Title
Methods to Study Intervention Sustainability Using Pre-existing, Community Interventions: Examples from the Water, Sanitation and Hygiene Sector

Author
Arnold, Benjamin Ford

Publication Date
2009

Peer reviewed|Thesis/dissertation

Methods to Study Intervention Sustainability Using Pre-existing, Community Interventions: Examples from the Water, Sanitation and Hygiene Sector

by

Benjamin Ford Arnold

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Epidemiology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor John M. Colford, Jr., Chair
Professor Alan E. Hubbard
Professor Kirk R. Smith

Fall 2009

Methods to Study Intervention Sustainability Using Pre-existing, Community Interventions: Examples from the Water, Sanitation and Hygiene Sector

Benjamin Ford Arnold

Abstract

Methods to Study Intervention Sustainability Using Pre-existing, Community Interventions: Examples from the Water, Sanitation and Hygiene Sector

by

Benjamin Ford Arnold

Doctor of Philosophy in Epidemiology

University of California, Berkeley

Professor John M. Colford, Jr., Chair

This dissertation presents a quasi-experimental study design to evaluate non-randomized, pre-existing community interventions not originally designed to gather rigorous data about sustainability and impact. The core components of the design include selection of a control group using propensity score matching with pre-intervention (baseline) secondary data and post-intervention follow-up in the field. The main advantages of the design include measurement of interventions implemented under actual field conditions (independent of scientific research), as well as the design's ability to gather information about the long term impacts and sustainability of interventions without years of costly prospective follow-up. Studies of non-randomized, pre-exisiting interventions must address threats to validity, principal among them: unmeasured confounding and informative censoring.

I outline the main strengths and weaknesses of the study design using simulation and empirical examples. I also apply the design in two sustainability field studies: a 3-year household water treatment and hygiene promotion intervention in rural Guatemala and a 5-year community led total sanitation, water supply and hygiene education intervention in rural India. In both studies, the design leads to samples of intervention and control groups with highly comparable baseline characteristics.

A principal finding of both field studies is poor initial impact and sustainability of the behavioral components of the interventions. In Guatemala, I find a small, five percentage point increase (8.7% vs. 3.3%) in the proportion of households that treat their water six months after the promotion intervention, but no differences in hygiene knowledge or practice, and no detectable differences in child health based on acute illness or growth. In India, I find a large, 33 percentage point increase (48% vs. 15%) in private toilet construction as a result of the intervention, but open defecation persists in 40% of households with functional private toilets. In India, diarrhea is rare in both intervention and control communities (1.8% over 14,259 child weeks), but most children show growth faltering by international standards (mean height-for-age Z-score: -1.98). Despite no signifiant differences in health between children living in intervention and control villages,

I observe important non-health benefits: private toilets increase privacy and safety during defecation for women and girls by 28 percentage points (81% vs. 53%), and private water taps reduce water collection time by a median of 25 minutes per day relative to public taps (50 vs. 75 minutes). I also find that hardware improvements are highly sustainable up to five years after implementation with more than 94% of private toilets and 96% of private water taps in use during repeated visits over one year.

Studies of non-randomized, pre-existing interventions are a rapid, low-cost alternative to prospective intervention studies for evaluating intervention sustainability. The study design and methodology developed in this dissertation are applicable to evaluating a broad range of pre-existing, community interventions beyond the water, sanitation and hygiene sector.

For Ellis and Oliver ☺

Contents

List of Tables

List of Figures

Acknowledgments

This work was only possible with the support from many close friends and colleagues. Jack Colford, my chair and mentor throughout graduate school, provided immense generosity and advice over the past five years. His contribution to my education extends beyond my ability to articulate it, but I can say that good epidemiology is impossible without good relationships and Jack is a great inspiration for cultivating both. I am eager and grateful to be working with him at Berkeley in the upcoming years. My other committee members, Alan Hubbard and Kirk Smith, provided valuable insight throughout the manuscript. Over the years Alan has patiently taught me the bulk of the statistical reasoning in this work. He is a gifted teacher with the rare ability to translate and communicate complex concepts into terms that mere mortals (like myself) can grasp. I am very thankful for Kirk's thoughtful and productive comments on the science and exposition herein. I would also like to thank Ira Tager for his mentorship throughout my time at Berkeley, and Subhrendu Pattanayak who inspired the study design while a visiting fellow at Berkeley.

The field studies were only possible because of the dedicated, hard work of my many collaborators, principal among them: Byron Arana, Ranjiv Khush, Alicia London, Kalpana Balakrishnan, Padmavathi Ramaswamy, Paramasivan Rajkumar, Rama Prabha and Daniel Mäusezahl. Two foundations, the Institute for Public Health and Water Research and the Open Square Foundation, funded the fieldwork, and the SODIS Foundation, WaterPartners International, and Gramalaya generously allowed us to evaluate their programs.

Above all, I want to thank my loving wife Ellis and our close network of family and friends who have kept me engaged in life beyond epidemiology. Ellis has been endlessly encouraging and patient, and I can only aspire to the generosity and support that she has given me throughout this process. Since the arrival of our son, Oliver, I would have completed fewer than 10 pages of this manuscript without considerable childcare help at all hours from Ellis, Alicia Mora and the Ford and Chisholm families (who actually built an office in our horse barn so that I could have the pleasure of working in relative peace while still at home with my wonderful family).

Chapter 1

Introduction and Background

1.1 Introduction

The goal of this dissertation is to motivate and address the research need for measuring the sustainability of water, sanitation and hygiene interventions in developing countries. The working definition of intervention *sustainability* for this manuscript is [1–3]:

> The capacity to maintain intervention services that will provide ongoing benefits to a target population for an extended period of time after the end of major financial, managerial and technical assistance from an external donor.

For example, in the context of a behavior-based household water treatment intervention, a measure of sustainability would be continuation of water treatment until a safe source (e.g., centrally treated and piped to houses) is available in the intervention population. Additional sustainability measures could include water quality and diarrhea prevalence.

Throughout the dissertation I will highlight some of the logistical and methodological difficulties involved in intervention sustainability research. First and foremost is the need to wait many months or years after the completion of an intervention to measure outcomes. This inherent logistical challenge of measuring intervention sustainability – combined with typically short research funding cycles [4, 5] – has created little incentive (or great difficulty) for scientific investigators to document the long term impacts of interventions. In the water, sanitation and hygiene sector this has created a dilemma for donors and governments who are poised to invest large sums on interventions to reduce diarrhea and other related disease burdens, but must base their decisions on evidence from short-duration efficacy trials. The primary methodologic contribution of my research is a quasi-experimental study design that shortens the waiting time for sustainability studies by taking advantage of pre-existing interventions that have been in place for many years. When approaching this problem, I found it necessary to use a novel application of propensity score matching using secondary baseline data, and I will illustrate the value of these techniques in general with specific applications to water, sanitation and hygiene interventions.

1.2 Specific aims

This dissertation has two specific aims:

1. Develop a quasi-experimental design methodology to evaluate non-randomized, pre-existing interventions.

2. Apply the methodology developed under Aim 1 to:

   (a) Measure the sustainability and impact six months after a three-year, combined household water treatment and handwashing behavioral intervention in rural Guatemala that ended in 2006.

   (b) Measure the sustainability and impact of a combined water supply, sanitation and hygiene education intervention in rural Tamil Nadu, India that was initiated in 2003 and concluded in 12 villages between 2004 and 2007.

Aim 2 draws on two field studies that I have completed with the help of colleagues over the past two years. The impacts that I consider focus on behavioral change and young child health outcomes: diarrhea, respiratory infections, height and weight. In the India field study I will also report additional non-health impacts that include time savings and perceptions of privacy and safety among women.

1.3 Structure of the dissertation

I have organized the dissertation into the following chapters:

- In this chapter I identify existing research gaps and make the case for intervention sustainability studies based on pre-existing interventions. I also highlight some of the key challenges that evaluations of pre-existing interventions raise for scientific validity.

- Chapter 2 addresses the first specific aim. It describes a general methodology for the evaluation of pre-exisisting interventions, and benchmarks the key design features against competing approaches based on simulation and empirical results.

- Chapter 3 presents behavioral and health impacts from a sustainability evaluation of a household water treatment and handwashing intervention in rural Guatemala that we conducted in 2007.

- Chapter 4 presents behavioral and health impacts from a sustainability study of a combined water supply, sanitation and hygiene intervention in rural Tamil Nadu, India that commenced in January 2008 and ended in March 2009.

- Chapter 5 draws summary conclusions from the main work and I propose natural extensions for future work.
- Appendices include detailed reviews of scientific evidence for intervention efficacy (A) and a detailed description of targeted maximum likelihood estimation for point treatment studies, which I will use repeatedly in Chapters 3 and 4 (B).

For readers interested mainly in the conclusions of the work, I recommend reading Chapters 1 and 5. Since I have included the majority of the topical background in the Introduction and Appendices, the analytic chapters focus less on background and more on methodology, results and interpretation.

1.4 Water, sanitation and hygiene interventions in developing countries

Preventable diseases that result from poor water quality, lack of sanitation, and absence of good hygiene behavior cause a tremendous disease burden among the world's poor. The greatest disease burden falls on infants and young children under five years old [6], and current estimates indicate that gastrointestinal illness (diarrhea) likely accounts for more than 17% of the 10 million annual deaths among young children [7]. By reducing normal food consumption and nutrient adsorption, diarrheal diseases are also a significant cause of malnutrition, leading to impaired physical growth and cognitive development [8–10], reduced resistance to infection [11], and, potentially, long-term gastrointestinal disorders [12]. Efficacy trials indicate that when properly implemented, water, sanitation and hygiene interventions can interrupt the transmission of infectious pathogens and reduce diarrhea by 20 to 40% in young children (see [13] for meta-analysis).

Substantial economic and quality of life related impacts result from water, sanitation and hygiene interventions. A recent World Health Organization (WHO) analysis estimates that the reduced time spent gathering water is one of the major economic benefits derived from water and sanitation improvements [14]. Briscoe summarizes data that indicate that without access to convenient water sources, family members (primarily women) often spend two to five hours collecting water each day [15]. Sizable economic impacts may also follow from missed work and school days [14]. Poor water, sanitation and hygiene conditions may also exact as-yet unquantified costs such as lack of privacy in the absence of adequate sanitation facilities or time spent seeking medical treatment for sick children.

Many intervention programs in developing countries have included structural and behavioral modifications to improve water supply and quality, sanitation, and hygiene. Figure 1.1 presents a model of gastrointestinal disease transmission and theoretical interruption points for water, sanitation and hygiene interventions. The visual representation of transmission pathways is useful because it highlights the complexity and context-specific nature of gastrointestinal disease transmission. The relative importance and impact of

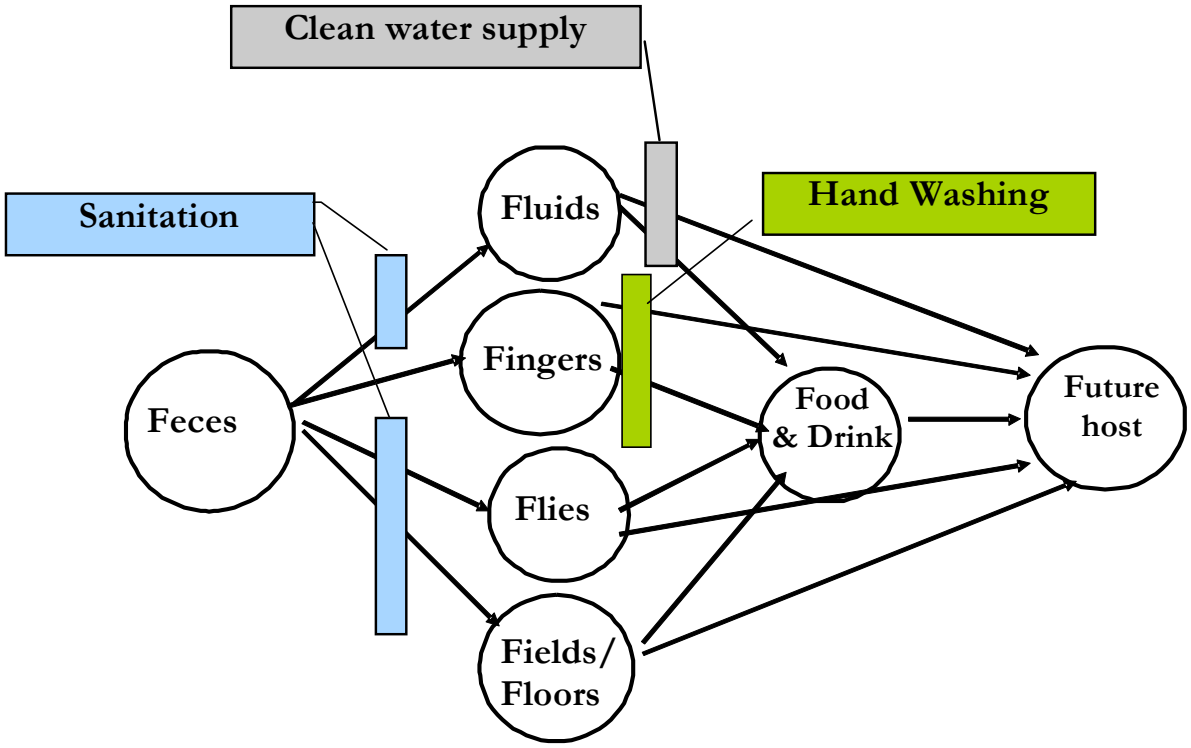


Figure 1.1: The "F Diagram" summarizes infectious gastrointestinal disease transmission and interruption in the context of water, sanitation and hygiene interventions. Adapted from Wagner and Laniox [16].

each intervention will depend on the specific conditions endemic to particular settings and populations.

Despite wide recognition of viable intervention points to reduce gastrointestinal disease transmission, concentrated efforts over the last 60 years have produced only modest progress in extending interventions to communities at highest risk of morbidity and mortality from gastrointestinal disease. The most current global estimates suggest that in 2006, 850 million people (13%) lacked access to improved water sources and 2.5 billion people (38%) did not have adequate sanitation [17] (Table 1.1). Globally, 420 million people need improved water access and 1.052 billion people need improved sanitation to meet the Millennium Development Goal 10 by 2015 [18]. Even water sources that are classified as "improved" may not be microbiologically safe [19], and installing private latrines does not guarantee that all household members will use them (see Chapter 4). The single most important barrier to wider availability of these interventions appears to be the high cost both to establish and to maintain interventions that effectively provide services and health benefits. Nevertheless, improvements in water, sanitation and hygiene have

been advocated for several decades. Countless programs have been implemented in developing countries throughout the world to improve these components using locally adapted interventions and low-cost technologies.

A 2007 *British Medical Journal* poll of 11,341 readers identified sanitation – defined as clean water and sewerage – as the most important medical advance since 1840 (it garnered 15.8% of votes, runners up were antibiotics (15%) and anesthesia (14%)) [20]. It is thus surprising and discouraging that there is a paucity of high quality, evidence-based strategies to guide these interventions [21]. Historically, in the late 19th and early 20th century in cities across the United States the introduction of filtration and or chlorination to centralized municipal water distribution systems was followed by a dramatic decline in typhoid fever specific mortality [22] and infant mortality [23]. Today, a major barrier to optimizing water and sanitary services in low income countries is the prohibitively high cost of providing optimal services, including large water treatment and distribution systems that would provide microbiologically and chemically safe water to all residents 24 hours a day, and sanitation systems that would consistently separate feces from all members of poor communities and the environment (especially from the water and food supply of at risk populations). Thus, all interventions in high need settings represent a tradeoff between what is theoretically ideal to achieve for water and sanitation and what can actually be achieved given the budgetary, technical, political and cultural constraints of a community.

Table 1.1: UNICEF/WHO Joint Monitoring Program (JMP) water supply and sanitation definitions [17].

Category	Description	% World Pop.
Water Supply		
Unimproved	Unprotected dug well, unprotected spring, cart with small tank/drum, tanker truck, and surface water (river, dam, lake, pond, stream, canal, irrigation channels), bottled water.	13
Other Improved	Public taps or standpipes, tube wells or boreholes, protected dug wells, protected springs and rainwater collection.	33
Piped Into Dwelling or Yard	Piped household water connection located inside the user's dwelling, plot or yard.	54
Sanitation		
Open Defecation	Defecation in fields, forests, bushes, bodies of water or other open spaces, or disposal of human feces with solid waste.	18
Unimproved	Facilities that do not ensure hygienic separation of human excreta from human contact. Unimproved facilities include pit latrines without a slab or platform, hanging latrines and bucket latrines.	12
Shared	Sanitation facilities of an otherwise acceptable type shared between two or more households. Shared facilities include public toilets.	8
Improved	Facilities that ensure hygienic separation of human excreta from human contact. They include: (i) Flush or pour-flush toilet/latrine to: piped sewer system, septic tank, or pit latrine. (ii) Ventilated improved pit (VIP) latrine. (iii) Pit latrine with slab. (iv) Composting toilet.	62

1.5 Sustainability of water, sanitation and hygiene interventions

In my opening remarks I proposed a broad, working definition of sustainability: "the capacity to maintain intervention services that will provide ongoing benefits to a target population for an extended period of time after the end of major financial, managerial and technical assistance from an external donor."[1] What constitutes an extended period of time depends on the specific nature of the intervention. In theory, behavior-based interventions are intended to be integrated into existing routines and permanently change participant behavior. If culturally integrated or institutionalized, new behaviors can persist beyond the lifespan of the original beneficiary population [1]. Structural interventions like reticulated water distribution may also have a benefit stream – namely clean and reliable water – that extends long into the future.

This definition applies to two main components of water, sanitation and hygiene interventions. First, there is the continued function and services of structural interventions like reticulated water systems, sewerage and latrines. Second, there is long-lasting behavior change, which includes behaviors like using latrines, handwashing and using household water treatment methods.

Intervention sustainability is important because interventions in the water, sanitation and hygiene sector are almost always motivated by the goal of creating a lasting improvement in the health and quality of life of beneficiaries. If they have no sustained impact, then they waste scarce and finite resources. Unsustained projects may also diminish community support and trust that can hinder future projects [1]. The ultimate goal for most donors and implementors in the sector is to invest in interventions that provide lasting benefits, and the scientific evidence base needs to address this fundamental issue.

Appendix A reviews in detail the current effectiveness evidence for water, sanitation and hygiene interventions in developing countries. In short, there have been few rigorous studies on health impacts from water supply improvements, source water treatment and sanitation improvements. There have been a large number of short-term efficacy[2] studies on behavioral interventions that focus on household water treatment and handwashing with soap, and these low-cost interventions appear to be as effective (or more effective) at reducing diarrhea than large capitalized improvements in water and sanitation infrastructure. Across all studies, the various interventions reduce child diarrhea by 20% to 40%. Evidence for additive or multiplicative health impacts from multiple interventions is sparse and conflicting (Appendix A includes a detailed review).

[1] There are additional branches of "sustainability theory" and "sustainability science," that focus on long term resource management [24].

[2] Consistent with common parlance in epidemiology, I draw a distinction between *efficacy* and *effectiveness*. Efficacy is intervention impact under ideal conditions (e.g., tightly controlled trials). Effectiveness is the intervention impact under real world conditions.

A recent cost-benefit analysis based on models that use secondary data finds that the majority of the benefits from water and sanitation interventions (hygiene interventions not considered) follow from time savings and not health improvements [14]. The costs of improved water supply and sanitation can be large, and their cost per Disability Adjusted Life Year (DALY) ranges between $200 and $2,500 US, depending on the population [25]. Household water treatment interventions have much greater cost effectiveness ratios (cost per DALY ranges between $20 and $680) [25]. It is less clear if the assumptions for these calculations are reasonable. The authors use a 10 year horizon for calculating benefits, yet intervention impacts are based on short-term efficacy trials that may not translate to program implementation [26–28] (see also Chapters 3 and 4).

Effectively providing benefits is a necessary condition for an intervention to have sustained impact, but it is not sufficient. If an intervention's goal is to create sustained health improvements in its target population, then in must be both effective and long-lasting. For structural interventions like reticulated water networks, beneficiaries must be able and motivated to maintain and repair systems or they will eventually break down. For behavioral interventions like household water treatment, individuals must permanently change their behavior so that they continue new practices after intensive promotion activities cease. The effectiveness of some interventions (e.g., latrines) requires both sustained technology function and behavior change.

1.5.1 What makes an intervention sustainable?

This dissertation focuses on methods and empirical examples of evaluating the sustainability of interventions. As such, I will present a summary of key points that relate to technology adoption and behavior change theory in this context rather than a detailed review (see [1–3, 29–37] for more details). The key conditions that contribute to sustainable water and sanitation technology adoption and use include [32–34]:

- cost (affordability)
- convenience / ease of use
- availability of replacement parts / materials
- knowledge of how to maintain and repair broken systems
- cultural and institutional relevance / appropriateness
- active participation and investment from stakeholders
- perceived benefit to users (in health, time savings, dignity or otherwise)

Many of these items are interrelated, and failure in any one area can derail the adoption of a new technology or practice. An often cited example (witnessed around the world)

involves beneficiary households using subsidized latrines for food storage because they perceive that to be a better use of the technology than for separating fecal waste from the environment [38]. Other common problems involve maintaining water and sanitation systems when they break down [32].

Behavior change campaigns that promote handwashing, household water treatment, or better defecation practices have historically used social marketing techniques. In its most simple form, social marketing invokes traditional commercial marketing strategies to change behaviors that relate to social goods. Social marketing campaigns are designed around three general areas [31]:

- *Opportunity*: institutional or structural conditions that influence an individual's chance to perform a behavior
- *Ability*: an individual's skills/proficiencies needed to perform a behavior
- *Motivation*: an individual's arousal or desire to perform a behavior.

All three conditions are necessary for a behavior to occur, and no subset of the conditions is sufficient. Social marketing campaigns attempt to influence people's behavior through these three leverage points: in theory, when the three elements are increased the probability of sustained behavior change increases. For example, if a household has a designated location to wash hands that includes a bar of soap, then the *opportunity* for handwashing is higher than if these conditions were absent. *Motivation* is the most broad, subjective component in the framework. It encompasses notions of attitude, belief, norms, and risk perception. Although interventions can change opportunity and ability using outside inputs, changes in motivation arise from within individuals and communities and are usually more difficult to change. Recent developments in hygiene and sanitation behavior change theory have suggested using people's innate disgust of dirty things to heighten their risk perception and increase their motivation to change behaviors (e.g., wash hands and avoid open defecation) [39, 40].

1.5.2 Current evidence for intervention sustainability

The best examples

To my knowledge there have been two intervention studies that followed participants over multiple years and included health outcomes [28, 41, 42]. The four-year intervention in rural Bangladesh studied by Hoque *et al.* included community tube wells, household latrines, and extensive hygiene education [42]. In a quasi-experimental design, Hoque *et al.* found that six years after the conclusion of the program, 66% of the intervention households still used the tube wells for water, down from 88% during the intervention period but far higher than the 4.9% in the control communities. The study also found sustained improvements in sanitation: 64% of latrines were functional, down from 98%

during the intervention period, and 83% of adults in intervention communities reported using the latrines (versus 7.5% in the control communities). Although the presence of ash for handwashing at the latrine was lower after 5 years that at the end of intervention activities (36% vs. 62%), it was still much higher than in control households (2%). At follow-up, children under five in intervention communities had less diarrhea than children in control communities (relative risk 0.625, 95% CI $0.36, 1.04$).

In a recently published paper from Luby *et al.* the investigators followed-up with a large cohort that participated in randomized trial of handwashing and water treatment in Karachi, Pakistan [28]. During the original trial in 2003 intervention households had 53% less diarrhea than control households [43]. The investigators re-enrolled 67% of the cohort 18 months after the conclusion of the trial and promotion activities and followed the population for an additional 63 weeks. At follow-up, intervention households were 1.5 times more likely to have a place with soap and water to wash hands (79% vs. 53%) and were 2.2 times more likely to demonstrate correct handwashing behavior (50% vs. 23%) compared to controls. Despite sustained behavior change, intervention households purchased soap at the same rate as control households, and there was no difference in the longitudinal prevalence of diarrhea over 236,110 child weeks of observation (longitudinal prevalence difference = -0.0015, 95%CI: $-0.0092, 0.0061$)

Handwashing promotion

In addition to the two more rigorous handwashing sustainability evaluations described above [28, 42], I am aware of two less rigorous handwashing sustainability studies. Wilson and Chandler revisited women who participated in a four month promotion campaign in Indonesia two years after its completion and found that 45 (79%) of 57 women still used hand soap [44]. The study did not include a control group and do not report baseline soap ownership so, while promising, the result is difficult to interpret. A second handwashing and hygiene sustainability evaluation focused on 10 panchayats[3] in Kerala, India where handwashing promotion interventions had been completed between 2 and 9 years earlier [45]. The interventions also included improved sanitation (latrines) and water supply. The follow-up study found that 297 (57.7%) of 515 women reported washing their hands with soap, and 225 (84.6%) of 266 respondents demonstrated correct handwashing techniques. In addition, 461 (89.5%) of 515 women reported that they always used a latrine for defecation. The authors report that the correct handwashing techniques were positively associated with remembering participating in health classes (odds ratio 2.04, 95% CI $1.05, 3.96$), and that handwashing prevalence was not associated with the time since intervention activities concluded. However, these results are difficult to interpret because the study did not include an adequate control group.[4]

[3] In India, panchayats are an administrative unit that typically includes $3-5$ villages and roughly 500 people.

[4] The study did include a single control panchayat selected through ad-hoc methods, where self-reported handwashing was <10% [45].

Household water treatment

Since 1999 there have been at least 35 efficacy trials on household water treatment and 14 on handwashing (see [46] and [47] for systematic reviews). The majority of these behavioral intervention trials have been short: they typically run for between three and six months. Arnold and Colford found that the median length of chlorine water intervention trials was six months (just two trials ran for 12 months or longer), and the treatment effect decreased with increasing study length [48].[5] To date, three studies have evaluated the adoption of household water treatment under non-trial conditions and all evaluated the Safe Water System (SWS), which combines chlorine disinfection with safe storage [49–51].

Makutsa *et al.* found that 58 (33.5%) of 173 households had detectable free chlorine in stored water (an indicator for chlorine disinfection use) six months into a CARE/Kenya social marketing campaign that promoted the SWS in conjunction with latrine provision and water supply improvements [49]. A follow-up study 18 months into promotion activities found that 43% of intervention households had free detectable chlorine and that intervention villages had 69% fewer episodes of diarrhea in children < 5 years [50].[6] Ram *et al.* evaluated a stand-alone SWS program implemented by CARE/Madagascar 18 months after the campaign started and found that 29 (54%) of 54 households had detectable free chlorine in their stored water during surprise visits; the investigators did not measure health outcomes [51].

Two recent studies have evaluated the sustainability of household water treatment with a sachet-based flocculant disinfectant (PUR) [26] and with ceramic filters [27]. Luby *et al.* conducted a follow-up survey six months after a year-long randomized trial of household water treatment with the PUR flocculant/disinfectant [26, 52]. The follow-up survey found that just 5% of households regularly treated their water. Brown *et al.* visited 506 randomly selected households that participated in a subsidized household ceramic water filter intervention in Cambodia between 2002 and 2006 [27]. Using retrospective measurement of filter use, the authors found that filter use declined by 2% per month after implementation with fewer than 10% of households using filters after 3.5 years (median length of use was approximately 2 years). The primary reason for disuse (63%) was filter breakage.

[5] There are multiple possible explanations for this. Participants may gradually abandon the new practices when faced with the increased daily burden of water treatment or handwashing. The decreased effect could also be an artifact of diarrhea seasonality. For example, in a 12-month, multi-arm chlorine, PUR, and handwashing trial, Luby *et al.* documented no differences between intervention and control groups during the peak of the rainy season, but found large differences during the dry season [43]. If shorter trials are (rationally) conducted during the time of the year when the interventions' impact is greatest to maximize statistical power, then the effect estimates would be larger compared to year-long trials that average the effect over the high and low impact seasons.

[6] The health estimates from this follow-up study could be confounded because the intervention was not randomized, the control villages were selected in an ad-hoc way using geographic proximity, and unfortunately the investigators provide minimal information about balance on potential confounders between intervention and control groups [50].

Water supply

There is a large literature on the sustainability of water supply projects in developing countries (see, for example [34–37]), but there are few studies that measure health impacts from water supply projects (see Waddington *et al.* [53] for a review). Water supply projects have had variable success from a sustainability standpoint. Some estimates put the failure rates of rural supply projects at 30% to 60% in Africa [37], and the study by Hoque *et al.* (described above) found just 66% of families using hand pumps after five years (although the authors document sustained reductions in child diarrhea, water supply was combined with sanitation and hygiene interventions) [42]. The only additional high quality health study of water supply identified by Waddington *et al.* is an observational study in India by Jalan and Ravallion [54]. The authors used national census data from 2001 and propensity score matching to estimate a 21% relative reduction in diarrhea among children under age five due to piped water. Although the authors provide estimates of effect modification by income and education, they do not provide information about the age of water supply infrastructure so it is unclear about whether the health impacts are sustainable over time. Clearly, a pre-requisite for sustainable health impacts is sustainable service. Recent development papers promote the success of community participation-based techniques to improve project longevity and sustainability [34–37, 55–57]. Early evidence from World Bank projects across Africa, Asia and Latin America suggests that the approach may work at scale [55–57], but most of the evidence is from young (< 1 year old) projects.

Sanitation

In addition to the study by Hoque *et al.* ([42], discussed above), there is some evidence that fully-subsidized latrines are sustainable over at least a one to five year time horizon. In the Gambia improved pit latrines were provided free of charge to 666 households in 32 villages. After 25 to 47 months each household was revisited; 77% of the provided latrines were still in use and 97% of latrines owners said they would make a new latrine when their current one was full [58]. In an evaluation of one of the Carter Center's subsidized latrine provision programs that included a random sample of 200 households across 50 villages in Niger, 86% of latrines were in regular use and 70% were clean after one year during unannounced visits [59].

Community Led Total Sanitation (CLTS) is a relatively new marketing approach that was pioneered in Bangladesh and has gained popularity worldwide [60]. CLTS aims to achieve universal sanitation coverage without subsidies by changing social norms and encouraging construction of low-cost latrines. Under CLTS, an external facilitator leads a community meeting and exercises designed to make residents aware of the magnitude of the sanitation problem, elicit feelings of disgust and shame, and create an impetus for collective action. Through an emphasis on the public nature of the problem, facilitators promote the goal of zero open defecation. Typically, communities are encouraged to

come up with their own latrine designs using locally available, low-cost materials that put sanitation within reach of even their poorest members [60].

I am aware of four peer-reviewed studies of CLTS interventions, and interestingly all focus exclusively or primarily on toilet construction and not on open defecation practices (the equally important behavioral component to sanitation improvements) or health. The most rigorous study by Pattanayak *et al.* evaluated a marketing and subsidy CLTS campaign in the state of Orissa, India [61]. They observed an 29% increase in private toilet construction relative to control villages, and an 26% increase in self-reported private toilet use over one year (the paper does not report open defecation practices). A case-control study conducted within a social marketing latrine promotion program in Ghana found that at the time of their survey only 60% of latrines (up to four years old) were functional and in use [62]. In Zimbabwe, a pilot study of the Participatory Hygiene And Sanitation Transformation (PHAST) campaign that combines social marketing with subsidized hardware increased latrine ownership to 43%, up from 2% coverage in (subjectively) matched control villages [63]. The authors report that all households with latrines no longer practice open defecation, but they treat the categories as mutually exclusive (which may be unreasonable, see Chapter 4). Finally an evaluation of the CLTS promotion campaign in Ethiopia that constructed over 89,000 latrines in 2004 found that 87% of 160 randomly selected participants had completed latrines and that 90% of these latrines were in use at the end of the campaign in December 2004 [64].

Summary comments on sustainability evidence

The quantity and quality of sustainability research in the water, sanitation and hygiene sector is inadequate to inform investments in the sector. Existing evidence conflicts, and there is genuine equipoise in the scientific community about whether the simple interventions are sustainable, especially with respect to health outcomes and other benefits. Donors, governments and policy makers have immediate practical use for intervention studies that incorporate longer time scales. Specifically, when they make cost-benefit calculations they need to make assumptions about the benefits that accrue over the life of the investment. If health benefits estimated from short-term efficacy studies inform the calculations, then they could both under-estimate the future benefit stream (if the intervention becomes more effective over time, e.g., through delayed adoption) or over-estimate the future benefit stream (if the intervention becomes less effective over time, e.g., due to poor maintenance or abandonment). The best estimates of the global cost of meeting the water component of the Millennium Development Goals is US$ 42 billion, and for sanitation it is US$ 142 billion [14]. The immense costs of water and sanitation infrastructure guarantee that they will compete for scarce resources with other health priorities. Realistic estimates of health and non-health benefits should inform spending decisions of this magnitude.

1.6 Motivation and challenges of studying pre-existing interventions

As described in the previous section, there is sparse scientific evidence for or against the sustainability of water, sanitation and hygiene interventions for periods longer than one year. A likely reason for the evidence gap in sustainability research (broadly) is that it requires years of prospective data collection using traditional longitudinal cohort designs. This can be both expensive and logistically difficult in a funding climate that rarely awards research grants with time horizons over four years and where it is rare to obtain sequential awards (NIH funding, for example [4, 5]).

A methodologically ideal study design for evaluating intervention sustainability would be to randomize a real-world implementation of the intervention, and then prospectively follow the study population for many years, collecting outcomes at baseline and throughout follow-up. Given a sufficient sample size, randomization will balance potentially confounding factors, and the causal effect of the intervention can be estimated without bias if participants exiting the study (censoring) do so independent of treatment and the outcome [65]. Even if intervention compliance were poor, that would be an important finding in itself. This design requires years of prospective data collection, which is both costly and logistically difficult. Luby *et al.* used a design similar to this in their handwashing sustainability evaluation in Pakistan (described above) [28]. An alternative to lengthy prospective data collection is to take advantage of pre-existing interventions to measure indicators of sustainability. (I define "pre-existing" interventions as those that were designed and deployed prior to a structured scientific study.) Evaluating such non-randomized, pre-existing interventions has advantages but poses several methodologic challenges.

Advantages of studying pre-existing interventions

There are at least two distinct advantages to studying pre-existing interventions to measure sustainability. First, studying pre-existing water, sanitation and hygiene interventions can yield results years faster and at much lower cost than if we started at baseline and followed intervention populations for many years.

A second advantage is that pre-existing interventions are implemented without careful monitoring by research staff. Much of the existing data on water, sanitation and hygiene interventions has been collected in the context of highly monitored populations that participate in field trials. In most cases, the study participants are visited each week (or month) and this style of intense monitoring can cause important behavior changes that we would not necessarily expect under non-study conditions where visits are less frequent (a version of the "Hawthorne Effect"[7]) (Several papers discuss self-reporting bias of di-

[7]Hawthorne Effects are effects caused by the act of conducting a scientific study that would not otherwise occur. An example is the phenomenon where people temporarily change their behavior or performance during a research study simply because they receive more attention than usual.

arrhea outcomes in the context of monitored populations [46, 48, 66–68].) In the end, it is unclear how well the impacts from the short and intense studies generalize to real development work.

Challenges of studying pre-existing interventions

Studying pre-existing interventions raises multiple important methodologic challenges. First, implementing organizations often target interventions to communities that are in most need and are a non-random sample of communities. This process makes the identification of an appropriate control group difficult and in many cases impossible. Second, accurate and detailed data on pre-intervention conditions rarely exist for the study population, yet are essential to select a control group and to establish baseline comparability between groups. Third, the intervention itself may not be standardized across communities, leading to imprecise definitions of treatment. Fourth, since the evaluation relies on retrospective measurement, there is no way to evaluate whether individuals that exit the study population do so because of some common effect of the treatment and the outcome (informative censoring). Finally, implementing organizations may have little incentive to cooperate with outside researchers because of concerns that their interventions may not be found to be sustainable or effective and thus not eligible for future funding.

Contributions to this topic

The quasi-experimental study design that I outline in Chapter 2 addresses many of these methodologic challenges by outlining necessary conditions for conducting evaluations of pre-existing interventions, and by introducing a novel application of propensity score matching to help reduce bias in estimating treatment effects. Chapters 3 and 4 implement the design in two separate studies of non-randomized, pre-existing interventions that include household water treatment and handwashing education (Guatemala) and water supply, sanitation and hygiene education (India).

Bibliography

[1] Shediac-Rizkallah MC, Bone LR. Planning for the sustainability of community-based health programs: conceptual frameworks and future directions for research, practice and policy. Health Educ Res. 1998;13(1):87–108.

[2] USAID. Sustainability of Development Programs: A Compendium of Donor Experience. Washington DC; 1988.

[3] LaPelle NR, Zapka J, Ockene JK. Sustainability of Public Health Programs: The Example of Tobacco Treatment Services in Massachusetts. Am J Public Health. 2006;96(8):1363–1369.

[4] Rajan TV, Clive J. NIH Research Grants: Funding and Re-funding. JAMA. 2000;283(15):1963–.

[5] NIH. Research Portfolio Online Reporting Tool (RePORT), Average Project Period Length (accesssed online, http://report.nih.gov/index.aspx?section=NIHFunding). National Institutes of Health; 2009.

[6] Murray CJ, Lopez AD. Mortality by cause for eight regions of the world: Global Burden of Disease Study. Lancet. 1997;349(9061):1269–76. Lancet.

[7] Bryce J, Boschi-Pinto C, Shibuya K, Black RE. WHO estimates of the causes of death in children. Lancet. 2005;365(9465):1147–52.

[8] Guerrant DI, Moore SR, Lima AA, Patrick PD, Schorling JB, Guerrant RL. Association of early childhood diarrhea and cryptosporidiosis with impaired physical fitness and cognitive function four-seven years later in a poor urban community in northeast Brazil. Am J Trop Med Hyg. 1999;61(5):707–13.

[9] Humphrey JH. Child undernutrition, tropical enteropathy, toilets, and handwashing. Lancet. 2009 Sep;374(9694):1032–1035.

[10] Checkley W, Buckley G, Gilman RH, Assis AM, Guerrant RL, Morris SS, et al. Multi-country analysis of the effects of diarrhoea on childhood stunting. Int J Epidemiol. 2008;37(4):816–30.

[11] Baqui AH, Sack RB, Black RE, Chowdhury HR, Yunus M, Siddique AK. Cell-mediated immune deficiency and malnutrition are independent risk factors for persistent diarrhea in Bangladeshi children. Am J Clin Nutr. 1993;58(4):543–8. The American journal of clinical nutrition.

[12] Schneider RE, Shiffman M, Faigenblum J. The potential effect of water on gastrointestinal infections prevalent in developing countries. Am J Clin Nutr. 1978;31(11):2089–99.

[13] Fewtrell L, Kaufmann RB, Kay D, Enanoria W, Haller L, Colford J J M. Water, sanitation, and hygiene interventions to reduce diarrhoea in less developed countries: a systematic review and meta-analysis. Lancet Infect Dis. 2005;5(1):42–52.

[14] Hutton G, Haller L, Bartram J. Global cost-benefit analysis of water supply and sanitation interventions. J Water Health. 2007;5(4):481–502.

[15] Briscoe ME. Research note: proxy responses in health surveys: a methodological issue. Sociology of Health and Illness. 1984;6(3):359–65.

[16] Wagner EG, Lanoix JN. Excreta disposal for rural areas and small communities. Monogr Ser World Health Organ. 1958;39:1–182.

[17] WHO. 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. WHO, UNICEF; 2008.

[18] Hutton G, Bartram J. Global costs of attaining the Millennium Development Goal for water supply and sanitation. Bull World Health Organ. 2008;86(1):13–9.

[19] Luby SP. Quality of drinking water. Bmj. 2007;334(7597):755–6.

[20] Ferriman A. BMJ readers choose the "sanitary revolution" as greatest medical advance since 1840. Bmj. 2007;334(7585):111–a–.

[21] Blum D, Feachem RG. Measuring the impact of water supply and sanitation investments on diarrhoeal diseases: problems of methodology. Int J Epidemiol. 1983;12(3):357–65.

[22] Melosi M. The Sanitary City. Baltimore: The Johns Hopkins University Press; 2000.

[23] Cutler D, Miller G. The role of public health improvements in health advances: the twentieth-century United States. Demography. 2005;42(1):1–22.

[24] Kennedy D. Sustainability. Science. 2007;315(5812):573–573.

[25] Haller L, Hutton G, Bartram J. Estimating the costs and health benefits of water and sanitation improvements at global level. J Water Health. 2007;5(4):467–80.

[26] Luby SP, Mendoza C, Keswick BH, Chiller TM, Hoekstra RM. Difficulties in bringing point-of-use water treatment to scale in rural Guatemala. Am J Trop Med Hyg. 2008;78(3):382–7.

[27] Brown J, Proum S, Sobsey MD. Sustained use of a household-scale water filtration device in rural Cambodia. J Water Health. 2009;7(3):404–12.

[28] Luby SP, Agboatwalla M, Bowen A, Kenah E, Sharker Y, Hoekstra RM. Difficulties in maintaining improved handwashing behavior, Karachi, Pakistan. Am J Trop Med Hyg. 2009 Jul;81(1):140–145.

[29] Chapman S. Evaluating Social Marketing Interventions, Ch 7. Thorogood M, Coombes Y, editors. Oxford: Oxford University Press; 2004.

[30] Greenberg MR. The diffusion of public health innovations. Am J Public Health. 2006;96(2):209–10.

