Causal inference methods to study nonrandomized, preexisting development interventions

Benjamin F. Arnold[a,1], Ranjiv S. Khush[b], Padmavathi Ramaswamy[c], Alicia G. London[b], Paramasivan Rajkumar[c], Prabhakar Ramaprabha[c], Natesan Durairaj[c], Alan E. Hubbard[a], Kalpana Balakrishnan[c], and John M. Colford, Jr.[a]

[a]School of Public Health, University of California, Berkeley, CA 94720-7358; [b]Aquaya Institute, San Francisco, CA 94129; and [c]Department of Environmental and Health Engineering, Sri Ramachandra Medical College and Research Institute, Porur, Chennai 600116, Tamil Nadu, India

Edited* by Kirk R. Smith, University of California, Berkeley, CA, and approved November 2, 2010 (received for review July 2, 2010)

Empirical measurement of interventions to address significant global health and development problems is necessary to ensure that resources are applied appropriately. Such intervention programs are often deployed at the group or community level. The gold standard design to measure the effectiveness of community-level interventions is the community-randomized trial, but the conditions of these trials often make it difficult to assess their external validity and sustainability. The sheer number of community interventions, relative to randomized studies, speaks to a need for rigorous observational methods to measure their impact. In this article, we use the potential outcomes model for causal inference to motivate a matched cohort design to study the impact and sustainability of nonrandomized, preexisting interventions. We illustrate the method using a sanitation mobilization, water supply, and hygiene intervention in rural India. In a matched sample of 25 villages, we enrolled 1,284 children <5 y old and measured outcomes over 12 mo. Although we found a 33 percentage point difference in new toilet construction [95% confidence interval (CI) = 28%, 39%], we found no impacts on height-for-age *Z* scores (adjusted difference = 0.01, 95% CI = −0.15, 0.19) or diarrhea (adjusted longitudinal prevalence difference = 0.003, 95% CI = −0.001, 0.008) among children <5 y old. This study demonstrates that matched cohort designs can estimate impacts from nonrandomized, preexisting interventions that are used widely in development efforts. Interpreting the impacts as causal, however, requires stronger assumptions than prospective, randomized studies.

impact evaluation | study design | propensity score matching | community-led total sanitation | open defecation

In 2000 the United Nations member states agreed upon the Millennium Development Goals (MDGs), which formalized the global community's renewed commitment to solve some of the world's most intractable health and development problems. The MDGs set aggressive targets for 2015 in core metrics, such as reducing by two-thirds the under 5 y population mortality rate and reducing by half the population without access to safe drinking water and basic sanitation. Governments, foundations, and nongovernmental organizations (NGOs) have subsequently increased investment in global health and development programs and have relied on the scientific community to help rigorously measure the impact and cost effectiveness of the interventions. Such empirical measurement is necessary to guarantee that resources are applied in the best possible way (1).

Many development programs use community interventions that deploy treatments at the group level, because they change the physical or social environment, because they cannot be delivered to individuals, or because they wish to capture group-level dynamics. The gold standard for inference in community interventions is a community-randomized trial because the design eliminates confounding bias (2). Bias from other sources can result from frequent measurement (3) or lack of blinding treatment (blinding is rarely possible for community interventions) (4, 5). Even if unbiased, trials must evaluate treatments that are amenable to randomization and typically estimate the average effect of an intervention under ideal conditions (delivery and compliance) in populations most likely to benefit; it is widely acknowledged that treatment effects estimated in such trials can differ from those obtained when the intervention is deployed in the general population (6). Measuring intervention sustainability using prospective trials can also be difficult due to logistical complexity, short funding cycles, and rare sequential awards (7).

The gap between the evidence generated by most community-randomized trials and information that is directly useful to policy makers suggests that studies of nonrandomized, preexisting community interventions implemented by governments and NGOs could contribute both unique and complementary data to inform evidence-based decisions. The sheer number of such community interventions, relative to randomized studies addressing the same issues, speaks to a need for a more rigorous methodology with which to evaluate the impact of the interventions used. We define "nonrandomized, preexisting" interventions as those that were designed and deployed before a structured scientific study.

In this article we draw on the potential outcomes model for causal inference (8–11) to motivate a matched cohort design that enables scientific learning from preexisting, real-world implementation programs under a reasonable set of conditions (described below). Causal inference methods have been well articulated in the statistics and economics literature for decades, but have only recently gained popularity in epidemiology. Here, we frame the matched cohort design in terms of potential outcomes—an extension of prior epidemiologic literature on the design (4, 12)—and tailor it to studies of preexisting, community interventions. The design we propose is most relevant for evaluations with a nonrandomized, predefined intervention group of communities, baseline (pretreatment) data on key confounding variables, and finite resources so that outcome measurement is not possible in all communities. The design naturally estimates the average treatment effect among those most likely to receive an intervention from providers who will actually deliver it—a policy-relevant quantity (1)—and the design enables rapid collection of data about intervention sustainability. We illustrate the usefulness and limitations of the approach with a village-level sanitation mobilization, water supply, and hygiene education intervention conducted in rural Tamil Nadu, India. We believe this general methodology will be useful to study a wide range of preexisting, development interventions beyond the sanitation, water, and hygiene sector.

