data collection is done either with paper instruments designed to be optically read, or via an electronic input device such as a mobile phone (for example Water For People's FLOW system for survey data collection using mobile phones; see Hayward 2011).

[11] In fact if we adjust Briscoe et al.'s values for inflation 1986-2011, we obtain a present-day range of $140,000–$2 million.

[12] For ethical reasons, any study involving analysis of parasites in stools will generally be judged to require provision of free treatment (i.e. a course of an anthelminthic such as ivermectin) to subjects found to be infected. This increases study cost (though certainly the increased cost can be considered easily justifiable, since it has direct health impact), and also complicates study design in a before-after study.

HIE based on an objective measure like stool analysis (see e.g. Barreto et al. 2010) has clear advantages over HIE based on reported diarrhoea, as already discussed: but despite the need for only a single visit to each household, additional costs (including laboratory and medication costs[12]) will generally mean that total costs may be higher, although by how much will depend on the study setting. In the ongoing LSHTM Orissa trial, worm sampling is not proving too expensive, suggesting that a BAC study using worm sampling could probably be undertaken within the $100,000-$200,000 ballpark budget range indicated above.

Time-frame may be an important constraint on integration of HIEs into WASH interventions, as previously noted by PREM (2006). Many urban WASH interventions run over periods of about 3 years, which may be insufficient for an HIE, especially bearing in mind that HIE as proposed in this paper requires a period of about a year before the intervention starts. For example, WSUP has long-term ongoing city-level programmes, working closely with local partners including municipalities and utilities, in 6 countries mostly in sub-Saharan Africa (see www.wsup.com).

But funding for most interventions within these programmes comes from grants from funding agencies that often extend for just 2 or 3 years from approval to closure and final reporting. Even within a 3-year project, an HIE of urban WASH intervention is difficult, particularly if this involves substantial infrastructure construction. For example, let us suppose we want to do an HIE of an intervention centred on construction of communal toilets and associated hygiene education in 3 districts of City X. From the date of project approval, we ideally need a) 3 months for HIE design, plus b) 12 months for the "before" evaluation, plus say c) 24 months for programme implementation and impact,[13] plus d) 12 months for the "after" evaluation, and finally e) 3 months for HIE analysis and write-up: this comes to a total of 54 months (i.e. 4 ½ years); note that the interventions did not start until month 15.

One option we would suggest to funders and implementers alike is to consider grant periods of say 5 years, in which the last 18 months involves no activity other than impact evaluation; i.e. all interventions are finished by month 42.

If long time-frames are not feasible, the possibility of reducing the time period over which data is collected should not be ruled out: for example, instead of collecting data for one year before the intervention and one year after, we might consider collecting data for just 3 months before and 3 months after, in both cases probably during the rainy season when diarrhoea incidences are likely to be highest.

Finally: at the outset of this section, we noted that we are here proposing procedures for HIE of already-defined WASH interventions, as opposed to deliberate design of WASH interventions in order to optimally facilitate HIE. However, this is not black-and-white, and even if the basic characteristics of the intervention are already defined, it will certainly make sense for funders and implementers to talk through the possibility of HIE at an early stage; alternatively, donors might consider incorporating an "HIE option" into calls for proposals, with special dispensations regarding the proportion of budget that can be allocated to monitoring and evaluation, and regarding time-frame and reporting requirements. It is also clearly critical for implementers to take HIE design into consideration during project planning: for example, in the model outlined above, the "before" evaluation requires that there be a period of 15 months between project approval and the first interventions.

[13] Carrying out an impact evaluation immediately after completion of an intervention is unlikely to be the most useful approach: in the case of an intervention centred around construction of communal toilets, for example, there will typically be a time-lag of several months between termination of the construction work and widespread community use of the new facility; and this behaviour change and associated improvement in local environmental hygiene will take some time to impact on disease burdens. So even if the construction and other "direct" components of an intervention are completed within a year, it will probably be more meaningful to wait another year before evaluating impacts.

> BAC studies provide an opportunity to collect data on other impacts

9. Is health impact the only important impact?

This paper has worked from the assumption that health impact should be the primary impact sought in government- or donor-financed WASH interventions.[14] Are we then suggesting that health improvements are the only relevant impact?