[31] Rothschild ML. Carrots, sticks, and promises: A conceptual framework for the management of public health and social issue behaviors. Journal of Marketing. 1999;63(4):24–37.

[32] Campos M. Making sustainable water and sanitation in the Peruvian Andes: an intervention model. J Water Health. 2008;6 Suppl 1:27–31.

[33] Sobsey MD, Stauber CE, Casanova LM, Brown JM, Elliott MA. Point of use household drinking water filtration: A practical, effective solution for providing sustained access to safe drinking water in the developing world. Environ Sci Technol. 2008;42(12):4261–7.

[34] Carter RC, Tyrrel SF, Howsam P. The Impact and Sustainability of Community Water Supply and Sanitation Programmes in Developing Countries. Water and Environment Journal. 1999;13(4):292–296.

[35] Gine R, Perez-Foguet A. Sustainability assessment of national rural water supply program in Tanzania. Natural Resources Forum. 2008;32(4):327–342.

[36] Gleitsmann BA, Kroma MM, Steenhuis T. Analysis of a rural water supply project in three communities in Mali: Participation and sustainability. Natural Resources Forum. 2007;31(2):142–150.

[37] Harvey PA, Reed RA. Community-managed water supplies in Africa: sustainable or dispensable? Community Dev J. 2007;42(3):365–378.

[38] Nekesa J. Many decades of sanitation promotion, but no change – where have we gone wrong? (presentation). Stockholm; 2008. .

[39] Curtis VA. A natural history of hygiene. Can J Infect Dis Med Microbiol. 2007;18(1):11–4.

[40] Curtis VA. Dirt, disgust and disease: a natural history of hygiene. J Epidemiol Community Health. 2007;61(8):660–4.

[41] Aziz KM, Hoque BA, Hasan KZ, Patwary MY, Huttly SR, Rahaman MM, et al. Reduction in diarrhoeal diseases in children in rural Bangladesh by environmental and behavioural modifications. Trans R Soc Trop Med Hyg. 1990;84(3):433–8. Transactions of the Royal Society of Tropical Medicine and Hygiene.

[42] Hoque BA, Juncker T, Sack RB, Ali M, Aziz KM. Sustainability of a water, sanitation and hygiene education project in rural Bangladesh: a 5-year follow-up. Bull World Health Organ. 1996;74(4):431–7.

[43] Luby SP, Agboatwalla M, Painter J, Altaf A, Billhimer W, Keswick B, et al. Combining drinking water treatment and hand washing for diarrhoea prevention, a cluster randomised controlled trial. Trop Med Int Health. 2006;11(4):479–89.

[44] Wilson JM, Chandler GN. Sustained improvements in hygiene behaviour amongst village women in Lombok, Indonesia. Trans R Soc Trop Med Hyg. 1993;87(6):615–6. Transactions of the Royal Society of Tropical Medicine and Hygiene.

[45] Cairncross S, Shordt K, Zacharia S, Govindan BK. What causes sustainable changes in hygiene behaviour? A cross-sectional study from Kerala, India. Soc Sci Med. 2005;61(10):2212–20.

[46] Clasen T, Schmidt WP, Rabie T, Roberts I, Cairncross S. Interventions to improve water quality for preventing diarrhoea: systematic review and meta-analysis. Bmj. 2007;334(7597):782. BMJ (Clinical research ed.

[47] Ejemot RI, Ehiri JE, Meremikwu MM, Critchley JA. Hand washing for preventing diarrhoea. Cochrane Database Syst Rev. 2008;(1):CD004265.

[48] Arnold BF, Colford J J M. Treating water with chlorine at point-of-use to improve water quality and reduce child diarrhea in developing countries: a systematic review and meta-analysis. Am J Trop Med Hyg. 2007;76(2):354–64. The American journal of tropical medicine and hygiene.

[49] Makutsa P, Nzaku K, Ogutu P, Barasa P, Ombeki S, Mwaki A, et al. Challenges in implementing a point-of-use water quality intervention in rural Kenya. Am J Public Health. 2001;91(10):1571–3. American journal of public health.

[50] Garrett V, Ogutu P, Mabonga P, Ombeki S, Mwaki A, Aluoch G, et al. Diarrhoea prevention in a high-risk rural Kenyan population through point-of-use chlorination, safe water storage, sanitation, and rainwater harvesting. Epidemiol Infect. 2008;136(11):1463–71. Epidemiology and infection.

[51] Ram PK, Kelsey E, Rasoatiana, Miarintsoa RR, Rakotomalala O, Dunston C, et al. Bringing safe water to remote populations: an evaluation of a portable point-of-use

intervention in rural Madagascar. Am J Public Health. 2007;97(3):398–400. American journal of public health.

[52] Reller ME, Mendoza CE, Lopez MB, Alvarez M, Hoekstra RM, Olson CA, et al. A randomized controlled trial of household-based flocculant-disinfectant drinking water treatment for diarrhea prevention in rural Guatemala. Am J Trop Med Hyg. 2003;69(4):411–9.

[53] Waddington H, Snilstveit B, White H, Fewtrell L. Water, sanitation and hygiene interventions to combat childhood diarrhoea in developing countries. Int Initiative for Impact Eval. 2009 Aug;Synthetic Review 001.

[54] Jalan J, Ravallion M. Does piped water reduce diarrhea for children in rural India? Journal of Econometrics. 2003;112(1):153–173.

[55] Isham J, Narayan D, Pritchett L. Does Participation Improve Performance? Establishing Causality with Subjective Data. World Bank Econ Rev. 1995;9(2):175–200.

[56] Prokopy LS. The relationship between participation and project outcomes: Evidence from rural water supply projects in India. World Development. 2005 Nov;33(11):1801–1819.

[57] Isham J, Kahkonen S. Institutional Determinants of the Impact of Community-Based Water Services: Evidence from Sri Lanka and India. Economic Development and Cultural Change. 2002;50(3):667–691.

[58] Simms VM, Makalo P, Bailey RL, Emerson PM. Sustainability and acceptability of latrine provision in The Gambia. Trans R Soc Trop Med Hyg. 2005;99(8):631–7. Journal Article Research Support, Non-U.S. Gov't England.

[59] Diallo MO, Hopkins DR, Kane MS, Niandou S, Amadou A, Kadri B, et al. Household latrine use, maintenance and acceptability in rural Zinder, Niger. Int J Environ Health Res. 2007;17(6):443–52.

[60] Kar K. Subsidy or self-respect? Participatory total community sanitation in Bangladesh. Institute of Development Studies, Working paper 184; 2003.

[61] Pattanayak SK, Yang JC, Dickinson KL, Poulos C, Patil SR, Mallick RK, et al. Shame or subsidy revisited: social mobilization for sanitation in Orissa, India. Bull World Health Organ. 2009;8:580 – 587.

[62] Rodgers AF, Ajono LA, Gyapong JO, Hagan M, Emerson PM. Characteristics of latrine promotion participants and non-participants; inspection of latrines; and perceptions of household latrines in Northern Ghana. Trop Med Int Health. 2007;12(6):772–82.

[63] Waterkeyn J, Cairncross S. Creating demand for sanitation and hygiene through Community Health Clubs: a cost-effective intervention in two districts in Zimbabwe. Soc Sci Med. 2005;61(9):1958–70.

[64] O'Loughlin R, Fentie G, Flannery B, Emerson PM. Follow-up of a low cost latrine promotion programme in one district of Amhara, Ethiopia: characteristics of early adopters and non-adopters. Trop Med Int Health. 2006;11(9):1406–15.

[65] Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004;15(5):615–25.

[66] Genser B, Strina A, Teles CA, Prado MS, Barreto ML. Risk factors for childhood diarrhea incidence: dynamic analysis of a longitudinal study. Epidemiology. 2006;17(6):658–67.

[67] Schmidt WP, Luby SP, Genser B, Barreto ML, Clasen T. Estimating the longitudinal prevalence of diarrhea and other episodic diseases: continuous versus intermittent surveillance. Epidemiology. 2007;18(5):537–43.

[68] Schmidt WP, Cairncross S. Household Water Treatment in Poor Populations: Is There Enough Evidence for Scaling up Now? Environmental Science & Technology. 2009;43(4):986–992.

Chapter 2

A Quasi-experimental Design to Evaluate Non-randomized, Pre-existing Community Interventions

2.1 Goals

Cook and Shadish [1] write: "Most of the scholars writing about the analysis of quasi-experimental data . . . are reluctant apologists for their work, preferring randomized experiments but realizing that they are not always possible." It is in this spirit that I propose the work herein: randomized experiments are not always possible and there is much to learn from carefully-conducted observational studies. In this chapter I outline minimum criteria that non-randomized, pre-existing interventions must meet for investigators to estimate valid treatment effects. I propose a quasi-experimental design and review its advantages, limitations and main threats to validity. I also demonstrate its utility compared to other approaches using didactic simulations and an empirical example from Guatemala.

2.2 Introduction

Studies of interventions that are deployed to independent groups of individuals for convenience or theoretical reasons make up the class of community (or group) intervention studies. In this chapter I restrict the focus to community interventions that are deployed to known geographical units, such as rural villages or discrete neighborhoods in urban areas. I make this restriction because the availability of secondary data is a core component of the design (described in detail below), and is typically available for administrative units with known geography (as in a national census).

Despite a broad literature on methods for prospective community intervention studies [2–4], I am unaware of a description of a study design to retrospectively evaluate non-randomized, pre-existing community interventions not originally designed to gather rigorous data about sustainability and impact. I define "non-randomized, pre-existing" interventions as those that were designed and deployed prior to a structured scientific study.

There are clear advantages that motivate studies of pre-existing interventions. First, evaluating pre-existing interventions avoids the frequent and intense monitoring that often accompanies prospective intervention studies, which can hamper measuring the true effectiveness of the interventions because the study itself can modify intervention deployment, participant behavior and outcome measurement ("Hawthorne effects"). Second, studies of pre-existing interventions can gather information about the long term impacts and sustainability of interventions without years of costly prospective follow-up. As a relatively inexpensive design, such studies can also provide key information that contributes to planning prospective randomized trials.

Evaluating non-randomized, pre-existing community interventions poses several methodologic challenges. First, implementing organizations often target interventions to communities that are in most need and are a non-random sample of communities. This non-random community selection makes the identification of an appropriate control group difficult and potentially impossible. Second, accurate and detailed data on pre-intervention

conditions rarely exist for the study population, yet are essential to select a control group and to establish baseline comparability between intervention and control groups. Third, the intervention itself may not be standardized across communities, leading to imprecise definitions of treatment. Finally, implementing organizations may have little incentive to cooperate with outside investigators because of concerns that their interventions may not be found to be sustainable or effective and thus not eligible for future funding.

In this chapter I propose a quasi-experimental design that selects control communities using matching techniques based on observed, pre-intervention characteristics. Under suitable conditions the study design addresses the methodologic challenges raised by attempting to measure intervention impacts without randomized assignment and prospective follow-up. Throughout the chapter I highlight both the strengths and limitations of the approach. Section 2.3 provides background on causal inference and a matching framework that I will use to motivate the quasi-experimental design. Section 2.4 outlines some of the practical considerations that a team of investigators should address prior to initiating a study, such as identifying an evaluation partner and deciding whether an intervention meets minimal necessary conditions for evaluation. In Section 2.5 I describe the quasi-experimental study design that uses propensity score matching to select intervention and control villages and statistical adjustment to estimate the parameters of interest. I then evaluate the performance of this approach relative to alternate designs using didactic simulations (Section 2.7) and an empirical example from Guatemala (Section 2.8). Section 2.9 concludes with a discussion.

2.3 Background: estimating treatment effects and matched designs

2.3.1 The Neyman-Holland-Rubin causal model

Nearly all causal inference problems in applied research frame the parameters of interest in the context of the Neyman-Holland-Rubin causal model [5–7]. See works by Freedman [8] and Sekhon [9] for comprehensive reviews and the history of the Neyman-Holland-Rubin causal model. The model conceptualizes causal inference in terms of potential outcomes under treatment and control, only one of which is observed. Let $Y_{i,a}$ denote the potential outcome for individual i with treatment level a for all treatment levels $a \in \mathcal{A}$. The treatment level a can be continuous, but for simplicity assume there are two levels of treatment, one for control ($Y_{i,0}$) and one for treatment ($Y_{i,1}$). Causal inference can be viewed as a missing data problem in this framework because it is possible to observe only one of the potential outcomes for each individual.

In the sections below, it will be useful to describe the data in terms of a full data distribution, in which all potential outcomes are realized, and an observed data distribution, in which only one of the outcomes is realized. It is necessary to first define parameters of

interest on the full data, and then estimate them using the observed data. Identifiability assumptions tie the estimator based on the observed data ($\hat{\psi}$) to the true parameter based on the full data (ψ).

Let $X = (W, Y_a : a \in \mathcal{A}) \sim P_0$ be the full data distribution, which includes all covariates (W) and treatment specific outcomes (Y_a) for all levels of treatment $a \in \mathcal{A}$. For simplicity of exposition, consider a binary treatment (control versus treated), so $a \in (0, 1)$. Let $O = (W, A, Y) \sim P$ be the observed data, where A is a treatment indicator equal to 1 when individual i is in the treated group and 0 when individual i is in the control group. A is analogous to a censoring indicator for Y, where $Y_i = A_i Y_{i,1} + (1 - A_i)Y_{i,0}$.

2.3.2 Estimating treatment effects

All of the treatment effects that I consider in this dissertation are mean differences (equivalent to risk differences for binary outcomes). Other parameters of interest can include ratios (e.g., risk ratios, odds ratios), but for simplicity and relevance the material below focuses on differences. The causal difference between treatment and control for individual i defined on the full data X is

$$\psi_i = Y_{i,1} - Y_{i,0} \tag{2.1}$$

This fundamental missing data problem cannot be solved at the level of an individual, and so the problem is typically reformulated at the population level. At the population level it is possible to estimate mean differences between treatment and control groups.

Randomized treatment

If the treatment assignment is randomized, then A is independent of the full data $\{A \perp\!\!\!\perp X\}$. Thus for $j = 0,1$, A is independent of the potential outcomes $\{A \perp\!\!\!\perp Y_j\}$:

$$E[Y_{i,a}|A_i = a] = E[Y_{i,a}] \tag{2.2}$$

Given these results, it is possible to derive an consistent estimate of the population level causal difference parameter based on the observed data O:

$$\hat{\psi} = E[Y_{i,1} - Y_{i,0}] = E[Y_{i,1}] - E[Y_{i,0}] = E[Y_i \mid A_i = 1] - E[Y_i \mid A_i = 0] \tag{2.3}$$

With randomized treatment, $\hat{\psi}$ is an unbiased estimator of ψ under the following assumptions:

1. The observed values Y_i are a realization of the potential outcomes Y_a [1]

[1] This is the Consistency Assumption that ties the observed data O to the full data X.

2. The treatment status of any unit is independent of potential outcomes for all other units, and treatment is the same for all units [2]

Non-Randomized treatment

In observational studies treatment assignment A is not randomized. When A is not randomized, estimating the causal difference ψ is more difficult because covariates, W, that are potentially related to the outcome are almost always unbalanced between treatment and control groups; thus, the association of A and Y is confounded (Figure 2.1) [11].

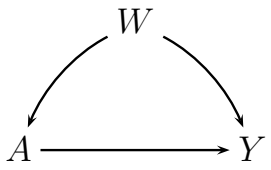


Figure 2.1: A directed acyclic graph of the most simple form of confounding between A and Y by W.

Unbiased estimation of ψ with non-randomized treatment requires two additional assumptions:

3. W must precede both the treatment A and the outcome Y (i.e., temporal ordering must be explicit)

4. All confounders W of the relationship between A and Y are measured[3]

All four assumptions are empirically unverifiable from the data. Using all four assumptions, observed treatment A is independent of the potential outcomes $(Y_a : a \in \mathcal{A})$ conditional on W: $\{A \perp\!\!\!\perp Y_a \mid W\}$, and it is possible to recover a consistent population level estimate of ψ:

$$\psi^{ATE} = E_W\{E[Y_i|A_i = 1, W_i] - E[Y_i|A_i = 0, W_i]\} \tag{2.4}$$

ψ^{ATE} is the treatment effect averaged over strata defined by W. (By assumption 4, A is effectively randomized within each strata of W). The ATE superscript is short for Average Treatment Effect, and will be used to differentiate it from an alternate estimator motivated below.

Further problems arise if individuals in the treatment and control groups are drawn from different populations and do not share common values for W. This scenario is

[2]This is the Stable Unit Treatment Value Assumption (SUTVA) [6, 10]

[3]This is often called the No Unmeasured Confounders Assumption (epidemiology), Ignorability Assumption (statistics), or the Selection on Observables Assumption (economics).

likely in studies of pre-existing interventions when the intervention is not randomized and instead targeted to a specific subpopulation. The essential problem is that $\hat{\psi}^{ATE}$ is only well-defined if there is good overlap (support) in W between the treatment and control groups. Another way to describe it is that the probability of receiving treatment is not close to 0 or 1 for all levels of W:

$$0 < P(A = 1 \mid W) < 1 \tag{2.5}$$

One approach to help improve overlap in W is to estimate a different parameter of interest. Until this point, I have discussed estimating the mean effect in the entire population (the Average Treatment Effect, or ATE). Consider a second parameter of interest: the average treatment effect among the treated (ATT):

$$\psi^{ATT} = E[Y_{i,1} - Y_{i,0} \mid A = 1] \tag{2.6}$$

Similar to the ATE in non-randomized designs, it is possible to estimate the ATT as a function of the observed data:

$$\psi^{ATT} = E_{W^*}\{E[Y_i|A_i = 1, W_i] - E[Y_i|A_i = 0, W_i] \mid A_i = 1\} \tag{2.7}$$

where the outer expectation is taken over $W^* = W \mid A = 1$ (i.e., the values of W in the treatment group). All of the previous assumptions are required for consistency, but Heckman *et al.* show that equation (2.5) only needs to hold for levels of W^* [12]. In other words, for the ATT estimator there does not need to be overlap between treatment and control groups in the entire distribution of W, but only the distribution of W among the treated group (W^*).

2.3.3 Exact matching in community intervention studies

Matching in epidemiologic studies fundamentally involves biased sampling from the study population [13]. The goal of matching in the design is to ensure covariate overlap between groups for covariates used to match and, in some cases, to improve statistical efficiency. In community level intervention studies, exact matching and its variants are used to help guarantee that one or more baseline covariates is balanced in the intervention and control groups [2]. In practice, communities are most commonly matched on size, geography or baseline measures of the primary outcome [3].

Community intervention studies typically have a small number of independent units (communities). The small number of independent units makes it more likely that there will be imbalance between treatment and control groups in potentially confounding covariates. In randomized designs there are essentially two approaches to help mitigate this problem: stratified randomization and exact matching [2]. Stratified randomization involves stratifying communities based on one or more covariates, and then randomizing

treatment within each strata. Exact matching involves assembling pairs of communities with identical values for one or more covariates, and then randomizing treatment within each pair.

In non-randomized community interventions, matching is the only design tool available to help improve covariate balance between intervention and control groups. An alternate way to frame the issue in non-randomized matched cohorts is to view matching as a strategy to help ensure that treatment and control communities come from the same population (based on matching covariates, at least). For example, if a water supply improvement intervention is targeted to the subgroup of communities that rely only on surface water sources, control communities need to be chosen from that same subgroup of communities who rely only on surface water.

The most simple type of matching in community intervention studies is exact matching using a single categorical covariate. Let W be a categorical covariate that is used to match a total of $2N$ communities into N pairs. In the randomized case, N community pairs are assembled such that within each pair the two communities have identical values for W. Then, treatment is randomized within each pair, so that one community receives treatment and the other serves as a control. A slight variant on this matching approach is when W is continuous. One approach for a continuous W is to discretize it into categories and then match communities as if it were a categorical variable. A second approach is to choose a caliper, α, such that communities are only matched if $|W_i - W_j| \leq \alpha$, for communities $i \neq j$.

In non-randomized designs, matching is slightly different because the treatment group – and its covariate distribution – is already defined. Matching involves identifying a set of potential controls ($\geq N$) and finding a match for each of the N treatment communities from the control set based on their covariate values, W. Several technical issues arise in this approach. First, as a practical matter, control communities are likely matched without replacement in community level intervention studies (i.e., a control community can only be paired to a single treatment community).[4] Second, matches may not be unique (multiple controls can match to a treatment community, and vice versa). A simple matching algorithm in this scenario, assuming one categorical W, is:

1. Randomly order the treatment communities from $1, \ldots, N$
2. For $i = 1, \ldots, N$, find a match in the control set based on W_i
3. If multiple controls can be matched to each treatment community, iterate over steps 1 and 2 until no more matches are identified

[4] In theory [14], controls can be matched with replacement (i.e., a single control could be matched to multiple treatment units and then upweighted in the analysis), but to my knowledge this has never been done in the context of matched community intervention studies [2, 3].

This simple matching algorithm increases in complexity if there are multiple covariates and if the covariates are continuous. Among other capabilities, the `Matching` package in R implements matching on multiple covariates [15].

An important limitation of matching is that it may be impossible to find matches for some communities – particularly if there is a small number of communities (in the case of randomized designs) or a small ratio of potential control communities to treatment communities (in the case of non-randomized designs). If unmatchable communities exist, investigators must decide whether to exclude them: if they do, then they change the population to which they can make inference.

This problem becomes intractable if investigators want to match on more than one or two covariates. In non-randomized interventions, this may be desirable if there are a large number of potentially confounding variables that influence whether a community receives the intervention. Matching on one or two covariates may be inadequate to ensure sufficient overlap across the entire set of potential confounders (i.e., the condition in equation 2.5 may not hold). Exact matching with a large set of covariates generally breaks down because in finite samples we run out of data: it becomes impossible to find a match for some (or many) communities.[5] The second didactic simulation (Section 2.7.2) illustrates that even for well-distributed, binary matching variables it becomes impossible to match all treated communities to control communities using more than three matching variables with sample sizes typical of community interventions. The next section describes propensity score matching, which is an extension to exact matching methods that accommodates a large number of potential confounders.

2.3.4 Propensity score matching

Propensity score matching is a common approach that empirical researchers use to match treatment and control units using high dimensional data [14, 16, 17]. The broad goal is to identify control units that are highly similar to each treatment unit based on a set of covariates W. If W includes just one or two categorical covariates, then exact matching is possible (see previous section). When higher dimensional matching is required, the propensity score approach simplifies the problem by collapsing a large set of covariates W into a single scalar – the propensity score – and then matching treatment and control units using a one-dimensional match on the propensity score.

Let $g(1|W) = P(A = 1|W)$ be the propensity score: it is the probability of receiving treatment ($A = 1$) given a set of covariates (W). It attempts to capture the treatment mechanism in probabilistic terms, which is known for randomized trials but typically unknown in observational studies [18]. Given $0 < P(A = 1|W) < 1$, Rosenbaum and

[5] Typically, an analysis can only split the data a few times before there are empty cells in one of the groups. This is sometimes referred to as the "curse of dimensionality."

Rubin [14] prove that:

$$\psi^{ATT} = E_{g(1|W)}\{E[Y_i|A_i = 1, g(1|W_i)] - E[Y_i|A_i = 0, g(1|W_i)] \mid A_i = 1\} \tag{2.8}$$

where the outer expectation is taken over the distribution of $g(1|W) \mid A = 1$. Notice that the only difference between equation (2.6) and equation (2.8) is that the vector of covariates W_i has been replaced by a single scalar $g(1|W_i)$.

The propensity score is unknown for observational studies and must be estimated, typically using logistic regression with maximum liklihood: $g(1|W) = 1/[1 + \exp(-\beta W)]$. The logistic regression returns predicted probabilities, $\hat{g}(1|W)$, but treatment and control units are typically matched using the linear predictor from the model, $logodds = \beta W$, because its range is not compressed near 0 and 1 and it tends to be more normally distributed [9].

There are numerous ways to match treatment and control units after estimating the propensity score. The most common include nearest-neighbor matching [14], optimal matching [19] and Mahalanobis distance matching [16]. For all of the analyses in this dissertation I have used nearest neighbor matching because (i) it is conceptually simple and (ii) in the village-level studies of pre-existing interventions the different matching methods tend to not differ in the matches chosen because there are very few units to match (on the order of 10 to 20). A nearest neighbor match selects the closest control unit to each treatment unit based on the linear predictor of the propensity score.

After creating the matched set of treatment and control units, it is possible estimate the ATT by simply averaging across the differences between matched pairs (equation 2.8). In the quasi-experimental design that I propose below, I use propensity score matching in the design stage to select villages into the study (attempting to "mimic" a randomized trail [18]). This selection in the design stage implies the ATT estimator (equation 2.8). The primary goal of matching in this design is to improve the likelihood that intervention and control groups will have good overlap in a large number of potentially confounding variables. It is likely that the match will not remove all counfounding: residual confounding could still exist because matching will likely be imperfect (assumption 5, above, will not hold), so additional adjustment using pre-treatment variables measured at follow-up may be necessary. I discuss this issue in more detail in sections 2.5 and 2.6.

2.4 Necessary conditions for studies of non-randomized, pre-existing community interventions

Many community level interventions that are planned outside of the scientific process have characteristics that make them difficult or impossible to evaluate rigorously. Before embarking on an evaluation of a non-randomized, pre-existing intervention, investigators should confirm that the intervention meets basic conditions that will enable a valid study.

Table 2.1 outlines six conditions that are necessary for a valid study of a pre-existing intervention under reasonable assumptions. I describe them in more detail below.

1. A Partnership with the Implementing Organization

A successful study of a pre-existing intervention necessarily relies on extensive input from the implementing organization. This condition is not limited to studies of pre-existing interventions: it applies to intervention studies where the intervention is not carried out by the scientific team. Details about the dates of intervention, activities performed, target population, geographic location, participation, and goals of the intervention are all important inputs to the study. Identifying a willing implementation partner is a necessary step in obtaining this critical information.

In theory, implementing organizations may have little incentive to cooperate with a formal scientific study because of concerns that negative study results could jeopardize their future funding. In practice, I have found that most implementing organizations enthusiastically support rigorous, independent studies because of their potential to elevate the status of their work, and because they fundamentally believe in evidence-based practice. The incentive structure may further favor scientific studies as funding agencies and organizations increasingly require independent evaluations of activities that they support.

2. Sufficient Intervention Scale

Interventions that are only deployed to a small number of communities (e.g., fewer than 10) will be difficult to evaluate due to the small number of independent units for analysis [2]. A large number of intervention communities is favorable both in the design and analysis of the study because it guarantees a sufficient number of independent units to have adequate statistical power.

3. Availability of Control Communities

A rigorous study needs to include control communities that have not received the intervention. There are some exceptions, but most parameters of interest (Section 2.6) require a counterfactual, which investigators estimate using a control group. As I will detail below, a larger set of potential control communities increases the likelihood of identifying a control population that is highly similar to the intervention population.

4. Independence of Communities

Standard definitions of treatment effects in a causal inference framework require the assumption that the treatment status of any unit (in this case, the community) is independent of potential outcomes for all other units [6, 20].[6] Statistical analyses of community-

[6]This is the Stable Unit Treatment Value Assumption (SUTVA).

level interventions also rely on the assumption of independence between the units of analysis [2, 3]. Valid designs need to address these assumptions by including sufficient numbers of independent communities.

5. Uniformity of the Intervention

Identifying treatment effects typically requires a precise definition of the intervention treatment. If the implementing organization varies the treatment in every community, then identifying a common parameter of interest for the intervention population becomes difficult. Investigators should be able to define and measure the treatment, and the process is simplified if the treatment is homogeneous across communities. The assumption of a uniform treatment across communities is most likely a simplifying assumption, but it is a conservative assumption with respect to inference because it combines the within-community variability with the between-community variability. In this chapter I do not address deviations from this condition because it is necessary to have sufficient numbers of repeated observations within each treatment type to have valid statistical inference.

6. Availability of Baseline (Pre-Intervention) Data

Baseline (pre-intervention) data are necessary for identifying an appropriate control group, and demonstrating that the intervention and control groups are comparable at baseline. There are two fundamental temporality criteria that the data should meet: (i) the data should be collected before the intervention was deployed (otherwise they reflect post-treatment conditions), and (ii) the data should approximate the conditions at the time that the implementing organization chose intervention communities. In environments or populations that are highly dynamic, the data need to be collected close to the actual intervention initiation. In addition, to be useful the dataset should encompass all intervention communities and a large number of potential control communities. It should include variables that relate to intervention community selection and potential confounding variables. If baseline data are available on the outcome(s) of interest, then a difference-in-differences (DID) parameter can be estimated, which is favorable to a post-only comparison (Section 2.6).

Even when a study meets all six conditions, estimating unbiased treatment effects still requires strong assumptions. Studies of non-randomized, pre-existing interventions that do not meet all six conditions require even stronger – and less realistic – assumptions. I view the first condition (a good partnership) as necessary for any study of a pre-existing intervention; without it, the study lacks essential contextual and scientific information. These six conditions will surface repeatedly in the sections below.

Table 2.1: Necessary conditions for studies of non-randomized, pre-existing community interventions

Condition	Main Rationale
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.
2. Sufficient intervention scale	Each community is a single unit of analysis and adequate numbers are needed for valid statistical analyses.
3. Availability of control communities	Control communities are necessary to provide a counterfactual comparison group.
4. Independence of communities	Theoretical and statistical constructs require that each community is independent with respect to the intervention and the outcome of interest.
5. Uniformity of the intervention across communities	A uniform intervention is necessary to define and estimate a common treatment effect across communities.
6. Availability of baseline (pre-intervention) data	Baseline data allow investigators to establish baseline comparability between intervention and control communities. Baseline data also provides information for informative sampling.

2.5 A quasi-experimental design

2.5.1 Overview and main steps

Outside of scientific studies interventions are rarely deployed at random, and communities that receive interventions are likely different, on average, from communities that do not. Community interventions in developing countries typically target the populations most in need (e.g., the most poor or the worst health). For example, Community Led Total Sanitation (CLTS) campaigns typically focus on villages or neighborhoods where latrine ownership is rare and open defecation is frequent [21]. The careful selection of treatment and control communities can help ensure that the two groups are similar based on baseline (pre-treatment) characteristics. Without baseline comparability, pre-treatment differences between treatment and control groups could lead to differences in the outcome of interest, independent of the intervention. This confounding results in biased estimates of treatment effects. Although easy to state in principle, operationalizing the selection process into a reproducible series of steps poses difficult challenges. In the context defining a study population, objective, repeatable community selection is required to accurately define the quantities measured in the study and the parameters that they can estimate.

If an intervention program meets the six conditions in Table 2.1, then there is potential to identify intervention impacts using an observational, quasi-experimental approach. Here, I propose a design that attempts to approximate the conditions of a randomized experiment by selecting control communities that are as similar as possible to intervention communities based on observable characteristics. It does this by selecting control communities using a propensity score match based on secondary data collected just before intervention implementation. (I define secondary data as data collected by organizations other than the research team, such as a national census.) I assume that post-intervention outcomes are not available for all of the communities identified in the baseline data, and that outcomes can only be measured in a subsample. Matching in the design stage helps improve the comparability of intervention and control communities that are selected into the study for outcome measurement. Ideally, the matching would remove all differences between intervention and control communities. Treatment effects could then be estimated using simple differences in means. In practice, this is unlikely because secondary data will be incomplete and will have more measurement error than a rigorous scientific study. The matching will, however, ensure that intervention and control communities have good overlap on a large set of potentially confounding characteristics. This leverages the study design to remove a large portion of the bias so that estimated effects rely less on statistical analyses and the assumptions they require [18, 22].

Table 2.2 summarizes the main steps in the design used to evaluate pre-existing interventions. Assuming that relevant, secondary data were collected close the beginning of the intervention and that the intervention is complete (steps 1 – 4 in Table 2.2), the research study begins by contacting the implementing organization and (perhaps sepa-

rately) obtaining the pre-intervention secondary dataset that includes information on both the intervention communities and a large set of potential control communities.

The next step is to identify the selection criteria that the implementing organization used to choose intervention communities. This criteria, combined with baseline secondary data, will help inform the model selection approach used to match intervention communities to control communities (more detail below). After matching and community selection, a field team collects post-intervention data on key pre-treatment characteristics (to re-assess balance) and outcomes.

Table 2.2: Summary of key activities in a quasi-experimental design to evaluate a non-randomized, pre-existing intervention, sorted by temporal order (steps). Since the intervention is pre-existing, the research team only participates in steps 5, 6 and 7.

Step	Intervention Activities	Evaluation Activities
1		Pre-intervention secondary data collected on a large set of communities (by implementing organization, national census, or other source)
2	Intervention communities selected	
3	Intervention begins	
4 *	Intervention ends	
5 *		Study is conceived. Investigators contact the implementing organization to establish a relationship and collect key information about the intervention. Investigators obtain secondary data (collected in step 1).
6		Intervention and matched control communities selected based on pre-intervention, secondary data
7		Post-intervention data collection in selected communities (cross-sectional, or prospective)

* Study could be conceived and begin before the intervention ends.

2.5.2 Matching to select a study population

Data requirements

Reconstructing the treatment mechanism and estimating the propensity score requires community-level data and information from the implementing organization. The data need to satisfy three criteria. First, the data should cover the set of communities that received the intervention and a set of communities from which investigators can select comparable controls. Good matches are more likely if there are many more potential control than treatment communities [23]. Second, the data need to include information on key characteristics that influence the outcome of interest and could confound the relationship between the intervention and the outcome. Finally, the data should reflect conditions at the time the intervention started (at baseline) – that is, they should be collected close to the start of the intervention and not after it was initiated. If the data were collected after the intervention commenced, then they reflect post-treatment conditions and may not represent conditions at the time that implementing organizations selected the communities.

Rosenbaum and Silber use the ethnographic term "thick description" to describe the process of obtaining a rich, detailed narrative of treated individuals (or communities) in the context of propensity score matching [24]. In studies of pre-existing interventions, scientific investigators partner with the implementing organization to document in detail the decision process that lead to community selection. This process helps define both the set of potential controls and the covariates needed to match controls to treatment communities. In many cases, implementing organizations use census data to target their efforts. If available to investigators, these data become an essential resource for community selection.

The most common secondary data sources for community-level baseline data are national census surveys, which collect data at the household level and provide data with identifying information aggregated to the level of community (villages in rural areas, neighborhoods in urban areas). If the intervention implementing organization or another party collected household-level baseline data prior to their intervention, then investigators should evaluate carefully whether the data collection effort complied with proper consent procedures.

Model selection approaches

The propensity score model is typically estimated using logistic regression $g(1|W) = 1/[1 + \exp(-\beta W)]$, where $g(1|W)$ is the propensity score and W is a set of covariates. Even while the implementing organization helps identify a minimal set of variables to include in W, the linear combination of W is usually unknown and must be chosen. In addition, more comparable control communities may be selected by including additional variables not identified by the implementing organization. For techniques that use a

propensity score in either matching or inverse weighting, bias reduction is only achieved if W includes covariates that relate to both the treatment and the outcome. Simulation studies have demonstrated that including covariates that only predict treatment and not the outcome will not reduce bias and increase the variability of any resulting estimators [25, 26].

The model form can be selected iteratively using specifications chosen by investigators using a systematic process [27]. The basic approach is to specify a treatment model, calculate balance statistics (next section) and modify the treatment model until the best balance is achived over potentially confounding covariates. Alternatives to this often tedious – and potentially inconsistent [9] – approach use machine learning to help exhaust the possibilities. For example, a super learner algorithm can be used to optimally predict treatment given a set of covariates using cross-validation and the L2 loss function [28]. Another (and in my opinion, more promising) example is the `GenMatch` algorithm implemented in the Matching package in R, which uses genetic optimization to search the parameter space for covariate weights that minimize a balance-based loss function – it is technically an alternative to propensity score matching, but can include the propensity score as one of many covariates [9, 15]. The advantage of `GenMatch` is that its loss function is usually some metric of covariate balance between groups, which is the ultimate goal of matching in the design.

2.5.3 Evaluating balance of intervention and control groups

After identifying a study population and selecting a sample it is important to assess whether or not treatment and control groups are similar (or balanced) with respect to potentially confounding characteristics. Although no single metric will be equally sensitive to all departures of balance, two metrics that have been proposed in the matching literature include the standardized difference in means and the variance ratio [18]. Sekhon [29] further suggests using metrics based on cumulative distribution functions of the covariates, such as the bootstrapped Kolmogorov-Smirnov test [30].

The standardized difference is calculated as:

$$\text{Standardized Difference} = (\mu_T - \mu_C) \ \div \ [(S_T^2 + S_C^2)/2]^{1/2} \ \times \ 100 \tag{2.9}$$

where μ_T and μ_C are the sample means of the treatment and control groups, respectively, and S_T^2 and S_C^2 are their sample variances. The variance ratio is calculated as S_T^2/S_C^2. Rubin suggests that standardized differences greater than 100 (1 standard deviation difference in the means) are too large to control for using linear regression unless the analyst specifies the model perfectly [18]. Rubin also notes that even if differences in the means are small, linear modeling adjustments are very sensitive to non-linearities in the relationship between the covariate and the outcome when the variance ratio of the covariate approaches 1/2 or 2. As a working measure, Rubin denotes variance ratios "of concern" if they are further from 1 than 4/5 or 5/4.

Using balance statistics is essential to evaluate the performance of matching during the community selection process. However, the balance metrics can also be useful when applied to data collected during the research study (post-intervention). If investigators can reasonably assume that the covariates could not be influenced by the intervention, then these comparisons can provide an additional robustness check for the similarities between intervention and control communities. For example, national census data will rarely include information about household involvement in community organizations, cultural perceptions and other measurable characteristics that may correlate strongly with unmeasurable confounders. If an intervention provided in-home toilets to households combined with a motivational behavior change message, then it would be reasonable to assume that the intervention would not change a household's participation in community activities. Yet, community participation may be an important marker of observable or unobservable characteristics that could confound the relationship between the in-home toilets and child diarrhea, and it would be important to check balance on that characteristic – even if measured post-intervention.

In addition to numeric summaries to asses balance graphical measures are a valuable tool – particularly for evaluating multivariate balance. Marginal balance across all individual confounders is necessary but not sufficient for joint balance between treated and control groups across the confounders. Graphical techniques can help somewhat in this regard. Rosenbaum and Rubin [27] evaluate balance on key binary covariates stratified by quintile of the estimated propensity score. They evaluate multivariate balance and support on the estimated propensity score by plotting box plots of the propensity score stratified by treatment group. Ho *et al.* propose comparing continuous covariates between groups using QQ plots for individual covariates, and provide other useful guidance for comparing balance [22]. Figure 2.2 demonstrates examples of other distributional plots of the estimated propensity score. The distributional plots of fitted propensity scores include at least three useful pieces of information. First, the analyst can evaluate whether the primary mass of the distributions overlap (an indication of overall similarity of the distributions). Second, the analyst can identify if there are both intervention and control observations at all levels of the covariates in the intervention group. For example, if the propensity score distribution of the intervention group is not completely overlapped by the propensity score distribution in the control group, then there are intervention communities with some combination of the covariates in the propensity score that are unique to that group. Under this condition, the ATT could be biased (equation 2.6). Finally, if the propensity score distributions approach 0 or 1 in either group, this also suggests that there are combinations of covariates perfectly (or nearly perfectly) predict treatment status (a violation of equation 2.5). It is possible to identify communities with extreme propensity scores, and determine whether further restrictions to the study population are necessary to validly estimate a parameter of interest.

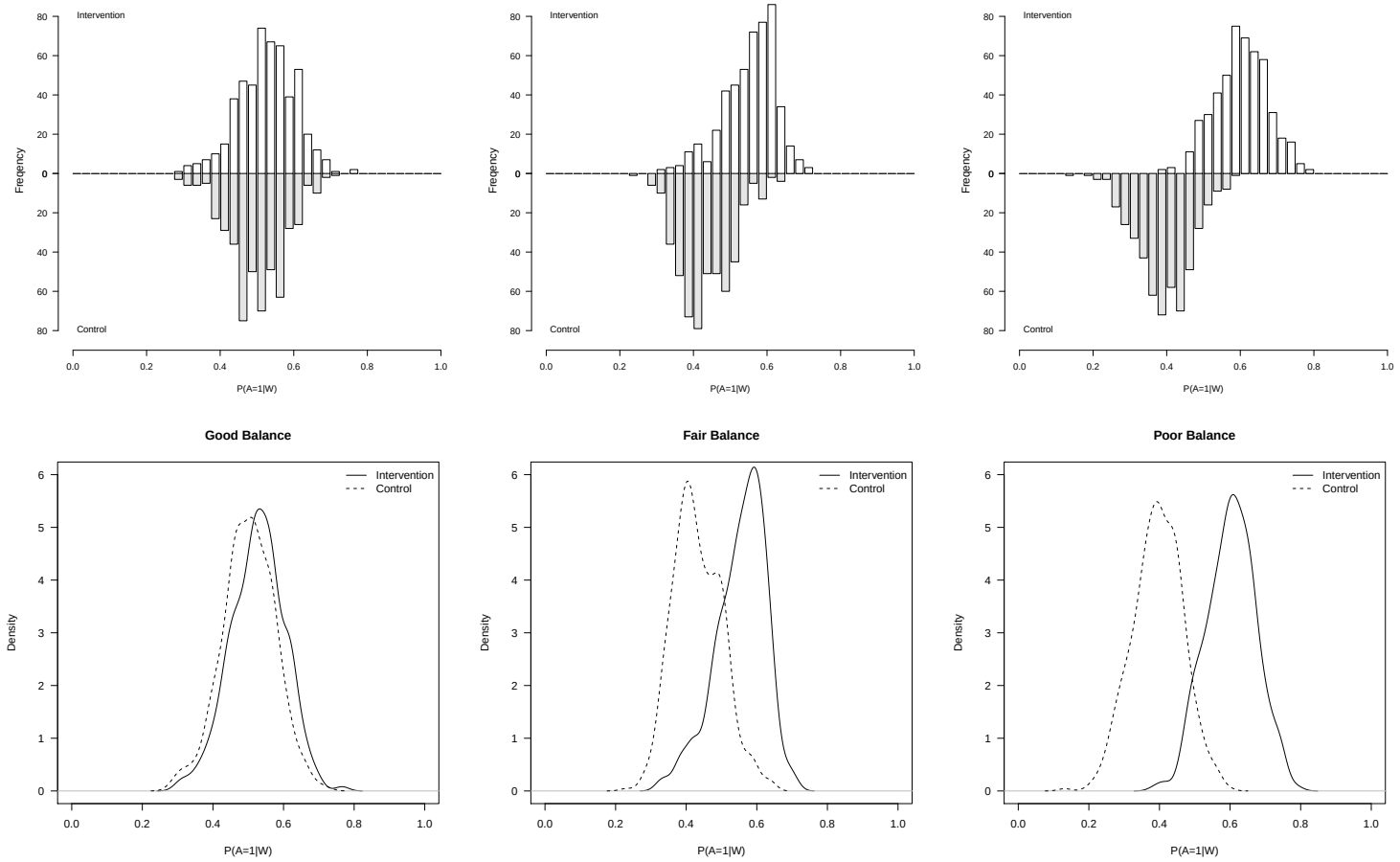


Figure 2.2: Graphical examples of balance on the estimated propensity score. The top panel includes histograms super-imposed by treatment group. The bottom panel includes kernel density smoother plots of the propensity score distributions by treatment group. Data in the left plots are from a randomized trial [31] and demonstrate good balance. Data in the center plots are from a quasi-experimental sample (Chapter 4) and demonstrate fair balance. Simulated data in the right plots demonstrate poor balance. N=1,000 (all plots).