Materials and Methods

Potential Outcomes Model for Causal Effects. Our approach to evaluating preexisting interventions is grounded in the Neyman–Rubin potential outcomes model (8–11). Let $Y_{i,1}$ denote the potential outcome for community i if the community receives an intervention (treatment), and let $Y_{i,0}$ denote its

Author contributions: B.F.A., R.S.K., P. Ramaswamy, A.E.H., K.B., and J.M.C. designed research; R.S.K., P. Ramaswamy, A.G.L., P. Rajkumar, P. Ramaprabha, and N.D. performed research; B.F.A., A.E.H., and J.M.C. analyzed data; and B.F.A., R.S.K., P. Ramaswamy, A.E.H., K.B., and J.M.C. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

[1]To whom correspondence should be addressed. E-mail: benarnold@berkeley.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1008944107/-/DCSupplemental.

potential outcome if it does not receive treatment. The treatment effect for community i is $\psi_i = Y_{i,1} - Y_{i,0}$, but only one potential outcome can ever be observed at the same time. If treatment is randomized, then the treatment assignment (A) is independent of the potential outcomes ($A \perp\!\!\!\perp Y_{i,1}, Y_{i,0}$), and the average difference between treatment groups in observed outcomes, Y_i, is an unbiased estimate of the individual community treatment effect: $\widehat{\psi} = E[Y_{i,1} - Y_{i,0}] = E[Y_{i,1}] - E[Y_{i,0}] = E[Y_i \mid A_i = 1] - E[Y_i \mid A_i = 0]$. Valid inference further requires that the units of intervention (communities) are independent—the treatment of one unit does not influence outcomes in another—and that those treated receive the same treatment (10).

In observational studies treatment assignment A is not random, and so the design does not guarantee that treatment is independent of the potential outcomes. There are usually characteristics (covariates) W that are common causes of both receiving treatment and the outcome and confound the unadjusted comparison of means. The potential outcomes model requires a "strong ignorability" assumption to identify unbiased treatment effects in observational studies (13). Strong ignorability states that all W are measured (no unmeasured confounding) and the treatment and control groups overlap for all combinations of W ($0 < P[A = 1 \mid W] < 1$). The assumption of no unmeasured confounding cannot be evaluated empirically and is a central problem for observational studies.

One approach to weaken the assumption slightly is to target a conditional average treatment parameter: The average treatment effect among the treated (ATT), $\psi_i^{\mathrm{ATT}} = Y_{i,1} - Y_{i,0} \mid A_i = 1$. Estimating the ATT weakens the covariate overlap assumption because it estimates the average effect in the subsample of treated units and thus requires that overlap exist between treatment and control groups at levels of $W \mid A = 1$ rather than the full distribution of W. Observational studies of preexisting interventions are usually constrained to estimating the conditional ATT parameter because without randomization interventions are usually targeted to, or adopted by, a self-selected group that is a nonrandom subset of the total population. However, the ATT is still a very policy-relevant parameter: when estimated for preexisting interventions, it is the average effect in the population most likely to receive the intervention given the providers who would actually deliver it.

Matching in the Design to Approximate a Randomized Experiment. Epidemiology and the social sciences have a long history of using matched cohort designs to study interventions and exposures that are not randomly assigned (14, 15). Recent efforts have used the design in prospective group-level intervention studies (16–18) and in preexisting intervention studies (19–21). In a typical scenario, investigators define a study population, and a subset of the population is selected to receive the intervention by a known or unknown process that is not random. Investigators have information about important confounders, but have not measured outcomes. Matched cohort studies incorporate nonrandom sampling from the study population so that the observed covariate distribution in the control group overlaps and closely matches the covariate distribution in the treatment group. In practice, the design naturally estimates the ATT and is consistent with a general approach of first assembling a control group that is as similar as possible to the treatment group using matching methods and then adjusting for any residual confounding using some form of regression (10, 11).

It is well established that exact matching methods fail to find matches for many treated units in finite samples because the dimension of the joint covariate distribution is too large for nonparametric inference (10). There are many multivariate matching approaches to help address this problem (11, 15, 22). One of the most common is propensity score matching, which simplifies the problem of matching on large numbers of covariates by collapsing the covariates into a single scalar—the propensity score—and then matching treatment and control communities using a one-dimensional match on the propensity score (23, 24). The propensity score is the probability of receiving treatment given a set of baseline covariates, $P(A = 1 \mid W)$, which is unknown for observational studies and must be estimated, usually with a logistic regression. There are numerous ways to match treatment and control units using functions of the propensity score, including nearest-neighbor matching, Mahalanobis distance matching, and optimal matching (11). Sekhon (15) discusses the limitations of propensity score matching in realistic scenarios, where covariates are poorly behaved and not linearly related to the outcome, and proposes a genetic matching algorithm that searches for a matched sample with optimal balance in W.