9.1 What about other types of benefit?

As noted, some researchers have attempted to assess the economic benefits of WASH improvements for low-income communities, considering all potential benefits including reduced mortality and morbidity, healthcare cost savings, time savings, and gain in productive time. Such approaches require monetisation of all benefits, so that different types of benefit can be meaningfully compared. One such study (Hutton et al. 2007) suggested that by far the most important benefit of WASH interventions was the convenience time saving due to easier access to water and sanitation facilities: for example, in the WHO's AFR-D subregion (West Africa), convenience time savings were estimated to account for 82% of total economic benefits, while health benefits (including healthcare cost savings) accounted for only 16%.[15]

However, the authors recognise that their methodology (notably their consideration only of impacts on diarrhoea) is likely to have systematically under-estimated health benefits. Furthermore, we believe that cost-benefit analyses of this type, necessarily based on numerous assumptions[16] about average benefits, may generate unreliable results. But certainly, and as noted in the previous section, BAC studies provide an excellent opportunity to collect data on other impacts at the same time as health impact collection.

[14] Note that we are here talking about the *justifications* for WASH investment. These are not the same as the *social stimuli* for WASH improvement at community level: for example, householders may be more likely to invest in a toilet because it improves their prestige or perceived wellbeing than for health reasons. Such stimuli are of critical importance for social marketing, but in the present context we are interested in *justifications*.

[15] Of which 9% value of deaths avoided, 3% value of baby sick days avoided, 4% health sector costs saved, and 0% patient health costs saved.

[16] For example, the authors make the basically arbitrary assumption that a switch from "unimproved" to "improved" sanitation will generate convenience time savings of 30 minutes per person per day; this seems to be based largely on the questionable assumption that ["unimproved" ->"improved"] equates to ["public toilet" ->"household toilet"]. For wider critiques of cost-benefit analysis, and in particular its heavy dependence on investigator assumptions and judgements, see for example Flyvbjerg (2008).

9.2 What about sustainability and pro-poorness?

If an urban WASH intervention, funded by donors and/or by national government, achieves a strong impact on health over a 3- or 5-year period, this does not necessarily mean that the improvement will be sustained over decades, or that it has benefitted the poorest people in the city, or that it will favour continued expansion and improvement of WASH services within the city. In other words, we consider that health impact should be the primary immediate goal of urban WASH interventions, but we do not consider that it should be the sole goal; indeed, an uncontextualised and short-termist focus on health impact would run the risk of creating a counter-productive structure of perverse incentives. We suggest that WASH programmes need to consider not only health, but also the following key issues:

- **Financial and operational sustainability of the intervention:** Any WASH intervention in a given district (for example a programme of construction of shared toilets coupled with hygiene education to reduce open defecation) needs to be sustainable over a reasonably long time period: a substantial health impact in year 3 is of questionable value if, 5 years later, the toilets are in ruins because users have not been willing or able to pay for their maintenance, or if people have reverted to open defecation because the hygiene education programme did not achieve a long-term change in attitudes and behaviours.
- **Financial and operational sustainability of ongoing improvements:** Related considerations apply in the wider context of the city. An intervention that is 100% funded by external donors may achieve a substantial health impact in year 3, but may not build local institutional capacity and may contribute to a culture of aid-dependency at the city level, reducing pressure on the municipal authorities to develop donor-independent mechanisms of financing and managing citywide WASH improvements.[17]
- **Social equity:** There are of course ethical grounds for expecting donor spending and government spending to focus on improving the welfare of the very poor (including homeless people and people in informal settlements), of women as well as men, and of disadvantaged groups (including the disabled, AIDS sufferers, and marginalised ethnic and religious communities). In fact, we suggest that the stronger emphasis on health impacts advocated by the present paper will tend to favour these groups (because forcing a focus on health impact will tend to force a focus on very poor communities with very poor baseline health); nonetheless, this cannot be assumed, and needs to be ensured at planning and evaluation.

[17] This ties in closely with WSUP's own strategic focus on working with local service providers (ranging from municipal government and large utilities down to local community groups and small-scale private operators), aiming to achieve a situation in which local capacity for revenue generation, management and planning reaches a level sufficient for donor-independent financing and expansion of WASH services.