2.5.4 Threats to validity

Unmeasured confounding

As described in detail in Section 2.3.2, if there are common causes of both the intervention treatment and the outcome that are unmeasured at baseline, then treatment effects will be biased. If unmeasured confounders are not included in the baseline secondary data, but can be measured in the post-intervention period by investigators, then adjusted estimates of the ATT are possible by conducting further statistical adjustment in the matched sample (see also Section 2.6). This is a central problem of non-randomized studies, and investigators should evaluate the plausibility of this assumption. I the discussion I outline some data-driven approaches to evaluating this assumption in observational data.

Informative censoring

In the time that elapses between the baseline measurement used to define the study population and the post-intervention measurement to assess outcomes, individuals or households within communities will exit the study population (commonly referred to as censoring). If censoring is independent of either the intervention treatment or the outcome then it is non-informative and will not bias the treatment effect. However, if censoring is a common effect of both the intervention treatment and the outcome of interest (even indirectly), then it is informative and will cause bias in an unknown direction [32]. Informative censoring is a potential source of bias in all study designs. In prospective designs, baseline characteristics of individuals who exit the study population are available to assess whether censoring is informative. However, in studies of pre-existing interventions the censored individuals are never measured and so investigators have no information about the magnitude of censoring or characteristics of those censored. Investigators must carefully evaluate the conditions for non-informative censoring to see if they may be reasonably met in their study population (see Hernan *et al.* and Hudson *et al.* for detailed discussions [32, 33]).

Measurement error

If outcomes or exposures are measured retrospectively in the post-intervention survey then there is potential they will be measured with more error than if they had been measured contemporaneously. The errors can be either independent of treatment (non-differential) or differential with respect to treatment. For example, participants in a sanitation marketing campaign may be able to recall the year that they built their latrine more accurately than individuals who built latrines in communities that were not part of the campaign – an example of non-differential measurement error. Measurement error will cause bias unless it is independent for exposures and outcomes and is non-differential [34]. Correlated, non-differential measurement error will typically bias treatment effects toward the null, but differential measurement error (correlated or independent) does not

bias treatment effects in a consistent direction. Measurement error can be reduced for some outcomes by limiting the recall period over which outcomes are measured, and by using objective measures rather than those that rely on self-report.

Sampling bias

Sampling bias is possible during community selection or in the selection of units below the community level (such as households) if outcomes are measured at a level below the community. It is common for investigators to be unable to include the entire target population in the study sample. In this case, investigators typically implement some type of random sampling to draw a representative sample and measure outcomes.

Sampling bias is possible if incomplete sampling frames exist. This is a potential problem when using secondary data to define the community sampling frame, and investigators must evaluate the completeness of the secondary data used to select communities. If outcomes must be sampled from within the community, then they should be collected from a random subset.

A more subtle form of community-level sampling bias is possible in this design if not all intervention communities are included in the study. For example, if the original intervention included 50 communities, but investigators only have resources to include half (n=25) of the intervention communities, then the study may not estimate overall treatment effect among the treated (the ATT). If exclusions are based on the propensity score (either through caliper matching [16] or excluding intervention communities off support [12]) then it changes the quantity estimated from the ATT to something that is difficult to define [9]. In this case, investigators know that they are estimating treatment effects for a subgroup of the intervention communities, but which subgroup? If sampling a subset of the intervention communities is necessary, then selecting those that have good overlap with control communities in their covariate distributions ensures greater internal validity (at the cost of external validity) [35]. This is analogous to randomized studies that enroll a non-representative subset of the potentially treated population, and investigators should evaluate which intervention communities are excluded and qualify their inference accordingly.

2.5.5 Practical considerations

Qualitative assessments should complement quantitative balance measures. Rapid-assessment "ground truth" visits can be particularly important when community matching and selection rely on secondary data. National census data quality varies greatly by country and even by region, and its accuracy should be confirmed in at least some of the potential study communities. This is akin to Rosenbaum and Silber's "thick description" exercise, whereby investigators verify how qualitatively comparable the selected communities are. If treatment and control communities differ greatly after the matching exercise, then the matching can be refined and repeated [24].

If the number of intervention communities is large and cannot all be included in the study then a subsample must be chosen, and one approach that can strengthen the internal validity of the study is to select those with the closest match to control communities based on the propensity score (see [35] for a review and an analysis of alternatives). If a small number of intervention communities is excluded based on the propensity score, then it may be possible to characterize them, but this is not guaranteed and the problem becomes difficult with large numbers of exclusions. A practical approach that will work in many applications is to restrict the population of intervention communities before the match to exclude those that are unlikely to have matches in the control set based on observed characteristics. This enables more clear definitions of the population of inference. For example, if some intervention communities are peri-urban and there are no peri-urban control communities (only rural), then investigators can exclude the peri-urban communities and restrict their inference to rural communities only. Population restrictions may be made on other grounds. Section 2.8 includes an empirical example with a pre-match restriction based on size, which was a practical design issue due to the difficulty of estimating community-level diarrhea prevalence without at least 20 children in the community.

2.6 Comments on analysis

2.6.1 Parameters of interest

In Section 2.3.2 I described the average treatment effect among the treated (ATT) and its relevance to matched designs (Simulation 1, below, provides additional detail and motivation). The ATT quantity can be estimated for either the level of the outcome post-intervention (a post-only estimate), or it can be estimated for the change in the outcome. Post-only estimates are the difference in mean outcomes after the intervention conditional on observed characteristics of the intervention communities (equation 2.6). This is appropriate if outcomes are not measured at baseline (e.g., in the secondary data used to match) or if they cannot be measured retrospectively with accuracy. For example, if the outcome of interest is child diarrhea prevalence and it is not measured at baseline, then it will be impossible to measure it retrospectively with accuracy beyond a few days [36].

In contrast, if the outcome is measured contemporaneously at baseline and is included in the secondary data, or if it can be measured retrospectively with accuracy, then it is possible to estimate a difference between intervention and control communities in the *change* in the outcome between baseline and follow-up. This is often called the difference-in-differences (DID) estimator [37]. Let $t = 0$ at baseline and 1 at follow-up (post-intervention). For individual i, and binary treatment $a = \{0, 1\}$, the counterfactual

outcomes in the full data are $Y_{i,a}(t)$. The DID is defined on the full data as:

$$\psi^{ATT,DID} = E\{[Y_{i,1}(1) - Y_{i,1}(0)] - [Y_{i,0}(1) - Y_{i,0}(0)] \mid A = 1\} \tag{2.10}$$

When outcomes are available at baseline and followup for each individual community, observed treatment is indicated by A, and the observed outcome at time t is indicated by $Y_{i,t}$, this parameter can be estimated as:

$$\begin{aligned} \psi^{ATT,DID} \quad = \quad & E_{W^*}\{(E[Y_i(1)|W_i, A_i = 1] - E[Y_i(0)|W_i, A_i = 1]) - \\ & (E[Y_i(1)|W_i, A_i = 0] - E[Y_i(0)|W_i, A_i = 0]) \mid A_i = 1\} \end{aligned} \tag{2.11}$$

where the outer expectation is taken over covariates $W^* = W|A = 1$. The DID parameter assumes that the intervention and control group outcomes would have followed parallel paths over time in the absence of the intervention (i.e., no interaction between treatment and time) [37]. Under this identifying assumption, the DID parameter is attractive because it removes unmeasured confounding that is time-invariant. The assumption of parallel outcome trajectories is more reasonable for groups that are highly comparable at baseline, but, like the identifying assumption of no unmeasured confounding in the post-only estimator, it is impossible to evaluate the assumption empirically.

2.6.2 Post-matching analysis and inference

Given that secondary baseline used to match intervention to control communities are typically incomplete or measured with error, it is likely that intervention and control groups will still differ on some confounding covariates. Additional statistical adjustment is necessary to remove bias if matching in the design did not balance all baseline covariates. This can occur through imperfect matching or if covariates were omitted from the baseline data. If omitted baseline covariates can be measured retrospectively in the post-intervention survey, then further adjustment is possible. For example, individual characteristics such as age, sex and education can be measured in the post-intervention period and used in statistical analysis for confounding reduction and increased precision for effect estimates.

Matching in the design stage imposes no constraints on parametric or semi-parametric statistical analyses conducted after data collection (see [22] for a more detailed discussion). To the extent that it improves the overlap on key confounding covariates, it will reduce the reliance of the study on statistical adjustment and make the findings less sensitive to parametric model miss-specifications.

Since matching is evaluated prior to measuring outcomes, the implications for inference are minimal [22], which would not be the case if they were considered concurrently [38]. Analysis methods such as linear regression or maximum likelihood condition on baseline (pre-treatment) covariates (W) and treatment (A) as fixed and exogenous. Since matching

in the design stage only modifies the sample in ways that are a function of W, then it is reasonable to continue to assume that W are fixed in the analysis [22].

If outcomes are assessed at units that are below the level of the community, such as households, then the usual problem of correlated measurements applies to inference. Huber-White robust ("sandwich") standard errors are viable if there are more than roughly 20 communities [39]. For smaller studies, investigators must rely on specified error distribution models, such as mixed models, which implicitly include correlation between observations within the same community [3, 40].[7]

2.7 Didactic simulation studies

2.7.1 Simulation 1: Valid parameters and estimators in non-randomized, pre-existing interventions

In this section, I use a simple simulation to illustrate parameters of interest that can and cannot be estimated from pre-existing community interventions. I also illustrate why a näive approach to community selection and analysis will fail to estimate a valid parameter of interest in non-randomized, pre-existing interventions without strong assumptions. This simulation, along with the design described in this chapter, assumes that outcomes are not observed at the time of the study design and that investigators measure outcomes in only a sample of intervention and potential control communities. Specifically, I show that matching in the design stage can recover an unbiased estimator of the average treatment effect among the treated (ATT) when the intervention is targeted to a non-random subset of the population. I also show that a study design that includes a random sample of control communities (a näive sample) can estimate the ATT without bias using regression if the model is correct and the estimator correctly constructed, but that the design is less efficient than a matched design because it does not benefit from careful control selection. In addition, I show that a design with näive sampling produces a biased estimate of the ATT in simple scenarios without correct model specification. I also show that even in simple scenarios it is impossible to recover the marginal average treatment effect (ATE) if a non-representative sample of the total population receives treatment because all approaches must rely on extrapolation beyond the limits of the data. The results are not new (see Heckman for a detailed discussion [12]), but they illustrate general lessons about quantities that can and cannot be estimated from targeted, non-randomized interventions. Given that most pre-existing community interventions are not randomized and that outcomes must be measured from a sample of communities, the results immediately apply to their evaluation.

[7] In applied work in this dissertation, under the supervision of my committee, I have used bootstrap resampling based on matched-village pairs to reflect the design (and potentially increase efficiency, since it retains information about which communities are most similar based on the match), but the correct coverage of this approach is still an area of research.

In this simulation, the observed treatment A is binary and the outcome Y is continuous. The counterfactual outcomes under different treatment regimens are denoted by Y_a, where Y_1 is the community-level outcome under treatment and Y_0 is the outcome without treatment. There is a single continuous covariate W. For example, Y_a could be the community mean weight-for-age Z-score for children under treatment $A = a$, and W could be the proportion of households in the community with a latrine. I generate the simulated data according to the following laws:

1. $W \sim U(0, 1)$
2. $Y_a = 0.25 \cdot W + 2 \cdot a - 2 \cdot a \cdot W + \epsilon$

where ϵ is an error term with $\epsilon \sim N(0, 0.25)$. After simulating counterfactual outcomes Y_0 and Y_1, I assign treatment with two scenarios:

1. Random assignment, where $P(A) = 0.5$
2. Non-random (targeted) assignment, where:
   $P(A|W \leq 0.5) = 0.5$ and $P(A|W > 0.5) = 0$

In this simulation, W is a confounder and effect modifier of the relationship between A and Y. Under the targeted assignment scenario there is, by construction, no support for $E[Y_1 - Y_0 \mid W]$ for $W > 0.5$. Figure 2.3 shows simulated data from one iteration of the simulation, and illustrates the problem of no support for the treatment effect conditional on W for values of $W > 0.5$ in the targeted assignment.

The parameter of interest in this simulation is the Average Treatment Effect among the Treated (ATT) (equation 2.7). In the simulation, the true value of the ATT under random assignment was 1.00, and the true value of the ATT under targeted assignment was 1.50. I consider three sampling and analysis scenarios for selecting control communities into the study:

1. A random sample of controls with incorrect regression model specification
2. A random sample of controls with correct regression model specification
3. A matched sample of controls based on a propensity score match with a simple difference in means between groups (no regression model)

All three approaches assume that the treatment effect can be estimated in an unbiased way conditional on observable, measured covariates, i.e. $(A \perp\!\!\!\perp Y_a)|W$, but this assumption is only met in the second and third scenarios.

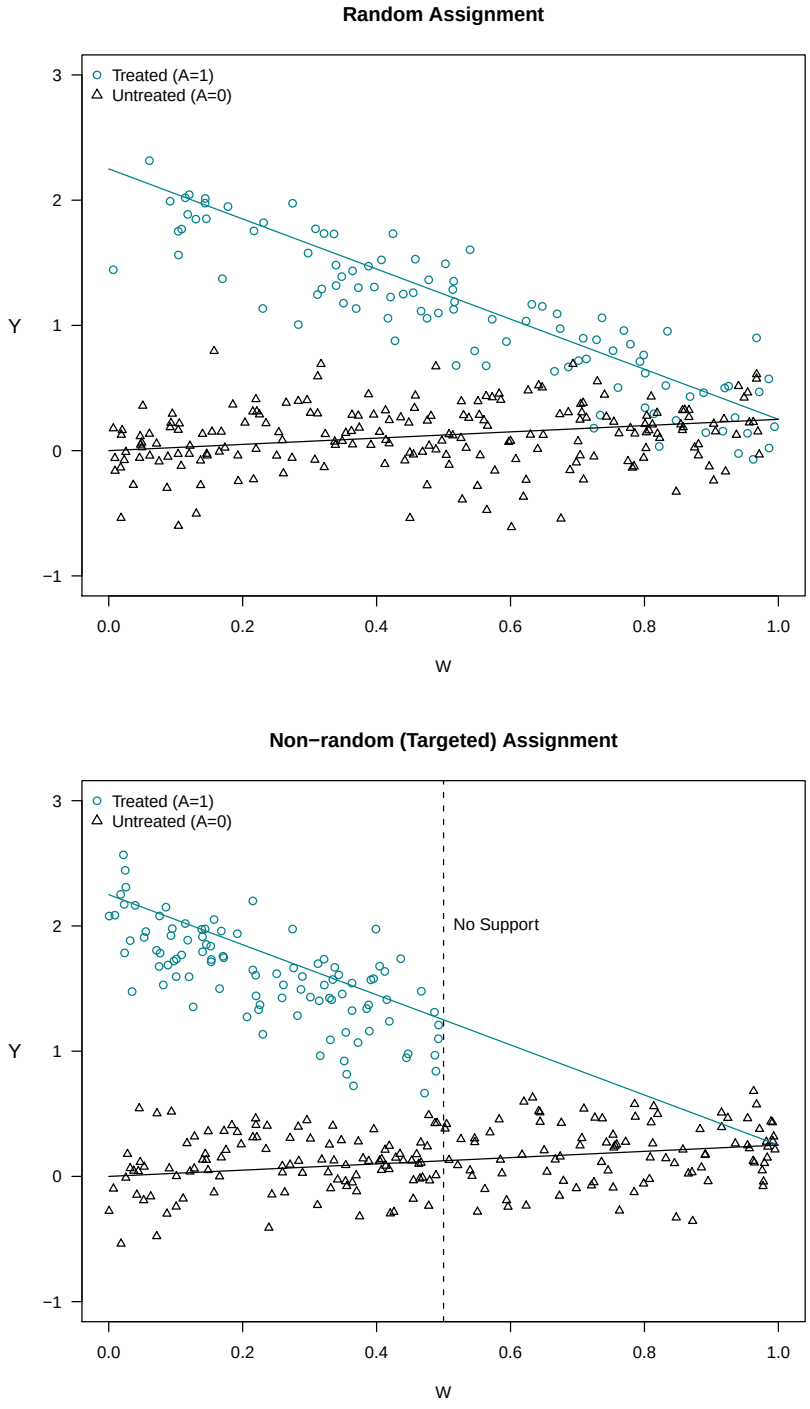


Figure 2.3: An example of simulated data from one iteration in the simulation. The underlying causal relationships are represented by lines. In the Random Assignment, treatment A is assigned independent of W. In the Targeted Assignment, treatment is not random: it depends on W and there are no treated communities for $W > 0.5$.

For the two scenarios with a random sample of controls, I model the expectation of Y using linear regression with a main effects only (incorrect) specification:

$$E(Y|A,W) = m(A,W) = \beta_0 + \beta_1 A + \beta_2 W \tag{2.12}$$

and using a correct specification that includes an interaction term between A and W:

$$E(Y|A,W) = m(A,W) = \beta_0 + \beta_1 A + \beta_2 W + \beta_3 A \cdot W \tag{2.13}$$

After fitting the two models, I impute the predicted the outcome Y for $A = 0$ and $A = 1$ only among treated units. The estimate of the effect is the difference between the two imputed values averaged over all strata in W, which is the G-computation estimator [55].

$$E_W\{E[Y|W, A = 1] - E[Y|W, A = 0] \mid A = 1\} \tag{2.14}$$

For the propensity score matching (PSM) design, I estimate the propensity score using logistic regression:

$$g(1|W) = Pr(A = 1|W) = 1/[1 + \exp(-\beta W)] \tag{2.15}$$

Using the predicted probabilities from the logistic regression model, I match each intervention unit to a single control (1:1) using the nearest neighbor distance without replacement. The PSM approach then estimates the ATT by averaging over the differences between each matched pair (equation 2.8).

In each simulation n communities receive treatment and there are $2n$ potential control communities. A sample of n control communities is included in the simulated study, and their selection is either random or based on PSM, depending on the scenarios above. Finally, I repeated the simulation for different numbers of communities in the treatment group: $n = 15$, 25, 50, 75 and 100.

Table 2.3 summarizes the bias, variance and root mean squared error (RMSE) of the three scenarios from 10,000 Monte Carlo simulations with random treatment assignment. As expected, the näive design that selects controls randomly is biased with the incorrect model specification (main effects only), but unbiased with the correct model specification. The scenario with PSM-based control selection has nearly identical bias and RMSE compared to a random sample with correct model specification for all design sizes.

Table 2.4 has the equivalent information from 10,000 Monte Carlo simulations with targeted, non-random treatment assignment. Again, the random sample with incorrect model specification is severely biased and the other two approaches are minimally biased. Figure 2.4 plots the distribution of estimates for $n = 75$. As expected, the variance of the designs decreases with increasingly large samples, but the PSM design is slightly more efficient for all sample sizes and thus has lower variance and RMSE (Figure 2.5). This is because PSM uses information about W to consistently select the control communities that provide the most useful information.

Finally, consider the scenario of an study of a targeted, non-randomized intervention that samples controls at random from the potential control set (as in scenarios 2 and 3). If the evaluators simply fit a regression model to the data and do not construct the correct G-computation ATT estimator (equation 2.14), which conditions the effect on $A = 1$, then they are attempting to estimate the marginal average treatment effect (ATE) in the population (equation 2.4).

The ATE can be estimated using the standard G-compuation estimator, which averages imputed values for Y with $A = 0$ and $A = 1$ over the entire population without conditioning on being in the treatment group ($A = 1$). However, since there is no support for estimating the treatment effect for $W > 0.5$, without conditioning the estimator on $A = 1$ this approach is biased (even with a correct model specification) and the estimates are not centered on either the ATE or the ATT (Figure 2.6).

This simulation has illustrated the following points. First, if an intervention is deployed to a non-random, targeted subgroup of the population, then it is only reasonable to attempt to estimate treatment effects in the subgroup of the population that shares the characteristics of the intervention group (the ATT, equation 2.7). While this point may appear trivial, the ATT does arise naturally from running a regression on an unmatched sample. The G-computation estimator, which imputes counterfactual outcomes for each individual, can estimate the ATT without bias as long as the regression is correctly specified and the effects are conditioned on the intervention population ($A = 1$) only.

In contrast, the matched design using PSM naturally estimates the ATT without the need to construct more complex estimators. PSM is favorable to a random sample with regression when evaluating pre-existing interventions for two reasons. If interactions exist between intervention treatment assignment (A) and confounding variables (W), then design that randomly samples controls must correctly specify these interactions in a regression model to recover an unbiased effect estimate. The PSM approach is free from this constraint because the matching step allows for arbitrarily complex relationships between A and W. Further, in small samples (e.g., $n = 15$ per treatment group), which are common in community-level intervention studies, the PSM approach will tend to be more efficient than random sampling because it uses baseline information to purposefully select control communities (Table 2.4, Figure 2.5). Heuristically, by purposefully selecting control communities, PSM is less likely to include communities in the sample that are completely different from the treated communities (and thus provide little or no information to the ATT).

Table 2.3: Simulation results for design and analysis scenarios with different numbers of treated units (n) with random treatment assignment. Bias and root mean square error (MSE) are calculated relative to the average treatment effect among the treated (ATT) over 10,000 Monte Carlo simulations.

Scenario	n=15	n=25	n=50	n=75	n=100
Bias					
1 Random sample, incorrect model	−0.4047	−0.3961	−0.4010	−0.4014	−0.4000
2 Random sample, correct model	0.0006	−0.0048	0.0010	−0.0001	0.0004
3 Propensity score match	0.0027	−0.0042	−0.0001	0.0003	−0.0002
Variance					
1 Random sample, incorrect model	0.0170	0.0107	0.0052	0.0034	0.0024
2 Random sample, correct model	0.0319	0.0202	0.0092	0.0053	0.0043
3 Propensity score match	0.0299	0.0201	0.0092	0.0053	0.0043
Root MSE					
1 Random sample, incorrect model	0.4251	0.4093	0.4073	0.4055	0.4030
2 Random sample, correct model	0.1784	0.1422	0.0954	0.0728	0.0652
3 Propensity score match	0.1727	0.1418	0.0950	0.0729	0.0651

Table 2.4: Simulation results for design and analysis scenarios with different numbers of treated units (n) with targeted (non-random) treatment assignment. Bias and root mean square error (MSE) are calculated relative to the average treatment effect among the treated (ATT) over 10,000 Monte Carlo simulations.

Scenario	n=15	n=25	n=50	n=75	n=100
Bias					
1 Random sample, incorrect model	0.0954	0.1041	0.0990	0.0983	0.1000
2 Random sample, correct model	−0.0110	0.0003	−0.0009	−0.0032	0.0002
3 Propensity score match	0.0029	0.0077	0.0039	0.0028	0.0034
Variance					
1 Random sample, incorrect model	0.0170	0.0107	0.0052	0.0034	0.0024
2 Random sample, correct model	0.0180	0.0112	0.0057	0.0035	0.0026
3 Propensity score match	0.0131	0.0085	0.0043	0.0027	0.0020
Root MSE					
1 Random sample, incorrect model	0.1613	0.1466	0.1221	0.1138	0.1114
2 Random sample, correct model	0.1345	0.1056	0.0751	0.0586	0.0505
3 Propensity score match	0.1145	0.0923	0.0655	0.0510	0.0443

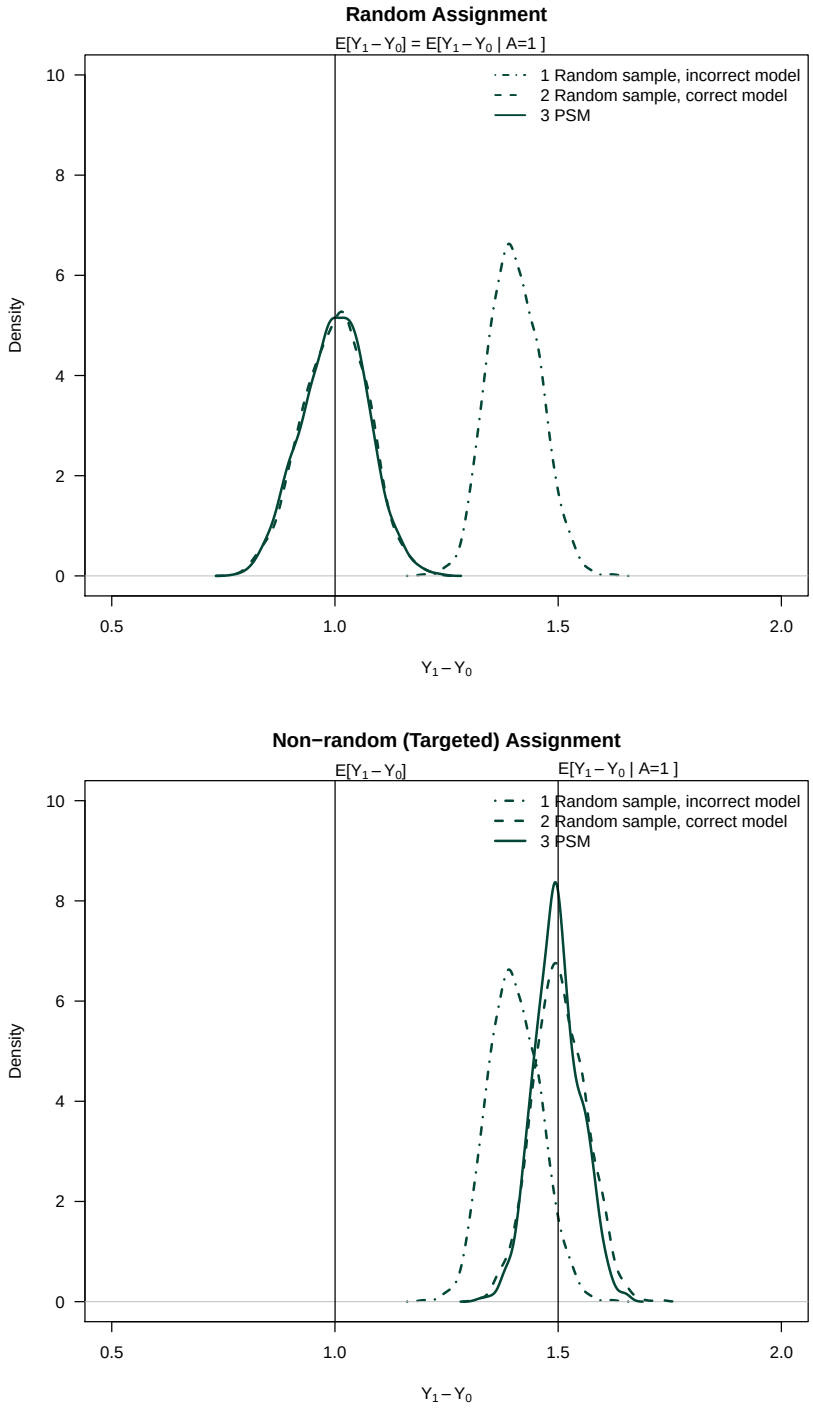


Figure 2.4: Kernel density smoothed distributions of effect estimates from three estimators in 1,000 Monte Carlo simulations with $n = 75$ treatment communities. In the top panel, treatment is assigned randomly, and in the bottom panel treatment is assigned in a targeted way based on a covariate W. Vertical lines mark the average treatment effect (ATE), $E[Y_1 - Y_0]$, and the average treatment effect among the treated (ATT), $E[Y_1 - Y_0|A = 1]$.

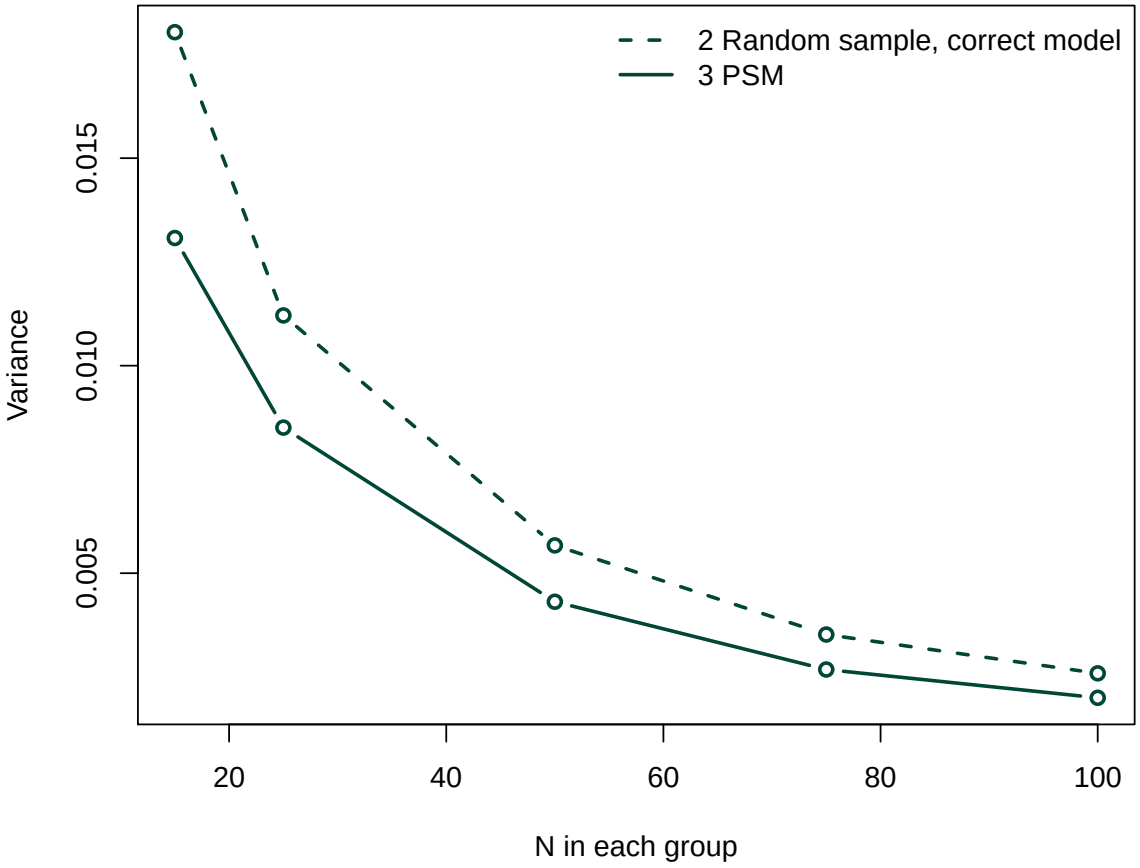


Figure 2.5: Variance of two estimators from the simulation at different sample sizes.

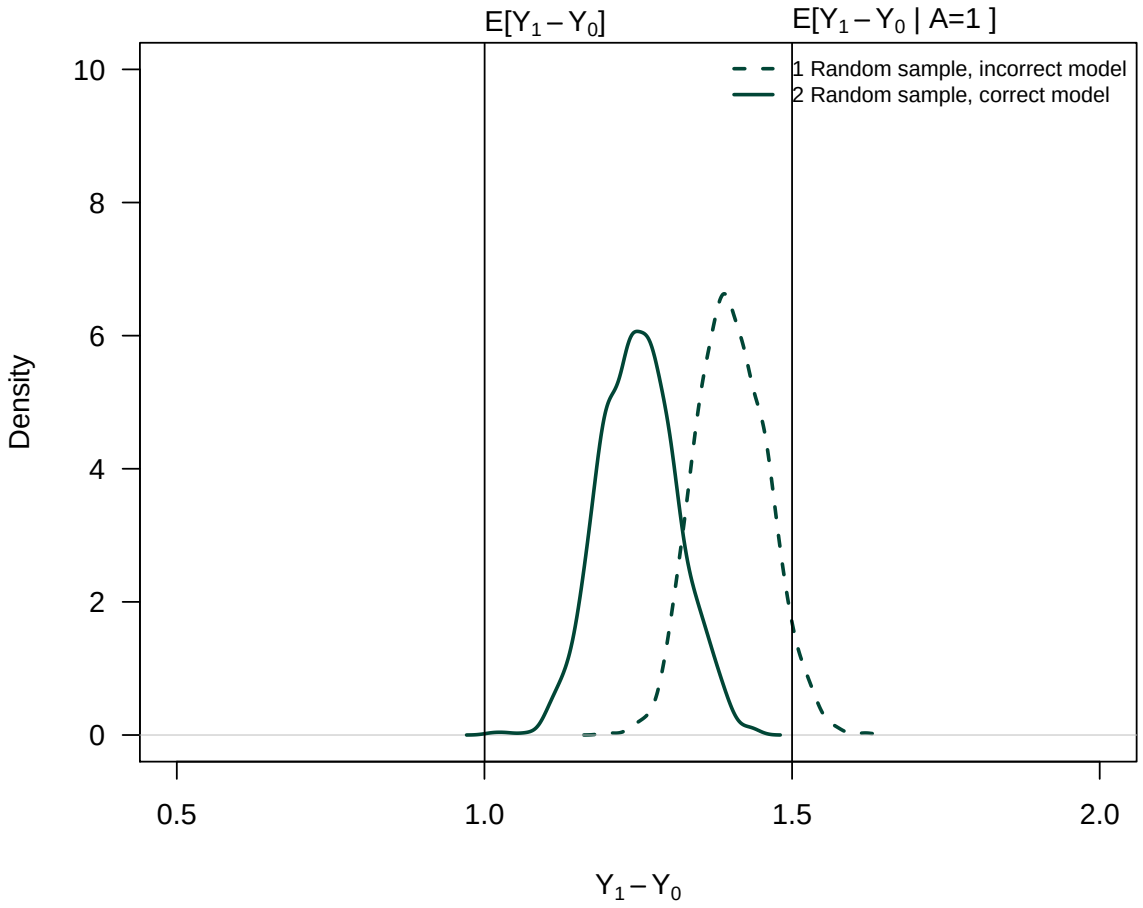


Figure 2.6: Kernel density smoothed distributions of effect estimates from the G-computation estimators of the average treatment effect (ATE) in 1,000 Monte Carlo simulations with targeted (non-random) treatment assignment and $n = 75$. Vertical lines mark the ATE, $E[Y_1 - Y_0]$, and the average treatment effect among the treated (ATT), $E[Y_1 - Y_0|A = 1]$.

2.7.2 Simulation 2: Multivariate matching and the curse of dimensionality

As I described earlier, matching and caliper matching using one or two covariates is a common design feature in cluster randomized trials [2, 3]. In studies of non-randomized, community level interventions it may be useful to match on multiple baseline covariates to help ensure similarity between groups in more than one or two covariates since treated communities may differ from potential control communities in many characteristics.

The problem with exact and caliper matching using a large set of covariates is that in practice there are typically not enough potential control communities to find matches for all treated communities. This often called the curse of dimensionality, and it prevents non-parametric estimation in many social science problems. Here, I demonstrate that even under very optimistic conditions, exact matching fails to match a significant proportion of intervention communities when the number of covariates exceeds three or four using sample sizes typical of community intervention studies.

In this simulation, treatment A is binary. I simulate 10 binary covariates $W = \{W_1, \ldots, W_{10}\}$ that are IID $\sim Binomial(n = 1, p = 0.5)$. I vary two parameters: the number of treated communities ($n = 15, 25, 50$) and the ratio of potential control communities to treated communities (2, 3 and 4). In each simulation, I attempt to match treated communities to control communities using exact matching without replacement on between 1 and 10 covariates from W, and calculate the proportion of treated communities that cannot be matched. I repeat the simulation 1,000 times and calculate the mean proportion of unmatched communities over the 10,000 Monte Carlo runs. All matching was conducted using the `Matching` package in R.

Figure 2.7 summarizes the mean proportion of treated communities that are unmatched using between 1 and 10 binary covariates to match. Across the scenarios I considered, matching on five covariates fails to find a match for between 6 and 32 percent of the treatment communities. Increasing the total sample size increases the number of covariates that can be used to match without losing data (Figure 2.7, top plot). Increasing the ratio of potential control communities to treatment communities (given a fixed number of treatment communities) also increases the number of covariates that can be used to match without losing much data.

This simulation demonstrates that for realistic sample sizes encountered in community-level interventions, exact matching with more than 3 or 4 covariates will result in a substantial fraction of unmatched treatment communities. This result is optimistic: the covariates I have used in this simulation are binary and evenly distributed in the two groups. Finding matches would be more difficult with continuous covariates and binary covariates that have proportions closer to 0 or 1. One remedy to this curse of dimensionality is to simplify the matching problem by collapsing a large set of covariates into a single scalar: the propensity score. The cost of this dimension reduction step is that it is necessary to impose parametric model assumptions on the treatment mechanism. Rosen-

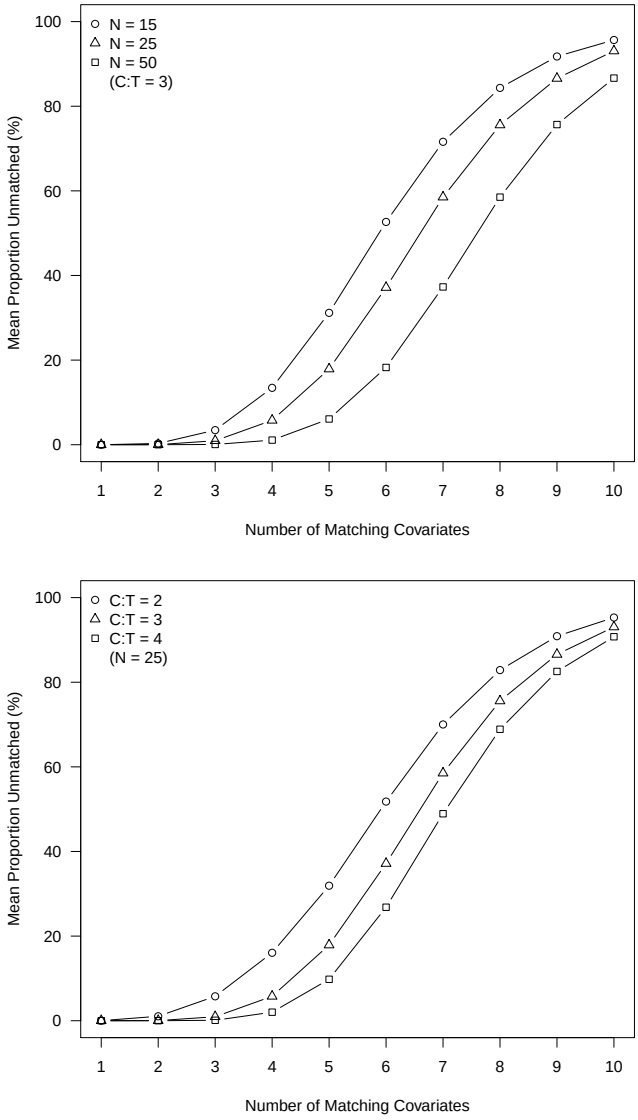


Figure 2.7: Summary of the mean proportion of treated communities that are unmatched over 10,000 Monte Carlo simulations that impose exact matching on between 1 and 10 binary covariates that are IID $\sim$ *Binomial*$(1, 0.5)$. The scenarios vary the number of treated communities (N) (top plot) and the ratio of the number of control to treated communities (C:T) (bottom plot).

baum and Rubin [42] provide a thorough discussion and analysis of the difficulties of exact matching, the bias that follows from discarding incomplete matches, and remedies based on propensity score matching.

2.7.3 Simulation 3: Strengths and limitations of propensity score matching for selecting control communities

In this simulation, I compare propensity score matching (PSM) to univariate and bivariate matching in the design stage. PSM is a convenient method to match using a large number of variables, where exact matching using the individual variables will typically run out of data (see Simulation 2). The metrics that I use to compare the different matching approaches include the standardized difference between intervention (treated) and control communities in four baseline covariates, the average bias in the average treatment effect among the treated (ATT), and the root mean squared error (RMSE) of each matching approach.