Whatever matching technique is used, matching does not solve the fundamental problem of unmeasured confounding in observational studies; however, selecting a matched control group in the design stage before measuring outcomes has important advantages (11). First, restricting field data collection to matched treatment and control communities is cost effective because it prevents outcome measurement in extraneous control communities that will not help estimate the ATT parameter. Second, matching helps guarantee that the observed covariate distributions in the treated group overlap with the control group, which enables the analysis to rely less on parametric statistical models and the assumptions they require (25). Matching also accounts for arbitrarily complex relationships between the treatment and covariates, which would need to be modeled explicitly if using regression alone (10). Finally, compared with post hoc statistical adjustment, matching can increase the statistical efficiency of difference parameters, which are useful contrasts for intervention studies (4, 12).

Matched Cohort Designs for Preexisting, Community Interventions. Fig. 1 provides an overview of the design. The innovative components of the design are its use of retrospective, baseline (preintervention) data at the community level to match intervention communities to control communities and its use of propensity score matching—or another alternative multivariate matching approach—to overcome the practical limitations of exact matching in finite samples.

A challenge of studying preexisting interventions is that investigators do not control the intervention, and many community-level interventions that are planned outside of the scientific process have characteristics that make them impossible to evaluate. Before evaluating a nonrandomized, preexisting intervention, investigators should confirm that the intervention meets basic conditions that will enable a valid study (Table 1). In this article we focus on community interventions that are deployed to known geographical units, such as rural villages or neighborhoods in urban areas. We make this restriction because the availability of baseline data collected for purposes other than the study at hand is a core component of the design. These data are typically available for administrative units with known geography (as in a national census), but in theory the design applies to any unit of intervention.

Threats to Validity. ***Unmeasured confounding.*** Because the matched design for preexisting interventions relies on data collected in the past—often independent from the study—it is likely that the data available to match will be incomplete or poorly measured. Matching will improve the balance for measurable characteristics, but is unlikely to remove all differences between treated and control communities so the strong ignorability assumption is unlikely to hold. Matching in the design does not preempt subsequent data analysis (11). Investigators can conduct additional statistical adjustment using data collected in the field study, but they must make a reasoned argument that adjustment covariates could not fall on the causal path between the intervention treatment and outcome of interest (10). If pretreatment outcomes can be measured retrospectively, then the change in the outcome can be compared between treatment and control groups. This "difference-in-differences" parameter removes time-invariant unmeasured confounding assuming the two groups would have had parallel outcome trajectories absent treatment (11). As a robustness check, we recommend falsification tests, where the analysis is repeated for outcomes that could not be influenced by the treatment to investigate whether other interventions or characteristics correlated with treatment could account for the results.

Informative censoring. In the time that elapses between the baseline measurement used to define the study population and the postintervention out-

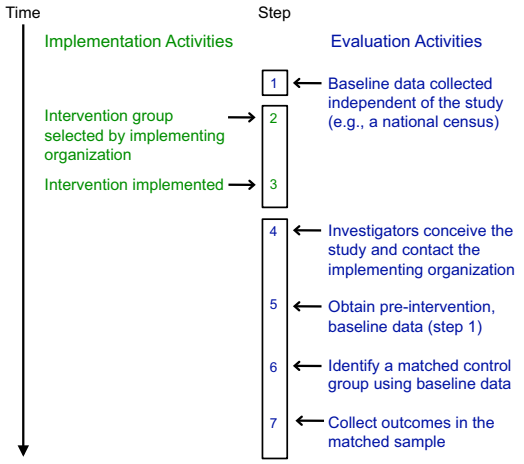

Fig. 1. Overview of the matched cohort design for preexisting interventions.

Table 1. Necessary conditions for matched cohort studies of nonrandomized, preexisting community interventions

	Condition, rationale, and example from this evaluation
1	A partnership with the implementing organization. The implementing organization is the key provider of information about the intervention components, how the intervention beneficiaries were selected, and the timeline and location of activities. Example: We partnered with Water.org and Gramalaya through their funding organization.
2	Sufficient intervention scale. Each community is the independent unit of intervention and it is unusual to have adequate power without at least 8–10 communities per group (2). Example: The intervention included 12 independent communities.
3	Uniformity of the intervention across communities. A relatively uniform intervention is necessary to define and estimate a common treatment effect across communities (in practice, implementation will often vary slightly across communities). Example: The NGOs implemented sanitation, water supply, and hygiene education improvements to raise all communities to a high level of coverage for all three components.
4	Availability of control communities. Control communities are necessary to provide a counterfactual comparison group. Ideally, there should be at least 2 potential control communities for every treatment community. Example: We started with 240 potential control communities from neighboring blocks.
5	Community independence. As with community randomized trials, all units of intervention must be independent with respect to effect of the intervention on outcomes (i.e., no spillover effects). Example: We selected control communities from separate administrative blocks to prevent spillover. We also ensured that communities were qualitatively independent during a rapid assessment following the match but before data collection.
6	Availability of baseline (preintervention) data. Baseline data that include key confounding covariates are used for matched sampling of communities. Baseline data provide a basis for judging baseline comparability of groups in the matched sample. They should reflect conditions at the time of intervention community selection. Example: Census 2001 and Tamil Nadu Water Supply and Drainage board 2003 data collected in the 2 y before the program started included key sanitation, water, and socioeconomic covariates at the community level.