10. And what if health impact evaluation isn't feasible?

We have here argued that HIEs should be carried out more frequently than is currently the case, and that this could have profound implications for the design of WASH interventions. However, we are certainly not suggesting that all WASH interventions should include HIEs: in many cases health impact evaluation will not be useful, sensible or feasible (Briscoe et al. 1986). In such cases, what are the implications of this paper's central argument (that health impacts should be the primary goal of urban WASH interventions)?

First, at the intervention appraisal and planning stages, we propose a stronger focus on districts with poor baseline health. This means that donors and implementers need to resist pressures pushing them in the direction of "easy pickings": for example, if donors push implementers towards sanitation marketing models based on leverage of householder finance to pay for construction of household latrines by local entrepreneurs (in principle not an unreasonable long-term strategy), this will tend to discourage selection of very poor intervention districts with very poor baseline health.

Second, and again at the appraisal and planning stages, we propose a stronger focus on the likely health impacts of WASH interventions. For example, we might reasonably suspect that a focused "total" intervention (e.g. construction of 100 shared toilet facilities serving all households within a low-income district of 3000 people, coupled with a programme of improved drainage) will have greater health impact than a more diffuse intervention involving construction of 100 shared toilets serving some households within a district of 30,000 people, and without any programme to improve drainage. [Clearly, this is just one illustrative example: how to achieve maximal health impact in a given district with a given budget clearly involves detailed location-specific analysis, beyond the scope of the present paper.]

Third, at the evaluation stage we propose a stronger focus on outcomes that are likely to be associated with health, as opposed to immediate project outputs (*see footnote 5, pg 3*): for example, toilet usage, as opposed to number of toilets constructed; or increased frequency of handwashing, as opposed to number of people exposed to hygiene education campaigns. In fact, this message (that effective programme evaluation requires data on "use", not "access") is widely accepted in the sector: however, we would suggest that even these indicators may often be rather poor proxies for health impact, particularly in the absence of consideration of "TotalUrban Sanitation".

In other words, achieving health impact is likely to require integrated consideration of all aspects of faecal-oral disease transmission: not just increased usage of improved water supplies and improved toilet facilities, but also diverse other contributing factors including improved handwashing, food hygiene and child faeces hygiene, minimisation of faecal contamination of the local environment, and flood protection or flood-proofing of sanitation facilities. We suggest that, in the absence of HIE, evaluation procedures should aim to assess improvements across all of these areas. It is beyond the scope of the present paper to assess how exactly this might be achieved: however, we would briefly mention the possibility of using measures of faecal contamination of the local environment, such as score-card based observational measures, or more objective measures based on microbiological analyses.

11. Conclusions

- This paper suggests that there is not a strong need for "proof-of-concept" health impact studies to "demonstrate" in a general sense that WASH has a positive impact on health.
- Instead, we suggest that there is a need for more frequent health impact studies of specific WASH programmes, primarily designed to assess the cost-effectiveness of each programme, thereby encouraging funders and implementers to focus on the best ways of achieving real health impact, rather than relying on proxy outcome indicators which may not in fact be closely related to health impact (e.g. "increased use of improved sanitation facilities").
- For this purpose, we argue that before-after concurrent-control (BAC) studies will often be the best option.[18] BAC studies cannot achieve the very high levels of proof that can be obtained by a purpose-designed large RCT study: however, they can achieve a high level of plausibility, and in the context of programme evaluation (as opposed to purpose-designed research) we suggest that they will often give stronger plausibility and generalisability than an RCT, making the findings potentially more useful for policy and practice.
- Based on the estimates given in this paper for the cost of BAC designs, we suggest that sufficiently rigorous health impact evaluation of urban WASH interventions (and indeed other WASH interventions) is not as prohibitively expensive as it is widely held to be. We suggest that, with strict cost control, properly designed studies will often be achievable for around $100,000–$200,000, and we suggest that this may be a cost-efficient expenditure for interventions above a certain size.
- Whether an RCT or BAC design is selected, there are strong arguments for using objective outcome measures (e.g. helminths in stools) as opposed to self-reported diarrhoea. In all cases, it is important for observers to be independent, i.e. neither in reality nor in appearance part of the programme team.
- We do not suggest that HIE should be carried out for all urban WASH interventions, but rather only for a subset of large-scale interventions; where HIE is not carried out, we nevertheless argue for stronger consideration of likely health impact at the design stage, and a stronger focus on more plausible proxy indicators of health impact in programme evaluation.
- In order to enable HIEs, we suggest that funding agencies might consider incorporating an "HIE option" into calls for proposals. Under this option, a) interventions would be required to demonstrate health impact and not judged on interim outputs (e.g. number of toilets built, or amount of householder finance leveraged); b) implementers would be allowed longer reporting periods; c) implementers would be able to allocate a higher proportion of total budget (perhaps 10%) to evaluation; and d) the HIE itself would be conducted by an agency independent of both funding agency and implementer.
- We suggest that the more widespread use of HIE would contribute to an evidence base of 'what works', and help identify the generalisable features of urban WASH interventions that are successful in delivering lasting health improvements.
- We suggest that more frequent HIE could bring about very beneficial changes in how WASH investments are allocated and targeted. A more routine requirement to demonstrate health impacts could drive real advances in intervention effectiveness and intervention efficiency.