The goals of this simulation are to evaluate PSM relative to univariate and bivariate matching under favorable conditions where the matching covariates are all well distributed and linearly related to the outcome. I highlight some of the limitations of the PSM approach by including scenarios where PSM fails to remove bias even when all confounders are observed and included in the propensity score model.

Like simulation 1, the observed treatment A is binary and the outcome Y is continuous. The counterfactual outcomes under different treatment regimens are denoted by Y_a, where Y_1 is the community-level outcome under treatment and Y_0 is the outcome without treatment. In this simulation there are four covariates $W = \{W_1, W_2, W_3, W_4\}$ that are all normally distributed, and W_1 and W_2 are highly correlated. I simulate data for a super population of 10,000 communities according to the following laws:

1. $W_i \sim N(0, 1)$, for $i = 1, \ldots, 4$
2. $\text{cor}(W_1, W_2) = 0.9$ and $\text{cor}(W_i, W_j) = 0$ for $i, j \notin \{1, 2\}$.
3. $Y_a = W_1 + W_2 + W_3 + W_4 + a + \epsilon$

where ϵ is an error term with $\epsilon \sim N(0, 0.25)$. After simulating counterfactual outcomes Y_0 and Y_1, I assign the observed treatment A as a binomial random variable with probability:

$$P(A = 1|W) = \begin{cases} [1 + \exp(-W_1 - W_2 - W_3)]^{-1} & \text{if } W_1 < 0.5 \text{ or } W_3 < 0.5 \\ 0 & \text{otherwise} \end{cases}$$

I then take a random sample of treated and potential control communities from the super population for different sample size scenarios. I consider three different sample sizes for the number of treated communities: $n_1 = 15, 25, 50$, and three different ratios of potential control communities to treatment communities $n_0/n_1 = 1.5, 2, 3$. These sample sizes reflect realistic scenarios often encountered in community intervention studies. Using

these samples, I match intervention to control communities using a 1:1 match and nearest neighbor matching with seven different matching approaches:

- Scenarios 1 – 3: Univariate matching on W_1, W_2 or W_3
- Scenarios 4 – 5: Bivariate matching on $\{W_1, W_2\}$ or $\{W_2, W_3\}$
- Scenario 6: PSM with a $P(A = 1|W) = [1 + \exp(-W_2 - W_3 - W_4)]^{-1}$
- Scenario 7: PSM with a $P(A = 1|W) = [1 + \exp(-W_1 - W_2 - W_3 - W_4)]^{-1}$

Comparison of the matching estimators

Table 2.5 summarizes results from the simulation with $n_1 = 50$ and $n_0 = 100$ ($n_0/n_1 = 2$). The standardized differences in W for the unmatched (random sample) estimator reflect the background imbalance in the covariates without matching. Note that by construction, only W_1, W_2 and W_3 confound the relationship between treatment A and the outcome Y because they are common causes of both. Thus, W_4 is always well balanced, no matter what method is used to select communities. Covariates W_1 and W_2 are stronger confounders than W_3 (indicated by greater imbalance in the unmatched sample). None of the univariate or bivariate matching approaches can fully reduce bias because they cannot match on all of the confounding covariates.

Consider a condition where W_1 is not observed, but W_2 (which is highly correlated with W_1) is measured and used to match. When control communities are selected using a bivariate match based on W_2 and W_3 (scenario 5), it improves the balance on W_1 relative to a random draw (standardized difference = 40.8 vs. 69.4), but it does not balance W_1 well enough to remove all of the bias in the ATT estimator. The PSM approach fares no better when W_1 is omitted from the treatment model (scenario 6): W_1 remains imbalanced (standardized difference = 31.3) and the ATT is biased. The confounding covariates are all balanced and the majority of bias removed only when all confounding covariates are included in the PSM treatment model.

Rubin [23] proves that even when all confounding covariates are observed, the PSM approach can fail to reduce bias when the pool of potential controls is too small to find good matches for each treated unit. Table 2.6 summarizes the simulation results for the PSM estimator that includes all covariates (scenario 7) across a range of different treatment and control sample sizes. For a fixed number of treated communities, increasing the pool of potential controls improves balance on the covariates after matching and leads to larger bias reductions. For a fixed ratio of control to treated communities, increasing the number of treated communities also improves balance and reduces bias, but the effect is much less dramatic. In this simulation, it is more important to have a large pool of controls that to have more treated units, and good balance and large bias reductions can be achieved even for small sample sizes (e.g., $n_1 = 15$).

Table 2.5: Summary of bias, root mean square error (RMSE) and standardized differences for the four covariates across 1,000 Monte Carlo simulations using univariate matching, bivariate matching and propensity score matching (PSM). For the PSM scenarios, the covariates listed were entered as main effects in the propensity score model. The true treatment effect (ATT) in the simulations is 1. In all scenarios there are 50 treated communities and 100 potential control communities used in the match.

Matching approach	Bias	RMSE	Standarized Difference			
			W_1	W_2	W_3	W_4
Unmatched (random sample)	1.175	1.241	69.4	59.5	17.5	-1.9
1 Univariate, W_1	0.241	0.382	6.5	21.3	10.1	-1.5
2 Univariate, W_2	0.472	0.579	31.1	6.2	11.3	-1.9
3 Univariate, W_3	1.118	1.181	75.5	64.2	1.3	-1.6
4 Bivariate, W_1, W_2	0.212	0.365	8.6	7.8	10.1	-1.7
5 Bivariate, W_2, W_3	0.504	0.586	40.8	13.8	-0.8	-1.8
6 PSM: W_2, W_3, W_4	0.346	0.434	31.3	6.8	-4.0	-0.8
7 PSM: W_1, W_2, W_3, W_4	0.074	0.204	7.5	7.9	-4.6	-0.7

This simulation has illustrated the following points. First, PSM is less biased and achieves better covariate balance relative to univariate or bivariate matching when there are more than two confounding covariates, all confounders are measured and the confounders are well behaved. Second, in conditions favorable to matching, PSM can fail to balance unobserved covariates even if they are highly correlated to observed covariates used in the match. Third, even with all confounders observed and correct model specification, PSM can perform poorly without a large pool of potential controls from which to find matches.

This simulation illustrates that the bias in the ATT estimator corresponds with imbalances (large standardized differences) in confounding covariates. Under the assumption that all confounders are measured, if the confounders are well balanced between groups then the design will estimate the ATT with little bias. However, if an unmeasured confounder exists, then a PSM approach can balance observed covariates, but still have highly biased treatment effects (scenario 6 in Table 2.5, with W_1 unmeasured). This is important, because in studies of non-randomized, pre-existing interventions (and all non-randomized studies) it is only possible to evaluate baseline covariate balance for observed variables. Groups may be well balanced on observed covariates, and the assumption of no unmeasured confounding covariates must be evaluated with care.

Table 2.6: Summary of average bias, root mean squared error (RMSE) and standardized difference in covariates for the fully-specified PSM estimator across 1,000 Monte Carlo simulations. Scenarios vary the ratio of potential control to treated communities (n_0/n_1) and the number of treated communities (n_1). The true treatment effect (ATT) in the simulations is 1.

n_0/n_1	n_1	Bias	RMSE	Standarized Difference			
				W_1	W_2	W_3	W_4
1.5	15	0.565	0.759	35.1	31.8	4.4	-0.2
	25	0.418	0.564	26.4	24.4	1.2	0.4
	50	0.358	0.454	22.2	20.2	1.5	0.3
2.0	15	0.266	0.492	18.8	17.2	-3.1	0.3
	25	0.159	0.341	12.5	12.4	-4.9	0.7
	50	0.074	0.204	7.5	7.9	-4.6	-0.7
3.0	15	0.059	0.395	6.8	6.7	-5.8	-0.1
	25	0.033	0.262	4.8	5.5	-6.2	0.8
	50	-0.016	0.173	1.6	3.1	-5.6	0.0

As a final note, the conditions in this simulation are favorable to matching because the covariates W are well distributed and linearly related to the outcome Y. Sekhon demonstrates that PSM can fail to remove bias with näive treatment model specifications in more realistic conditions where covariates are poorly distributed and not linearly related to the outcome [29]. Indeed, it is easy to construct scenarios where PSM makes balance worse on some covariates while improving the balance in others (see Section 2.8, below). In realistic circumstances using machine learning (e.g. Genetic Matching, [9]) to match treated and control communities may perform better than PSM, but will still not solve the problem of unmeasured confounding.

2.8 Empirical example

In this section I describe a study of a non-randomized, pre-existing household water treatment and handwashing intervention that took place in rural Guatemala (see Chapter 3 for details and results) [43]. The intervention and conditions surrounding it meet the six necessary conditions that I outlined in Section 2.4. Specifically:

1. We had access to people and information from the implementing organizations

2. The intervention was deployed to a large number of communities
3. There existed a large number of potential control communities
4. We could reasonably assume the communities were independent
5. The intervention was standardized across communities
6. We had access to pre-intervention (baseline) village-level census data

Between October 2003 and September 2006, two non-governmental organizations, Caritas and Catholic Relief Services, implemented a large household water treatment and handwashing campaign in approximately 90 villages in rural Guatemala. The intervention was not randomized. Instead, the NGOs selected beneficiary communities based on need using Census data that included information about primary water sources and sanitation coverage. They also used illness information (vaguely defined) from the municipal health post. The non-random village selection makes it likely that on average the villages that received the intervention were different at baseline than villages that did not.

Together with my colleagues (see Chapter 3), we conducted an evaluation with the goal of measuring the intervention's impact on household water treatment and handwashing practices and on diarrhea, respiratory infections and growth in children under age five. To help inform control village selection, we obtained data from Guatemala's 2002 national Census that provided village-level information on a variety of characteristics including water sources, sanitation, housing materials, unemployment, occupations, education and demographics [44]. Importantly, this information was collected before the intervention, and enabled us to purposefully select a control group that was more similar to the intervention group than a simple random sample.

In the original study, we selected villages based on a propensity score match (PSM). Here, I return to the original census data to compare the PSM method with five additional sampling methods. I consider the performance of the following six village selection schemes:

1. Random selection
2. Random selection with probability in proportion to size (PPS)
3. Matched selection based on the proportion of households with tap water
4. Matched selection based on village size, measured by number of children < 5 years
5. Matched selection based on village size (as above) and geographic proximity
6. Matched selection based on a propensity score estimated using 12 baseline covariates

I compare the methods using two performance metrics constructed from baseline covariates in the different selected study populations: the standardized difference in means and the variance ratio [29].

Sampling details

In this analysis I use pre-intervention data from the 2002 Census for 88 villages in the study region, and limit the selection to villages with at least 50 children under age five (this restriction was made for logistical reasons in the study to guarantee sufficient sample sizes in each village). This restriction excluded 7 intervention villages and 30 potential control villages, leaving 23 intervention villages and 28 potential control villages for the match. After a rapid assessment of the study region, the original study excluded two additional potential control villages. The two excluded villages were located close to the municipal center, had a large fraction of residents living in the United States sending remittances, and consequently were qualitatively wealthier than other villages in the region. This wealth was newly generated and was not reflected in the 2002 census. The final sample of candidate study villages included 23 intervention villages and 26 potential controls. Table 2.7 summarizes the baseline means for the entire sample, the restricted sample and the propensity score matched sample. The majority of the bias reduction resulted from the matching step and not the initial restriction based on village size.

In all of the matching scenarios (below), I match intervention to control villages using a 1:1 match based on the nearest neighbor distance with a random start. I sample without replacement, and in the matching scenarios I select the 15 closest matches for comparison. I limit the total sample size to 30 villages to mimic the actual study design. All matching was conducted using the `Matching` package in R[15].

1. *Random selection.* The random selection draws a random sample of 15 intervention and 15 control villages from the two groups without replacement.

2. *Random selection with probability in proportion to size (PPS).* For selection with probability in proportion to size (PPS), I select a random sample of 15 intervention villages and 15 control villages, with each village weighted by its population of children under age five. Naturally, this sampling scheme favors larger villages. The rationale for including the non-matching PPS selection method is its frequent use in large surveys. The sample is self-weighting, which is convenient if the study seeks to make inference to the total target population (not necessarily a goal of an intervention study) [45].

3. *Univariate match based on % of households with tap water.* I match villages based on the proportion of households that had tap water at baseline.

4. *Univariate match based on the number of children* < 5 *years.* I match villages using the number of children under five, a measure of village size.

5. *Bivariate match based on geography and number of children* < 5 *years.* I first stratify the villages by seven major water drainages in the study region, and then perform the nearest neighbor matching routine using the number of children under five.

6. *Propensity score match based on 12 covariates.* I calculate the probability of receiving the intervention conditional on baseline covariates (the propensity score), and matched intervention to control villages using the linear predictor of the propensity score model. The functional form for the propensity score model (equation 2.15) is unknown, and so I used the Deletion/Substitution/Addition (DSA) machine learning algorithm to search the space of candidate estimators and select the model with the lowest average cross-validated loss [46]. The DSA-selected propensity score model included the covariates as main effects with no interactions after allowing for up to two-way interactions, a maximum quadratic order for each term, and up to 15 total terms. Finally, I match each intervention village to a control village without replacement using the linear predictor of the model with nearest neighbor matching. This step resulted in 19 matched pairs. Due to time-in-field constraints imposed by the study, we included the 15 pairs with the closest match in our study.

Results

When compared to alternate village selection methods, the PSM approach creates better balance across the set of key covariates than selection based on random sampling (with and without PPS) and univariate or bivariate matching based on tap water access, village size and geography (Figures 2.8 and 2.9). The PSM approach produces standardized differences < 38 (all well below one standard deviation) and only one variance ratio > 1.2 ("of concern" based on Rubin's analysis) [18]. PPS sampling performs poorly compared to all matching methods, and, in general, matching on at least one variable improves the balance compared to a simple random sample. Although the univariate matching samples based on the proportion of households with tap water and the number of children under 5 creates better balance for some covariates compared to PSM, neither approach produces a sample that is as well balanced across all of the baseline covariates. Stratifying by geography and then matching on village size leads to worse balance than matching on village size alone. This is likely due to restricting the pool of potential control villages for each intervention village.

Table 2.7: Summary of baseline covariate means by intervention group and their standardized difference (SD) in three samples: all villages in the region, a restricted sample excluding villages with fewer than 50 children < 5, and the study sample after restriction and propensity score matching.

	All Villages			Restricted Sample			Study Sample		
Mean	Control	Intervention	SD*	Control	Intervention	SD*	Control	Intervention	SD*
Male (%)	51.2	50.7	-1.0	51.9	50.5	-2.7	52.2	51.6	-1.3
Children < 5 (%)	18.8	19.5	1.7	19.4	19.7	0.9	19.5	19.4	-0.2
Age 7 - 14 that work (%)	11.1	11.1	0.1	12.7	11.1	-5.3	12.5	11.3	-3.7
Female literacy (%)	42.4	41.8	-1.2	38.4	41.4	6.1	38.2	41.6	7.0
Work in agriculture (%)	85.8	89.6	12.5	88.3	90.4	7.0	88.1	90.5	8.0
Total households	54.5	91.0	66.6	81.6	110.3	60.5	85.8	97.9	28.3
Houses with tap water (%)	52.5	73.3	47.0	47.8	73.1	57.2	59.8	67.8	17.2
Houses with latrine (%)	52.6	61.3	17.8	50.9	60.5	19.6	45.4	63.3	37.2
Houses with electricity (%)	45.3	60.8	31.8	43.9	62.3	37.9	46.9	55.1	16.5
Houses with soil floors (%)	75.0	80.5	14.0	78.9	80.8	4.8	79.2	79.4	0.5
Houses with thatched roofs (%)	44.8	49.8	9.8	47.8	49.6	3.6	47.6	48.9	2.8
Dist. to municipal center (km)	18.9	16.2	-28.0	18.5	16.7	-18.9	19.0	15.7	-32.7
Number of Villages	58	30		26	23		15	15	
Number of Households	3160	2731		2121	2538		1287	1469	

* The standardized difference is equal to the difference in standard deviations of the mean (I–C) multiplied by 100. It is calculated as $(\mu_I - \mu_C) \div [(S_I^2 + S_C^2)/2]^{1/2} \times 100$. For example, a difference of 1 SD is equal to 100. A value of zero indicates equality of the means.

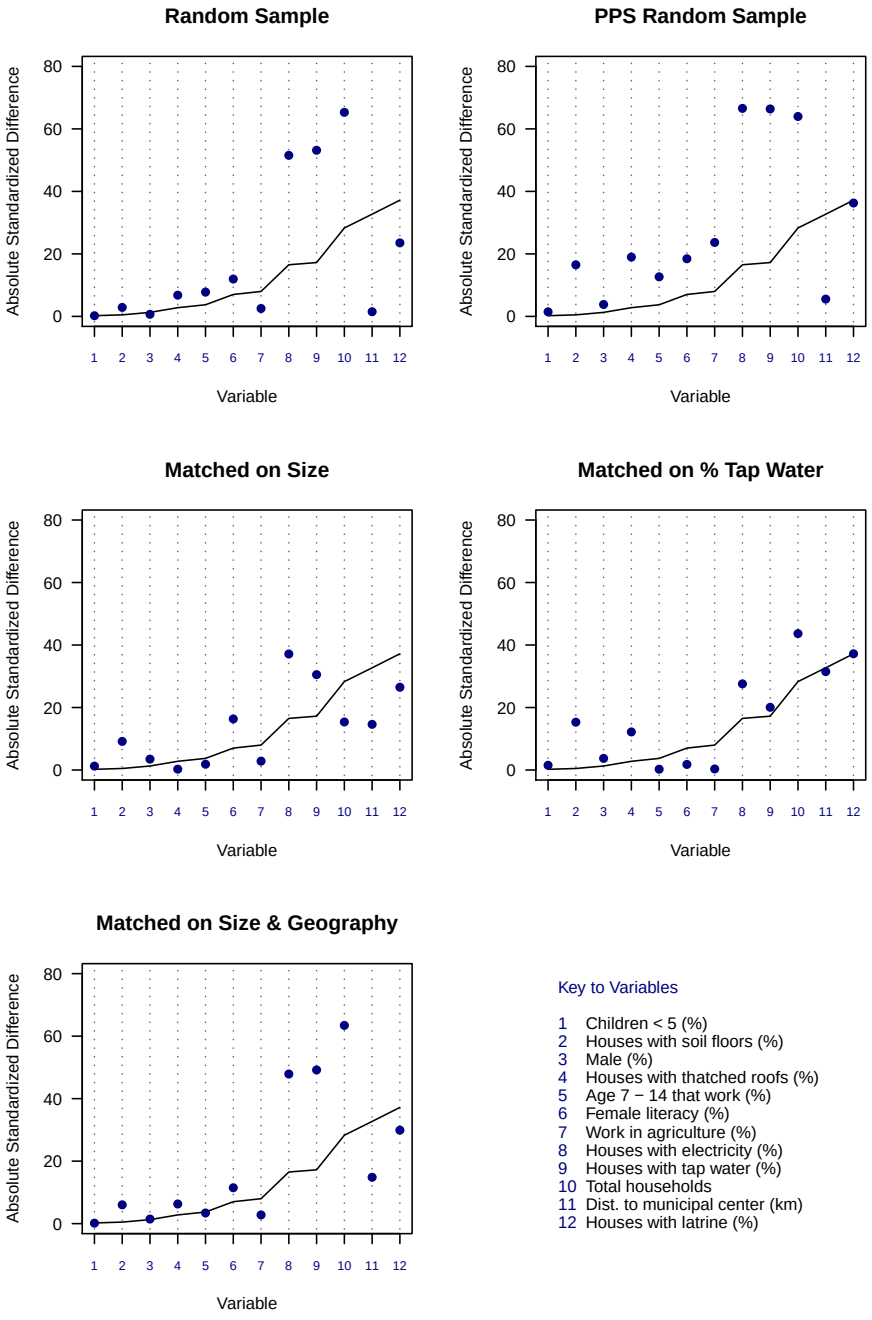


Figure 2.8: Comparison of propensity score matching (PSM) to alternate village selection approaches based on the standardized difference (equation 2.9). Absolute values of the standardized difference are presented for ease of comparison. Covariates are sorted from lowest to highest absolute standardized difference in the PSM sample. The PSM values are represented by a solid line. Dots above the line indicate worse balance than PSM (larger differences). Values that fall beyond the scale of the plot are indicated with an arrow.

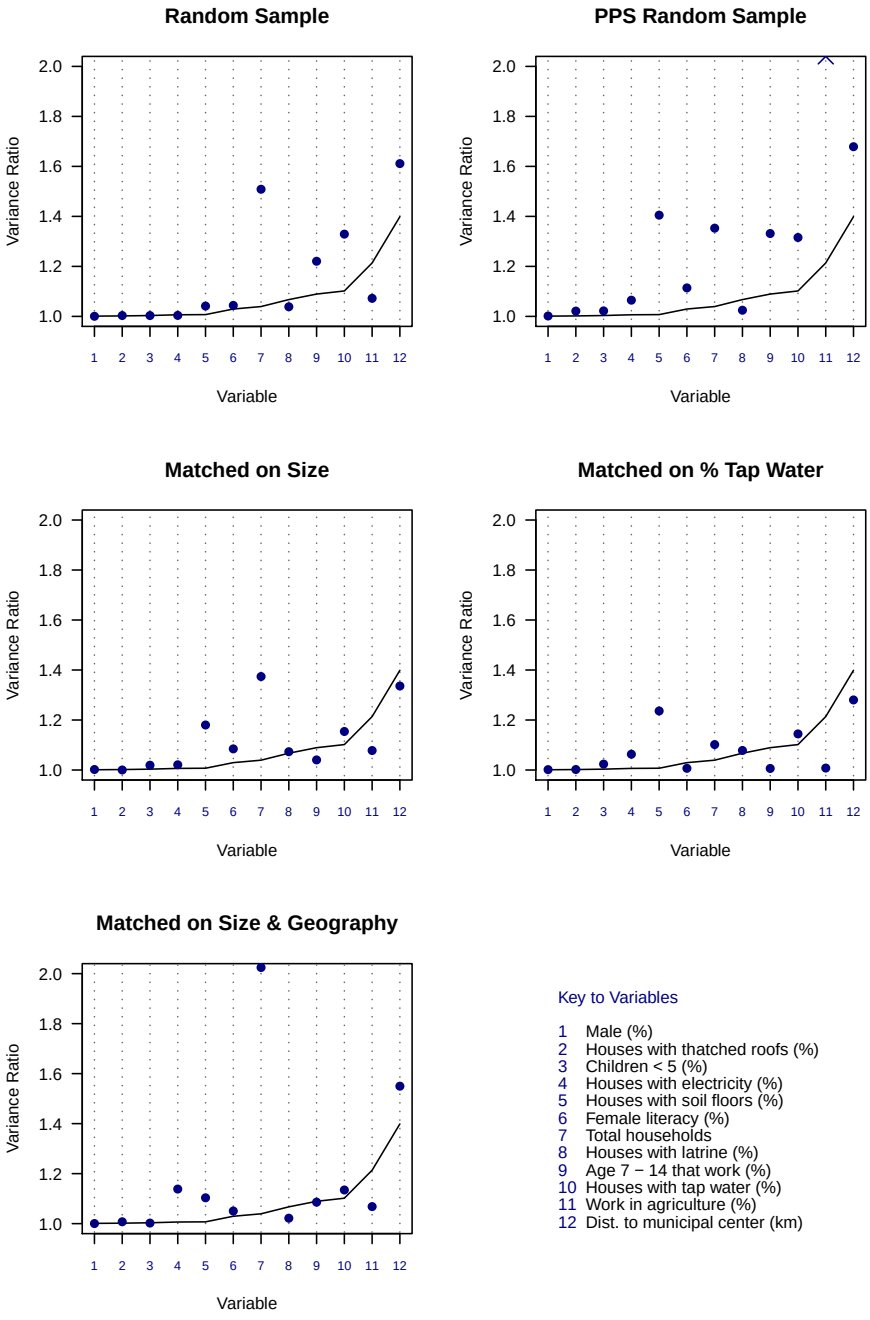


Figure 2.9: Comparison of propensity score matching (PSM) to alternate village selection approaches based on the variance ratios of Intervention and Control groups. Variance ratios < 1.0 have been inverted for ease of comparison. In each plot, covariates are ranked from lowest to highest variance ratio in the PSM sample. The PSM values are represented by a solid line. Dots above the line indicate worse balance than PSM (larger variance imbalance). Variance ratios that fall beyond the scale of the plot are indicated with an arrow.

2.9 Discussion

General comments on the design

In this chapter I have described a design that can collect valid information about treatment effects in non-randomized, pre-existing community interventions under suitable conditions and assumptions that are typical of rigorous observational studies. Propensity score matching in the design stage helps select a sample of control communities that more closely reflect the intervention group. By leveraging the design to reduce bias, the study can rely less on model specification and statistical analyses using the post-intervention data [18, 22]. Unlike studies of randomized interventions, this design relies on a key identifying assumption: conditional on observed baseline characteristics the intervention treatment effects are unbiased (or unconfounded). This is a strong assumption, but it is the assumption of most observational analyses. If outcomes can be measured retrospectively with accuracy, then DID estimators are possible. The advantage of the DID estimators is that they additionally remove time-invariant, unmeasured confounding by assuming that absent the treatment effect, the change in the outcome over time would be the same for intervention and control communities [37, 47]. Although this design applies in principal to household- or individual-level interventions, baseline secondary data for matching is rarely (if ever) available with necessary identifier information for households or individuals. Alternate quasi-experiemental approaches such as regression discontinuity (RD) or instrumental variables (IV) may be useful if the intervention deployment follows a natural experiment (both approaches have "as-if" random components that can be difficult to justify except under very specific applications) [48].

Related designs

If outcomes are measured retrospectively, the design I have proposed is a cross-sectional cohort design [33] with the addition of intentional sampling to assemble the cohort. This design is also related to earlier work on prospective quasi-experimental designs by Rubin [18], Preisser *et al.* [4] and Pattanayak *et al.* [49]. All three papers describe matching-based, quasi-experimental designs similar to what I have proposed here, but they assume prospective follow-up. In their designs, matching is done using pre-intervention secondary data after intervention communities are selected, but before the intervention starts. This general approach has great utility for interventions that cannot be randomized, such as in community-demand-driven interventions. Selecting a matched control set before intervention implementation (as opposed to after the intervention, as I describe here) enables the research team to collect baseline outcome data that serve to (i) validate the matching exercise and (ii) construct DID estimators that do not rely on retrospective measurement. Shadish *et al.* provide a thorough review of additional, related quasi-experimental designs [48].

Motivating applications for the design

The strong assumptions that studies of non-randomized, pre-existing interventions require ensures that they cannot replace randomized, prospective studies for making causal statements about intervention impacts, but they can complement them. There are at least two main reasons that make studies of pre-existing interventions compelling. First, the act of including an intervention and its population in a scientific study can change the way an implementing organization delivers an intervention and it can change how the intervention is perceived by participants. Scientific measurement, such as surveys or other intrusive activities, can itself change behavior and self-reported outcomes (Several papers discuss self-reporting bias of diarrhea outcomes in the context of monitored populations [50–54].) Studies of pre-existing interventions introduce scientific measurement only after the intervention is complete, and so are free from this potential effect (sometimes called the Hawthorne Effect).

Second, evaluating pre-existing interventions can yield information about medium to long-term sustainability of the intervention without years of prospective follow-up. However, using this design for sustainability research requires additional scientific nuance, particularly for behavioral outcomes or highly dynamic outcomes like acute illnesses. For example, if a sustainability study uses a design that matches intervention to control communities using baseline characteristics, but includes only a single follow-up measurement in the post-intervention period (after intervention activities cease), it becomes difficult to interpret a finding of "no difference" between intervention and control. For example, consider the hypothetical intervention impact trajectories plotted in Figure 2.10. A finding of no difference during the post-intervention period is consistent with both scenario 2 where the intervention initially has large impacts, but those impacts are unsustained and scenario 3 where the intervention has limited impact initially, so there is little to sustain. Measuring outcomes at the conclusion of the intervention in addition to outcomes at follow-up solves this identification problem. Note that this problem does not change the validity of the post-intervention follow-up comparison between treatment and control, but it does have implications for how to interpret the results and how the results inform future interventions (i.e., is the problem no impact in the first place, sustaining impacts over the long term, or a combination of both?) Table 2.8 summarizes the three key measurements needed to establish intervention sustainability and their rationale.

Limitations of the approach

The process of selecting a control group using baseline secondary data will only reduce bias if the data used to match intervention communities to control communities contain the major confounding variables, are measured close before the start of the intervention and contain relatively little error. If the baseline data omit key confounding variables, then PSM is unlikely to improve balance on those variables. Resulting estimates of the ATT may be severely biased, even if omitted variables are highly correlated with observed

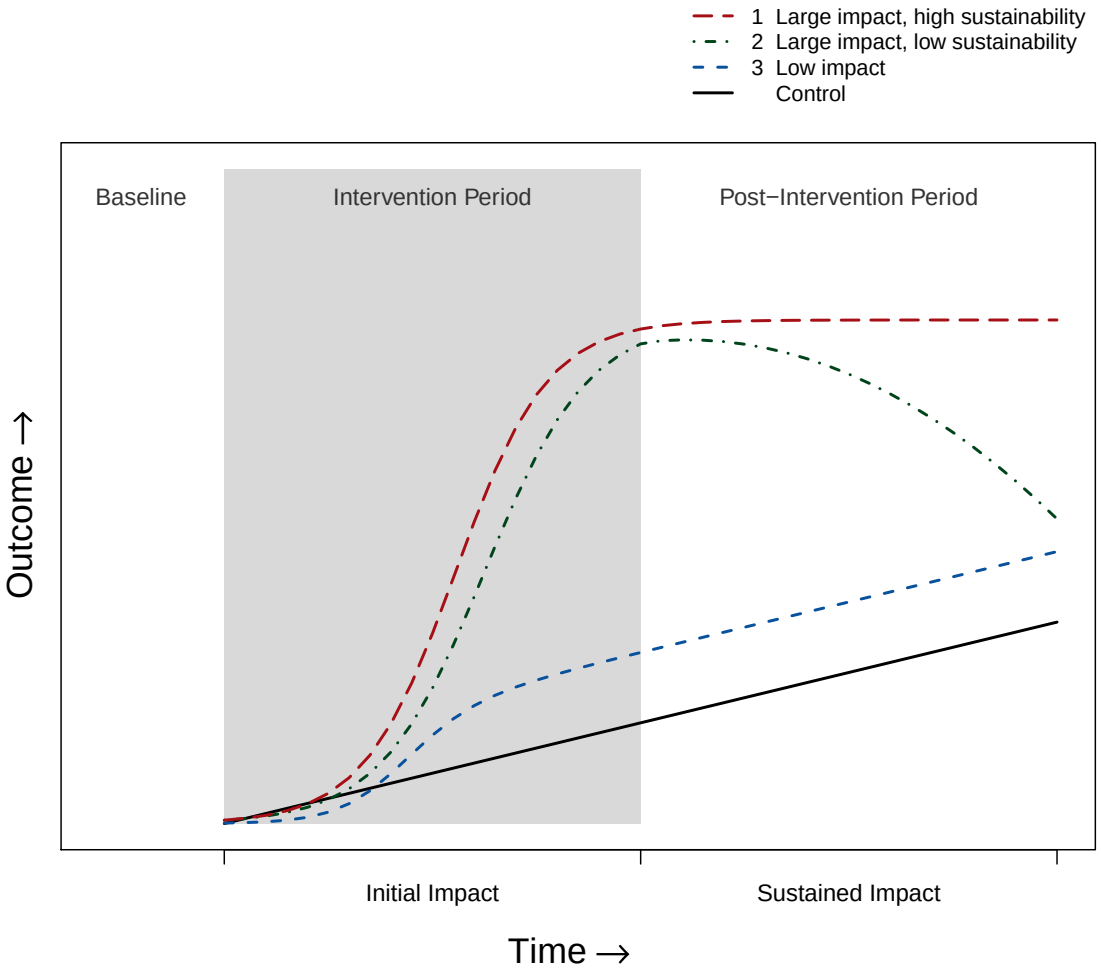


Figure 2.10: Hypothetical scenarios for intervention impact over three periods: baseline, intervention, and post-intervention. Without measurements at the end of the intervention and at follow-up, a large impact/low sustainability intervention is indistinguishable from a low impact intervention.

Table 2.8: Summary of measurements needed to document intervention sustainability.

	Measurement	Rationale
1	Pre-intervention (baseline)	Establish comparability of intervention & control groups
2	Intervention conclusion	Demonstrate initial intervention impact
3	Follow-up (post-intervention)	Demonstrate sustained impact

variables (Simulation 3). In some cases it will be possible to supplement the baseline data with additional information collected post-intervention. If the study collects supplemental variables in a post-intervention survey, then these variables may be used in statistical adjustment if investigators can justify that they could not possibly be influenced by the intervention, and if there is overlap between groups in the variables' distributions. The problem of omitted variable bias is a limitation of all non-randomized designs, and studies of non-randomized, pre-existing community interventions must evaluate this assumption carefully.

Robins presents an approach to bounding the bias from unmeasured confounding using a sensitivity analysis [55]. He recommends conducting simulations using actual data on observed covariates W, the exposure A, the outcome Y combined with a hypothetical unmeasured confounder U. Sensitivity of the treatment effect to unmeasured confounding can be evaluated by varying the strength of association between the unmeasured confounder and both the treatment and outcome. Sekhon proposes a more applied strategy whereby investigators conduct "placebo tests" in their observational data [9]. The basic concept is to find a stratum of the population defined by observed covariates for which the treatment effect is known with some certainty (for example, identifying a subgroup in which there could be no possible treatment effect), and then apply the observational analysis approach to that subgroup to see if it can recover the result that is known a priori. Evaluating the sensitivity of results to potential unmeasured confounding should be included in observational studies.

Since the design mimics a prospective study by identifying the study population retrospectively with community-level baseline data, investigators will not typically know which individuals exit the population between baseline and follow-up and are censored. If censoring is a common effect of the treatment and outcome then it can lead to bias [32]. In actual prospective studies, investigators will typically collect baseline data on individuals, and can assess whether censoring is independent of treatment. Since individuals are

not measured until follow-up, this design cannot empirically evaluate whether censoring is informative and must evaluate its potential using plausibility arguments and indirect measures. For example, it may be possible to estimate community-level migration rates during the post-intervention period, which in some contexts may be a reasonable proxy for migration rates between baseline and follow-up.

For some pre-existing interventions, the selected intervention communities will be so different from all potential control communities that good matches cannot be found. Larger numbers of potential control communities improve the chances of finding a good match [23], but good matches are not guaranteed. The success of matching will be highly context-specific. In some cases, good matches can be found even in small samples. In the empirical example from Guatemala (Section 2.8, see also Chapter 3), I was able to find good matches with just 30 intervention communities and 58 potential control communities using restriction and propensity score matching. Fortunately, the exercise of matching intervention and control communities using secondary data can be completed without sending a team to the field. The research team can determine whether the ATT is even likely to be estimable for a given intervention and population, and if not, they avoid spending costly resources to collect primary field data.

A final limitation of this design is that it does not guarantee contrasts on key exposures of interest in dynamic populations. To illustrate this point, I will use a concrete example from the intervention evaluated in Chapter 4. The original intervention included a combination of private toilet construction, public and private water supply improvements and hygiene education in 12 villages. Using the design described in this chapter, we selected 13 matched control villages and our field team collected post-intervention information on key exposures and outcomes. The matching data were collected two years prior to the initiation of the intervention, and the follow-up data were collected seven years from baseline (between 5 months and 4 years after intervention activities ceased, depending on the village). We found that in the interim period between baseline and follow-up, control villages made similar improvements to their water infrastructure independent from the intervention. The two groups differed greatly in their access to private toilets, but not in water supply, so the study could not measure the full impact of the intervention because no counterfactual population without improved water supply improvements existed. The practical and ethical problem of maintaining a "pure control" group over long periods of time applies to any intervention study intervention (prospective or retrospective, randomized or not), and should be considered carefully when designing studies of pre-existing interventions that took place years prior to a follow-up measurement. A valid length of time between baseline and follow-up will depend on how dynamic the population is during the period. Rapid assessments conducted after community selection but before main field activities can provide information about whether differences in key exposures still exist between intervention and control communities.

Conclusions

In this chapter I have outlined the necessary conditions and assumptions required for studies of non-randomized, pre-existing interventions. I proposed a quasi-experimental design to estimate intervention impacts, and I demonstrated why the ATT is the only promising parameter to estimate in this context. Using propensity score matching to select a control group leverages the design to reduce bias and increase statistical efficiency, while imposing no constraints on additional statistical adjustment using variables collected retrospectively in the post-intervention period. The method is not fool-proof, and I outlined its main limitations in didactic simulations and in the discussion.

Sekhon notes that while rigorous observational studies are important and needed, the only designs that can be mass produced with relative success rely on random assignment [9]. In general, inference from observational studies requires a more complicated, thoughtful, theory-dependent process of argument construction than inference from randomized experiments [1, 48]. This chapter provides general guidelines for designing studies of non-randomized, pre-existing interventions, but many components of the design will vary depending on specific applications, and there are conditions where it can fail. In this sense, this design is not "mass producible" like many randomized designs, and requires investigators to evaluate the threats to validity in each application. Observational designs necessarily require additional care compared to randomized designs because their inference relies on stronger assumptions. Yet, with careful planning and under suitable conditions, studies of non-randomized, pre-existing interventions can provide important information that is difficult to obtain with prospective studies.

Bibliography

[1] Cook TD, Shadish WR. Social Experiments: Some Developments over the Past Fifteen Years. Annual Review of Psychology. 1994 Jan;45(1):545–580.

[2] Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. Am J Public Health. 2004;94(3):423–32.

[3] Campbell MJ, Donner A, Klar N. Developments in cluster randomized trials and Statistics in Medicine. Stat Med. 2007;26(1):2–19.

[4] Preisser JS, Young ML, Zaccaro DJ, Wolfson M. An integrated population-averaged approach to the design, analysis and sample size determination of cluster-unit trials. Stat Med. 2003;22(8):1235–54.

[5] Splawa-Neyman J, Dabrowska DM, Speed TP. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. Statistical Science. 1990;5(4):465–472.

[6] Holland PW. Statistics and Causal Inference. Journal of the American Statistical Association. 1986;81(396):945–960.

[7] Rubin DB. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. Journal of Educational Psychology. 1974;66(5):688–701.

[8] Freedman DA. Statistical models for causation - What inferential leverage do they provide? Evaluation Review. 2006;30(6):691–713.

[9] Sekhon JS. Opiates for the Matches: Matching Methods for Causal Inference. Annual Review of Political Science. 2009;12:487–508. ISI Document Delivery No.: 471PU Times Cited: 0 Cited Reference Count: 136.

[10] Rubin DB. [On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.] Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies. Statistical Science. 1990;5(4):472–480.

[11] Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999;10(1):37–48.

[12] Heckman JJ, Ichimura H, Todd P. Matching as an econometric evaluation estimator. Review Of Economic Studies. 1998 Apr;65(2):261–294.

[13] Rothman K, Greenland S. Modern Epidemiology. 2nd ed. Philadelphia: Lippincott-Raven Publishers; 1998.

[14] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55.

[15] Sekhon J. Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching package for R. Journal of Statistical Software. Forthcoming;.

[16] Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. American Statistician. 1985;39(1):33–38.

[17] D'Agostino J R B. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med. 1998;17(19):2265–81.

[18] Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. Stat Med. 2007;26(1):20–36. Statistics in medicine.

[19] Rosenbaum PR. Optimal matching for observational studies. Journal of the American Statistical Association. 1989;84(408):1024–1032.

[20] Rubin DB. Bayesian Inference for Causal Effects: The Role of Randomization. Annals of Statistics. 1978;6(1):34–58.

[21] Kar K. Subsidy or self-respect? Participatory total community sanitation in Bangladesh. Institute of Development Studies, Working paper 184; 2003.

[22] Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political Analysis. 2007;15(3):199–236. Times Cited: 46.

[23] Rubin DB. Multivariate Matching Methods that are Equal Percent Bias Reducing, I: Some Examples. Biometrics. 1976;32(1):109–120.

[24] Rosenbaum PR, Silber JH. Matching and thick description in an observational study of mortality after surgery. Biostatistics. 2001;2(2):217–32. Journal Article England.

[25] Brookhart MA, van der Laan MJ. A semiparametric model selection criterion with applications to the marginal structural model. Computational Statistics & Data Analysis. 2006;50(2):475–498. 0167-9473 doi: DOI: 10.1016/j.csda.2004.08.013.

[26] Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable Selection for Propensity Score Models. Am J Epidemiol. 2006;163(12):1149–1156.

[27] Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association. 1984;79(387):516–524. ISI Document Delivery No.: TK975 Times Cited: 708 Cited Reference Count: 18.

[28] van der Laan MJ, Polley EC, Hubbard AE. Super learner. Stat Appl Genet Mol Biol. 2007;6:Article25.

[29] Sekhon J. Alternative balance metrics for bias reduction in matching methods for causal inference (available at: http://sekhon.berkeley.edu/); 2007.

[30] Abadie A. Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models. Journal of the American Statistical Association. 2002;97(457):284–292.

[31] Colford JM, Hilton JF, Wright CC, Arnold BF, Saha S, Wade TJ, et al. The Sonoma Water Evaluation Trial: A Randomized Drinking Water Intervention Trial to Reduce Gastrointestinal Illness in Older Adults. Am J Public Health. 2009 Sep;Advance access DOI: 10.2105/AJPH.2008.153619.