come measurement, individuals within communities will exit the study population (commonly referred to as censoring). If censoring is a common effect of both the intervention treatment and the outcome, then it is informative and will cause bias (26). Informative censoring is a potential source of bias in all study designs, but in prospective designs, characteristics of individuals who exit the study population are available to assess whether censoring is informative. In studies of preexisting interventions, the censored individuals are never measured and so investigators have no direct information about the magnitude of censoring or characteristics of those censored.

Measurement error. If outcomes or exposures are measured retrospectively in the postintervention survey, then they will likely be measured with more error than if they had been measured contemporaneously. Measurement error will cause bias unless it is independent of both treatment status and outcomes (27). Limiting the recall period over which outcomes are measured and using objective outcomes rather than those that rely on self-report can reduce measurement error.

Sampling bias. Sampling bias is possible during community selection or, if outcomes are measured below the community level, in the selection of units below the community level. Investigators should evaluate the completeness of baseline data used to select communities, as incomplete sampling frames could lead to systematic bias. If outcomes are sampled from within the community, then they should be collected from a random sample.

Application of the Design

Sanitation, Water Supply, and Hygiene Intervention in India. Between 2003 and 2007 two NGOs, Water.org and Gramalaya, implemented a combined environmental intervention in 12 rural villages near the city of Tiruchirappalli in Tamil Nadu, India. The intervention combined water supply improvements and repairs with sanitation and hygiene behavior change campaigns that used similar demand mobilization to India's Total Sanitation Campaign. Intervention details varied slightly by village (Table S1), and its intent was to bring all villages to a high level of water supply, sanitation access, and hygiene knowledge. *SI Materials and Methods* includes details of the intervention and study location. The primary objective of the field study was to revisit households after the conclusion of intervention activities to assess outcomes compared with a control group matched on preintervention characteristics. Outcomes included sanitation, water and hygiene conditions and behavior, and health in children <5 y old measured by caregiver-reported diarrhea and anthropometric growth. Diarrhea and child weight measure acute illness in young children, whereas height measures cumulative effects of acute diarrheal illness and chronic intestinal enteropathy caused by repeated exposure to gastrointestinal pathogens (28–30).

Control Selection and Outcome Measurement. The intervention was not randomized and was deployed in villages that were purposely selected by the NGOs. To help reduce potential bias due to differences between intervention and control villages at baseline, we selected control villages with a combination of restriction, propensity score matching on baseline characteristics, and rapid assessment in late 2007. Fig. 2 summarizes the selection process, and *SI Materials and Methods* includes a more detailed description. We enrolled a random sample of up to 50 households per village with children <5 y old. Between January 2008 and April 2009 we visited each participating household once per month for a total of 12 visits. All data collection followed protocols approved by the institutional review boards at the University of California, Berkeley, and Sri Ramachandra Medical College, Chennai, India, and all participants provided informed consent. *SI Materials and Methods* includes details of our exposure and outcome measurement methodology, as well as our statistical analyses using the matched cohort sample. Briefly, we collected detailed information about sanitation conditions and practices, water sources and water quality, and hygiene indicators and handwashing knowledge. In each visit, we collected symptoms of diarrhea, respiratory illness, and general illness in children <5 y old (7 d recall). In the first and last surveys we collected anthropometric growth measurements for children <5 y old. We measured the change in private toilet ownership and water supply between 2003 and 2008 on the basis of household reports in 2008 (retrospective recall for 2003). For all other outcomes we compared groups using postintervention outcomes measured in the 2008–2009 field visits. All estimates are conditional on the matching process using baseline confounders (Fig. 2 and *SI Materials and Methods*). We conducted adjusted

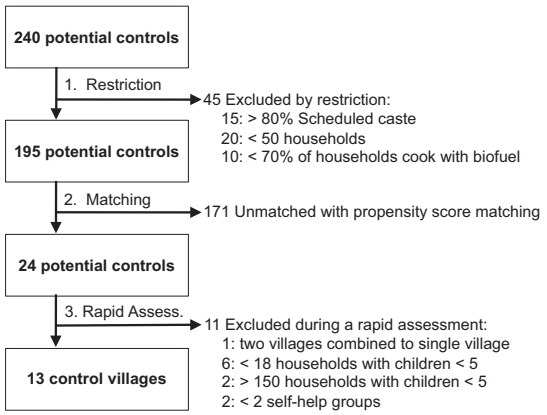

Fig. 2. Control village selection process in the Tamil Nadu study.

analyses using a marginal, *g*-computation estimator with a village-level stratified bootstrap for inference (31).