[18] Though course RCTs remain appropriate for other purposes: notably, in in-depth explorations of the aetiology of infectious diseases, designed a priori as research studies rather than as evaluations of existing interventions, RCTs certainly remain the design of choice. Likewise, for evaluation of household-level interventions such as household water treatment, RCTs may be perfectly feasible.

References

- Alam N, Henry FJ & Rahaman MM (1989) Reporting errors in one-week diarrhoea recall surveys: experience from a prospective study in rural Bangladesh. *Int. J. Epidemiol*. 18: 697-700.
- Arnold B, Arana B, Mausezahl D, Hubbard A & Colford JM (2009) Evaluation of a pre-existing, 3-year household water treatment and handwashing intervention in rural Guatemala. *Int. J. Epidemiol*. 38: 1651-1661.
- Arnold BF, Khush RS, Ramaswamy P, London AG, Rajkumar P, Ramaprabha P, Durairaj N, Hubbard AE, Balkrishnan K & Colford JM (2010) Causal inference methods to study nonrandomized, preexisting development interventions. PNAS 107: 22605-22610.
- Azurin JC & Alvero M (1974) Field evaluation of environmental sanitation measures against cholera. *Bull. World Health Organ* 51: 19-26.
- Barreto ML, Genser B, Strina A, Teixeira MG, Assis AMO, Rego RF, Teles CA, Prado MS, Matos SMA, Santos DN, Dos Santos LA & Cairncross S (2007) Effect of city-wide sanitation programme on reduction in rate of childhood diarrhoea in northeast Brazil: assessment by two cohort studies. *Lancet* 370: 1622-1628.
- Barreto ML, Genser B, Strina A, Teixeira MG, Assis AM, Rego RF, Teles CA, Prado MS, Matos SM, Alcantara-Neves NM & Cairncross S (2010) Impact of a citywide sanitation program in northeast Brazil on intestinal parasites infection in young children. *Environ Health Perspect* 118: 1637-42.
- Bhandari M, Busse JW, Jackowski D, Montori VM, Schünemann H, Sprague S, Mears D, Schemitsch EH, Heels-Ansdell D & Devereaux PJ (2004) Association between industry funding and statistically significant pro-industry findings in medical and surgical randomized trials. *CMAJ*. 170: 477-480.
- Biran A, Schmidt WP, Wright R, Jones T, Seshadri M, Isaac P, Nathan NA, Hall P, McKenna J, Granger S, Bidinger P & Curtis V (2009) The effect of a soap promotion and hygiene education campaign on handwashing behaviour in rural India: a cluster randomised trial. Trop. Med. Int. Health 14: 1303-1314.
- Blum D & Feachem RG (1983) Measuring the impact of water supply and sanitation investments on diarrhoeal diseases: problems of methodology. *Int. J. Epidemiol*. 12: 357-365.
- Boerma JT, Black RE, Sommerfelt AE, Rutstein SO & Bicego GT (1991) Accuracy and completeness of mothers' recall of diarrhoea occurrence in pre-school children in demographic and health surveys. *Int. J. Epidemiol.* 20: 1073-1080.
- Briscoe J, Feachem RG & Rahaman MM (1986) Evaluating health impact: Water supply, sanitation and hygiene education. International Development Research Centre (IRDC) TS248e, Ottawa.
- Brown CA & Lilford RJ (2006) The stepped wedge trial design: a systematic review, BMC Medical Research Methodology 6: 54.
- Byass P & Hanlon PW (1994) Daily morbidity records: recall and reliability. *Int. J. Epidemiol*. 23: 757-763.
- Cepeda MS, Boston R, Farrar JT & Strom BL (2003) Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am. J. Epidemiol*. 158: 280-287.
- Chavasse DC, Shier RP, Murphy OA, Huttly SR, Cousens SN & Akhtar T (1999) Impact of fly control on childhood diarrhoea in Pakistan: community-randomised trial. *Lancet* 353: 22-25.
- Clasen TF, Bostoen K, Schmidt WP, Boisson S, Fung IC, Jenkins MW, Scott B, Sugden S & Cairncross S (2008) Interventions to improve excreta disposal for the prevention of diarrhoea (protocol for a Cochrane Review). The Cochrane Library.
- Clasen T, Bartram J, Colford J, Luby S, Quick R, & Sobsey M (2009) Comment on "Household Water Treatment in Poor Populations: Is There Enough Evidence for Scaling up Now?" *Environ. Sci. Technol.* 2009, DOI: 10.1021/es9008147.
- Clasen T, Schmidt WP, Rabie T, Roberts I & Cairncross S (2007) Interventions to improve water quality for preventing diarrhoea: systematic review and meta-analysis. *BMJ* 334: 782.
- Clasen T, Bostoen K, Schmidt WP, Boisson S, Fung IC, Jenkins MW, Scott B, Sugden S & Cairncross S (2010) Interventions to improve disposal of human excreta for preventing diarrhoea. *Cochrane. Database Syst. Rev*. 6:CD007180.
- Clasen T (2008) Scaling Up Household Water Treatment: Looking Back, Seeing Forward. Public Health and the Environment, World Health Organization, Geneva.
- Cochrane Collaboration (2011) Cochrane Handbook for Systematic Reviews of Interventions. www.cochrane-handbook.org
- Compernolle p & de Kemp a (2009) Methodology for Evaluations of Budget Support Operations at Country Level. Tools for "Step 2": The Evaluation of the Impact of Government Strategies. The Hague.