[32] Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004;15(5):615–25.

[33] Hudson JI, Pope J H G, Glynn RJ. The cross-sectional cohort study: an underutilized design. Epidemiology. 2005;16(3):355–9.

[34] Hernan MA, Cole SR. Invited Commentary: Causal Diagrams and Measurement Bias. Am J Epidemiol. 2009 Oct;170(8):959–962.

[35] Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. Biometrika. 2009;96(1):187–199.

[36] Zafar SN, Luby SP, Mendoza C. Recall errors in a weekly survey of diarrhoea in Guatemala: determining the optimal length of recall. Epidemiol Infect. 2009;p. 1–6. Journal article.

[37] Meyer BD. Natural and quasi-experiments in economics. Journal of Business & Economic Statistics. 1995;13(2):151–161.

[38] Abadie A, Imbens GW. Large sample properties of matching estimators for average treatment effects. Econometrica. 2006;74(1):235–267.

[39] Feng Z, Diehr P, Peterson A, McLerran D. Selected statistical issues in group randomized trials. Annu Rev Public Health. 2001;22:167–87. Journal Article Review United States.

[40] Donner A, Klar N. Design and Analysis of Cluster Randomization Trials in Health Research. London, England: Arnold; 2000.

[41] Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000;11(5):550–60.

[42] Rosenbaum PR, Rubin DB. The bias due to incomplete matching. Biometrics. 1985;41(1):103–116. Times Cited: 79.

[43] Arnold B, Arana B, Mausezahl D, Hubbard A, Colford J John M. Evaluation of a pre-existing, 3-year household water treatment and handwashing intervention in rural Guatemala. Int J Epidemiol. 2009;p. DOI:10.1093/ije/dyp241.

[44] INE. Censos Nacionales XI de Poblacion y VI de Habitacion, Guatemala 2002. Instituto Nacional de Estadistica; 2002.

[45] Kish L. Survey Sampling. New York: John Wiley & Sons; 1965.

[46] Sinisi SE, van der Laan MJ. Deletion/substitution/addition algorithm in learning with applications in genomics. Stat Appl Genet Mol Biol. 2004;3:Article18.

[47] Abadie A. Semiparametric difference-in-differences estimators. Review of Economic Studies. 2005;72(1):1–19.

[48] Shadish WR, Cook TD, Campbell DT. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston: Houghton Mifflin Company; 2002.

[49] Pattanayak SK, Poulos C, Yang JC, Patil SR, Wendland KJ. Of taps and toilets: quasi-experimental protocol for evaluating community-demand-driven projects. J Water Health. 2009;7(3):434–51. Journal Article England.

[50] Clasen T, Schmidt WP, Rabie T, Roberts I, Cairncross S. Interventions to improve water quality for preventing diarrhoea: systematic review and meta-analysis. Bmj. 2007;334(7597):782. BMJ (Clinical research ed.

[51] Arnold BF, Colford J J M. Treating water with chlorine at point-of-use to improve water quality and reduce child diarrhea in developing countries: a systematic review and meta-analysis. Am J Trop Med Hyg. 2007;76(2):354–64. The American journal of tropical medicine and hygiene.

[52] Genser B, Strina A, Teles CA, Prado MS, Barreto ML. Risk factors for childhood diarrhea incidence: dynamic analysis of a longitudinal study. Epidemiology. 2006;17(6):658–67.

[53] Schmidt WP, Luby SP, Genser B, Barreto ML, Clasen T. Estimating the longitudinal prevalence of diarrhea and other episodic diseases: continuous versus intermittent surveillance. Epidemiology. 2007;18(5):537–43.

[54] Schmidt WP, Cairncross S. Household Water Treatment in Poor Populations: Is There Enough Evidence for Scaling up Now? Environmental Science & Technology. 2009;43(4):986–992.

[55] Robins JM. Association, causation, and marginal structural models. Synthese. 1999;121(1-2):151–179.

Chapter 3

Evaluation of a pre-existing, 3-year household water treatment and handwashing intervention in rural Guatemala

Preface

The work present in this chapter is the product of myself and a large number of collaborators. For this reason I use the pronoun "we" rather than "I" throughout. Here, I describe what "we" implies. I secured funding for the project from the Institute for Public Health and Water Research (`www.ipwr.org`) in late 2006 and designed the study with input from Jack Colford (UC Berkeley), Alan Hubbard (UC Berkeley) and Byron Arana (Universidad del Valle de Guatemala – UVG). Along with Byron, members of Fundación SODIS Matthais Saladin and Alvaro Solano assisted with selecting the study site and population in Guatemala. Nazario Lopez (UVG), Rodrigo Gramajo Rodriguez (UVG), Andri Christen (Swiss Tropical Institute – STI) and Daniel Maeusezahl (STI) helped me with the questionnaire design and translation. Nazario Lopez (UVG) was my fieldwork supervisor who managed our small team of very dedicated field staff: Carlos Miguel Loyo, Manuel Morales, Juan Mendoza and Pedro Joaquin. I was present in the field for roughly 75% of the data collection (mainly helping with logistics, planning and quality control). Maricruz Alvarez (UVG) conducted all the water quality analyses. I conducted all of the analyses and drafted all of the text herein. In addition to my dissertation committee, Byron and Daniel provided comments on the content of this chapter in the process of submitting it to peer-review.

The majority of the material that I present in this chapter was published in the *International Journal of Epidemiology* with my coauthors Byron Arana, Alan Hubbard, Daniel Maeusezahl, and Jack Colford [1].

3.1 Goals

In this chapter I apply the methods developed in Chapter 2 to evaluate a pre-existing household water treatment and handwashing intervention in Guatemala. The design combines historic, pre-intervention secondary data collected in the 2002 Census, with cross-sectional outcome measurement in 2007 (collected 6 months after the completion of the intervention). The study illustrates how useful information about longer-term intervention outcomes can be collected quickly using pre-existing interventions that were implemented outside of formal research activities.

3.2 Background

Between October 2003 and September 2006, two non-governmental organizations (NGOs), Caritas and Catholic Relief Services, implemented a large household water treatment and handwashing campaign in approximately 90 villages across three municipalities in rural eastern Guatemala. We conducted our evaluation in the municipality of Camotán be-

cause both the NGO records were more complete for that municipality compared to others (Jocotán and San Juan Ermita). The implementing organizations had oversight from the SODIS Foundation (www.fundacionsodis.org), who provided input into the training materials, social marketing messages and general implementation strategy. The NGOs promoted three water treatment methods: boiling, solar disinfection (SODIS), and chlorination using dilute bleach. Based on our exchanges with some of the Caritas technicians, the campaign likely emphasized the SODIS method over chlorination and boiling, but they encouraged families to use their own preferred method (or combination of methods). Handwashing education and social marketing included demonstrations of correct technique that emphasized using soap or detergent and scrubbing thoroughly. The promotion also emphasized critical times to wash hands that included: before cooking, before eating, before feeding children, after defecation and after changing babies.

All villages received the same intervention package and all activities were initiated at the same time (October 2003). The intervention program used a "train the trainer" model, where NGO technicians trained local community women to promote the behavior change through social marketing and household visits. The NGOs recruited approximately one community promoter per 25 participating households. The trained health promoters later visited households with children under age three or with pregnant mothers to promote water treatment and handwashing with soap. The visits occurred monthly or bi-monthly and lasted approximately 30 minutes each.

Promoters educated mothers about proper nutrition for their children, and at the end of each visit gave the family a small ration of rice, beans and oil. This nutritional component to the intervention was implemented at a regional scale in concert with many additional NGOs, UNICEF and Guatemala's National Plan for the Reduction of Chronic Malnutrition (a response to a drought and subsequent famine in 2001 that struck Camotán and adjacent Jocotán). This component was not unique to intervention villages in our sample (indeed, we confirmed that all villages in our sample – intervention and control – received food aid during the study period).

There exists no formal record of the proportion of eligible households that participated, but technicians on the ground suggest that the majority of eligible households participated. At the conclusion of the intervention the implementing organization conducted a survey of participating households and recorded water treatment behavior based on self-report. The survey estimated that 70% of participating households regularly used some method of household water treatment (village level self-reported treatment range: 29% - 100%). The SODIS Foundation provided these data at the start of our evaluation.

The primary objective of this study was to revisit households six months after the conclusion of the intervention to assess water treatment behavior, basic hygiene knowledge and practices, and child health. We measured child health using self-reported symptoms of acute diarrheal and respiratory illness. We also used anthropometric measurements that have demonstrated utility as outcome measures for water and sanitation interventions [2–4]. To our knowledge these outcome measures have been reported in water supply and

sanitation studies, but not in water quality or handwashing studies.

3.3 Methods

3.3.1 Setting

This study was conducted in the Camotán municipality in the mountainous state of Chiquimula, Guatemala near the eastern border with Honduras. Camotán is a mountainous region with 94 rural villages located between 2 and 37 km from the municipal center, and typically accessed by dirt roads. The primary occupation is agriculture; corn and beans are the main crops with some coffee grown at higher elevations. Recent government surveys report that 89% of people live in moderate or extreme poverty [5]. Water is obtained from mountain springs and surface water. Community and household taps, where available, are connected to gravity-fed spring networks, and water sources are typically contaminated with fecal organisms[6].

3.3.2 Study Design

Village selection

We implemented a cross-sectional cohort design with a seven-day retrospective risk period [7]. All data collection followed protocols approved by the institutional review boards at the University of California, Berkeley and the Universidad del Valle de Guatemala, and all participants provided informed consent.

Since the intervention was non-randomized and villages were purposely selected by the implementing organizations, intervention villages were likely different, on average, from other villages in the study region. Pre-treatment differences between intervention and control villages could lead to differences in water and hygiene practices and child health, independent of the intervention. To help increase comparability between intervention and control villages, and to reduce confounding by observable characteristics, we used restriction and propensity score matching [8] based on pre-intervention characteristics to select intervention and control villages. All study villages – intervention and control – were selected in 2007, after the intervention ended. We adapted the selection approach from a series of recent prospective, non-randomized, community-level intervention studies of underage drinking interventions [9], education programs [10], and water supply and sanitation improvements [11] (also see Chapter 2).

We obtained village-level 2002 census data that contained detailed information about demographics, education, housing conditions, water sources, and sanitation for 88 villages in the study region [12]. To guarantee sufficient numbers of children and to improve comparability between intervention and control villages, we restricted our sample to villages with at least 50 children under 5 living in them. This restriction excluded 7 intervention

villages and 30 potential control villages, leaving 23 intervention villages and 28 potential control villages for the match. After a rapid assessment of the study region, we excluded two additional potential control villages. The two excluded villages were located close to the municipal center, had a large fraction of residents living in the United States sending remittances, and consequently were qualitatively wealthier than other villages in the region. This wealth was newly generated and was not reflected in the 2002 census. Our final sample for the match included 23 intervention villages and 26 potential controls.

We modeled the probability of participation in the behavior change intervention using a logit model: logit $Pr(A = 1|W) = \alpha' W$, where the logit function is $\log[p/(1-p)]$, A is an indicator variable equal to 1 if a village participated in the intervention and 0 otherwise, and W is a vector of characteristics that included the percentages in each village of: males; children under age five; literate females; individuals employed in agriculture; households with private water taps; households with private wells; households with private latrines; households with electricity; and households with soil floors. Additionally we measured the number of households, people per household, and distance to the municipal center. Importantly, the covariates in W were selected after detailed discussions with program technicians in the implementing organizations, and include information that the organizations used to select intervention villages.

In both the propensity score model used in our design and the targeted maximum likelihood estimation used in our adjusted analyses, one must estimate regressions that are not of direct interest (nuisance parameters), but are necessary to estimate the parameter of interest. The consistency of our estimates is contingent on the consistency of these nuisance parameter estimates. To estimate the nuisance parameters we used the Deletion/Substitution/Addition (D/S/A) algorithm, which is a flexible model-selection approach that fits polynomial terms and their tensor products using cross-validation [13]. Our D/S/A-selected propensity score model included the covariates as main effects with no interactions after allowing for up to two-way interactions, a maximum quadratic order for each term, and up to 15 total terms [13]. We excluded intervention and control villages outside the region of common support (overlap) on the propensity score [14]. Finally, we matched each intervention village to a control village without replacement using the linear predictor of the model with nearest neighbor matching. This step resulted in 19 matched pairs. Due to time-in-field constraints, we included the 15 pairs with the closest match in our study.

Household selection

We selected households within each village using a stratified systematic sample. Our team used village sketch maps from the municipal planning department to split each village into two geographic strata with roughly equal numbers of houses. Within each stratum, the field supervisor chose a random start, and the interviewer teams visited every third house until 10 houses were sampled. The inclusion criteria for the study were: (i) at least one

child under age five living in the home and (ii) the family had lived in the village since 2003 or earlier (the time of intervention start). If a selected household met our inclusion criteria but the primary caretaker was away, the field team returned two additional times before choosing a replacement household.

3.3.3 Data collection

A team of four trained fieldworkers and a field supervisor from the Universidad del Valle de Guatemala with 12 years of research experience conducted household interviews during the dry season between April and June of 2007. The survey instrument was pre-tested and validated over a two-week period in nearby, non-study villages. In all cases, the child health and behavioral questions were answered by the child's primary caregiver (nearly always the mother).

3.3.4 Water sample collection and analysis

We collected household water samples in a random sample of 48 households from 8 study villages (4 intervention and 4 control). Water samples were collected in 100 ml Whirl-Pack™ bags in a fashion that mimicked each household's water retrieval practices: by either dipping a household cup into the vessel to transfer the water, pouring water from the storage container into the sample container, or, if a household did not store drinking water, retrieving water directly from the tap. The field team transported the water samples in a cooler to the laboratory at the Universidad del Valle de Guatemala for culturing within 20 hours of collection. Samples were processed using the Colilert Quantitray 1000 kit (IDEXX Laboratories), and we used a most probable number (MPN) table to quantify Escherichia coli.

3.3.5 Outcome definition and measurement

The primary health outcomes of this study were diarrhea, acute lower respiratory-tract infections (ALRI), and child growth measured by height, weight and mid upper arm circumference.

Self-reported health outcomes

During the household interviews, field staff collected self-reported illness symptoms from each child's mother using a health calendar modeled after Goldman *et al.* [15]. We defined diarrhea as three or more loose or watery stools in 24 hours, or a single stool with blood or mucus [16]. We also recorded symptoms for a previously published measure of highly credible gastrointestinal illness (HCGI), which includes any of the following four conditions: vomiting, watery diarrhea, soft diarrhea and abdominal cramps, or nausea and abdominal cramps [17]. We defined clinical ALRI according to the WHO clinical

case definition: cough or difficulty breathing with a raised respiratory rate (>60 breaths per minute in children younger than 60 days old, >50 breaths per minute for children aged 60 – 364 days, >40 per minute for children aged 1 – 5 years) [18]. Fieldworkers recorded the number of breaths over 30 seconds using a wristwatch (our field supervisor standardized breathing rate measurements during the week-long field worker training and pilot study, but did not record formal measures of inter-rater reliability).

Anthropometrics

All fieldworkers were standardized on anthropometric measurement techniques over two days of training, and they collected measurements in teams of two (we followed standard protocols from the Demographic and Health Survey [19]). During the two days of anthropometry training, all fieldworkers measured the same child and their measurements were checked against the field supervisor's measurement. By the end of the anthropometry training period (n=8 children) all measurements were within each instrument's the margin of error compared to the field supervisor's measurement (treated as a gold standard). The field team collected all anthropometric measurements at the participants' homes at the time of the interview. Fieldworkers weighed children under age two in the lying position and children aged 2 to 5 standing on infant scales accurate to 0.1 kg (Tanita 1380). Fieldworkers measured the length of children under age two in the reclining position and children aged 2 to 5 in the standing position using portable wooden stadiometers (420 Measure All) accurate to 0.1 cm. Upper arm circumference was measured for children aged 6 months and older at the mid point of the upper right arm using an elastic tape accurate to 0.1 cm.

Knowledge and practices

We collected a series of knowledge and practice (KAP) outcomes that included water treatment, water storage, handwashing, and general hygiene practices. Fieldworkers measured water treatment practices using self-reported behavior. Families that reported treating their water were classified as "confirmed" if they (i) reported treating their water in the previous seven days, (ii) had treated water in their home at the time of the interview, and (iii) could produce the materials they used to treat water. Fieldworkers evaluated the presence of treated water based on self-reported information and a sample (not tested) provided by the family. Treatment materials included a designated pot and storage container for water boiling, plastic polyethylene terephthalate (PET) bottles for SODIS, and liquid bleach or chlorine tablets and a designated storage container for chlorine treatment. Fieldworkers collected self-reported handwashing behavior by asking an open question to mothers about when they washed their hands in the past 24 hours and coding answers using five critical times: before cooking, eating, or feeding children and after defecation or changing the baby. Fieldworkers collected information about hygiene and water storage with discrete spot check observations during the interview.

3.3.6 Statistical Methods

Sample size estimation

We based the sample size calculation on diarrhea because it was the primary outcome for which we had the most information from a similar population in Guatemala.

We used simulation (described below) to estimate the statistical power of alternate design scenarios based on the longitudinal prevalence of diarrhea. Simulation was a natural choice for our analysis because we had access to raw data from a large water intervention trial conducted in a similar population in Guatemala from 2001-2002 (the population was in the Department of San Juan Sacatepequez [20]). The trial data enabled us to obtain estimates of village- and individual-level variance in diarrhea prevalence. We model the probability of diarrhea in comparison and treatment groups using the formula:

$$p = \frac{1}{1 + \exp - (\beta_0 + b_{0i} + b_{0ij} + \beta_1 \cdot A)} \tag{3.1}$$

where β_0 is the baseline prevalence of diarrhea, b_{0i} is the random effect in the baseline prevalence of diarrhea in the ith cluster, b_{0ij} is the random effect in the baseline prevalence of diarrhea in the jth individual in the ith cluster, β_1 is log(odds ratio), and A is an indicator variable for treatment (vs. a comparison group).

In our power calculations, we assumed 15% baseline prevalence of diarrhea (β_0), which was consistent with the data used for variance estimation. We estimated the variances of the village- and individual-level random effects (β_{0i} and β_{0ij}) by fitting an intercept-only random effects on the training data set and using the model estimates of the variances of the random effects. In each simulation, we assumed that cluster- and individual-level random effects were normally distributed with mean zero and standard deviations SD(b_{0i})=0.14 and SD(b_{0ij})=0.667 (obtained from the random effects model).

In each simulation iteration, we distribute a hypothetical population into two groups – one that receives treatment and one that does not. All calculations assume 15 villages in each treatment group due to time-in-field constraints. The presence of diarrhea for each child at each follow-up visit is estimated as a binomial random variable using the calculated probabilities from equation (3.1). The difference in the prevalence of diarrhea between treatment groups is estimated using a generalized linear model and robust standard errors [21]. We repeat the simulation process 1000 times for each design scenario. Power is estimated as the fraction of the 1000 iterations in which we find a statistically significant difference between the two treatment groups at the 0.05 level (i.e., assuming a one-sided alpha of 0.05).

The power simulations indicated that with 15 villages in each treatment group the study would have sufficient power (80%) to detect prevalence differences of between 5.5% and 6.5% (Table 3.1).

Table 3.1: Power $(1-\beta)$ from simulations for various treatment effects and village (cluster) sizes. All estimates assume 15% baseline prevalence and 15 villages in each treatment group. LPD is short for Longitudinal Prevalence Difference.

LPD	Children per Village			
	20	25	30	35
-1.0%	0.139	0.129	0.144	0.140
-1.5%	0.145	0.180	0.177	0.181
-2.0%	0.217	0.216	0.235	0.266
-2.5%	0.225	0.246	0.290	0.329
-3.0%	0.260	0.338	0.365	0.429
-3.5%	0.352	0.432	0.468	0.501
-4.0%	0.407	0.491	0.548	0.605
-4.5%	0.498	0.571	0.643	0.706
-5.0%	0.581	0.679	0.706	0.766
-5.5%	0.656	0.727	0.785	0.848
-6.0%	0.727	0.812	0.850	0.894
-6.5%	0.819	0.880	0.912	0.933
-7.0%	0.834	0.910	0.940	0.982
-7.5%	0.899	0.946	0.975	0.986

Measures of self-reported illness

Gastrointestinal and respiratory outcomes were measured using daily longitudinal prevalence [22],[1] a disease measure that is more strongly correlated with child mortality than incidence [23]. We limited the longitudinal prevalence data to a two day recall window after identifying under-reporting of symptoms for recall periods longer than two days, a finding that has been documented in other studies of self-reported diarrheal illness [24, 25].

Unadjusted outcome analyses

Using daily symptoms reported at the time of the interview we reconstructed the 48-hour retrospective risk period for each child in the study. Valid inference in a cross-sectional cohort study requires that there is no recall bias and that there is no informative censoring: two key assumptions that hold in these data [7].

The parameter of interest for all outcomes (both unadjusted and adjusted) is the marginal treatment effect conditional on selection into the study based on restriction and

[1] Longitudinal prevalence is calculated by dividing the number of days with illness by the total days of observation.

propensity score matching. We estimate the parameter as:

$$E(Y|A=1,W^*) - E(Y|A=0,W^*) \tag{3.2}$$

where Y is the outcome of interest, A is an indicator equal to 1 if a child lives in an intervention village and 0 otherwise, and W^* is the set of characteristics among intervention villages in the study sample ($W^* = W|A=1$, >49 children under 5). Thus, our inference is limited to the set of intervention villages for which there is a comparable control village based on the village selection method: this is an average treatment among the treated (ATT) estimator (see Chapter 2 for background and derivation).

For self-reported health outcomes we calculated the difference in the daily longitudinal prevalence between the intervention and control groups. We converted the anthropometric measurements to age- and sex-specific Z-scores using a publicly available Stata algorithm (*WHO Anthro*) that references the 2006 WHO Growth Standards [26], and calculated the difference in Z-score means. For binary KAP outcomes, we calculated the difference in prevalence of each outcome between the intervention and control groups. For all unadjusted estimates we calculated percentile-based 95% confidence intervals using a bootstrap with matched village pairs as the sampling unit (to reflect the design) and 1,000 iterations [27].

We have reported risk differences because additive risk is generally more useful for estimating the magnitude of public health problems than relative risk, which is more useful for disease etiology. There is a large literature on this that is rooted in Rothman's sufficient/component cause model, where he postulates that additive risk is the most "natural" scale to look for interaction in public health interventions [28]. The basic rationale for choosing between additive versus relative association measures is summarized by Leite *et al.* [29]:

> Fitting different [absolute vs. relative] models implies qualitatively different measures of exposure effect. In a prospective study, rate difference is a measure of the absolute effect or excess incidence of infection experienced by the exposed. This statistic is useful for estimating the magnitude of the public health problem represented by the exposure. On the other hand, rate ratio is a unit-free measure of how much more likely it is for the exposed to become infected than the unexposed. As this statistic is an indicator of the strength of an association, it is the usual measure in etiologic research.

Leite *et al.* refer to rates, but it is an equivalent argument for risk/prevalence.

Adjusted outcome analyses using targeted maximum likelihood

For self-reported health and anthropometric outcomes we adjusted for potentially confounding variables using recently developed targeted maximum likelihood estimation (MLE) [30], which is described in detail in Appendix B. The targeted MLE approach

is similar to standard maximum likelihood regression, but it targets the likelihood to a specific parameter of interest: in this case, either the longitudinal prevalence difference (diarrhea, HCGI, respiratory symptoms) or difference in means (anthropometrics). All analysis was conducted using R software (www.r-project.org).

As before, let Y be an outcome of interest, A be the intervention status equal to 1 if a child lives in an intervention village and 0 otherwise, and $W^* = W|A = 1$ be a set of covariates in the treated population that are potential confounders of the relationship between A and Y. Following notation from Appendix B, we calculated adjusted estimates using the following estimator:

$$\theta_n^{T-MLE} = \frac{1}{n}\sum_{i=1}^{n} Q^k(1, W_i) - Q^k(0, W_i) \tag{3.3}$$

where k is the number of iterations for the update, and $Q^k(A, W)$ is the predicted outcome for each intervention status A given covariates W (see Appendix B for complete details and derivations). Evaluate the updated regression at $A = 1$ and $A = 0$ to get two predicted outcomes for each child. The estimator θ_n^{T-MLE} is the empirical mean of the difference across the population for two predicted outcomes, one under treatment and the other under no treatment, conditional on selection into the study. It is a reformulation of the estimator in equation (3.2).

In our initial estimate of $Q^0(A, W^*)$ there are a potentially large number of covariates in W^* and the models' functional form is unknown. We initially considered village-level indicators and characteristics that were unlikely to have been affected by the intervention (i.e., they were pre-treatment). The covariates that we considered are listed in Table 3.2.

We eliminated the number of children < 5 and < 15 due to collinearity with the total number of persons in each household. We also eliminated minutes per day retrieving water and satisfaction with water quantity due to collinearity with water source. We restricted the covariate set to those that were considered to be potential confounders by the authors [31] and had a strong association with the outcome based on a previously published backward deletion approach [28, 32]. The backward deletion approach selects variables that, when removed from a multivariable specification including all candidate covariates, change the treatment coefficient by 5% or more.

After this dimension reduction step, we chose final model specifications using the Deletion/Substitution/Addition (D/S/A) selection algorithm allowing for two-way interactions, quadratic terms, and 15 total terms [13]. We ran the backward deletion and D/S/A model selection separately for each health outcome.

We also selected terms for the treatment mechanism, $g^0(A|W)$ by initially including covariates that had an absolute standardized mean difference greater than 20 (intervention minus control) and could not reasonably be influenced by the intervention. We chose a threshold standardized difference of 20 because it roughly corresponded to univariate t-statistic-based p-values of 0.01, and because it reduced the covariate set sufficiently

Table 3.2: Covariates considered in targeted maximum likelihood estimation.

Covariate	Abbreviation
Child's characteristics	
Sex	sex
Age (months)	age
Total months breast fed	bftot
Mother's characteristics	
Age (years)	mage
Works for money (yes/no)	mwork
Literate (yes/no)	mlit
Leaves village $\geq$ once per week (yes/no)	trips
Household characteristics	
Total persons living in home	totp
Num. children < 15 years	num15
Num. children < 5 years	num5
Electricity (yes/no)	elec
Dirt floor (yes/no)	dirt
Thatched roof (yes/no)	palm
Home ownership (yes/no)	homeown
Land ownership (yes/no)	landown
Use banking services (yes/no)	bank
Have relatives in USA (yes/no)	relus
Have relatives in the Capital (yes/no)	relguat
Travel time by car to the municipal capital (min)	ttime
Durable good ownership (yes/no)	
Refrigerator	refri
Radio	radio
Television	tv
Mobile phone	cell
Bicycle	bike
Automobile	car
Water supply	
Primary water source (factor)	watsource
Private tap	
Public tap	
Public well	
Spring	
Surface water (river/lake)	
Minutes per day retrieving water	wattime
Satisified with water quantity (yes/no)	watsat
Sanitation	
Latrine ownership (yes/no)	latrine
Animals in living vicinity of house (yes/no)	
Pigs	pigs
Chickens/ducks	birds
Dogs/cats	dogscats
Cows/horses/mules/donkeys	stock

for model selection. We selected our final treatment model using the D/S/A algorithm allowing for two-way interactions, quadratic terms, and 15 total terms. We used the same treatment model for all outcomes.

Identical to the unadjusted approach, we calculated percentile-based 95% confidence intervals for the estimates using a bootstrap with matched village pairs as the sampling unit (to reflect the design) and 1,000 iterations [30].

For binary outcomes, the estimates are the marginal, population-averaged difference in the longitudinal prevalence for Intervention minus Control. For continuous Z-score outcomes, the estimates are the marginal, population-averaged difference in Z-score for Intervention minus Control. In each case we report the final covariates considered in $Q^0(A, W)$ and $g^0(A|W)$. We also summarize the distribution of $g^0(1|W)$ because the distribution of the predicted probability of treatment helps identify whether there exists common support on the covariates W between the treatment groups. This is also referred to as the experimental treatment assumption [33]. Specifically, the parameter of interest is only well-defined if: $0 < \text{P(A=1|W)} < 1$, which states that variation in treatment exists for each stratum of W [30].

A re-analysis of the data using self-reported participation as treatment

As an extension to the primary analysis, we summarize the treatment effect estimates after defining the intervention population as the 147 households (49% of the intervention group) who reported participating in the intervention. In this analysis, we re-allocated the remaining 153 intervention households to the control group. This is analogous to a "treatment actually received" analysis in a randomized trial, and so there may be important self-selection into the treatment group that can lead to confounding bias. However, we present these exploratory results following the recommendations of Victora *et al.* [35], who advocate presenting treatment effects among those who actually received treatment adjusted for as many potentially confounding characteristics as possible.

Our method of confounding adjustment in this analysis is identical to the primary targeted MLE approach described in Section 3.3.6 above, although we repeated all model selection routines for $g^0(A|W)$ and $Q^0(A, W)$ using the alternate treatment definition (self-reported participation). Since this alternate treatment definition varies at the household-level, our inference relies on a bootstrap that resamples children at the household level. This approach assumes that households are independent, which is a stronger assumption than in our primary analysis where we assume that children in separate villages are independent.

3.4 Results

3.4.1 Village selection and pre-intervention characteristics

The village selection process improved the comparability of intervention and control villages across a range of important, pre-intervention characteristics. Table 3.4.1 summarizes pre-intervention covariate means for control and intervention villages and their standardized difference (SD), which is equal to the difference in means in standard deviations (see equation 2.9 on page 38). Before restriction and matching, intervention villages had more households on average (91 versus 55; SD = 67) and had a greater proportion of households with tap water (73% versus 53%; SD = 47), latrines (61% versus 53%; SD = 18), and electricity (61% versus 45%; SD = 32). After restriction and matching, balance improved for nearly all covariates, however, the balance of private latrine ownership worsened after the selection (SD increased from 18 to 37).

3.4.2 Population characteristics

Interviewers visited a total of 660 households across 30 villages. Of these, 60 (9.0%) households refused to participate in the study: 27 (8.3%) in intervention communities and 33 (9.9%) in control communities. The final sample included 600 households, 929 children under age five and 1,858 child-days of observation after restricting the recall period to 2 days. Fieldworkers obtained complete anthropometric measurements for 872 children in the sample.

Consistent with pre-intervention conditions, intervention and control villages remained well balanced across a wide range of potentially confounding variables in 2007 (Table 3.4). Water sources were balanced across the two groups: 67% of control households and 66% of intervention households had private taps; 19% of intervention households and 15% of control households used springs or other surface water as their primary water source. We did observe some imbalance in socioeconomic covariates: households in intervention villages were more likely to have electricity (55% versus 44%), a mobile phone (35% versus 28%), a bicycle (15% versus 7%) and a latrine (64% versus 55%). Overall, this suggests that if anything intervention villages are quantitatively better of than control villages. In intervention villages 147 (49%) of study households reported participating in the CRS/Caritas intervention.

Table 3.3: Summary of pre-intervention characteristics before and after village selection, Camotán, Guatemala, Census 2002

Mean	All Villages			Study Sample		
	Control	Inter-vention	Std. Diff.*	Control	Inter-vention	Std. Diff.*
Male (%)	51.2	50.7	-1.0	52.2	51.6	-1.3
Children < 5 (%)	18.8	19.5	1.7	19.5	19.4	-0.2
Age 7 - 14 that work (%)	11.1	11.1	0.1	12.5	11.3	-3.7
Female literacy (%)	42.4	41.8	-1.2	38.2	41.6	7.0
Work in agriculture (%)	85.8	89.6	12.5	88.1	90.5	8.0
Total households	54.5	91.0	66.6	85.8	97.9	28.3
Houses with tap water (%)	52.5	73.3	47.0	59.8	67.8	17.2
Houses with latrine (%)	52.6	61.3	17.8	45.4	63.3	37.2
Houses with electricity (%)	45.3	60.8	31.8	46.9	55.1	16.5
Houses with soil floors (%)	75.0	80.5	14.0	79.2	79.4	0.5
Houses with thatched roofs (%)	44.8	49.8	9.8	47.6	48.9	2.8
Dist. to municipal center (km)	18.9	16.2	-28.0	19.0	15.7	-32.7
Number of Villages	58	30		15	15	
Number of Households	3160	2731		1287	1469	

* The standardized difference is equal to the difference in standard deviations of the mean (I–C) multiplied by 100. It is calculated as $(\mu_I - \mu_C) \div [(S_I^2 + S_C^2)/2]^{1/2} \times 100$. For example, a difference of 1 SD is equal to 100. A value of zero indicates equality of the means.

Table 3.4: Summary of post-intervention characteristics and measures of balance in the study population, Camotán, Guatemala, 2007

Mean	Control	Intervention	Standardized Difference *	Variance Ratio †
Child characteristics	N=455	N=474		
Male (%)	50.3	54.4	8.21	0.99
Age (months)	30.5	28.2	-13.56	0.98
Total months breastfed	16.4	15.1	-17.22	0.91
Mother's characteristics	N=300	N=300		
Age (years)	30.5	29.7	-9.64	0.83
Currently pregnant (%)	11.0	12.7	5.15	1.13
Works for money (%)	11.3	16.0	13.59	1.34
Literate (%)	25.0	28.7	8.27	1.09
Leaves village $\geq$ once/week (%)	38.5	28.1	-10.90	0.84
Household characteristics	N=300	N=300		
Total persons living in home	6.8	7.0	8.41	1.25
Num. children $<$ 15 years	3.7	4.0	11.74	1.26
Num. children $<$ 5 years	1.5	1.6	10.00	1.05
Electricity (%)	43.7	55.3	23.46	1.00
Dirt floor (%)	67.0	61.7	-11.13	1.07
Thatch roof (%)	25.3	25.0	-0.77	0.99
Home ownership (%)	95.3	92.7	-11.23	1.53
Land ownership (%)	51.3	49.7	-3.33	1.00
Use banking services (%)	8.7	7.3	-4.91	0.86
Have relatives in USA (%)	13.3	11.0	-7.13	0.85
Have relatives in the Capital (%)	17.0	16.3	-1.79	0.97
Durable good ownership (%)				
Refrigerator	9.7	6.3	-12.29	0.68
Radio	81.0	83.7	6.98	0.89
Television	14.3	13.3	-2.89	0.94
Mobile phone	28.3	35.3	15.04	1.13
Bicycle	7.0	15.0	25.74	1.96
Automobile	3.7	4.0	1.73	1.09

Table 3.4 – continued on next page

Table 3.4 – continued from previous page

Mean	Control	Intervention	Standardized Difference *	Variance Ratio †
Water supply				
Primary water source (%)				
Private tap	67.3	66.0	-2.82	1.02
Public tap	8.7	9.0	1.17	1.03
Public well	8.0	4.7	-13.70	0.60
Spring	12.3	10.7	-5.22	0.88
Surface water (river/lake)	3.0	8.3	23.18	2.63
Other	0.7	1.3	6.69	1.99
Minutes per day retrieving water	26.9	29.5	3.84	1.09
Satisified with water quantity (%)	81.0	73.3	-18.31	1.27
Sanitation				
Latrine ownership (%)	55.3	64.0	17.71	0.93
Animals in living area (%)				
Pigs	42.3	56.0	27.55	1.01
Chickens/ducks	94.7	95.0	1.50	0.94
Dogs/cats	92.7	93.3	2.61	0.92
Cows/horses/mules/donkeys	14.3	9.0	-16.64	0.67

* The standardized difference is equal to the difference in standard deviations of the mean (I–C) multiplied by 100. It is calculated as $(\mu_I - \mu_C) \div [(S_I^2 + S_C^2)/2]^{1/2} \times 100$. For example, a difference of 1 SD is equal to 100. A value of zero indicates equality of the means.

† The variance ratio is equal to S_I^2/S_C^2.

3.4.3 Water quality samples

Of the 48 stored water samples that we analyzed, nearly all contained E. coli: only 2 (4%) samples had MPN <1 per 100 ml, and the mean (SD) log10 E. coli concentration per 100 ml was 1.975 (0.870) in the control and 2.292 (1.033) in the intervention group. Although *E. coli* concentrations were slightly higher on average in intervention villages, the means were not statistically significantly different (two-sample t-test for difference in means: $t_{46} = -1.15, p = 0.256$) and their distributions were similar (Figure 3.1).

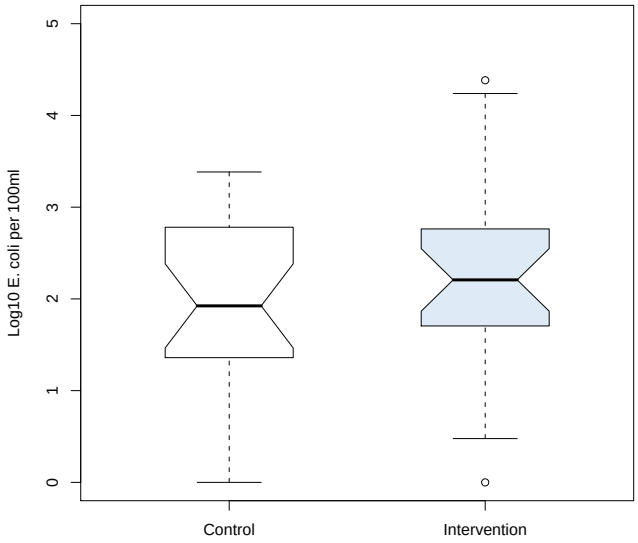


Figure 3.1: Notched box plot of $\log_{10}$ *E. coli* concentrations per 100 ml in household water samples from intervention (n=24) and control (n=24) households. The two samples are not statistically different because the notches overlap.

3.4.4 Knowledge and practice outcomes

Overall, 85% of study households were satisfied with their drinking water quality, but only 65% of respondents believed that their drinking water was clean. Among families that reported treating their water, the main motivating reasons they gave for treating their water included: water is contaminated (51%), improves health (34%) and improves the taste (6%). The primary reason families gave for not treating their water was that it was already clean (48%) followed by: bad taste (14%), not interested (11%) and no time (7%). Among families that reported treating their water, perceived water treating efficacy and ease of use were the most important factors contributing to their choice of water treatment method (Table 3.4.4). Cost appears to be a strong factor for families choosing to use SODIS (18% vs. only 4% among families boiling water or using chlorine). A majority of families that use chlorine are motivated by its ease of use (54%, Table 3.4.4).

Table 3.5: Reasons for using different household water treatment methods among self-reported users (responses to an open-ended question: *Why do you like to use that method?*).

Reason for treating water	Any Method (%)	Boiling (%)	SODIS (%)	Chlorine (%)
Cleans water the best	28	33	26	35
Easiest to use	37	34	37	54
Has the best taste	15	15	18	8
Least expensive	6	4	18	4
Only option / do not know others	12	16	8	4
Other reasons / do not know	4	5	3	0
N	163	110	39	26

In both intervention and control households there was a large discrepancy between self-reported water treatment behavior, and confirmed water treatment behavior (Figure 3.4.4). Both self-reported water treatment and confirmed water treatment behavior was substantially lower than the 70% of families using some water treatment method reported by the Caritas and the SODIS Foundation at the completion of the intervention, six months prior to our survey.

Intervention households were more likely to treat their water than control households based on self-reported activity (33.3% versus 21.0%; Risk Difference (RD) = 0.12, 95% CI 0.01 – 0.24), and based on confirmed water treatment activity at the time of the visit (8.7% versus 3.3%; RD = 0.05, 0.01 – 0.10) (Table 3.6).

Overall, 509 (85%) of the 600 households in the study reported receiving information about handwashing at some point in the previous three years. Table 3.4.4 summarizes

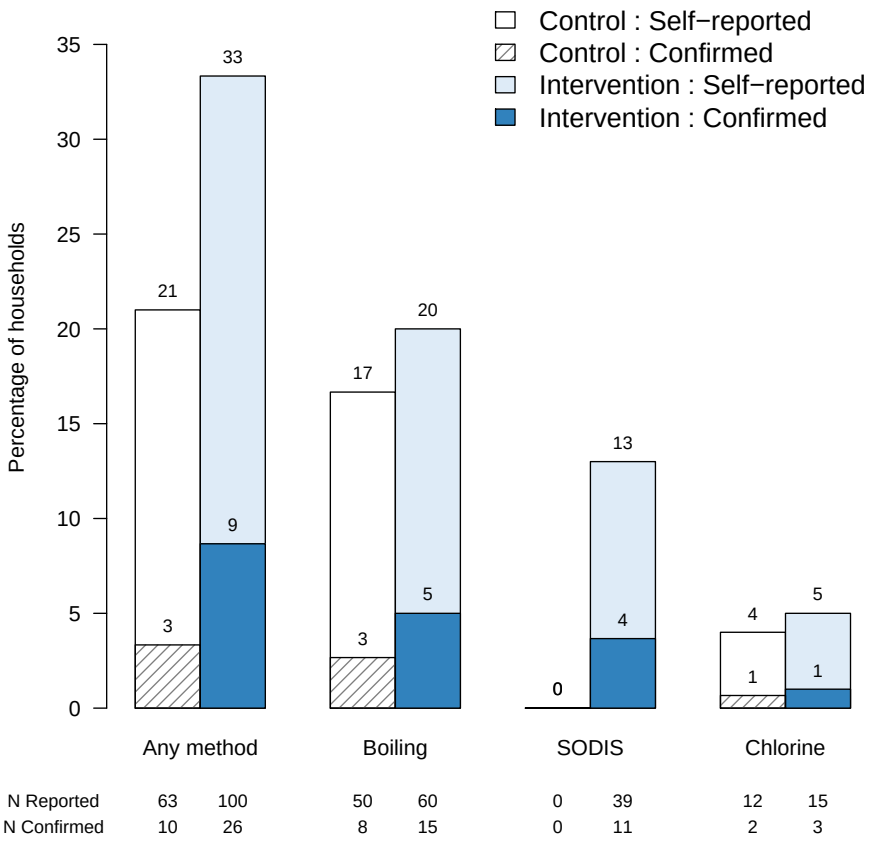


Figure 3.2: Summary of self-reported and confirmed water treatment behavior by type of treatment. Detailed definitions and differences between groups with 95% confidence intervals are reported in Table 3.6 on page 97.