Matched Cohort Characteristics. The matched cohort design led to a set of matched intervention and control villages that were very similar at baseline, and the approach improved balance greatly for the most imbalanced characteristics (income, biofuel use, sanitation, and water supply) (Table 2). Household characteristics such as durable goods ownership, community participation, and education level were also very similar in our intervention and control groups, postintervention in 2008 (Table S2). Intervention villages were slightly more agricultural than control villages and consequently there were small differences in home ownership, housing materials, and literacy. Our field study sample included 456 control and 444 intervention households [totals exclude 17 control households and 33 intervention households that either moved between listing and enrollment ($n = 44$) or refused]. The 900 households included 1,284 children <5 y old (648 control and 636 intervention). Of these, 612 (94%) control and 608 (96%) intervention children completed the 12-mo follow-up.

Household Sanitation and Water Infrastructure. At baseline in 2003, intervention and control groups were highly similar in toilet and water infrastructure on the basis of census data (Table 2) and retrospective measurement (Fig. 3). Between 2003 and 2008, intervention households were far more likely to build a new private toilet than controls (change in toilet ownership 2003–2008: 48% vs. 15%, $P < 0.0001$, Fig. 3). The intervention increased toilet coverage greatly in the most socially and economically marginalized households (*SI Materials and Methods* and Fig. S1). Of the private toilets in the sample, 89% were pour–flush toilets with a water seal, 5% were ventilated improved pit latrines, and 5% were unimproved concrete slab pit latrines. Toilets were new: 83% were constructed during the 5-y intervention period (since 2003) and 94% were constructed in the last 10 y. Of the 374 households with private toilets, 94% were classified as functional and in use during inspections over the 12-mo period.

Gains in private and public taps were more modest, and increases between 2003 and 2008 did not differ significantly between intervention and control villages (Fig. 3). All households in the study had improved water sources on the basis of the WHO/Unicef Joint Monitoring Program definition (32), and 93% obtained water from public or private taps fed by ground water-supplied overhead tanks. We observed some fecal contamination in household drinking water samples: 27% (120/441) had ≥10 *Escherichia coli* colony-forming units (cfu) per 100 mL. We also found evidence of microbial contamination from general environmental sources in household drinking water: 84% (2,551/3,026) of samples tested positive for hydrogen sulfide (H_2S)-producing bacteria and 91% (2,755/3,015) of samples had ≥100 total coliform cfu per 100 mL (Table S3). Households with private taps spent a median 50 min per day gathering water vs. a median 75 min for households with public taps.

Table 2. Summary of preintervention characteristics before and after village selection

	All villages		Study sample	
Mean	Control	Intervention	Control	Intervention
Demographic				
Total households	170	161	181	161
Persons per household	5	5	5	5
Scheduled caste, %	19	12	15	12
Children ≤5 y old, %	12	12*	12	12
Female literacy, %	52	48	49	48
Socioeconomic				
Employment rate, %	81	78	79	78
Cultivators, %	27	28	31	28
Agricultural laborers, %	24	33	21	33
Marginal workers, %	19	22	21	22
Females work, %	74	69	71	69
Panchayat income (Rp/person)	12,255	7,470***	7,143	7,470
Per-capita cattle ownership	4	4	5	4
Use banking services, %	29	25	25	25
Use biofuel for cooking, %	91	97**	96	97
Own radio, %	43	43	38	43**
Own television, %	21	16	17	16
Own scooter/moped, %	10	10	9	10
Sanitation and water				
Private toilet/latrine, %	15	8**	9	8*
Open defecation, %	85	92**	91	92*
Tap water (private/public), %	75	76	75	76
Hand pump, %	12	14**	18	14
Other water source, %	13	10*	7	10
Persons per hand pump	260	302	240	302
Persons per deep bore well	437	679**	510	679
Water supply level (lpcd)	12	15**	14	15
No. of villages	240	12	13	12

Authors' calculations using India National Census 2001 and Tamil Nadu Water Supply and Drainage 2003 surveys are shown. lpcd, liters per capita per day; Rp, rupees. Scheduled castes include historically disadvantaged, low rank Indian castes, which are currently under government protection. Kolmogorov–Smirnov test for differences in distribution between control and intervention groups: *$P < 0.1$; **$P < 0.05$; ***$P < 0.01$.

Sanitation and Hygiene Behavior. Households in intervention villages were 11 percentage points less likely to report practicing open defecation (77% vs. 88%) than control households (Table S4). Adult open defecation in intervention villages, which had all been declared "open defecation free," ranged between 35% and 83%. Reductions in open defecation were largest among women and smallest among children <5 y old (Table S4). Households that practiced open defecation reported that adult sites were outside the village (98%), but 91% of sites for children <5 y old were within the village. In households with private toilets, 39% reported that adults practice daily open defecation and 52% reported that children <5 y old practice daily open defecation. The most common answers to an open-ended question about the reasons for continuing to practice open defecation despite owning a toilet were no choice (50%), privacy (26%), convenience (25%), and safety (9%). Discrete hygiene spot checks collected by interviewers show overall moderate hygiene conditions, and intervention households fare the same or worse across a large number of indicators (Table S5). Overall, self-reported hand-

washing with soap was rare: Women reported washing their hands after defecation in 24% of 2,657 caregiver interviews (Table S5).