- Cowling BJ, Chan KH, Fang VJ, Cheng CK, Fung RO, Wai W, Sin J, Seto WH, Yung R, Chu DW, Chiu BC, Lee PW, Chiu MC, Lee HC, Uyeki TM, Houck PM, Peiris JS & Leung GM (2009) Facemasks and hand hygiene to prevent influenza transmission in households: a cluster randomized trial. *Ann. Intern. Med*. 151: 437-446.
- De Allegri M, Pokhrel S, Becher H et al. (2008) Step-wedge cluster-randomised community-based trials: an application to the study of the impact of community health insurance. *Health Res. Policy Syst*. 6: 10.
- Emerson PM, Lindsay SW, Walraven GE, Faal H, Bøgh C, Lowe K & Bailey RL (1999) Effect of fly control on trachoma and diarrhoea. *Lancet* 353: 1401-1403.
- Fewtrell L, Kaufmann RB, Kay D, Enanoria W, Haller L & Colford JM (2005) Water, sanitation, and hygiene interventions to reduce diarrhoea in less developed countries: a systematic review and meta-analysis. *Lancet Infect. Dis*. 5: 42-52.
- Flyvbjerg B (2008) Public Planning of Mega-projects: Overestimation of Demand and Underestimation of Costs, in H Priemus, B Flyvbjerg & B van Wee (eds), Decision-Making On Mega-Projects. Cost-benefit Analysis, Planning and Innovation. Edward Elgar Publishing, Inc., Cheltenham, UK and Northamton, MA, USA, pp. 120-144.
- Global Fund (2009) Framework for Operations and Implementation Research in Health and Disease Control Programs. www.theglobalfund.org/en/me/guidelines_tools
- Grosskurth H, Mosha F, Todd J, Mwijarubi E, Klokke A, Senkoro K, Mayaud P, Changalucha J, Nicoll A, ka-Gina G, et al. (1995) Impact of improved treatment of sexually transmitted diseases on HIV infection in rural Tanzania: randomised controlled trial. *Lancet* 346: 530-536.
- Hasan KZ, Briend A, Aziz KM, Hoque BA, Patwary MY & Huttly SR (1989) Lack of impact of a water and sanitation intervention on the nutritional status of children in rural Bangladesh. *Eur. J. Clin. Nutr*. 43: 837-843.
- Hayes RJ & Moulton LH (2009) Cluster Randomised Trials, Chapman & Hall/CRC, Boca Raton.
- Hayward T (2011) GIS & mapping tools for water and sanitation infrastructure. WSUP Practice Note, www.wsup.com/sharing/PracticeNotes.htm
- Hoque BA, Juncker T, Sack RB, Ali M, & Aziz KM (1996) Sustainability of a water, sanitation and hygiene education project in rural Bangladesh: a 5-year follow-up. *Bull. World Health Organ* 74: 431-437.
- Humphrey J (2009) Child undernutrition, tropical enteropathy, toilets and handwashing. *The Lancet* 374: 1032-1035.