Table 3.6: Water storage and treatment practices following a three-year point-of-use water treatment and handwashing intervention, Camotán, Guatemala, 2007. N=300 intervention and N=300 control households.

Outcome	Control		Intervention		Risk Difference	
	%	(N)	%	(N)	(95% CI) *	
Water storage practices						
Stores drinking water in home	80.3	(241)	80.7	(242)	0.003	(-0.07, 0.08)
Excl. covered or narrow mouth	60.7	(182)	62.7	(188)	0.020	(-0.06, 0.10)
Exclusively covered	58.0	(174)	62.3	(187)	0.043	(-0.04, 0.12)
Self-reported water treatment						
Any method	21.0	(63)	33.3	(100)	0.123	(0.00, 0.24)
Boiling	16.7	(50)	20.0	(60)	0.033	(-0.07, 0.14)
SODIS †	0.0	(0)	13.0	(39)	0.130	(0.07, 0.19)
Chlorine	4.0	(12)	5.0	(15)	0.010	(-0.02, 0.04)
Confirmed water treatment ‡						
Any method	3.3	(10)	8.7	(26)	0.053	(0.02, 0.09)
Boiling	2.7	(8)	5.0	(15)	0.023	(0.00, 0.05)
SODIS	0.0	(0)	3.7	(11)	0.037	(0.01, 0.06)
Chlorine	0.7	(2)	1.0	(3)	0.003	(-0.01, 0.02)

* Standard Errors and 95% confidence intervals calculated by bootstrap resampling matched village pairs with 1000 iterations.

† SODIS: Solar Disinfection.

‡ Water treatment was confirmed if the family (i) self-reported treating water in the previous 7 days, (ii) had treated water at the time of the interview, and (iii) could produce the materials they used to treat water.

responses to an open-ended question about sources of handwashing information. Households reported receiving handwashing information from friends and family (32%) as often as from their local health post (33%) with virtually no difference between intervention and control villages (Table 3.4.4). Households in intervention villages reported receiving information from health promotors more often than households in control villages (46% v.s 35%), but the data suggest that handwashing promotion activities are common in the entire study population.

We did not observe differences between intervention and control groups in self-reported handwashing behavior, or spot check observations of hygienic conditions (Table 3.8). Both groups reported washing hands frequently before cooking (77% in intervention, 81% in control), but less frequently around contact with feces: 51% in intervention and 52% in control reported washing hands after defecation, and just 13% of mothers in both groups reported washing hands after changing their baby (Table 3.8). Soap was present in most homes (90%), but similar in intervention and control villages (RD = 0.03, -0.05 – 0.11). Interviewers observed animal or human feces in the vast majority of households in both intervention (77%) and control (70%).

Table 3.7: Self-reported handwashing information sources.

Information source	Total (%)	Control (%)	Intervention (%)
Health promoter	41	35	46
Local health post	33	34	32
Private doctor	0	0	0
Friends and family	32	33	30
Radio	3	4	3
Other	5	6	5
N	509	250	259

3.4.5 Child health outcomes

Model selection

The model selection process successfully identified covariates to use in adjusted specifications, summarized in Table 3.9. In no case did the D/S/A algorithm select interactions of covariates in W with the treatment.

The predicted probabilities from $g^0(A|W)$ indicate that there is common support for the covariates selected in W. The probabilities $g^0(1|W) = P(A = 1|W)$ are bounded away from 0 and 1, and range from 0.19 to 0.94 (median = 0.49, interquartile range = 0.44 – 0.69). Figure 3.3 plots a histogram of the predicted probabilities of receiving the intervention for control and intervention children, and demonstrates good overlap in the distributions. This result helps confirm the usefulness of carefully selecting control villages in the design stage, and indicates that our parameters of interest (equation 3.2) are well-defined in this dataset.

Figure 3.3: Histogram of predicted probabilities of receiving treatment, $g^0(1|W) = P(A = 1|W)$, for children in intervention and control villages using the specification in Table 3.9. Bin width is 0.05.

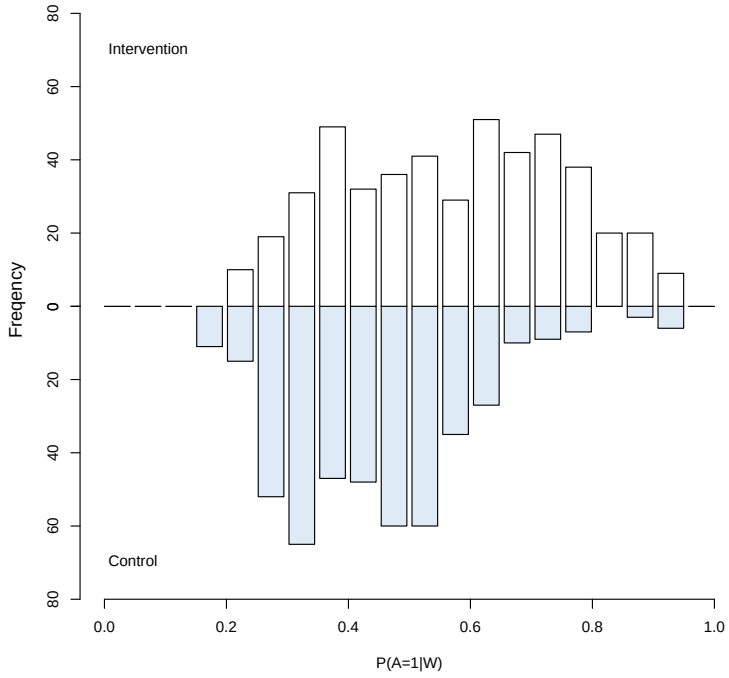


Table 3.8: Handwashing and hygiene conditions following a three-year point-of-use water treatment and handwashing intervention, Camotán, Guatemala, 2007. N=300 intervention and N=300 control households.

Outcome	Control		Intervention		Risk Difference	
	%	(N)	%	(N)	(95% CI) *	
Self-reported handwashing †						
Before cooking	81.0	(243)	77.3	(232)	-0.037	(-0.11, 0.04)
Before eating	33.3	(100)	33.7	(101)	0.003	(-0.09, 0.09)
Before feeding children	20.3	(61)	16.3	(49)	-0.040	(-0.14, 0.06)
After defecation	52.3	(157)	50.7	(152)	-0.017	(-0.12, 0.09)
After changing baby	12.7	(38)	12.7	(38)	0.000	(-0.10, 0.10)
Spot check observations						
Mother's hands are clean	90.3	(271)	89.0	(267)	-0.013	(-0.07, 0.04)
Mother's nails are clean	73.3	(220)	72.0	(216)	-0.013	(-0.10, 0.08)
Can produce a bar of soap	88.7	(266)	91.7	(275)	0.030	(-0.05, 0.11)
Bar soap is in plain view	56.7	(170)	59.0	(177)	0.023	(-0.07, 0.12)
Food is covered	53.3	(160)	55.7	(167)	0.023	(-0.11, 0.16)
Garbage present inside home	57.7	(173)	47.3	(142)	-0.103	(-0.23, 0.02)
Feces observed in living area	70.3	(211)	77.0	(231)	0.067	(-0.05, 0.18)

* Standard Errors and 95% confidence intervals calculated by bootstrap resampling matched village pairs with 1000 iterations.

† Responses to an open-ended question about handwashing in the 24 hours before the interview.

Acute health outcomes

Acute gastrointestinal illness and respiratory infections were prevalent in the study population, and all acute illness outcomes were slightly more prevalent in the intervention group (Figure 3.4). There was large variability in the village level prevalence of diarrhea, ranging from 1.9% to 28.6% (Figure 3.5). Some of the variability across villages is likely due to the relatively small number of children per village (mean=31), and the cross-sectional measurement. The between-village variability would likely shrink with repeated measures.

In children under five, the daily longitudinal prevalence of diarrhea and HCGI during the measurement period was 11.9% and 12.6%, respectively. Intervention and control groups did not differ in diarrhea (Longitudinal Prevalence Difference (LPD) = 0.004, 95%

Table 3.9: Summary of covariates included in targeted maximum likelihood estimation models for child health outcomes (abbreviations are on page 88).

Model, outcome	Covariates included in W
$g^0(A\|W)$	ttime, ttime^2, bike, cell, elec, latrine, watsource, pigs, stock
$Q^0(A,W)$, Diarrhea	age, bftot, dirt, watsource, age*watsource(public well), bftot*dirt, bftot*age
$Q^0(A,W)$, HCGI *	age, age^2, watsource, latrine, refri
$Q^0(A,W)$, Cough or diff. breathing	age
$Q^0(A,W)$, Congestion or coryza	age, age^2, dirt
$Q^0(A,W)$, ALRI †	mlit
$Q^0(A,W)$, Weight	age, age^2, tv, elec, palm, relus, watsource
$Q^0(A,W)$, Height	age, age^2, dirt, relguat
$Q^0(A,W)$, Weight-for-height	age, tv
$Q^0(A,W)$, Mid-upper arm circ.	age, age^2, tv, dirt, watsource

*Highly Credible Gastrointestinal Illness. The definition is included in the main text.

† Clinical Acute Lower Respiratory Infections. The definition is included in the main text.

Figure 3.4: Longitudinal prevalence of acute illness in 929 children under age 5. Camotán, Guatemala 2007. Table 3.10 includes symptom definitions.

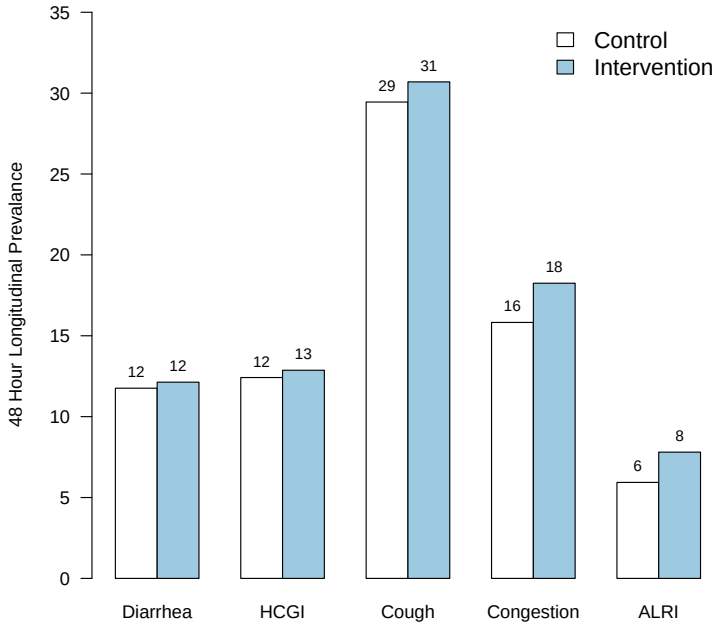
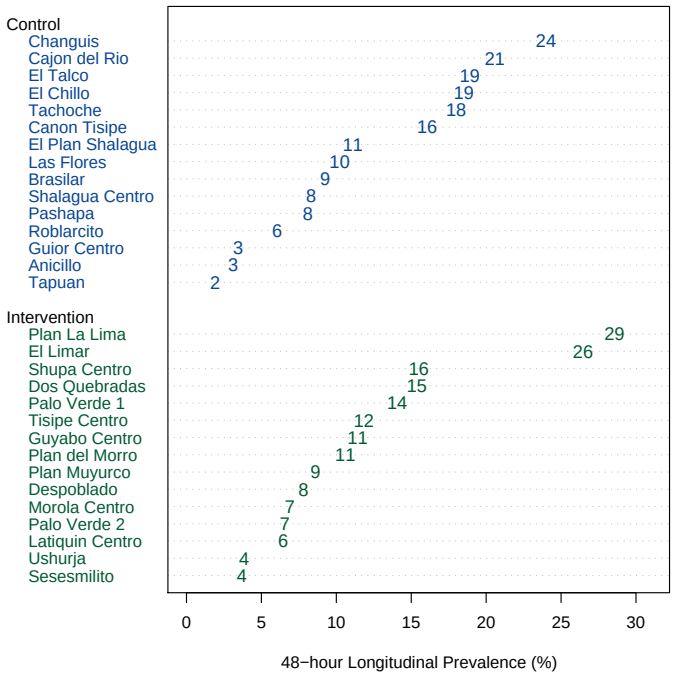


Figure 3.5: Longitudinal prevalence of diarrhea in the 30 study villages. Camotán, Guatemala 2007.



CI –0.051−0.058) or HCGI (LPD = 0.005, –0.054−0.065) (Table 3.10). Respiratory illness was common among children in the study: the daily longitudinal prevalence of cough or difficulty breathing was 30.0% and clinical ALRI was 6.9%. We observed no differences between the intervention and control groups in the longitudinal prevalence of cough or difficulty breathing (LPD = 0.012, –0.097−0.137) or ALRI (LPD = 0.019, –0.028−0.078) (Table 3.10). Adjusted estimates using targeted MLE did not differ from the unadjusted estimates, although the standard errors are 12% and 9% smaller for diarrhea and HCGI, respectively (Table 3.12).

Child growth

Study children were generally well-nourished but, consistent with our acute self-reported health outcomes, we found no differences in anthropometric measures between children living in intervention and control villages (Figure 3.6, Table 3.11). Differences in Z-score means for height, weight, height-for-weight and mid-upper-arm circumference were all less than 0.07 standard deviations. Adjustment for a large set of potential confounding variables using targeted maximum likelihood did not change the unadjusted results, and

if anything suggest that child growth in intervention villages lags slightly behind child growth in control villages. Adjusted estimates have between 13% and 43% smaller standard errors than the unadjusted estimates for child growth outcomes (Table 3.12).

Table 3.10: Unadjusted and adjusted difference in longitudinal prevalence of illness in 929 children under age 5 following a three-year household water treatment and handwashing intervention, Camotán, Guatemala, 2007. Adjusted values were estimated using targeted maximum likelihood.

Outcome	Control	Intervention	Unadjusted			Adjusted		
	Days Ill / Observed	Days Ill / Observed	LPD	SE *	(95% CI) *	LPD	SE *	(95% CI) *
Diarrhea	107/910	115/948	0.004	0.0288	(-0.051, 0.058)	0.007	0.0254	(-0.037, 0.059)
HCGI †	113/910	122/948	0.005	0.0308	(-0.054, 0.065)	0.010	0.0282	(-0.042, 0.068)
Cough or diff. breathing	268/910	291/948	0.012	0.0597	(-0.097, 0.137)	0.003	0.0592	(-0.111, 0.117)
Congestion or coryza	144/910	173/948	0.024	0.0249	(-0.026, 0.071)	0.023	0.0249	(-0.022, 0.075)
ALRI ‡	54/910	74/948	0.019	0.0278	(-0.028, 0.078)	0.018	0.0285	(-0.031, 0.077)

∗ Standard Errors and 95% confidence intervals calculated by bootstrap resampling matched village pairs with 1000 iterations.
† Highly Credible Gastrointestinal Illness. The definition is included in the main text.

‡ Clinical Acute Lower Respiratory Infections. The definition is included in the main text.

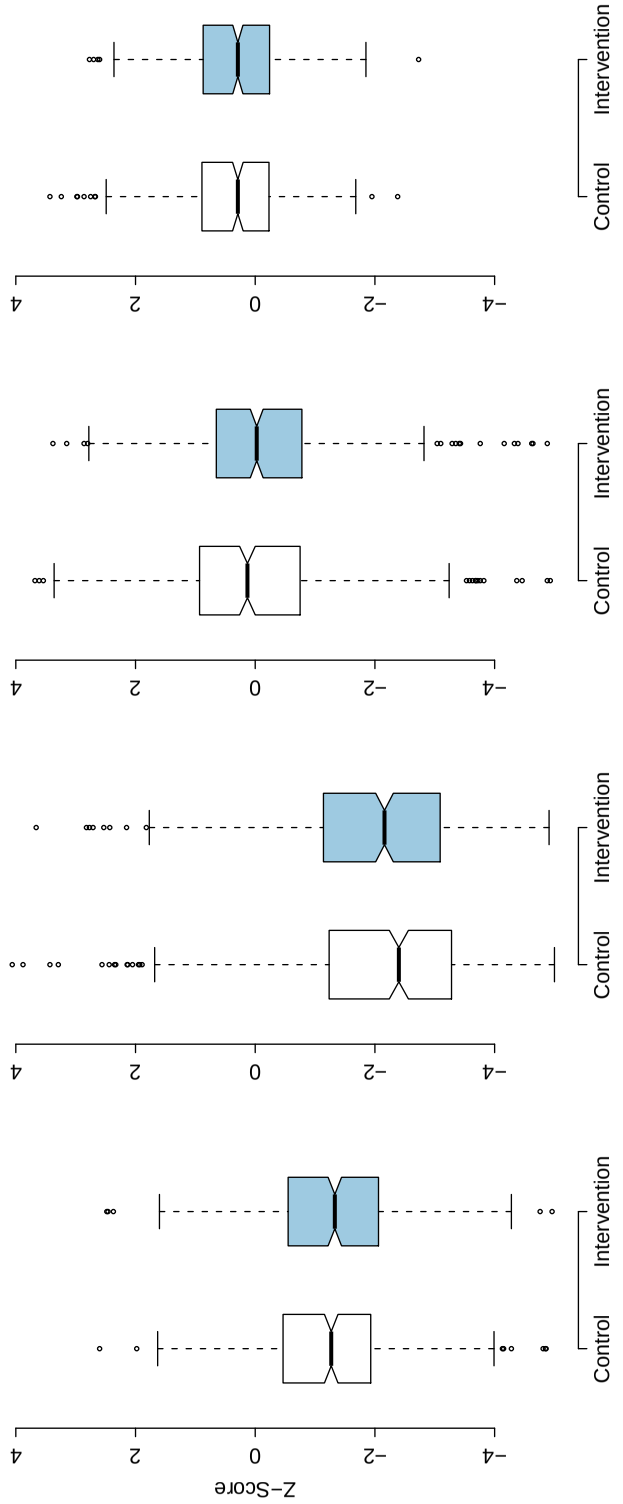


Figure 3.6: Z-scores distributions for child growth measures in 872 children under age five. Z-scores were calculated using the 2006 WHO International Growth Standards [26]. Camotán, Guatemala 2007.

Table 3.11: Unadjusted and adjusted difference in anthropometric Z-scores in children under age 5 following a three-year household water treatment and handwashing intervention, Camotán, Guatemala, 2007. Adjusted values were estimated using targeted maximum likelihood.

Z-Score *	Control			Intervention			Unadjusted			Adjusted		
	N	Mean	SD	N	Mean	SD	Diff.	SE	(95% CI) †	Diff.	SE	(95% CI) †
Weight	423	-1.312	1.325	453	-1.365	1.219	-0.053	0.1368	(-0.331, 0.206)	-0.111	0.0768	(-0.254, 0.050)
Height	424	-2.177	1.880	453	-2.136	1.596	0.041	0.1605	(-0.305, 0.326)	-0.055	0.1338	(-0.332, 0.177)
Weight-for-height	421	-0.122	1.728	451	-0.187	1.421	-0.066	0.0967	(-0.248, 0.124)	-0.019	0.0837	(-0.174, 0.145)
Mid-upper-arm circ.	401	0.348	0.884	426	0.335	0.825	-0.014	0.0806	(-0.166, 0.145)	-0.057	0.0657	(-0.183, 0.079)

∗ Z-scores were calculated using a standard WHO Stata algorithm and 2006 world reference data.
† Standard Errors and 95% confidence intervals calculated by bootstrap resampling matched village pairs with 1000 iterations.

Table 3.12: Relative efficiency of unadjusted and targeted maximum likelihood estimation (t-MLE) for treatment effects. The relative efficiency is calculated as the percentage change from the unadjusted to the t-MLE standard errors.

Outcome	SE (Unadj.)	SE (t-MLE)	Rel. Eff (%)
Diarrhea	0.0288	0.0254	-11.8
HCGI *	0.0308	0.0282	-8.6
Cough or diff breathing	0.0597	0.0592	-0.9
Congestion or coryza	0.0249	0.0249	0.1
ALRI †	0.0278	0.0285	2.4
Weight ‡	0.1368	0.0768	-43.9
Height ‡	0.1605	0.1338	-16.6
Weight-for-height ‡	0.0967	0.0837	-13.4
Mid-upper-arm circ. ‡	0.0806	0.0657	-18.4

* Highly Credible Gastrointestinal Illness. The definition is included in the main text.

† Clinical Acute Lower Respiratory Infections. The definition is included in the main text.

‡ Z-score

Intra-cluster correlation estimation

We estimated the intra-cluster correlation (ICC) of binary child health outcomes using the `aod` package in R with restricted maximum likelihood (REML) and 1000 Monte Carlo replicates to estimate the 95% confidence intervals. We estimated the intra-cluster correlation of continuous Z-scores using the `loneway` command in Stata (www.stata.com). Table 3.13 summarizes the household- and village-level ICCs. Respiratory outcomes are highly correlated within households (ICC range: 0.25 - 0.39), as are child Z-scores (ICC range 0.05 - 0.29). Diarrhea (ICC=0.07) and HCGI (ICC=0.08) were less-correlated within households than the other child health outcomes. As expected, all outcomes were less correlated at the village level than at the household level.

3.4.6 Re-analysis using self-reported participation as treatment

A re-analysis of the data that defined the 147 (49%) intervention households who reported participating in the intervention as the treatment group did not alter our findings. Like the primary analysis, intervention households were more likely to practice water treatment (9.5% vs. 4.9%, risk difference = 0.05, 95% CI $0.00-0.10$) (Table 3.14), but did not report

Table 3.13: Household- and village-level intra-cluster correlation estimates. The 95% confidence intervals for binary outcomes are based on Monte Carlo simulation. The 95% confidence intervals for Z-scores are based on asymptotic standard errors.

Outcome	Household-Level		Village-Level	
	ICC	(95% CI)	ICC	(95% CI)
Diarrhea	0.073	(0.016, 0.278)	0.015	(0.003, 0.063)
HCGI *	0.084	(0.022, 0.257)	0.026	(0.009, 0.076)
Cough or diff. breathing	0.389	(0.307, 0.475)	0.084	(0.043, 0.163)
Congestion or coryza	0.250	(0.166, 0.358)	0.015	(0.003, 0.078)
ALRI †	0.342	(0.260, 0.446)	0.055	(0.026, 0.123)
Weight ‡	0.293	(0.185, 0.401)	0.061	(0.015, 0.107)
Height ‡	0.176	(0.057, 0.295)	0.053	(0.010, 0.096)
Weight-for-height ‡	0.052	(0.000, 0.182)	0.012	(0.000, 0.035)
Mid upper arm circ ‡	0.291	(0.175, 0.406)	0.051	(0.008, 0.094)

* Highly Credible Gastrointestinal Illness. The definition is included in the main text.
† Clinical Acute Lower Respiratory Infections. The definition is included in the main text.
‡ Z-scores

washing their hands more at critical times (Table 3.15). Unlike the primary analysis, households who reported participating in the intervention were more likely to have soap in their home (95.9% vs. 88.3%, risk difference = 0.08, 95% CI $0.04 - 0.12$).

Table 3.14: Water storage and treatment practices following a three-year point-of-use water treatment and handwashing intervention, Camotán, Guatemala, 2007. N=147 intervention and N=453 control households. Unlike the primary analysis, intervention treatment is assigned based on self-reported participation in the intervention.

Outcome	Control		Intervention		Risk Difference		
	%	(N)	%	(N)	(95% CI) *		
Water storage practices							
Stores drinking water in home	80.4	(364)	81.0	(119)	0.006	(-0.07, 0.08)	
Excl. covered or narrow mouth	61.6	(279)	61.9	(91)	0.003	(-0.09, 0.10)	
Exclusively covered	59.8	(271)	61.2	(90)	0.014	(-0.08, 0.09)	
Self-reported water treatment							
Any method	23.8	(108)	37.4	(55)	0.136	(0.05, 0.23)	*
Boiling	17.9	(81)	19.7	(29)	0.018	(-0.06, 0.10)	
SODIS †	1.1	(5)	23.1	(34)	0.220	(0.15, 0.29)	
Chlorine	5.1	(23)	2.7	(4)	-0.024	(-0.06, 0.01)	
Confirmed water treatment ‡							
Any method	4.9	(22)	9.5	(14)	0.047	(0.00, 0.10)	
Boiling	3.8	(17)	4.1	(6)	0.003	(-0.03, 0.04)	
SODIS	0.4	(2)	6.1	(9)	0.057	(0.02, 0.10)	
Chlorine	1.1	(5)	0.0	(0)	-0.011	(-0.02, 0.00)	

95% Confidence Intervals calculated by bootstrap resampling households with 1000 iterations.

† SODIS: Solar Disinfection.

‡ Water treatment was confirmed if the family (i) self-reported treating water in the previous 7 days, (ii) had treated water at the time of the interview, and (iii) could produce the materials they used to treat water.

Table 3.15: Handwashing and hygiene conditions following a three-year point-of-use water treatment and handwashing intervention, Camotán, Guatemala, 2007. N=147 intervention and N=453 control households. Unlike the primary analysis, intervention treatment is assigned based on self-reported participation in the intervention.

Outcome	Control		Intervention		Risk Difference	
	%	(N)	%	(N)	(95% CI) *	
Self-reported handwashing †						
Before cooking	77.7	(352)	83.7	(123)	0.060	(-0.02, 0.13)
Before eating	35.1	(159)	28.6	(42)	-0.065	(-0.15, 0.02)
Before feeding children	18.8	(85)	17.0	(25)	-0.018	(-0.09, 0.06)
After defecation	52.3	(237)	49.0	(72)	-0.033	(-0.12, 0.06)
After changing baby	12.4	(56)	13.6	(20)	0.012	(-0.05, 0.08)
Spot check observations						
Mother's hands are clean	90.1	(408)	88.4	(130)	-0.016	(-0.08, 0.04)
Mother's nails are clean	72.2	(327)	74.1	(109)	0.020	(-0.06, 0.10)
Can produce a bar of soap	88.3	(400)	95.9	(141)	0.076	(0.04, 0.12)
Bar soap is in plain view	56.1	(254)	63.3	(93)	0.072	(-0.02, 0.16)
Food is covered	53.6	(243)	57.1	(84)	0.035	(-0.05, 0.12)
Garbage present inside home	54.1	(245)	47.6	(70)	-0.065	(-0.16, 0.03)
Feces observed in living area	73.5	(333)	74.1	(109)	0.006	(-0.08, 0.08)

* 95% Confidence Intervals calculated by bootstrap resampling households with 1000 iterations.

† Responses to an open-ended question about handwashing in the 24 hours before the interview.

As described in the methods, we repeated all model selection exercises and adjusted analyses using targeted MLE for child health outcomes. The predicted probability of receiving treatment using self-reported participation as the outcome was well distributed in the two groups, however there is a region of the distribution in the treatment group with no support in the control group (for $\mathrm{P}(A = 1|W) > 0.8$, Figure 3.4.6). Although this does not completely invalidate the adjusted analyses, it does suggest that there may be some bias.

On average, children in intervention households had higher prevalence of acute illness symptoms, though we did not identify any differences that were statistically significant at the 95% confidence level. There was no detectable difference in the longitudinal prevalence of diarrhea (adjusted longitudinal prevalence difference (aLPD) = 0.012, 95% CI -0.032 − 0.060) or HCGI (aLPD = 0.013, 95% CI -0.041 − 0.068). Similarly, we did not identify detectable differences for acute respiratory symptoms (Table 3.16), or child growth measures (Table 3.17) using the alternate treatment definition.

Figure 3.7: Histogram of predicted probabilities of receiving treatment, $g^0(1|W) = P(A = 1|W)$, for children in intervention and control villages. Unlike the primary analysis, intervention status was assigned based on self-reported participation. The predicted probabilities range from 0.21 to 0.88 (median = 0.52). Bin width is 0.05.

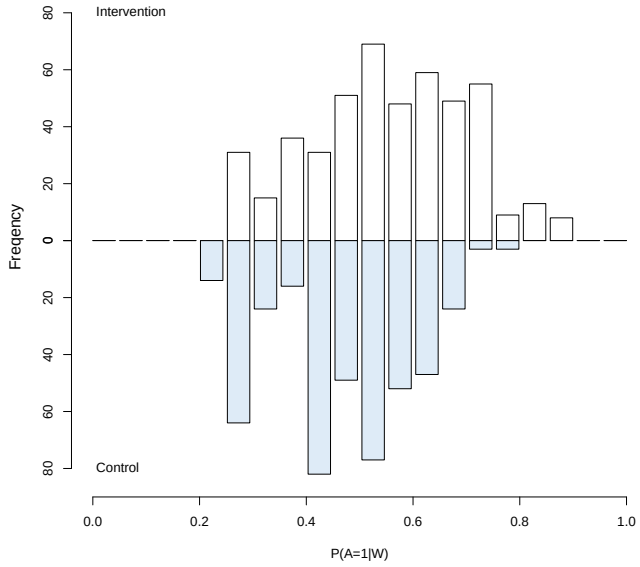


Table 3.16: Unadjusted and adjusted difference in longitudinal prevalence of illness in 929 children under age 5 following a three-year household water treatment and handwashing intervention, Camotán, Guatemala, 2007. Adjusted values were estimated using targeted maximum likelihood.Unlike the primary analysis, intervention treatment is assigned based on self-reported participation in the intervention.

Outcome	Control	Intervention	Unadjusted			Adjusted		
	Days Ill / Observed	Days Ill / Observed	LPD	SE *	(95% CI) *	LPD	SE *	(95% CI) *
Diarrhea	174/1404	48/454	-0.018	0.0242	(-0.064, 0.028)	0.012	0.0246	(-0.032, 0.060)
HCGI †	186/1404	49/454	-0.025	0.0238	(-0.069, 0.022)	0.013	0.0281	(-0.041, 0.068)
Cough or diff. breathing	405/1404	154/454	0.051	0.0395	(-0.029, 0.130)	0.003	0.0604	(-0.112, 0.119)
Congestion or coryza	220/1404	97/454	0.057	0.0346	(-0.012, 0.130)	0.034	0.0258	(-0.015, 0.087)
ALRI ‡	83/1404	45/454	0.040	0.0247	(-0.005, 0.090)	0.018	0.0285	(-0.031, 0.076)

∗ 95% Confidence Intervals calculated by bootstrap resampling households with 1000 iterations.
† Highly Credible Gastrointestinal Illness. The definition is included in the main text.

‡ Clinical Acute Lower Respiratory Infections. The definition is included in the main text.

Table 3.17: Unadjusted and adjusted difference in anthropometric Z-scores in children under age 5 following a three-year household water treatment and handwashing intervention, Camotán, Guatemala, 2007. Adjusted values were estimated using targeted maximum likelihood.Unlike the primary analysis, intervention treatment is assigned based on self-reported participation in the intervention.

Z-Score *	Control			Intervention			Unadjusted			Adjusted		
	N	Mean	SD	N	Mean	SD	Diff.	SE	(95% CI) †	Diff.	SE	(95% CI) †
Weight	659	-1.345	1.291	217	-1.322	1.208	0.024	0.1070	(-0.195, 0.236)	-0.071	0.0968	(-0.264, 0.116)
Height	659	-2.192	1.755	218	-2.048	1.686	0.144	0.1472	(-0.146, 0.434)	-0.094	0.1156	(-0.322, 0.136)
Weight-for-height	656	-0.149	1.646	216	-0.175	1.347	-0.026	0.1244	(-0.276, 0.204)	-0.037	0.0823	(-0.197, 0.111)
Mid-upper-arm circ.	623	0.341	0.853	204	0.342	0.856	0.000	0.0701	(-0.136, 0.143)	-0.061	0.0671	(-0.178, 0.084)

∗ Z-scores were calculated using a standard WHO Stata algorithm and 2006 world reference data.
† 95% Confidence Intervals calculated by bootstrap resampling households with 1000 iterations.

3.5 Discussion

Results in context

To our knowledge this is the second [36] post-intervention follow-up study of a combined household water treatment and handwashing behavior change intervention, and the first to extend propensity score matching and targeted maximum likelihood estimation to the design and analysis of a pre-existing intervention. The absence of child health impacts is consistent with the modest improvement we observed in water treatment behavior (Table 3.6), no detectable differences in handwashing behavior, and highly contaminated living environments (Table 3.8). These findings are consistent with efficacy trials of household water treatment that have found that health impacts are contingent on compliance [37, 38].

Our health findings are also consistent with a recently published sustainability study of a combined household water treatment and handwashing intervention in Pakistan [36, 39]. The sustainability assessment focused on the handwashing component of the intervention, and found 18 months after the conclusion of intervention activities that intervention households 1.5 times more likely to have a place with soap and water to wash hands (79% vs. 53%) and were 2.2 times more likely to demonstrate correct handwashing behavior (50% vs. 23%) compared to controls. Despite these improved behaviors, there was no difference between intervention and control groups with in regard to soap purchases (2.3 vs. 2.2 bars per week) or child diarrhea (longitudinal prevalence difference = -0.0015 [-0.0092, 0.0061]).

Our confirmed water treatment adoption in intervention households (9%) is lower than water treatment adoption reported after a CARE/Madagascar Safe Water System (SWS) campaign, which promoted chlorine treatment with safe storage. Ram *et al.* found that 54% (29/54) of households had detectable free chlorine in their stored water 18 months after the campaign [40]. Parker *et al.* also report higher sustained adoption after a clinic-based SWS and handwashing intervention: 71% (36/51) of households had detectable free chlorine one year after the intervention [41]. Our water treatment behavior results are consistent with Luby *et al.* , who found 5% (22/462) of households regularly treating their water six months after the completion of a year-long household flocculent-disinfectant intervention trial in Guatemala [42].

The lack of sustained water treatment behavior is consistent with a recent assessment of household water treatment methods [43], which asserts that SODIS is difficult to use compared to other household water treatment methods (note: the review did not include boiling). Our finding that 210 (48%) of 437 households that did not treat their water believed that the water was already clean underscores the difficulty of promoting household water treatment in populations with at least moderately good access to tap water. In this study population, 75% of households have private or public taps as their primary water source (Table 3.4). Based on our water quality tests, households with private taps do have more clean water than households with other sources, but it is still not clean: the geometric mean *E. coli* concentration of the 35 samples from private taps was 102 per

100 mL, versus a geometric mean of 295 per 100 mL in the 13 samples from other sources ($t_{46} = 1.50, p = 0.14$). Part of the contamination likely occurs during storage: 81% of the population stores their water (32 of the 35 private tap water samples were from stored water), and it is well documented that pathogen concentrations can increase in stored water due to in-home contamination from hands or dirty containers [44, 45]. All of our water samples are taken from household samples and so we cannot elucidate whether the contamination occurs at the source (e.g., from leakage into the reticulation systems) versus in the home (but see Chapter 4 for more detailed analysis of this issue in the Indian context).

Our handwashing and hygiene findings suggest that the presence of soap is common in all villages in the region, but that self-reported handwashing remains infrequent around all key activities except cooking (Table 3.8). This finding contrasts with two earlier studies that report sustained handwashing behavior change many years after short-duration interventions, though neither study included an adequate control group [46, 47]. That we did not observe large differences in self-reported handwashing and observed hygiene practices is not surprising given the apparent existence of additional handwashing and hygiene promotion present in the control villages (Table 3.4.4). Although the Caritas activities did lead to a marginal increase in the number of mothers that report visits by health promoters (46% versus 35% in control villages), this additional information and motivation has not led to sustained behavior change.

Comments on methodology

Our results demonstrate that with available pre-intervention secondary data the careful selection of a study population in the design stage can greatly improve the baseline comparability of intervention and control groups in the evaluation of a pre-existing intervention. Our design was feasible because the implementing organizations provided essential information about the intervention, pre-intervention census data were available, and the organizations used a standardized intervention. Without these conditions, it would be difficult or impossible to identify an adequate control group and define meaningful treatment effects. Prospective, randomized designs have implemented pair matching on one or two variables such as baseline illness or community size to help improve the comparability of treatment arms [48]. The limitation of one- or two-variable matching in non-randomized designs is that implementing organizations usually rely on many (or ill-defined) characteristics to choose intervention recipients, and matching on one or two covariates is unlikely to balance a large set of potential confounders. Propensity score matching simplifies multivariate matching by accommodating continuous covariates and reducing a large set of matching characteristics to a single scalar. Restriction and matching limit inference to the population ultimately included in the study, and the treatment effect estimated is average treatment effect among the treated (not the entire population). However, when interventions are targeted to a subset of the population making inference to segments of

the population that do not share characteristics with those treated must rely on extrapolation beyond the limits of the data [14] (see Chapter 2, Section 2.7.1 for simulation results).

The large difference between self-reported and confirmed water treatment (Table 3.6) suggests that self-reported water treatment behavior overestimates actual practice. Schmidt and Cairncross recently outlined the problems of self-reported health outcomes in non-blinded studies of household water treatment [49]. Our self-reported health outcomes likely suffer from less reporting bias because we do not have frequent, repeated visits, we used a health calendar to collect symptoms, and we minimized recall to 48 hours. Our objective anthropometric outcomes are an important complement to the self-reported outcomes, and the null treatments effect is consistent across all outcomes.

Study limitations

There are limitations to our study. Our design does not include baseline outcome measurement. It is possible that intervention villages were in worse health than controls before the intervention, and that their health improved to control levels by 2007. We think this scenario is unlikely given the limited behavior change we observed and the comparability of intervention and control villages across a broad range of demographic, socioeconomic, and environmental characteristics in both 2002 and 2007.

Second, it is possible that there exists residual confounding that is masking the intervention effect. However, the small covariate imbalances between the groups in 2002 and 2007 suggest that children in intervention villages live in slightly wealthier homes, but experience similar water, sanitation and hygiene conditions to children living in control villages (Tables 3.4.1 and 3.4). If bias from unmeasured confounding exists, we would expect child illness in intervention communities to be biased downward, away from the null. This potential bias would be a larger concern if we observed better child health in intervention villages, and if anything it is worse (Tables 3.10 and 3.11), an unexpected finding that does not provide evidence in support of the intervention.

Third, we only measured outcomes at one point in time, and it is possible that we misclassified families with respect to behavior and illness since these characteristics likely vary over time. For example, families might only treat their water seasonally or when their tap water supply is out of service. We attempted to reduce misclassification by using measures of water treatment and hygiene that did not change rapidly over time, and by supplementing acute child health outcomes with anthropometric measurements.

Fourth, only 49% of intervention households reported participating in the intervention. This modest participation rate may have diluted the treatment effect sufficiently to lead to a null finding with respect to effectiveness but is itself an important finding with respect to future implementation. Comparing the subgroup of participating intervention households to non-participants in unadjusted and adjusted analyses did not change our conclusions (Section 3.4.6).

A final limitation is that our cross-sectional measurement does not ultimately resolve whether the intervention was sustainable. Two scenarios are consistent with our results: (i) the intervention successfully increased water treatment behavior among participating families, but the new behaviors were not sustained after intervention completion, or (ii) the intervention never led to behavior change and there was nothing to sustain. The only available reference point to evaluate these scenarios was an end-of-intervention survey conducted by the implementing organization in which 70% of participating households in our study villages reported consistent household water treatment. While this estimate is likely biased upward, in our survey six months after the intervention 33% of intervention village households self-reported that they treat their water, a measurement prone to similar upward bias (Table 3.8). Taken together, these measurements suggest that water treatment likely tapered off after activities ceased. Future studies could address sustainability more rigorously by collecting measurements at the end of the intervention period followed by identical measures later to capture changes over time.

Conclusion

Six months after the end of a three-year intervention in rural Guatemala we observed minimal sustained water treatment and handwashing behavior of the types promoted by the intervention, which consequently led to no impacts on acute gastrointestinal, respiratory, or anthropometric measures. Our findings highlight the difficulty of achieving sustained new behavior adoption in the context of non-research intervention campaigns. Future research in this sector should focus on identifying techniques to improve and sustain behavior adoption that implementing organizations can use in development programs. Our study design provides a useful template for effectiveness evaluations of pre-existing intervention campaigns initiated outside of formal research activities.

Bibliography

[1] Arnold B, Arana B, Mausezahl D, Hubbard A, Colford J John M. Evaluation of a pre-existing, 3-year household water treatment and handwashing intervention in rural Guatemala. Int J Epidemiol. 2009;p. DOI:10.1093/ije/dyp241.

[2] Esrey SA, Habicht JP, Casella G. The complementary effect of latrines and increased water usage on the growth of infants in rural Lesotho. Am J Epidemiol. 1992;135(6):659–66.

[3] Esrey SA. Water, waste, and well-being: a multicountry study. Am J Epidemiol. 1996;143(6):608–23.

[4] Checkley W, Gilman RH, Black RE, Epstein LD, Cabrera L, Sterling CR, et al. Effect of water and sanitation on childhood health in a poor Peruvian peri-urban community. Lancet. 2004;363(9403):112–118.

[5] Mapas de pobreza en Guatemala al 2002. Guatemala City; 2006.