Privacy and Safety for Women and Girls. Private toilet owners were 28 percentage points more likely to report that women and girls feel safe while defecating during the day or night compared with households without private toilets (81% vs. 53%). Overall, the intervention increased the perception of privacy and safety for women and girls during defecation by 13 percentage points compared with controls (72% vs. 59%, Table S4).

Child Health. We identified 259 diarrhea cases from 14,259 child weeks of observation (mean prevalence 1.8%). The mean diarrhea prevalence was slightly higher in intervention villages than in control villages (1.96% vs. 1.67%), and the two groups differed primarily during the summer months (Fig. S2). In unadjusted analyses, we did not observe differences in diarrhea between children in intervention and control villages [longitudinal prevalence difference (LPD) = 0.003, 95% confidence interval (CI) = –0.002, 0.008]. Adjusted estimates, which account for a large set of potentially confounding characteristics (Table S6), also showed no difference between groups (LPD = 0.003, 95% CI = –0.001, 0.008). Despite low diarrhea prevalence, 53% of the children were stunted, 47% were underweight, and 19% were wasted on the basis of weight-for-height. Over 37% were both stunted and underweight (definitions in *SI Materials and Methods*). Mean *Z*-scores were low for both height (mean = –1.96, SD = 1.69) and weight (mean = –1.86, SD = 1.16). We observed no difference in anthropometric *Z*-scores between intervention and control groups [adjusted difference (adj. diff.) in height = 0.01, 95% CI = –0.15, 0.19; and weight = 0.03, 95% CI = –0.11, 0.17; Fig. S3]. Impacts on height are most likely before age 24 mo (33). Restricting the analysis of height-for-age to children who were most likely to benefit from the intervention (<12 mo old at the conclusion of intervention activities, n = 1,093) did not change our findings (adj. diff. = 0.04, 95% CI = –0.28, 0.36).

Discussion

Evaluations of Preexisting Interventions. In this article we have drawn on causal inference theory to develop an evaluation method for nonrandomized, preexisting interventions. Traditionally, such interventions are often evaluated with a before–after comparison in the intervention group alone or with a postintervention, cross-sectional survey in intervention and comparison groups. Before–after comparisons lack a counterfactual comparison and cannot address what would have happened in the absence of intervention. A postintervention, cross-sectional survey neither demonstrates baseline comparability between intervention and comparison communities nor guarantees overlap between groups in important confounding characteristics. In contrast, under appropriate conditions (Table 1), the matched cohort design that we have proposed can demonstrate baseline comparability between intervention and control groups and ensure overlap for observable baseline characteristics so that a valid counterfactual comparison is possible. The attractive features of the design are that it naturally estimates the average effect of an intervention deployed by actually implementing organizations in populations most likely to receive it and yields information about intervention sustainability without years of prospective follow-up. Our motivating example adds to three previous applications of similar methods (to our knowledge) to evaluate preexisting interventions in development settings (19–21), but prior work has not clearly articulated the design's underlying framework, assumptions, and threats to validity. In *SI Materials and Methods* and Fig. S4, we discuss the details of interpreting intervention sustainability in the context of this design. Although we have framed the design in the context of community interventions, in principle it could apply to any unit of intervention (e.g., households or individuals) if appropriate baseline data are available and the study meets the conditions in Table 1.

Our evaluation from Tamil Nadu illustrates many of the strengths and weaknesses of the design. The use of restriction, propensity score matching, and rapid assessment to select control villages (Fig. 2) led to highly similar intervention and control groups on the basis of key exposures and socioeconomic characteristics at baseline (Table 2 and Fig. 3). Although it remains possible that unmeasured confounding has masked the intervention effect, the extremely good overlap in observable confounding characteristics between groups at baseline and follow-up makes this scenario unlikely (Table 2 and Table S2). As a robustness check, we repeated the analyses using caregiver-reported fever in the previous 7 d among children <5 y old as our outcome. Fever is a nonspecific outcome that should not be influenced by the intervention and thus serves as a falsification test. We found no difference between groups in fever (combined prevalence = 11.8%; adj. LPD = 0.008; 95% CI = –0.006, 0.021). Nonetheless, we relied on matched, postintervention differences for all outcomes besides toilet and tap construction; evaluations that use difference-in-differences estimators by comparing the changes in outcomes from baseline to postintervention could be more robust to unmeasured confounding if baseline outcome measures are available (11).

For matching in the design to reduce bias, baseline data must be accurate and complete with respect to key confounders (Table 1, condition 6). Without meeting this condition, matching is unlikely to improve the comparability of intervention and control groups. For sanitation, water, and hygiene interventions, the major confounding variables and intermediate outcomes are often available from national census data, but this condition may not hold for some development research questions. If investigators use secondary data to match groups, as we did in this study, we recommend a brief qualitative and quantitative rapid assessment to validate the data in matched communities before the full field study. This exercise is consistent with integrating ethnographic "thick description" into the selection process (34)—using checks to ensure that intervention units that appear comparable on the basis of computer records are comparable if observed directly.