- Hutton G, Haller L & Bartram J (2007) Global cost-benefit analysis of water supply and sanitation interventions. *J Water Health* 5: 481-502.
- Imai K, King G & Nall C (2009) The essential role of pair matching in cluster-randomized experiments, with application to the Mexican Universal Health Insurance Evaluation. *Statistical Sciences* 24: 29-53.
- Ioannidis JP (2005) Why most published research findings are false. *PLoS. Med*. 2: 124.
- Joshi D, Fawcett B & Mannan F (2011) Health, hygiene and appropriate sanitation: experiences and perceptions of the urban poor. *Environment and Urbanization* 23: 91-111.
- Kaufman JS, Cooper RS, & McGee DL (1997) Socioeconomic status and health in blacks and whites: the problem of residual confounding and the resiliency of race. *Epidemiology* 8: 621-628.
- Kolahi AA, Rastegarpour A & Sohrabi MR (2009) The impact of an urban sewerage system on childhood diarrhoea in Tehran, Iran: a concurrent control field trial. *Trans. R. Soc. Trop. Med. Hyg*. 103: 500-505.
- Luby SP, Agboatwalla M & Hoekstra RM (2011) The variability of childhood diarrhea in Karachi, Pakistan, 2002 – 2006. *Am. J. Trop. Med. Hyg*. 84: 870-7.
- Messou E, Sangare SV, Josseran R, Le Corre C & Guelain J (1997) [Impact of improved sanitary conditions and domestic hygiene on the incidence of ascaridiasis and ancylostomiasis in children two to four years old in the rural zones of Ivory Coast]. *Bull. Soc. Pathol. Exot*. 90: 48-50.
- Morris SS, Cousens SN, Kirkwood BR, Arthur P & Ross DA (1996) Is prevalence of diarrhea a better predictor of subsequent mortality and weight gain than diarrhea incidence? *Am. J. Epidemiol*. 144, 582-588.
- Nanan D, White F, Azam I, Afsar H & Hozhabri S (2003) Evaluation of a water, sanitation, and hygiene education intervention on diarrhoea in northern Pakistan. *Bull. World Health Organ* 81: 160-165.
- Norman G, Pedley S & Takkouche B (2010) Effects of sewerage on diarrhoea and enteric infections: a systematic review and meta-analysis. Lancet Infect Dis. 10: 536-44.
- Norman G & Pedley S (2011) Exploring the negative space: evaluating reasons for the failure of pro-poor targeting in urban sanitation projects. *Journal of Water, Sanitation and Hygiene for Development* 1: 86-101.
- Pearson M (2011) Results based aid and results based financing: What are they? Have they delivered results? HLSP Institute, www.hlsp.org/Home/Resources/Resultsbasedaid.aspx