[6] Vaides Lopez O. Plan municipal de agua y saneamiento, Camotan, Chiquimula, 2005 - 2020; 2005.

[7] Hudson JI, Pope J H G, Glynn RJ. The cross-sectional cohort study: an underutilized design. Epidemiology. 2005;16(3):355–9.

[8] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55.

[9] Preisser JS, Young ML, Zaccaro DJ, Wolfson M. An integrated population-averaged approach to the design, analysis and sample size determination of cluster-unit trials. Stat Med. 2003;22(8):1235–54.

[10] Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. Stat Med. 2007;26(1):20–36. Statistics in medicine.

[11] Pattanayak SK, Poulos C, Yang JC, Patil SR, Wendland KJ. Of taps and toilets: quasi-experimental protocol for evaluating community-demand-driven projects. J Water Health. 2009;7(3):434–51. Journal Article England.

[12] INE. Censos Nacionales XI de Poblacion y VI de Habitacion, Guatemala 2002. Instituto Nacional de Estadistica; 2002.

[13] Sinisi SE, van der Laan MJ. Deletion/substitution/addition algorithm in learning with applications in genomics. Stat Appl Genet Mol Biol. 2004;3:Article18.

[14] Heckman JJ, Ichimura H, Todd P. Matching as an econometric evaluation estimator. Review Of Economic Studies. 1998 Apr;65(2):261–294.

[15] Goldman N, Vaughan B, Pebley AR. The use of calendars to measure child illness in health interview surveys. Int J Epidemiol. 1998;27(3):505–12.

[16] Baqui AH, Black RE, Yunus M, Hoque AR, Chowdhury HR, Sack RB. Methodological issues in diarrhoeal diseases epidemiology: definition of diarrhoeal episodes. Int J Epidemiol. 1991;20(4):1057–63.

[17] Colford J J M, Wade TJ, Sandhu SK, Wright CC, Lee S, Shaw S, et al. A randomized, controlled trial of in-home drinking water intervention to reduce gastrointestinal illness. Am J Epidemiol. 2005;161(5):472–82.

[18] Gove S. Integrated management of childhood illness by outpatient health workers: technical basis and overview. The WHO Working Group on Guidelines for Integrated Management of the Sick Child. Bull World Health Organ. 1997;75 Suppl 1:7–24.

[19] ORC-Macro. Demographic and Health Survey Interviewer's Manual. MEASURE DHS Basic Documentation No. 2. ORC Macro; 2006.

[20] Reller ME, Mendoza CE, Lopez MB, Alvarez M, Hoekstra RM, Olson CA, et al. A randomized controlled trial of household-based flocculant-disinfectant drinking water treatment for diarrhea prevention in rural Guatemala. Am J Trop Med Hyg. 2003;69(4):411–9.

[21] Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics. 1986;42(1):121–30. Biometrics.

[22] Schmidt WP, Luby SP, Genser B, Barreto ML, Clasen T. Estimating the longitudinal prevalence of diarrhea and other episodic diseases: continuous versus intermittent surveillance. Epidemiology. 2007;18(5):537–43.

[23] Morris SS, Cousens SN, Kirkwood BR, Arthur P, Ross DA. Is prevalence of diarrhea a better predictor of subsequent mortality and weight gain than diarrhea incidence? Am J Epidemiol. 1996;144(6):582–8.

[24] Alam N, Henry FJ, Rahaman MM. Reporting errors in one-week diarrhoea recall surveys: experience from a prospective study in rural Bangladesh. Int J Epidemiol. 1989;18(3):697–700.

[25] Heuveline P, Goldman N. A description of child illness and treatment behavior in Guatemala. Soc Sci Med. 2000;50(3):345–64.

[26] WHO. WHO Anthro Software. Geneva: WHO; 2008. Available from: `http://www.who.int/childgrowth/software/en/`.

[27] Freedman DA. Statistical Models: Theory and Application. New York: Cambridge University Press; 2005.

[28] Rothman K, Greenland S. Modern Epidemiology. 2nd ed. Philadelphia: Lippincott-Raven Publishers; 1998.

[29] Leite ML, Nicolosi A, Osella AR, Molinari S, Cozzolino E, Velati C, et al. Modeling incidence rate ratio and rate difference: additivity or multiplicativity of human immunodeficiency virus parenteral and sexual transmission among intravenous drug users. Northern Italy Seronegative Drug Addicts Study. Am J Epidemiol. 1995;141(1):16–24. American Journal of Epidemiology.

[30] van der Laan M, Rubin D. Targeted Maximum Likelihood Learning. Int J Biostatistics. 2006;2(1):1–38.

[31] Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999;10(1):37–48.

[32] Colford J J M, Wade TJ, Schiff KC, Wright CC, Griffith JF, Sandhu SK, et al. Water quality indicators and the risk of illness at beaches with nonpoint sources of fecal contamination. Epidemiology. 2007;18(1):27–35.

[33] Mortimer KM, Neugebauer R, van der Laan M, Tager IB. An application of model-fitting procedures for marginal structural models. Am J Epidemiol. 2005;162(4):382–8.

[34] Cole SR, Hernan MA, Robins JM, Anastos K, Chmiel J, Detels R, et al. Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. Am J Epidemiol. 2003;158(7):687–94. American journal of epidemiology.

[35] Victora CG, Habicht JP, Bryce J. Evidence-based public health: moving beyond randomized trials. Am J Public Health. 2004;94(3):400–5.

[36] Luby SP, Agboatwalla M, Bowen A, Kenah E, Sharker Y, Hoekstra RM. Difficulties in maintaining improved handwashing behavior, Karachi, Pakistan. Am J Trop Med Hyg. 2009 Jul;81(1):140–145.

[37] Clasen T, Schmidt WP, Rabie T, Roberts I, Cairncross S. Interventions to improve water quality for preventing diarrhoea: systematic review and meta-analysis. Bmj. 2007;334(7597):782. BMJ (Clinical research ed.

[38] Arnold BF, Colford J J M. Treating water with chlorine at point-of-use to improve water quality and reduce child diarrhea in developing countries: a systematic review and meta-analysis. Am J Trop Med Hyg. 2007;76(2):354–64. The American journal of tropical medicine and hygiene.

[39] Luby SP, Agboatwalla M, Painter J, Altaf A, Billhimer W, Keswick B, et al. Combining drinking water treatment and hand washing for diarrhoea prevention, a cluster randomised controlled trial. Trop Med Int Health. 2006;11(4):479–89.

[40] Ram PK, Kelsey E, Rasoatiana, Miarintsoa RR, Rakotomalala O, Dunston C, et al. Bringing safe water to remote populations: an evaluation of a portable point-of-use intervention in rural Madagascar. Am J Public Health. 2007;97(3):398–400. American journal of public health.

[41] Parker AA, Stephenson R, Riley PL, Ombeki S, Komolleh C, Sibley L, et al. Sustained high levels of stored drinking water treatment and retention of hand-washing knowledge in rural Kenyan households following a clinic-based intervention. Epidemiol Infect. 2006;134(5):1029–36. Epidemiology and infection.

[42] Luby SP, Mendoza C, Keswick BH, Chiller TM, Hoekstra RM. Difficulties in bringing point-of-use water treatment to scale in rural Guatemala. Am J Trop Med Hyg. 2008;78(3):382–7.

[43] Sobsey MD, Stauber CE, Casanova LM, Brown JM, Elliott MA. Point of use household drinking water filtration: A practical, effective solution for providing sustained access to safe drinking water in the developing world. Environ Sci Technol. 2008;42(12):4261–7.

[44] Wright J, Gundry S, Conroy R. Household drinking water in developing countries: a systematic review of microbiological contamination between source and point-of-use. Trop Med Int Health. 2004;9(1):106–17. Journal Article Meta-Analysis Research Support, Non-U.S. Gov't Review England Tm & ih.

[45] Levy K, Nelson KL, Hubbard A, Eisenberg JN. Following the water: a controlled study of drinking water storage in northern coastal Ecuador. Environ Health Perspect. 2008;116(11):1533–40. Environmental health perspectives.

[46] Wilson JM, Chandler GN. Sustained improvements in hygiene behaviour amongst village women in Lombok, Indonesia. Trans R Soc Trop Med Hyg. 1993;87(6):615–6. Transactions of the Royal Society of Tropical Medicine and Hygiene.

[47] Cairncross S, Shordt K, Zacharia S, Govindan BK. What causes sustainable changes in hygiene behaviour? A cross-sectional study from Kerala, India. Soc Sci Med. 2005;61(10):2212–20.

[48] Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. Am J Public Health. 2004;94(3):423–32.

[49] Schmidt WP, Cairncross S. Household Water Treatment in Poor Populations: Is There Enough Evidence for Scaling up Now? Environmental Science & Technology. 2009;43(4):986–992.

Chapter 4

Evaluation of a pre-existing, combined sanitation, water and hygiene intervention in rural Tamil Nadu, India

Preface

As in Chapter 3, the work I present in this chapter is the product of myself and a large number of collaborators and I will continue to use the pronoun "we" rather than "I" to reflect this joint effort. Together with Ranjiv Khush (Aquaya Institute) and Jack Colford (UC Berkeley), we secured funding for the project from the Open Square Foundation in 2007. I designed the study with input from Jack, Ranjiv, and Alan Hubbard (UC Berkeley). Our collaborators in Chennai are members of Dr. Kalpana Balakrishnan's group at the Sri Ramachandra Medical College (SRMC), introduced to us by Kirk Smith (UC Berkeley). In addition to Kalpana, Dr. Padmavathi Ramaswamy, Dr. Padma Srikanth, Paramasivan Rajkumar ("Raaj") and Rama Prabha ("Ram") helped supervise the research activities. Padmavathi organized and managed all of the field research. Padma assisted Ranjiv develop the water quality testing protocol, and Raaj and Ram supervised all field activities. A dedicated field team of 10 interviewers from Trichy collected and entered all of the data. Alicia London (Aquaya Institute) managed the project from the US side. All team members provided extensive input into the questionnaire. Alicia, Ranjiv and myself spent 2 weeks in Trichy to help with the interviewer training and piloting. Alicia and Ranjiv both returned multiple times over the year – I did not have the fortune to return. The majority of my involvement in this study has focused on the science, and reflects a gradual transition away from running study field logistics. I have conducted all of the analyses and drafted all of the text herein. The results that I present here are part of a broader analysis of the socio-economic impacts of the intervention, which will include school attendance and health care expenditures.

4.1 Goals

In this chapter I extend the methods used in Chapter 3 to evaluate a pre-existing, combined community-led total sanitation (CLTS), water supply and hygiene intervention in rural India. Like the Guatemala study, in this chapter I use a quasi-experimental design that draws on historic, pre-intervention data collected in the 2001 Indian Census to select control villages. Here, I extend the design and analysis to include longitudinal measurement of health and observational outcomes, and retrospective measurement of less dynamic outcomes (e.g., toilet construction). This study provides a second example of using pre-existing interventions to measure the sustainability of interventions that were implemented outside of formal research activities. This chapter also highlights some of the limitations of using pre-existing interventions, and how it can be difficult to find a suitable control group in highly dynamic populations that make improvements independent of the intervention. Although the intervention program was a combined suite of intervention activities, we observe large differences between control and intervention villages in their access to private toilets, but not in their access to water sources or hygiene practices

and knowledge. With this in mind, the evaluation is primarily one of CLTS, against a backdrop of improved water and hygiene practices typical of developing countries.

4.2 Background

Between 2003 and 2007, WaterPartners International (WPI) and their local partner Gramalaya implemented a combined intervention campaign in 12 rural villages in the Tiruchirappalli district in the state of Tamil Nadu, India. The intervention combined public water supply improvements and repairs with hygiene and sanitation social marketing campaigns that used Community-Led Total Sanitation (CLTS) methods [1–3]. CLTS shows great promise at increasing access to and use of basic sanitation facilities. This new approach, pioneered in Bangladesh by the Village Education Resource Centre and WaterAid, aims to achieve universal sanitation coverage without subsidies by changing social norms and encouraging construction of low-cost latrines. Under CLTS, an external facilitator leads a community meeting and exercises designed to make residents aware of the magnitude of the sanitation problem, elicit feelings of disgust and shame, and create an impetus for collective action. Through an emphasis on the public nature of the problem, facilitators promote the goal of zero open defecation. Typically, communities are encouraged to come up with their own latrine designs using locally available, low-cost materials that put sanitation within reach of even their poorest members [4]. Public signs declaring a community free of open defecation are sometimes posted once all residents have access to sanitation facilities, as a reminder of the new social norm.

In addition to the water supply and social marketing components, in 8 of the 12 villages WPI/Gramalaya implemented an innovative microcredit scheme to enable families to borrow money from local Self Help Groups (SHGs) to construct private latrines, toilets, bathing facilities, water connections and stand posts. Gramalaya provided the loans directly to SHGs and the SHG members distributed the loans to individual borrowers in their village. Each SHG was responsible for repaying the entire loan in full, thus harnessing the communal responsibility for a single loan. By December 2007, Gramalaya had disbursed $98,883 in loans in the intervention villages for 496 water-related loans and 1,177 sanitation-related loans (average loan size: $59) [5]. Arney *et al.* provide additional details of the WPI/Gramalaya micro-credit program in the intervention villages [5].

Specific details of the intervention varied slightly by village (Table 4.1). All villages participated in CLTS campaigns and hygiene social marketing campaigns. Gramalaya renovated public water facilities (hand pumps, public stand pipes) and school sanitary blocks on an as-needed basis depending on whether the facilities were inadequate or in disrepair. The intervention's intent was to lead to comprehensive improvements in water supply, sanitation access and hygiene knowledge.

The primary objective of this study was to revisit households after the conclusion of intervention activities to assess water sources, water quality, sanitation access and practices, hygiene knowledge and practices and child health compared to a matched control

group of similar villages. We measure child health using self-reported gastrointestinal illness and anthropometric growth measurements in children under the age of 5 years.

4.3 Methods

4.3.1 Setting

This study was conducted in the Tiruchirappalli (Trichy) district in the state of Tamil Nadu, India. Intervention villages were located in the subdistricts of Thottiyam, Thuraiyur and Thathaiyangarpet. Control villages were located in adjacent subdistricts of Manachanallur and Uppiliyapuram. Villages are between 17 and 55 kilometers from the city of Tiruchirappalli, and are accessed primarily by paved roads (median walking distance to an all-weather road is three minutes). The climate is tropical, hot and subject to heavy rains during the monsoon season (August – December). During the study period the maximum temperature ranged between 23.0 and 40.7 degrees celsius, and there were 17 days with more than 25 mm (1 inch) of rain (Figure 4.1). All villages are rural and the primary occupation is rice agriculture and cultivation (66% of the working adults in our sample). Other major occupations include self-employed businesses (8.5%), truck drivers (6.0%), factory workers (2.7%) and skilled artisans (2.4%).

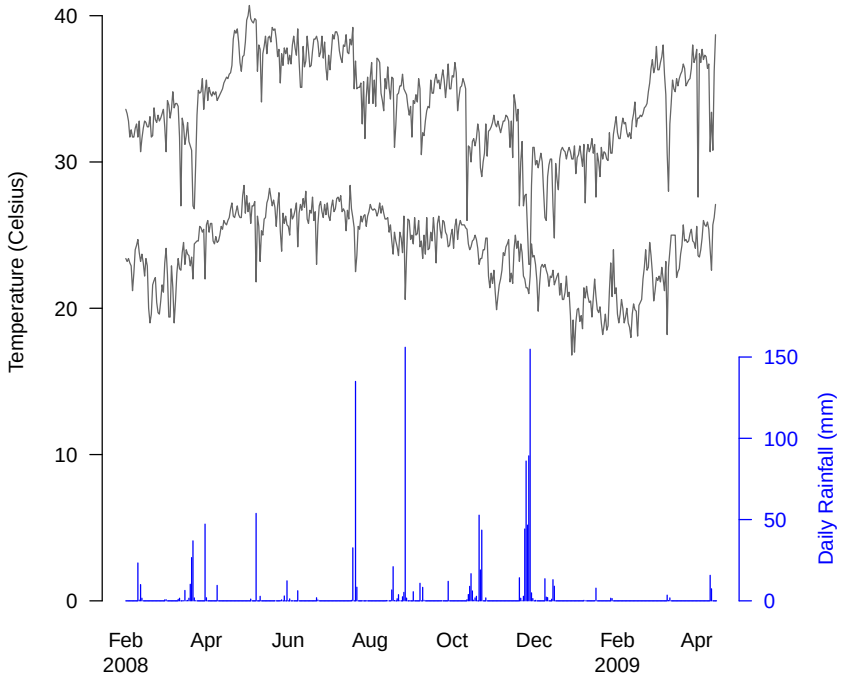


Figure 4.1: Daily minimum and maximum temperature (gray lines) and daily rainfall (vertical blue lines) recorded at the Tiruchirappalli airport during the study period.

Table 4.1: Summary of major intervention components in the 12 study villages reported by WaterPartners International and Gramalaya. Horizontal lines separate implementation projects. Villages 5-12 had access to microcredit loans for private household water and sanitation improvements. The age in months is the time elapsed from the intervention completion to middle of the first round of data collection (February 2008)

N	Village Name	N Households (Census 2001)	Project Dates	Age (mths)	Brief Intervention Description	
1	Keelakarthigaipatti	194	04/03 – 03/04	47	*Water:*	Community tube wells capped by hand pumps
2	Sakkampatti	140			*Sanitation:*	220 HH latrines, school sanitary block
					Hygiene	Child education ("health ambassadors") , hygiene promotion
3	Mettupatti	70	04/03 – 04/04	46	*Water:*	Community hand pumps
4	Periyanachipatti	80			*Sanitation:*	> 63 HH toilets, school sanitary block
					Hygiene:	Hygiene promotion, school health clubs
5	Ponnusangampatti	290	01/04 – 12/04	38	*Water:*	≈ 279 HH taps
6	Melakothampatti	90			*Sanitation:*	273 HH toilets
7	Theverappampatti	125			*Hygiene:*	Hygiene education campaign
8	Ayinapatti	114	01/05 – 03/06	23	*Water:*	≈ 45 HH taps, new school water taps
					Sanitation:	97 HH toilets, renovated school toilets
					Hygiene:	Hygiene and sanitation education training
9	Melakarthikaipatti	289	01/05 – 03/06	23	*Water:*	≈ 21 HH taps, 1 hand pump renovated, new school tap
					Sanitation:	370 HH toilets, renovated school toilets
					Hygiene:	Hygiene education (social health clubs)
10	Melanaduvalur	160	10/05 – 10/06	17	*Water:*	≈ 50 HH taps, 1 hand pump renovated, 14 public stand posts renovated
					Sanitation:	118 HH toilets installed, renovated school toilets
					Hygiene:	Village-wide hygiene education campaign
11	Kanganipatti	160	10/05 – 10/06	17	*Water:*	50 HH taps, 2 hand pumps renovated, new school tap
					Sanitation:	115 HH toilets, renovated school toilet facilities
					Hygiene:	Village-wide hygiene education campaign
12	Kollapatti	220	10/06 – 09/07	5	*Water:*	100 HH taps, restored/repaired 4 existing hand pumps & school water facilities
					Sanitation:	118 HH toilets, renovated school toilet facilities
					Hygiene:	Village-wide hygiene education campaign

4.3.2 Study design

We conducted a community-based, quasi-experimental prospective cohort study in 12 intervention and 13 matched control villages. We enrolled up to 50 households per village with children under age 5 years, and all study participants were visited monthly over one year (12 visits total). In the first visit we collected a large set of information that included demographic, economic and environmental variables. In each follow-up visit we assessed the health of children under age five and collected key environmental exposure information. 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.

Village selection

Similar to the intervention described in Chapter 3, the combined intervention from WPI and Gramalaya was not randomized and was deployed in villages that were purposely selected by the implementing organizations. Intervention villages were likely different, on average, from non-intervention villages. To help reduce potential bias due to differences between intervention and control villages at baseline, we used a combination of restriction and propensity score matching[6] based on pre-intervention characteristics to purposely select control villages. We further refined the selection based on pre-intervention characteristics with a rapid assessment in late 2007.

Data sources: We obtained panchayat-level data from the 2001 Indian national census.[1] Since the unit of intervention was the village (i.e., below the panchayat), we supplemented the 2001 census data with higher-resolution 2003 Tamil Nadu Water Supply and Drainage board (TWAD) survey data that included details about population, water supply and cattle ownership at the village (habitation) level.

Sampling frame definition and restriction: Our sampling frame included villages in two administrative blocks (Manachanallur and Uppiliyapuram) that neighbor intervention village blocks. We did not include control villages from the same blocks as the intervention because of heightened water and sanitation activity among non-governmental organizations (primarily Gramalaya with funding from WaterPartners International and WaterAid) in those blocks. There were 240 potential control villages in the original frame. As a first step to ensure that control villages were similar to intervention villages, we excluded villages that had $> 80\%$ scheduled caste population (N=15), had < 50 total households (N=20) and had $< 70\%$ of the households using biofuel for cooking (N=10).

Propensity score matching: With the remaining 195 potential control villages and the 12 intervention villages, we modeled the probability of receiving the intervention (A) conditional on a large set of covariates (W), using a logit model: logit $Pr(A = 1|W) = \alpha' W$, where the logit function is: logit $(p) = log[p/(1-p)]$. After estimating the model, we

[1] A panchayat is a group of villages that typically total around 500 people.

matched two control villages to each intervention village using a nearest-neighbor match based on the linear predictor from the model (the log-odds of participation in the intervention). We used an iterative approach to selecting covariates in W by re-estimating the model for different specifications and selecting the specification that selected a control group most closely balanced with the intervention group (based on a difference in means and nominal p-values). The final model included main effects only for the following covariates: number of households in the village, per-capita cattle ownership in the village, the panchayat-level income, and the proportions of: scheduled caste population, households with access to in-home or public tap water, literate female population, and households that use banking services.

Rapid assessment and additional exclusions: After selecting the 24 potential control villages, in late 2007 we conducted a rapid assessment of all intervention and control villages to measure basic information about the number of active self-help groups (SHGs), school and administrative facilities, primary livelihoods, car and tractor ownership, and basic water infrastructure. During the rapid assessment we also created a comprehensive listing of households with children under age five. The goal of the exercise was to reduce the control sample to one matched village per intervention village based on current conditions.

Our team found that two of the 24 villages were a single, contiguous village and so they were treated as one village in the selection process (conveniently they were matched to the same intervention village). The smallest intervention village in our sample had 18 households with children under five, and we eliminated 6 potential control villages that had fewer than 18 households with children under five. The rapid assessment determined that two potential control villages were substantially larger than the intervention villages, and that the larger size led to qualitative differences in village characteristics (e.g., they have their own hospitals) so we excluded villages with more than 150 households with children under 5. Finally, we found that all intervention villages had at least two self-help groups and so we excluded villages with fewer than 2 on the grounds that they may be less socially organized than the intervention villages. There were no additional characteristics that created obvious outliers in the control village set so we proceeded with 13 control villages (one village retained both of its matched controls).

Household selection

We listed all households in each village that had at least one child under age five years. From this sampling frame, we selected a random sample of 50 households per village. If a village had fewer than 50 households, then all households from that village were included in the final sample.

4.3.3 Data collection

A team of 10 locally-hired interviewers and two supervisors from SRMC conducted household interviews between January 2008 and April 2009. The survey instruments were pre-tested and validated during a three-week period in nearby, non-study villages. The training and pre-testing was directed by senior members of the team from SRMC, Aquaya and UC Berkeley. In all cases, the child health and behavioral questions were answered by the child's primary caregiver (usually the mother).

4.3.4 Water sample collection and analysis

Beginning in the third survey round field staff collected water samples from all village sources and household drinking water. The 25 villages in our study have between one and seven village sources, and all village sources were tested in 10 survey rounds. Field staff collected 125 ml of water from village sources in a fashion that mimicked villager water retrieval practices. Field staff also collected drinking water samples from participant households during follow-up survey rounds. Participant households were randomly allocated into four groups. Two of the groups were measured in rounds 3 and 5, and the other two groups were measured in round 4 and 6. In survey rounds 7 through 12, households in one of the four groups were tested. Each household's drinking water was tested between 1 and 4 times over the study period.

Water samples were collected in 125 ml sterilized plastic bottles in a fashion that mimicked each household's water retrieval practices: by either dipping a household cup into the vessel to transfer the water, pouring water from the storage container into the sample container, or, if a household did not store drinking water, retrieving water directly from the tap. Along with the water samples, field staff recorded basic characteristics of the water conditions at the time of collection (such as storage container type).

The field team transported the all water samples in a cooler to a laboratory at nearby university for culturing within 24 hours. Sample water was passed through at 0.45 μ.m membrane filter, diluted at a ratio of 1:100 and incubated on HiCrome M-Tec Agar (HiMedia M1571) at 44.5°C for 24 hours. The number of purple colonies were counted and recorded as coliform bacteria. The number of blue-green colonies were counted and recorded as E. coli. If no coliforms were identified in a sample, then the sample was re-analyzed at a dilution of 1:10 and, if still no colonies grew, with no dilution. Samples without detectable coliforms or E. coli were set to 0.1 prior to analysis of quantitative counts.

Each sample was also analyzed for H_2S producing bacteria using the HiH2S test kit (HiMedia K020). Samples were left to incubate at room temperature for 24 hours, and if room temperature fell below 30°C, for an additional 12 hours. Samples were recored as positive for H_2S if they turned black.

4.3.5 Outcome definition and measurement

Sanitation and open defecation practices

During the first round survey field staff recorded detailed information about sanitation and defecation practices in each household. The survey collected information about four primary defecation categories: i) open defecation, ii) community toilets, iii) neighbor's toilets and iv) private toilets. For each household interviewers recorded self-reported defecation practices in all four categories, as well as details that included frequency of use, location, reasons for use and who in the household (men, women, children < 5) practiced each type of defecation. We also collected information from female respondents in the household about their perceived safety and privacy while defecating. In follow-up surveys, field staff collected information about new private toilet construction in study households. If given permission by the family, field staff inspected private toilets in each visit to determine whether it appeared that they were in regular use. If a household owned a private toilet, we asked household members to estimate its age.

Water sources

During the first round interview field staff collected detailed information about each household's water source access and use. The survey recorded the use of eight different water sources and reasons for a household's use or non-use of each source. If a household reported using a water source, then field staff collected details about the source including distance, number of trips per day, use of the water from each source and perceived safety, reliability and quality. Respondents were also asked to identify their primary water source, and when they began to use their primary water source (an estimate of its age).

Hygiene and handwashing

During each home visit, field staff collected a large number of spot check observations of household environments using objective criteria. Handwashing observations included details about whether a household had a dedicated handwashing station and whether it was stocked with water and soap. Private toilets (if owned) were inspected to collect information about cleanliness and the availability of toilet paper, soap and water for handwashing after defecation. Interviewers collected observations of animals and their feces in the home living area during the interviews and general cleanliness measures such as the presence of garbage in the home. Interviewers also collected measurements of child cleanliness for children under age 5 that were present at the time of the interview.

In addition to objective spot check observations, interviewers asked primary caregivers an open-ended question about their handwashing practices during the prior 24 hours. The interviewers coded responses into 12 critical times and whether the respondent reported washing with water alone or water with soap. Given the likely bias in self-reported hand-

washing measures [7], we collected this information primarily as a measure of handwashing knowledge.

Diarrhea and gastrointestinal illness

During the household interviews, field staff collected self-reported illness symptoms over the previous 14 days from each child's caregiver using a health calendar modeled after Goldman *et al.* [8]. The calendar records each day that the child has each individual symptom. We defined diarrhea as three or more loose or watery stools in 24 hours, or a single stool with blood or mucus [9]. We also recorded symptoms for a measure of highly credible gastrointestinal illness (HCGI), which includes any of the following four conditions: vomiting, diarrhea, soft stool and abdominal cramps, or nausea and abdominal cramps [10].

Child growth

During the pre-test phase, senior team members from SRMC and UC Berkeley standardized all fieldworkers on anthropometric measurement techniques over three full days of training. Field workers collected measurements in teams of two (we followed standard protocols from the Demographic and Health Survey [11]).The field team collected anthropometric measurements at the participants' homes during the first and last interviews (rounds 1 and 12).

Fieldworkers weighed children in the standing position when possible. They weighed children that were too young to stand in their caregiver's arms and re-weighed the caregiver separately (the values were later subtracted during the analysis). We measured weight using scales accurate to 0.1 kg (Tanita 1631), and the scales were tested for accuracy each morning with a standardized 10 kg weight. Fieldworkers measured the length of children under age two in the reclining position and children aged 2 to 5 in the standing position using portable stadiometers accurate to 0.1 cm (Seca 214). Upper arm circumference was measured for children aged 6 months and older at the mid point of the upper right arm using an elastic tape accurate to 0.1 cm.

4.3.6 Statistical methods

Sample size estimation

We powered the study around child diarrhea because that was the outcome for which we had the most information about expected effect sizes and variability to inform the calculations. We estimated study sample size and power using standard methods for the comparison of two proportions with repeated measures [12, 13]. Since we anticipated multiple levels of correlation in the data (individual, household and village), we calculated a

combined design effect for different study designs using estimates of intra-cluster correlation (ICC) for diarrhea derived from a Guatemalan cohort of 952 children < 5 followed for 52 weeks [14]. The ICC estimates we used were: individual (0.15), household (0.09), village (0.008), which fell in the range of reasonable values compared to other developing country estimates [15]. We assumed a baseline diarrhea prevalence of 10% based on an earlier large intervention study in Tamil Nadu [16]. We also assumed 12 intervention villages and 12 control villages, with 1.3 children under age 5 per household. Table 4.2 summarizes power under different assumptions. With between 30 and 50 households per village the study is well-powered to detect differences of 2.5 percentage points in diarrhea prevalence.

Table 4.2: Summary of power $(1 - \beta)$ estimated under different sample size and effect size assumptions. All scenarios assume 12 villages in each group, 12 visits, and 10% prevalence in the control group (p_0).

Households	Effect size $(p_0 - p_1)$				
per Village	0.030	0.025	0.020	0.015	0.010
20	0.841	0.682	0.484	0.295	0.153
25	0.902	0.762	0.560	0.346	0.176
30	0.939	0.821	0.623	0.392	0.198
35	0.961	0.865	0.676	0.433	0.218
40	0.975	0.897	0.721	0.470	0.236
45	0.984	0.920	0.758	0.504	0.254
50	0.989	0.938	0.789	0.534	0.270
55	0.993	0.951	0.815	0.562	0.286
60	0.995	0.961	0.837	0.587	0.300

Measures of self-reported illness

We quantified diarrhea and HCGI using weekly longitudinal prevalence[2] [17], a disease measure that is more strongly correlated with child mortality than incidence [18]. We limited the longitudinal prevalence data to a seven day (1 week) recall window after identifying under-reporting of symptoms for recall periods longer than seven days.

[2]Longitudinal prevalence is calculated by dividing the number of weeks with illness by the total weeks of observation.

Unadjusted outcome analyses

The parameter of interest for all outcomes (both unadjusted and adjusted) is the marginal treatment effect conditional on selection into the study based on restriction and propensity score matching. We estimate the parameter as:

$$E(Y|A=1,W^*) - E(Y|A=0,W^*) \tag{4.1}$$

where Y is the outcome of interest, A is an indicator equal to 1 if a child lives in an intervention village and 0 otherwise, and W^* is the set of characteristics among intervention villages in the study sample ($W^* = W|A=1$, $\leq 80\%$ households are scheduled caste, $\geq 70\%$ households use biofuel, ≥ 50 households). Thus, our inference is limited to the set of intervention villages for which there is a comparable control village based on the village selection method: this is an average treatment among the treated (ATT) estimator (see Chapter 2 for background and derivation).

For child diarrhea we calculated the difference in the weekly longitudinal prevalence between the intervention and control groups. We converted the anthropometric measurements to age- and sex-specific Z-scores using a publicly available Stata algorithm that references the 2006 WHO Growth Standards [19], and calculated the difference in Z-score means. For binary sanitation, water and hygiene outcomes, we calculated the difference in prevalence (risk difference) of each outcome between the intervention and control groups.

In addition to calculating mean differences in 2008, for private toilet and tap construction, which we assumed could be reasonably well-estimated by households retrospectively, we calculated the difference between intervention and control villages in newly constructed toilets and taps during the five year intervention period. For these two outcomes, this difference in the change in private amenities is a difference-in-difference (DID) estimator that removes residual time-invariant confounding between groups [20].

For all unadjusted estimates we calculated percentile-based 95% confidence intervals using a bootstrap with matched village pairs as the sampling unit (to reflect the design) and 1,000 iterations [21].

Adjusted outcome analyses: child growth

We calculated adjusted estimates of the intervention on child growth using targeted maximum likelihood estimation (MLE) [22]. Appendix B includes an introduction to targeted MLE in the point treatment setting (also implemented in Chapter 3). Although we measured child anthropometry at two points for most children in the study, for the purpose of this analysis we analyze the data as point-treatment data with additional repeated measures within village. Let Y be a child's Z-score for an anthropometry measure and let A be an indicator variable equal to 1 if a child lives in an intervention village and 0 otherwise. Finally, let W be a set of covariates that could potentially confound or modify the relationship between A and Y. The parameter of interest is the marginal difference

in Y if all individuals in the study population were not treated (A=0) versus if they were all treated (A=1):

We calculated adjusted estimates using the k-step targeted MLE estimator:

$$\hat{\psi}_{T-MLE} = \frac{1}{n}\sum_{i=1}^{n} Q^k(1, W_i) - Q^k(0, W_i) \tag{4.2}$$

(see Appendix B for details and derivations). Identical to the unadjusted analyses, we estimated the standard error for adjusted effects using a bootstrap with matched village pairs at the unit of resampling and 1000 iterations.

Adjusted outcome analyses: gastrointestinal illness

For our adjusted analysis of gastrointestinal outcomes (diarrhea and HCGI), we consider targeted MLE with a longitudinal data structure. Here, we use a variation on the point treatment method called the Reduced Data-Targeted MLE (R-TMLE), first proposed by by van der Laan (pages 165 - 178) [23]. For parsimony we will simply refer to this estimator as the targeted MLE estimator. The advantage of the reduced data formulation is that it simplifies the estimation procedure in longitudinal data, while maintaining many of the advantages of a fully specified targeted MLE estimator.

The data are longitudinal with at most 12 monthly measurements. The outcome of interest $Y_a(t)$, is an individual-level indicator of a new episode of diarrhea or HCGI in month t (for $t = 0, \ldots, 11$) for children living in an intervention village ($a = 1$) or control village ($a = 0$). We follow child level outcomes even though village is the unit of treatment assignment because there are household- and individual-level covariates that may influence the outcome. The dataset includes a set of covariates W that contains a broad set of demographic and socio-economic characteristics that could not reasonably be influenced by the intervention.

Consistent with causal inference models, we define our parameter of interest in terms of potential outcomes (Section 2.3). In the methods below, it will be useful to describe the data in terms of a full data distribution, in which all potential outcomes are realized, and an observed data distribution, in which outcomes are realized on a subset of the population. We define our parameters of interest on the full data, and then estimate them using the observed data. Identifiability assumptions tie the parameter based on the observed data distribution to that of the full data distribution.

Let $X = (\bar{L}_0(11), \bar{L}_1(11)) \sim P_0$ be the full data distribution, where $\bar{L}_a(11)$ is an abbreviaion for $(L_a(0), \ldots, L_a(11))$. $L_a(0)$ includes all baseline covariates (W). $L_a(t)$ includes time-dependent outcomes $Y_a(t)$ and time-dependent covariates observed when each child is located in an intervention village ($a = 1$) or a control village ($a = 0$).

Let $O = (A, \bar{L}(11)) \sim P_0$ be the observed data, where we observe one potential outcome from the full data subject to the treatment assignment $A = I$(intervention village).

We assume that draws from O are identically distributed and that observations are independent between matched village pairs (but not within pairs).

Our parameter of interest is the marginal risk difference, defined on X:

$$\psi = P(Y_1 = 1) - P(Y_0 = 1) \tag{4.3}$$

where the probability of the outcome $P(Y_a = 1)$ is averaged over the entire follow-up period ($t = 0, \ldots, 11$). In this analysis, we assume that the data satisfy the conditional randomization assumption $\{A \perp\!\!\!\perp X \mid L(0)\}$; that is, conditional on observed baseline covariates $L(0)$ the treatment A is unconfounded. We also assume that censoring (loss to follow-up) is random, and that treatment does not change over time (a child either lives in an intervention village or does not).

Given our observed data $O = (A, \bar{L}(t))$, we define a reduced data structure that ignores any time-dependent covariates other than the month of follow-up, t, a child's age at each visit $age(t)$ and the outcome $Y(t)$: $O^r = (A, \bar{R}(t), \bar{Y}(t))$, where: $\bar{R}(t) = \{L(0), t, a\bar{g}e(t)\}$. We include the month of follow-up to control for seasonality in the outcome, and we include time-varying age because of its strong association with the outcome. Since, the probability of receiving treatment depends only on baseline covariates, it is constant over time. Thus, the probability of the reduced observed data random variable O_i^r for individual i is:

$$P(O_i^r) = P(L(0)_i) \times P(A_i|L(0)_i) \times \prod_{t=0}^{11} P(Y_i(t)|A_i, R_i(t)) \tag{4.4}$$

Note that in the above likelihood we assume there is no time-dependent process in the covariates: $P(Y(t)|A, \bar{R}(t)) = P(Y(t)|A, R(t))$. This assumption is satisfied in this analysis because calendar time and child age are deterministic processes that are not influenced by past conditions. We use the empirical distribution for $P(L(0)_i)$, but we need to estimate both $P(A_i|L(0)_i)$ and $P(Y_i(t)|A_i, R_i(t))$. Below we describe the process of obtaining an initial estimate of these quantities, and then updating them with the one-step targeted MLE. For notational convenience, let $L(0) = W$.

We can obtain a targeted estimate of the risk difference using the following steps:

1. Estimate probability of $Y(t)$ given A and $R(t)$ with maximum likelihood using a logistic regression model over all individuals and all time periods observed. We denote the predicted probabilities from this regression as $\hat{P}(Y_i(t)|A, R(t)) = \hat{Q}^0(A, R(t))$, where:

$$\hat{Q}^0(A, R(t)) = \frac{1}{1 + \exp -\hat{m}^0(A, R(t))} \tag{4.5}$$

   for some function $\hat{m}^0$ of the covariates A, and $R(t)$.

2. Estimate the probability of living in an intervention village $A = 1$ given baseline covariates W with maximum likelihood using a logistic regression model and data

at $t = 0$. We denote the predicted probabilities from this regression as $\hat{P}(A|W) = \hat{g}(A|W)$.

3. For each individual, calculate a covariate based on her observed values for A and W. We denote this covariate $h(A, W)$, where:

$$h(A, W) = \frac{I(A = 1)}{\hat{g}(1|W)} - \frac{I(A = 0)}{\hat{g}(0|W)} \tag{4.6}$$

4. Update the original regression by adding the covariate $h(A, W)$, and estimate the corresponding coefficient by maximum likelihood, holding the remaining coefficient estimates at their initial values. In practice, this is achieved by estimating a logistic regression of $Y(t)$ on $h(A, W)$ with $\hat{m}^0(A, R(t))$ as an offset with coefficient constrained to one. Let ϵ_n be the coefficient on $h(A, W)$. We denote this one-step updated regression $Q^1(A, R(t))$ where:

$$\hat{Q}^1(A, R(t)) = \frac{1}{1 + \exp -(\hat{m}^0(A, R(t)) + \hat{\epsilon}_n h(A, W))} \tag{4.7}$$

5. After estimating the updated conditional probability $\hat{Q}^1(A, R(t))$, simulate a dataset drawing matched village pairs with replacement from the original dataset to simulate the joint distribution of W and complete follow up (all 12 months). In practice, one can simulate any sample size, but in this analysis we simply drew the number of matched village pairs in the original data ($n = 12$). We then impute predicted counterfactual values for $Y_0(t)$ and $Y_1(t)$ with $Y_a(t) \sim Binomial(1, \hat{Q}^1(a, R(t)))$ using the simulated data. This is equivalent to simulating the likelihood in equation 4.4 with two treatment conditions (intervention and control).

6. Calculate the average probability of the outcome under each treatment a in the simulated data. Calculate the risk difference from the simulated counterfactual population averaged over all individuals and times as: $\hat{P}(Y_1 = 1) - \hat{P}(Y_0 = 1)$.

7. Repeat steps 1-6 for $B = 1,000$ times and calculate the mean and standard deviation over the B replicates. The mean over the replicates is the targeted ML estimator of the parameter of interest:

$$\hat{\psi}_{T-MLE} = \frac{1}{B} \sum_{b=1}^{B} \hat{P}(Y_{1,b} = 1) - \hat{P}(Y_{0,b} = 1) \tag{4.8}$$

and the standard deviation is a bootstrapped estimate of its standard error.

Note that a difference in means over values imputed from $Y_a(t) \sim Binomial(1, \hat{Q}^0(a, R(t)))$ (i.e., the predicted counterfactuals without the 1-step

update) is the standard G-computation estimator [22, 24]. Also note that due to the reduced data structure that excludes the possibility for time-dependent confounding, the targeted ML estimator can be obtained by simply taking the mean difference of predicted probabilities of illness over all times: $E_{R(t)}[\hat{Q}^1(1, R(t)) - \hat{Q}^1(0, R(t))]$, rather than simulating binary outcomes and then calculating the mean over the 1,000 replicates.