An additional weakness of this design is its vulnerability to bias from nonrandom subgroups of the population leaving between the intervention and the evaluation (informative censoring). Because such losses are difficult or impossible to measure retrospectively, the evaluation must rely on plausibility arguments. In

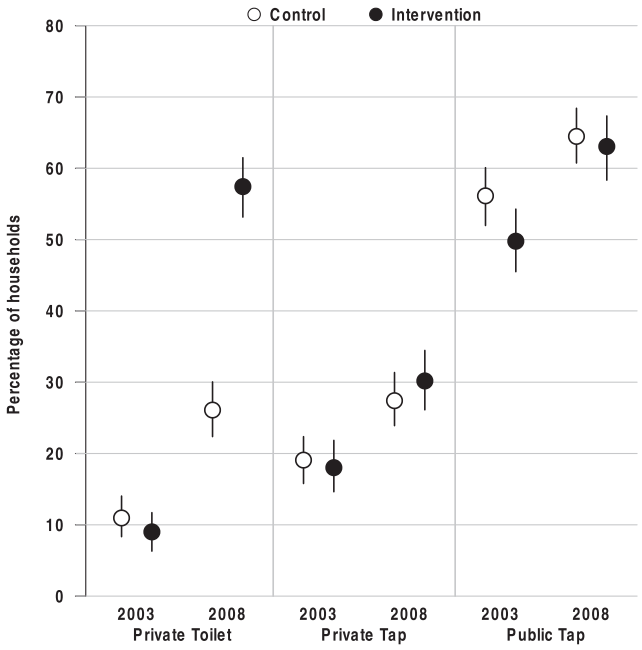

Fig. 3. Population access to private toilets, private water taps, and public water taps in 2003 and 2008. Vertical lines mark bootstrapped 95% confidence intervals. n = 456 control and n = 444 intervention households.

our study sample, just 4.4% of households were lost to follow-up and they were highly similar to those that remained (Table S7). We infer that informative censoring is not a major source of bias in this evaluation.

Combined Interventions, Child Diarrhea, and Growth. Child diarrhea was rare in this population without improved sanitation: 88% of control households practiced open defecation, yet their weekly diarrhea prevalence was just 1.67% over 12 mo. Although we were surprised by this low prevalence, the weekly diarrhea prevalence of all cases reported in the 13 control village health clinics was 1.36% over the same period. (We use the total number of children <5 y old from our original sampling frame as a denominator for the surveillance data. In our control village sample, 80% of diarrhea cases reported visiting the health clinic: $0.8 \times 1.67\% = 1.34\%$, which is very close to the 1.36% prevalence estimated through passive surveillance.)

Our results have a number of implications for government and NGO programs in the sector. They provide evidence that in some populations it is not necessary to combine improvements in water supply, sanitation, and hygiene conditions to achieve very low levels of child diarrhea (35, 36). We infer (although have not tested) that field open defecation is not a primary transmission pathway of diarrhea-causing pathogens for children <5 y old in this population. This study shows that in some rural Indian environments costly sanitation improvements are not guaranteed to have large health benefits, but do improve the perception of privacy and safety for women.

The study also shows that severe growth faltering can persist in populations with rare diarrhea. Poor nutrition is likely a key reason for this (30), but it remains possible that some faltering results from bacterial exposure that is insufficient to cause symptomatic diarrhea, but is sufficient to cause intestinal enteropathy in young children. Enteropathy is hypothesized to cause growth faltering through poor nutrient absorption and low-level immune system stimulation (29). Nutritionists have hypothesized that toilet provision and handwashing with soap could reduce enteropathy and improve growth (37). Our findings indicate that the environmental improvements observed in this study have been insufficient to measurably improve growth (Fig. S3).

Conclusions. Empirical evaluations of interventions that address the most significant global health and development problems are necessary to ensure that resources are applied most responsibly. It is often difficult or impossible to use randomized studies to measure such impacts. If nonrandomized studies are to be used, they require a more nuanced process of study design and interpretation than randomized studies. In this article we have summarized this process for preexisting interventions, and we expect the methodology could be used to study many types of development programs.

ACKNOWLEDGMENTS. We thank Water.org and Gramalaya for providing us with critical program information and allowing us to evaluate their intervention. We also thank the project field team for their invaluable contribution to collecting the data. This study was funded by a grant from the Open Square Foundation to the Aquaya Institute.