- PREM (2006) A Guide to Water and Sanitation Sector Impact Evaluation. Washington DC: World Bank.
- Ross DA, Wight D, Dowsett G, Buve A & Obasi AI (2006) The weight of evidence: a method for assessing the strength of evidence on the effectiveness of HIV prevention interventions among young people. *World Health Organ Tech. Rep. Ser.* 938, 79-102.
- Rubin DB (1997) Estimating causal effects from large data sets using propensity scores. *Ann. Intern. Med.* 127: 757-763.
- Schmidt WP & Cairncross S (2009) Household water treatment in poor populations: is there enough evidence for scaling up now? *Environ. Sci. Technol*. 43: 986-992.
- Schmidt WP, Aunger R, Coombes Y, Maina PM, Matiko CN, Biran A & Curtis V (2009a) Determinants of handwashing practices in Kenya: the role of media exposure, poverty and infrastructure. *Trop. Med. Int. Health* 14: 1534-1541.
- Schmidt WP, Boisson S, Genser B, Barreto ML, Baisley K, Filteau S & Cairncross S (2009b) Weight-for-age z-score as a proxy marker for diarrhoea in epidemiological studies. *J. Epidemiol. Community Health* 64: 1074-9
- Schmidt WP, Genser B, Barreto ML, Clasen T, Luby SP, Cairncross S & Chalabi Z (2010) Sampling strategies to measure the prevalence of common recurrent infections in longitudinal studies. *Emerg. Themes Epidemiol*. 7: 5.
- Schmidt WP, Luby SP, Genser B, Barreto ML & Clasen T (2007) Estimating the longitudinal prevalence of diarrhea and other episodic diseases: continuous versus intermittent surveillance epidemiology. *Epidemiology* 18: 537-43.
- Shadish WR, Cook TD & Campbell TD (2001) *Quasi-Experimental Designs that Use Both Control Groups and Pretests: In Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Wadsworth Publishing.
- Trémolet S (2011) Results-based financing for sanitation: is there a case for it? Can it work? Blog post at www.tremolet.com/blog/results-based-financing-sanitation-there-case-it-can-it-work
- Waddington H, Snilstveit B, White H & Fewtrell L (2009) Water, sanitation and hygiene interventions to combat childhood diarrhoea in developing countries. New Delhi, India, International Initiative for Impact Evaluation (3ie).
- WHO. Combating waterborne disease at the household level. World Health Organization, Geneva, Switzerland. 2007.
- Wright JA, Gundry SW, Conroy R, Wood D, du Preez M, Ferro-Luzzi A, Genthe B, Kirimi M, Moyo S, Mutisi C, Ndamba J & Potgieter N (2006) Defining episodes of diarrhoea: results from a three-country study in Sub-Saharan Africa. *J. Health Popul. Nutr*. 24: 8-16.
- WSUP/ODI (2011) Progress-linked finance: A study of the feasibility and practicality of a proposed WASH financing approach. Water & Sanitation for the Urban Poor (WSUP) and Overseas Development Institute (ODI), London.
- Young B & Briscoe J (1988) A case-control study of the effect of environmental sanitation on diarrhoea morbidity in Malawi. *J. Epidemiol. Community Health* 42: 83-88.
- Zafar SN, Luby SP & Mendoza C (2010) Recall errors in a weekly survey of diarrhoea in Guatemala: determining the optimal length of recall. *Epidemiol. Infect*. 138: 264-269.
- Zwane AP, Zinman J, Van Dusen E, Pariente W, Null C, Miguel E, Kremer M, Karlan DS, Hornbeck R, Giné X, Duflo E, Devoto F, Crepon B & Banerjee A (2011) Being surveyed can change later behavior and related parameter estimates. *Proc.Natl.Acad.Sci. USA* 108: 1821-1826.

***Credits**: This work was jointly conceived and written by WS (LSHTM) and GN (WSUP). OC (LSHTM) made substantial and important inputs to later versions of the manuscript. Opinions expressed herein are those of the authors, and do not necessarily reflect the positions of their respective organisations. Review inputs: Helen Pankhurst, Alison Parker and Kevin Tayler. Coordination: Gemma Bastin. Design: AlexMusson.com. [Version 1, October 2011.]*

This Discussion Paper is a WSUP/SHARE co-publication. WSUP **(www.wsup.com) is a tri-sector partnership between the private sector, civil society and academia with the objective of addressing the increasing global problem of inadequate access to water and sanitation for the urban poor and the attainment of the Millennium Development Goal targets, particularly those relating to water and sanitation. SHARE (www.shareresearch.org) generates rigorous and relevant research to ensure new and existing sanitation and hygiene solutions are adopted at scale in developing countries; it is a five-year research consortium funded by the UK Department for International Development, and its partners are LSHTM, ICDDR,B, IIED, SDI and WaterAid.**