1. Murray CJL, Frenk J (2008) Health metrics and evaluation: Strengthening the science. *Lancet* 371:1191–1199.
2. Murray DM, Varnell SP, Blitstein JL (2004) Design and analysis of group-randomized trials: A review of recent methodological developments. *Am J Public Health* 94:423–432.
3. McCarney R, et al. (2007) The Hawthorne Effect: A randomised, controlled trial. *BMC Med Res Methodol* 7:30.
4. Rothman K, Greenland S (1998) *Modern Epidemiology* (Lippincott–Raven, Philadelphia).
5. Wood L, et al. (2008) Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: Meta-epidemiological study. *BMJ* 336:601–605.
6. Horton R (2000) Common sense and figures: The rhetoric of validity in medicine (Bradford Hill Memorial Lecture 1999). *Stat Med* 19:3149–3164.
7. Rajan TV, Clive J (2000) NIH research grants: Funding and re-funding. *JAMA* 283:1963.
8. Splawa-Neyman J, Dabrowska DM, Speed TP (1990) On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Stat Sci* 5:465–472.
9. Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol* 66:688–701.
10. Morgan SL, Winship C (2007) *Counterfactuals and Causal Inference* (Cambridge University Press, Cambridge, UK).
11. Imbens GW, Wooldridge JM (2009) Recent developments in the econometrics of program evaluation. *J Econ Lit* 47:5–86.
12. Greenland S, Morgenstern H (1990) Matching and efficiency in cohort studies. *Am J Epidemiol* 131:151–159.
13. Rubin DB (1978) Bayesian inference for causal effects: The role of randomization. *Ann Stat* 6:34–58.
14. Cochran WG (1953) Matching in analytical studies. *Am J Public Health Nations Health* 43:684–691.
15. Sekhon JS (2009) Opiates for the matches: Matching methods for causal inference. *Annu Rev Polit Sci* 12:487–508.
16. Preisser JS, Young ML, Zaccaro DJ, Wolfson M (2003) An integrated population-averaged approach to the design, analysis and sample size determination of cluster-unit trials. *Stat Med* 22:1235–1254.
17. Pattanayak SK, Poulos C, Yang JC, Patil SR, Wendland KJ (2009) Of taps and toilets: Quasi-experimental protocol for evaluating community-demand-driven projects. *J Water Health* 7:434–451.
18. Newman J, et al. (2002) An impact evaluation of education, health, and water supply investments by the Bolivian Social Investment Fund. *World Bank Econ Rev* 16: 241–274.
19. Arnold BF, Arana B, Mäusezahl D, Hubbard A, Colford JM, Jr (2009) Evaluation of a pre-existing, 3-year household water treatment and handwashing intervention in rural Guatemala. *Int J Epidemiol* 38:1651–1661.
20. Cattaneo MD, Galiani S, Gertler PJ, Martinez S, Titiunik R (2009) Housing, health, and happiness. *Am Econ J Econ Policy* 1:75–105.
21. Pradhan M, Rawlings LB (2002) The impact and targeting of social infrastructure investments: Lessons from the Nicaraguan Social Fund. *World Bank Econ Rev* 16: 275–295.
22. Iacus S, King G, Porro G (2009) cem: Software for coarsened exact matching. *J Stat Softw* 30:1–27.
23. Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55.
24. Rosenbaum PR, Rubin DB (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Stat* 39:33–38.
25. Ho DE, Imai K, King G, Stuart EA (2007) Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit Anal* 15:199–236.
26. Hernán MA, Hernández-Díaz S, Robins JM (2004) A structural approach to selection bias. *Epidemiology* 15:615–625.
27. Hernán MA, Cole SR (2009) Invited commentary: Causal diagrams and measurement bias. *Am J Epidemiol*, 170:959–962 discussion, 963–964.
28. Schmidt W, et al. (2010) Weight-for-age z-score as a proxy marker for diarrhoea in epidemiological studies. *J Epidemiol Community Health*, 10.1136/jech.2009.099721.
29. Lunn PG (2000) The impact of infection and nutrition on gut function and growth in childhood. *Proc Nutr Soc* 59:147–154.
30. Black RE, et al.; Maternal and Child Undernutrition Study Group (2008) Maternal and child undernutrition: Global and regional exposures and health consequences. *Lancet* 371:243–260.
31. Ahern J, Hubbard A, Galea S (2009) Estimating the effects of potential public health interventions on population disease burden: A step-by-step illustration of causal inference methods. *Am J Epidemiol* 169:1140–1147.
32. UNICEF, WHO (2008) *World Health Organization and United Nations Children's Fund Joint Monitoring Programme for Water Supply and Sanitation (JMP). Progress on Drinking Water and Sanitation: Special Focus on Sanitation* (UNICEF and WHO, New York and Geneva).
33. Victora CG, de Onis M, Hallal PC, Blössner M, Shrimpton R (2010) Worldwide timing of growth faltering: Revisiting implications for interventions. *Pediatrics* 125:e473–e480.
34. Rosenbaum PR, Silber JH (2001) Matching and thick description in an observational study of mortality after surgery. *Biostatistics* 2:217–232.
35. Briscoe J (1984) Intervention studies and the definition of dominant transmission routes. *Am J Epidemiol* 120:449–455.
36. Eisenberg JNS, Scott JC, Porco T (2007) Integrating disease control strategies: Balancing water sanitation and hygiene interventions to reduce diarrheal disease burden. *Am J Public Health* 97:846–852.
37. Humphrey JH (2009) Child undernutrition, tropical enteropathy, toilets, and handwashing. *Lancet* 374:1032–1035.