Model selection for nuisance parameters

In all adjusted analyses for child growth and gastrointestinal illness, targeted MLE estimators require that we specify models for the nuisance parameters $Q^0(A, W)$ and $g(A|W)$. The functional forms of the nuisance parameter models are unknown, and could be a complex combination of child-, household- and village-level covariates. Yet, the consistency of the targeted MLE estimators rely on the correct specification of these models.

To reduce potential bias from model mis-specification we used a flexible machine algorithm called Super Learner that calculates predicted outcomes given a large set of covariates [25]. Super Learner is implemented in R in the `SuperLearner` package. Super Learner is a meta-learning algorithm that uses V-fold cross-validation to combine individual candidate learners into a single prediction using optimal weights. Each individual candidate learner is fit using V-fold cross validation. We included the following candidate learners in the Super Learner: generalized linear models with main effects, elastic net regression (a hybrid of lasso and ridge regression) [26], and generalized additive models [27]. All model selection algorithms were applied within each bootstrap iteration, so the standard errors of the estimates include variability from both sampling and model selection.

Table 4.3 includes baseline covariates included in model selection for all models. For each specific outcome we subset these covariates to those that had a univariate positive association with the outcome to improve the efficiency of the targeted MLE estimator [28, 29]. We defined a positive association as a univariate association with $p \leq 0.20$, or an odds ratio $\leq 0.83, \geq 1.2$ (binary gastrointestinal outcomes) or difference of 0.2 standard deviations (continuous anthropometry outcomes).

Subgroup analyses

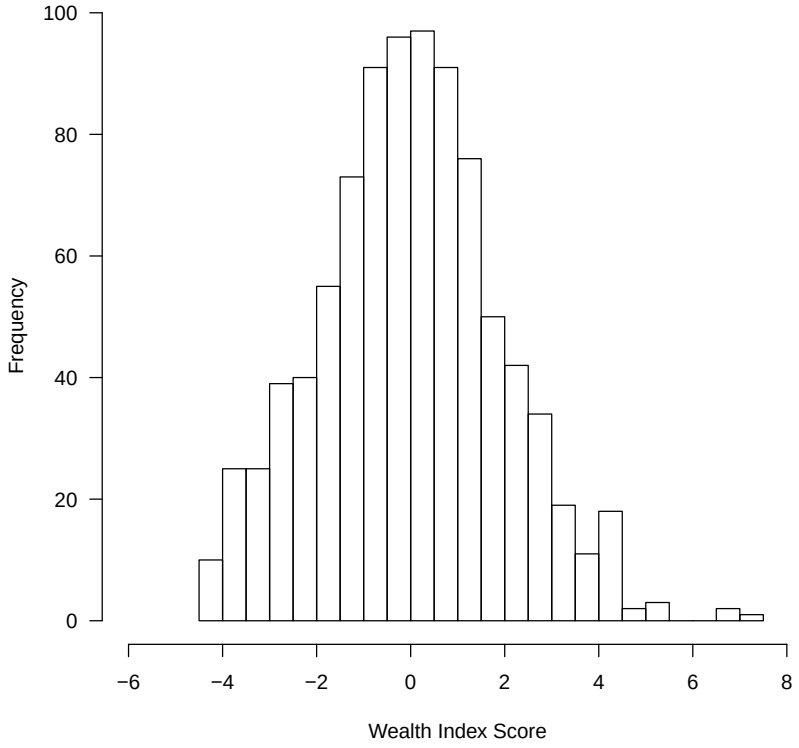
We explored whether the intervention's impact on new private toilet and private tap construction as well as child diarrhea and height varied by household wealth and by scheduled caste status. We created a household wealth index using principal components analysis based on housing characteristics and asset ownership (Table 4.5) [30, 31]. We used the first component (eigenvector) from the analysis, which has been used as a wealth index score in developing country studies (see [31] for an overview and [32–34] for examples in water and hygiene studies). The first component's eigenvalue was 3.74 and it explained 18.4% of the variability in household materials and assets. The wealth index was unimodal and approximately gaussian (Figure 4.2), and so we categorized households into quintiles

Table 4.3: Covariates used in model selection for adjusted analyses.

Category	Covariate
Child	Sex
	Age (months)
	Currently breast feeding (at baseline)
Adults	Primary caregiver's education level (factor)
	Participate in a community group
	At least one parent works in agriculture
	Scheduled caste
	Use banking services
	Mother works
	Mother's age
Household	Soil floor (versus concrete or tile)
	Thatch roof (versus improved materials)
	Household has electricity
	Family owns their home
	Family owns their land
	Total persons living in the home
Durable good ownership	Television
	Mobile phone
	Motorcycle or scooter
	Bicycle
	Mosquito net

of the wealth index for stratified analyses (N=180 households per category). Field staff recorded caste status during household interviews.

Figure 4.2: Histogram of wealth index scores derived from a principal components analysis using household characteristics and assets.



4.4 Results

4.4.1 Village selection and pre-intervention characteristics

Our selection of control villages led to intervention and control groups that were similar at baseline (pre-intervention) across a broad range of characteristics (Table 4.4). The restriction and matching led to better balance in many characteristics, such as female literacy, panchayat income, use of banking services, use of biofuel for cooking and the proportion of the population who were scheduled caste. The restriction and matching also led to greater imbalances in some covariates, including the proportion of households owning a private latrine and the proportion of households that use a handpump. After selection, intervention villages had a larger fraction of agricultural workers than control villages in 2001 (33% vs. 21%), and were more likely to own a private toilet (15% vs. 9%). Although the 2001 census data indicate some imbalance in private toilet ownership, we estimate that in our actual sample the groups were well balanced in private toilet ownership in 2003 (Section 4.4.3, below).

Table 4.4: Summary of pre-intervention characteristics before and after village selection. India National Census 2001 and Tamil Nadu Water Supply and Drainage (TWAD) 2003 surveys.

Mean	All Villages				Study Sample			
	Control	Interv.	SD*	p-value†	Control	Interv.	SD*	p-value†
Panchayat-level characteristics (Census 2001)								
Male (%)	49.8	49.9	9	0.693	49.6	49.9	43	0.570
Children <6 (%)	11.9	11.6	−20	0.049	12.2	11.6	−40	0.078
Female literacy (%)	52.4	47.5	−82	0.114	49.4	47.5	−32	0.331
Females work (%)	73.6	69.5	−19	0.345	70.9	69.5	−6	0.293
Cultivators (%)	26.8	28.4	9	0.688	31.2	28.4	−16	0.323
Agricultural laborers (%)	24.1	33.4	52	0.161	21.2	33.4	68	0.058
Marginal workers (%)	19.3	22.1	17	0.216	21.4	22.1	4	0.278
Panchayat income (100s Rupees per capita)	122.5	74.7	−186	0.002	71.4	74.7	13	0.359
Tap water, private + public (%)	74.7	76.2	10	0.387	75.3	76.2	6	0.215
Hand pump (%)	12.3	14.0	23	0.013	17.8	14.0	−53	0.079
Private toilet/latrine (%)	14.5	15.4	10	0.200	9.2	15.4	77	0.041
Use banking services (%)	29.0	24.9	−36	0.134	25.3	24.9	−4	0.537
Use biofuel for cooking (%)	90.8	96.7	283	0.009	95.7	96.7	50	0.264
Own radio	42.6	42.7	4	0.241	37.8	42.7	138	0.005
Own television	20.9	16.2	−68	0.120	17.2	16.2	−14	0.116
Own scooter/moped	10.1	10.2	1	0.559	8.7	10.2	29	0.578

(continued on the next page)

Table 4.4: *(continued)*

Mean	All Villages				Study Sample			
	Control	Interv.	SD*	p-value†	Control	Interv.	SD*	p-value†
Village-level characteristics (TWAD 2003)								
Total households	169.8	161.0	−12	0.440	181.2	161.0	−27	0.928
Persons per household	5.0	4.6	−32	0.413	4.5	4.6	7	0.803
Scheduled caste (%)	19.2	12.1	−39	0.580	15.0	12.1	−16	0.345
Per-capita cattle ownership	4.3	3.6	−27	0.627	4.8	3.6	−45	0.688
Population served per hand pump	259.5	301.8	11	0.980	240.0	301.8	16	0.647
Population served per borehole	437.0	678.8	60	0.004	509.6	678.8	42	0.228
Water supply required (liters per capita per day)	27.1	21.8	−42	0.099	30.3	21.8	−66	0.659
Water supply level (liters per capita per day)	12.4	14.8	25	0.005	14.2	14.8	7	0.550
Number of villages	240	12			13	12		
Number of households	40,759	1,932			2,356	1,932		

* SD: The standardized difference is equal to the difference in standard deviations of the mean (I–C) multiplied by 100. It is calculated as $(\mu_I - \mu_C) \div [(S_I^2 + S_C^2)/2]^{1/2} \times 100$. For example, a difference of 1 SD is equal to 100. A value of zero indicates equality of the means.

† Nominal boostrapped Kolmogorov-Smirnov p-values [35, 36].

4.4.2 Population characteristics

Our sample included 472 control households and 481 intervention households. Of these, 16 control households and 37 intervention households were not enrolled, primarily (n=41) because the families had moved away in the 2 months between listing and enrollment. Our final household enrollment included 456 control and 444 intervention households. Of these, 433 (95%) control and 424 (95%) intervention households completed all 12 months of follow-up.

Our sample included 1,173 children under five years old at the beginning of data collection. An additional 112 children were born into the cohort over the 12 month follow-up, for total samples of 648 control and 637 intervention children. Of these 612 (94%) control and 609 (96%) intervention children completed follow-up. Our final sample includes 14,259 person-weeks of observation.

In our first round of data collection in 2008, intervention and control households remain balanced on a large number of potentially confounding characteristics (Table 4.5). Similar to baseline, the intervention villages have a larger proportion of adults who work in agriculture (46% vs, 35%). Consistent with intervention villages being slightly more agricultural than control villages, intervention households are also more likely to have a soil floor (35% vs 28%), to have a thatched roof (28% vs. 21%) and to own their home (97% vs. 88%) or land (98% vs. 92%). Despite these differences, the two groups are highly similar in community participation, scheduled caste status, use of banking services, and female education. As a check for balance on the joint distribution of the covariates and potentially high-order combinations of the covariates, we modeled the probability of living in an intervention village A conditional on the adult and household level covariates W in Table 4.5, with $P(A = 1|W)$ predicted using the Super Learner machine learning meta-learner (see Statistical Methods, above). Figure 4.3 plots the predicted probabilities by treatment group. The probabilities are bound between 0.21 and 0.72. There is reasonably good support in the joint distribution of the covariates for estimating treatment effects. That the distributions do not line up perfectly indicates that there are imbalances in the joint distribution of the covariates, and that the unadjusted estimates could be biased.

Figure 4.3: Histogram of predicted probabilities of receiving the intervention, $\hat{P}(A = 1|W)$, for households in intervention and control villages based on predictions from the Super Learner run on adult and household level covariates in Table 4.5. Bin width is 0.025.

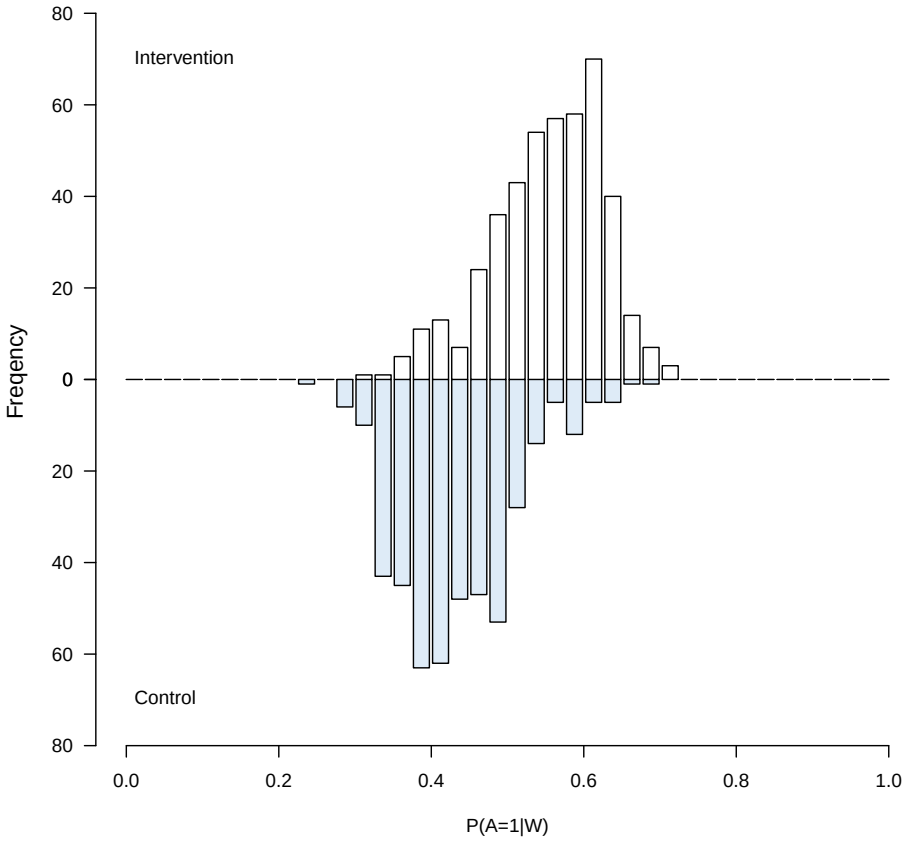


Table 4.5: Summary of post-intervention characteristics at the beginning of data collection in 2008. Standard errors (SEs) are adjusted for clustering at the village level. P-values are nominal.

	Control			Intervention			
	N	Mean	SE	N	Mean	SE	p-value
Children < 5 characteristics							
Female	596	0.515	0.018	577	0.473	0.019	0.108
Age in months	596	30.399	0.816	577	31.718	0.774	0.241
Ever breastfed	596	0.987	0.005	577	0.991	0.004	0.466
Currently breastfeeding	596	0.275	0.016	577	0.241	0.018	0.153
Adult characteristics							
Works for income*	1453	0.793	0.014	1465	0.825	0.016	0.147
Agriculture	1453	0.347	0.040	1465	0.457	0.024	0.017
Non-agriculture	1453	0.446	0.037	1465	0.368	0.016	0.050
Women work for income*	764	0.619	0.028	769	0.680	0.032	0.156
Agriculture	764	0.279	0.044	769	0.406	0.036	0.025
Non-agriculture	764	0.340	0.041	769	0.274	0.018	0.140
Male literacy	742	0.794	0.019	741	0.735	0.024	0.060
Female literacy	834	0.698	0.014	806	0.649	0.021	0.053
Female education							
No education	834	0.210	0.018	806	0.241	0.022	0.276
Primary school	834	0.265	0.022	806	0.249	0.011	0.530
Middle school	834	0.207	0.018	806	0.223	0.017	0.520
High school	834	0.203	0.015	806	0.202	0.010	0.982
Higher secondary or more	834	0.115	0.016	806	0.081	0.012	0.079
Mother's age (years)	446	26.978	0.235	438	26.760	0.228	0.506

(continued on the next page)

Table 4.5: *(continued)*

	Control			Intervention			
	N	Mean	SE	N	Mean	SE	p-value
Household characteristics							
Scheduled caste	456	0.140	0.037	444	0.119	0.068	0.787
Participates in a committee or group	456	0.482	0.037	444	0.450	0.057	0.640
Women participate in credit/finance/SHG†	456	0.351	0.041	444	0.338	0.040	0.821
Soil floor	456	0.279	0.034	444	0.351	0.022	0.071
Thatched roof	456	0.208	0.030	444	0.282	0.031	0.095
Total persons living in house	456	4.763	0.090	444	4.784	0.050	0.842
Total rooms in house	456	2.662	0.128	444	2.725	0.107	0.706
Sleeping rooms in house	456	1.794	0.069	444	1.761	0.064	0.729
Electricity	456	0.919	0.018	444	0.881	0.032	0.293
Home ownership	456	0.888	0.023	444	0.966	0.009	0.002
Land ownership	456	0.919	0.018	444	0.975	0.007	0.004
Bank account	456	0.221	0.017	444	0.209	0.028	0.714
Refrigerator	456	0.039	0.009	444	0.014	0.006	0.017
Radio	456	0.592	0.028	444	0.525	0.032	0.109
Television	456	0.728	0.058	444	0.577	0.059	0.066
Mobile phone	456	0.322	0.032	444	0.331	0.032	0.848
Motorcycle/scooter	456	0.270	0.028	444	0.236	0.021	0.338
Bicycle	456	0.737	0.037	444	0.791	0.031	0.267
Mosquito net	456	0.123	0.019	444	0.142	0.015	0.429

* Working populations exclude individuals reported to be too young to work or retired.

† SHG: Self-help group.

4.4.3 Improvements in household sanitation and water infrastructure

The WPI/Gramalaya program has greatly expanded access to private toilets and improved water sources (Figure 4.4). Intervention households are more than 3 times as likely to have constructed a new private toilet between 2003 and 2008 than control households (48% vs. 15%). Gains in new water sources (mainly private and public taps) have been more modest but are still substantial: 26% of intervention households versus 18% of control households report a new water source between 2003 and 2008. Intervention households are 1.5 times more likely to have installed a new private tap in their house over the period than control households (12% in intervention vs. 8% in control).

By subtracting the estimates of new toilet and tap construction from current estimates of toilet and tap ownership, we estimate the proportion of households that had private toilets and taps at baseline (before the intervention). Intervention and control villages were highly similar in private toilet ownership (9% vs. 11%) and private tap ownership (18% vs. 19%) before the intervention (Figure 4.5).

4.4.4 Sanitation and open defecation

Consistent with the large gains that resulted from the intervention, by 2008 intervention households are 2.2 times more likely to own a private toilet than control households (57% vs. 26%, Figure 4.5, Table 4.6). Over 89% of the private toilets in the study population are flush toilets, 5% are ventilated improved pit latrines, and 5% are unimproved concrete slab pit latrines. Over 83% of toilets were constructed within the last 5 years (since 2003) and 94% were constructed in the last 10 years. Of the 374 households with private toilets, 353 (94%) were classified as functional and in use during interviewer inspections over the 12 month period.

Households in intervention villages are 1.2 times less likely to report adults practicing open defecation (69% vs. 84%) than control households (Table 4.6). One component of the intervention was to declare villages "open defecation free". Although the majority of adults still report practicing open defecation, 98% of adults from all villages in our sample report that defecation sites fall outside of the village boundaries. Reductions in open defecation have been largest among women and smallest among children under 5 (Figure 4.6). Women living in intervention villages are 20 percentage points (61% vs. 81%) less likely to report practicing open defecation than women living in control villages. Accross all study villages, 82% of children < 5 practice open defecation and 91% of these defecation events occur within the village.

Ownership of a private toilet has not eliminated open defecation practices among adults in the study. Just under 40% of study households that have a private toilet report that adults practice open defecation daily (Figure 4.7). In households with private toilets, over 52% report that children < 5 still practice open defecation daily. Among households

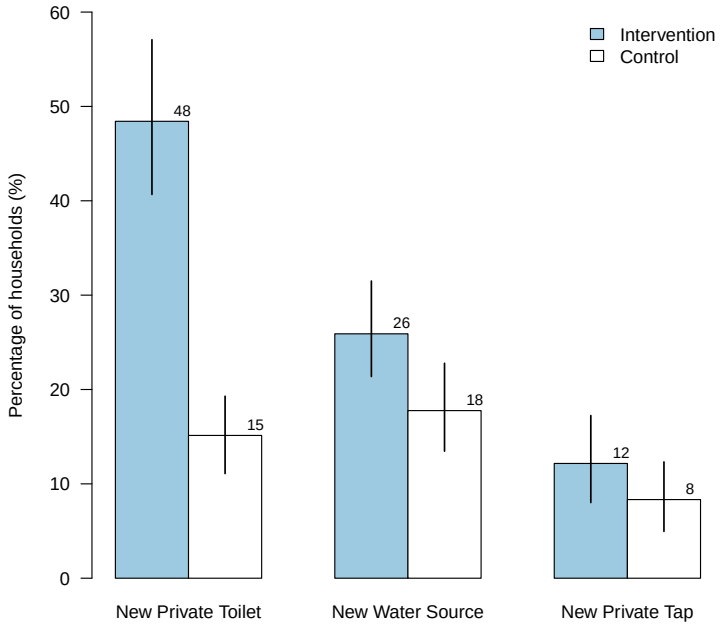


Figure 4.4: New private toilet and water sources in the five years that cover the intervention period (2003-2008). New private water taps are a subset of any new water source. N=456 control and N=444 intervention households. Vertical lines indicate 95% confidence intervals that were estimated by bootstrap resampling matched village pairs.

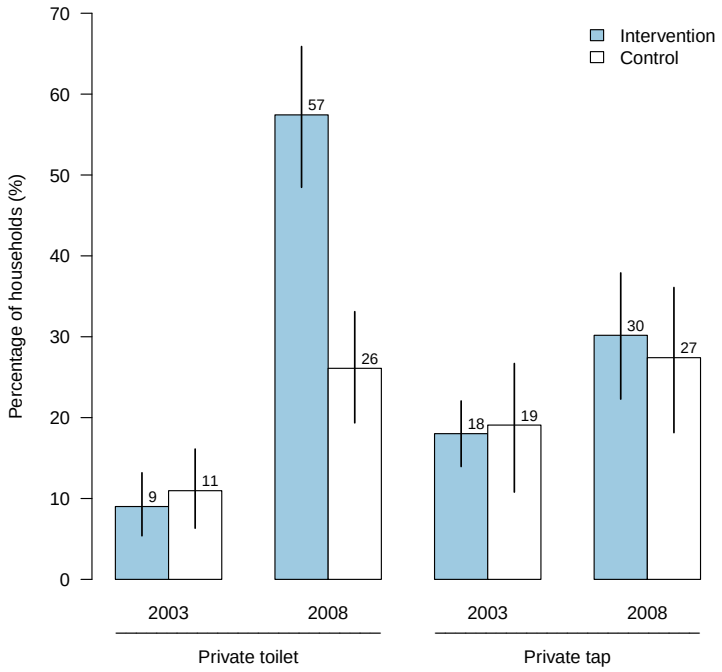


Figure 4.5: Private toilet and tap ownership before (2003) and after (2008) the intervention. We obtained 2003 estimates retrospectively by subtracting newly constructed facilities from those existing in 2008. N=456 control and N=444 intervention households. Vertical lines indicate 95% confidence intervals that were estimated by bootstrap resampling matched village pairs.

that own private toilets, the primary reasons given for continuing to practice open defecation are: no choice (50%), privacy (26%), convenience (25%) and safety (9%). Working in agriculture (and likely defecating in the fields) may contribute to the persistence of adult open defecation among toilet owners, but does not entirely explain it: among toilet owning households, if a member of the household works in agriculture the family is 14.9 percentage points (95% CI $3.5\%, 26.3\%$) more likely to report that adults practice open defecation daily than if the household does not have anybody who works in agriculture (44.4% vs. 29.5%).

On average, private toilets increase the perception of privacy and safety for women and girls during defecation. Private toilet owners are 1.5 times more likely (81.3% vs. 53.4%) to report that women and girls feel safe during defecation during the day or night than households that do not own private toilets (difference = 27.8%, 95% CI: 18.3%, 36.6%). The increase in private toilets in intervention villages has increased the overall perception of privacy and safety among women in intervention villages (Table 4.6). The intervention increased the perception of privacy and safety for women and girls during defecation all by 13 percentage points compared to control households (72% vs. 59%).

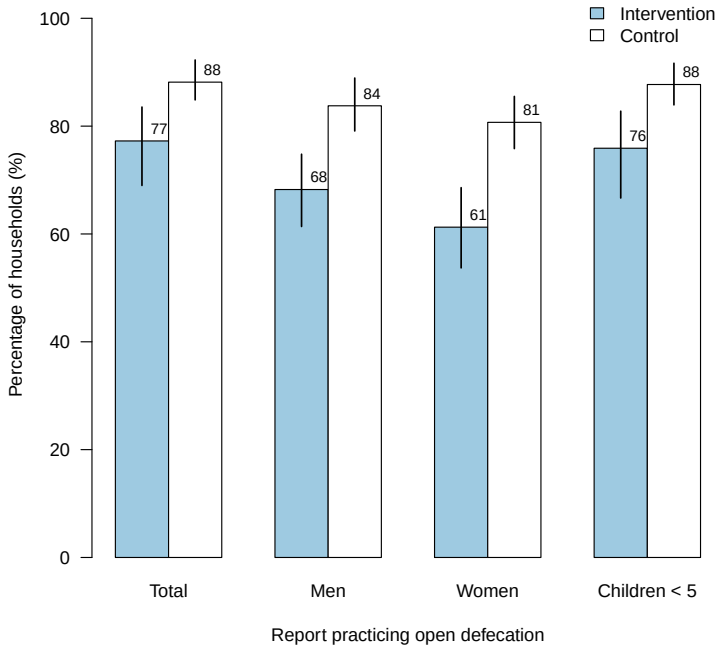


Figure 4.6: Open defecation practices among men, women and children under 5 years old in intervention and control households. N=456 control and N=444 intervention households. Vertical lines indicate 95% confidence intervals that were estimated by bootstrap resampling matched village pairs.

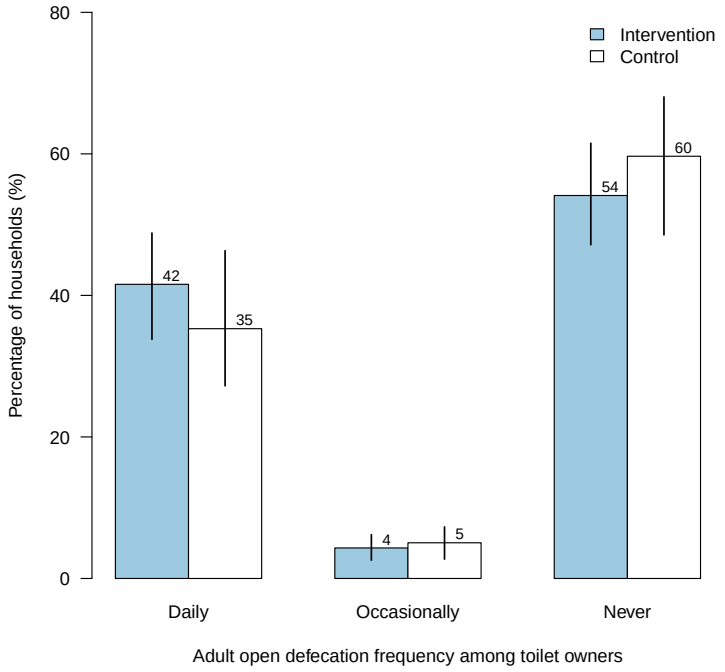


Figure 4.7: Adult open defecation frequency in households that have a private toilet in intervention and control households. N=119 control and N=255 intervention households. Vertical lines indicate 95% confidence intervals that were estimated by bootstrap resampling matched village pairs.

Table 4.6: Summary of open defecation (OD), private toilets and perceived privacy/safety for women and girls in control and intervention households. N=456 control and N=444 intervention households.

Outcome	Control		Intervention		Risk Difference	
	N	(%)	N	(%)	(95% CI)*	
Open Defecation						
Any OD	402	(88.2)	343	(77.3)	−0.109	(−0.208, −0.036)
Any adult OD	382	(83.8)	306	(68.9)	−0.149	(−0.229, −0.076)
Adult men OD	382	(83.8)	303	(68.2)	−0.155	(−0.236, −0.081)
Adult women OD	368	(80.7)	272	(61.3)	−0.194	(−0.264, −0.129)
Children < 5 OD	400	(87.7)	337	(75.9)	−0.118	(−0.217, −0.038)
Private toilets						
Have toilet in 2008	119	(26.1)	255	(57.4)	0.313	(0.233, 0.399)
New toilet since 2003	69	(15.1)	215	(48.4)	0.333	(0.266, 0.410)
Perceived privacy/safety for women & girls during defecation						
Women/girls have privacy	269	(59.0)	320	(72.1)	0.131	(0.021, 0.230)
Defecation safe, daytime	267	(58.6)	318	(71.6)	0.131	(0.013, 0.235)
Defecation safe, nighttime	267	(58.6)	317	(71.4)	0.128	(0.018, 0.234)

* Standard Errors and 95% confidence intervals calculated by bootstrap resampling matched village pairs with 1000 iterations.

4.4.5 Water sources and water quality

The study population primarily uses public and private taps as their water sources, and none report rainwater harvesting or purchasing water from a tanker or vender (Table 4.7). Although 10% of the study population reports that they have access to surface water, fewer than 1% of households report using it, and if used it is mainly for washing clothes (Table 4.7). The vast majority of households report using a single source: of 900 households, only 89 (10%) report using more than one source. Of these multi-source users, the majority (46%) use a mix of public taps and public wells. Intervention and control households are similar in their primary water sources (Figure 4.8). Switching from a public tap to a private tap saves a household a median 25 minutes each day: households that have a private tap as their primary water source spend a median of 50 minutes per day gathering water, and households with public taps spend a median of 75 minutes per day gathering water (Figure 4.9). Overall, households spend a median of 60 minutes per day fetching water (mean = 83 minutes). Consistent with highly similar water sources in intervention and control households, the time spent gathering water is not different between the two groups (median in both groups is 70 minutes per day, Figure 4.10).

Consistent with more recent improvements in public water sources in intervention villages, village source water quality is better in intervention villages than in control villages (Table 4.8). Intervention village water sources had less E. coli per 100 ml (difference in $\log_{10} = -0.29$, 95% CI: $-0.56, -0.05$) and the proportion of samples testing positive for E. coli (8.5% vs 16.6%, difference = -8.1%, 95% CI: $-1.0\%, 015.2\%$). Measures of total coliform concentrations and H_2S tests were also better in intervention villages, but were not statistically different from control.

By all water quality measures, household water samples were more contaminated than village source samples (Table 4.8, Figures 4.11, 4.12). For example, 15.4% of village overhead tank water samples tested positive for E. coli, but 22.6% of samples from households who use a public tap and 20.6% water samples from households who use a private tap tested positive for E. coli. More than 99% of household drinking water samples were from stored water (not collected directly from the tap).

Although household water samples in intervention villages were consistently cleaner than samples from control villages, differences between the groups were smaller than differences at the village source level, and none of the differences are statistically significant at the 95% confidence level (Table 4.8).

Table 4.7: Proportion (%) of households that use various water sources. The first two data columns summarize the proportion of households that could potentially use a source and the proportion that actually use a source. A primary source is the source that the household reports using most often. N=900 households.

Water source	Could use	Ever use	Primary	Activities			
			Source	Drinking	Cooking	Bathing	Washing
Private tap	29.9	29.3	28.8	27.6	27.7	28.8	29.0
Public tap	81.9	68.0	63.8	66.0	65.8	65.9	64.6
Private well (tube/bore/dug)	4.7	4.2	3.1	4.1	3.3	4.0	4.1
Public well (tube/bore/dug)	26.9	6.9	4.3	4.9	4.8	3.2	3.7
Neighbors (that give water away)	1.2	0.8	0.0	0.8	0.7	0.8	0.8
Surface water (river/stream/spring/lake)	9.9	0.6	0.0	0.0	0.0	0.1	0.6
Tanker/vender	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Rainwater	0.1	0.0	0.0	0.0	0.0	0.0	0.0

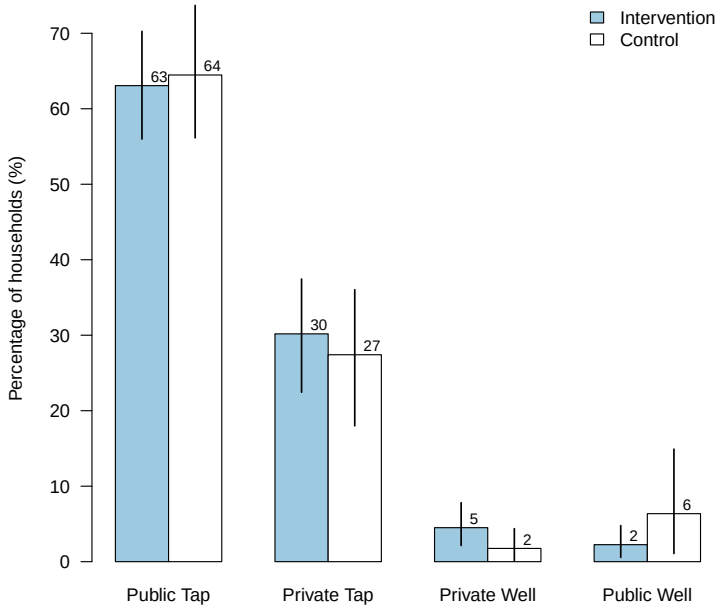


Figure 4.8: Primary water sources by intervention group. N=456 control and N=444 intervention households. Vertical lines indicate 95% confidence intervals that were estimated by bootstrap resampling matched village pairs.

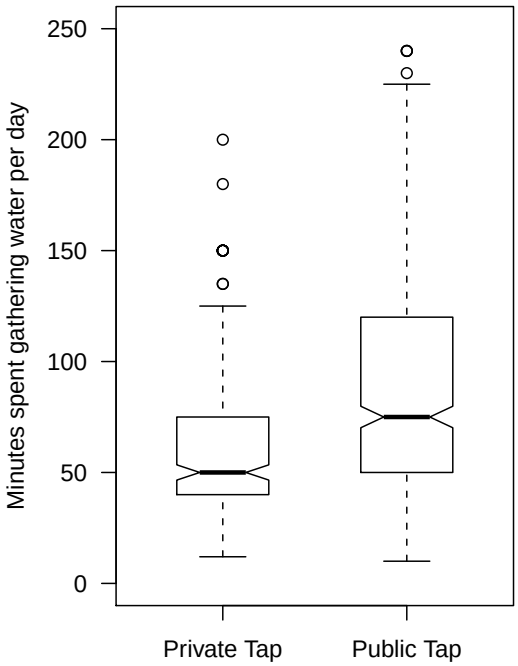


Figure 4.9: Time spent gathering water each day in households with private and public taps. The heavy lines and wedges indicate median values, the ends of the boxes are 25th and 75th percentiles, and the whiskers are 1.5 times the interquartile range. Outliers are indicated with dots.

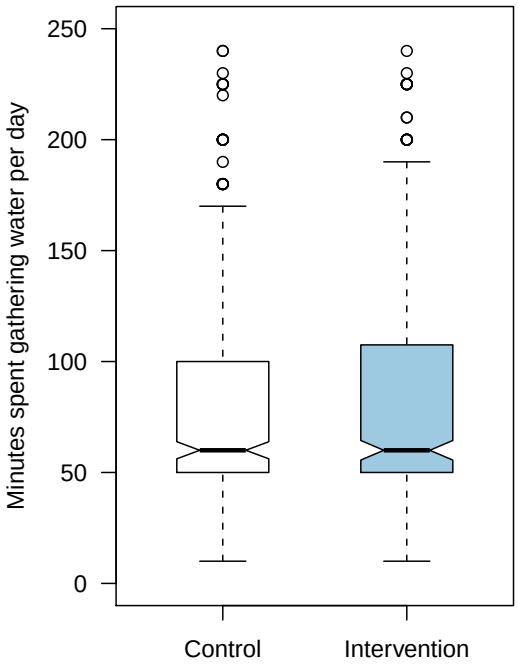


Figure 4.10: Time spent gathering water each day in intervention and control households. The heavy lines and wedges indicate median values, the ends of the boxes are 25th and 75th percentiles, and the whiskers are 1.5 times the interquartile range. Outliers are indicated with dots.

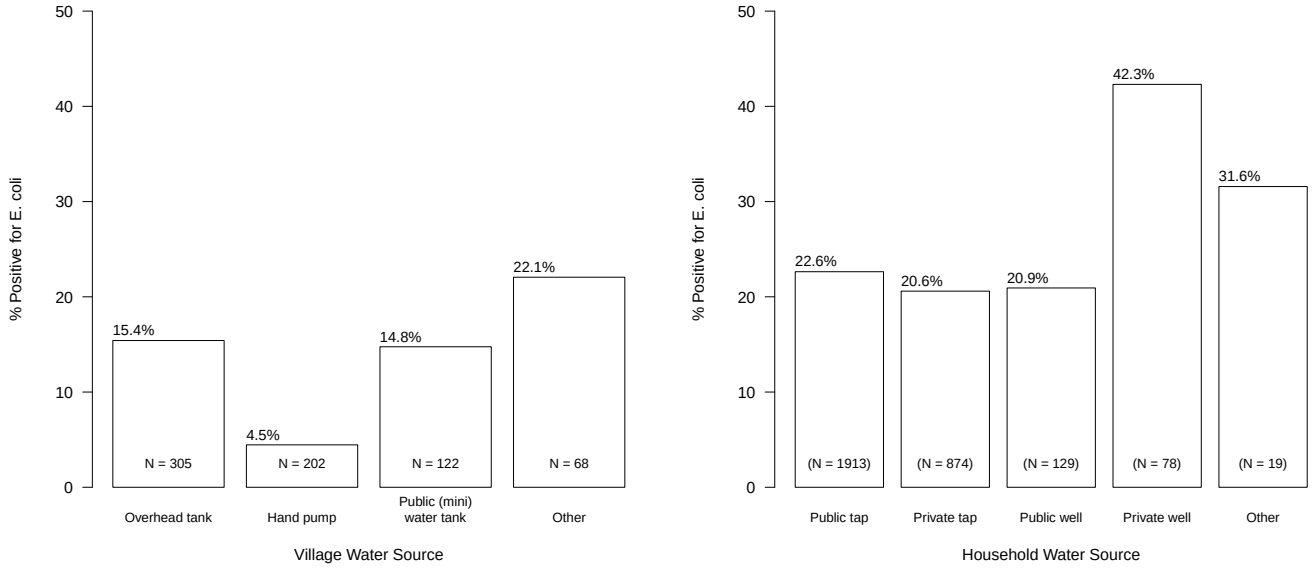


Figure 4.11: Proportion of water samples testing positive for E. coli. The left plot summarizes village water source samples, and the right plot summarizes household water samples.

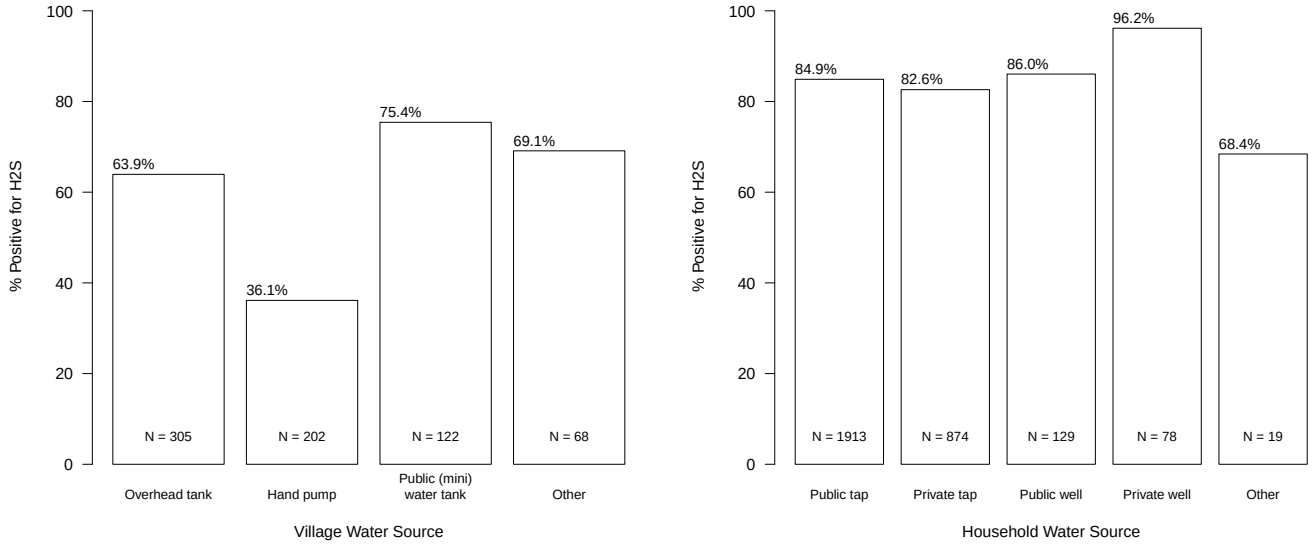


Figure 4.12: Proportion of water samples testing positive for H_2S producing bacteria. The left plot summarizes village water source samples, and the right plot summarizes household water samples.

Table 4.8: Summary of mean water quality measures for village source samples and household water samples measured over the study period. Mean $\log_{10}$ coliform and E. coli concentrations are per 100 ml.

Water quality measure	Control		Intervention		Difference	
	N	Mean	N	Mean	(95% CI)*	
Village source samples						
Log_{10} Total Coliforms	366	2.628	330	2.549	-0.078	(−0.534, 0.357)
Log_{10} E. coli	367	-0.427	330	-0.718	-0.291	(−0.553, −0.058)
Positive for E. coli (%)	367	0.166	330	0.085	-0.081	(−0.154, −0.014)
Positive for H_2S (%)	367	0.627	330	0.536	-0.090	(−0.218, 0.027)
Household water samples						
Log_{10} Total Coliforms	1269	3.304	1187	3.211	-0.092	(−0.215, 0.022)
Log_{10} E. coli	1277	-0.182	1197	-0.299	-0.117	(−0.256, 0.037)
Positive for E. coli (%)	1277	0.235	1197	0.211	-0.024	(−0.067, 0.021)
Positive for H_2S (%)	1271	0.859	1194	0.836	-0.023	(−0.061, 0.012)

* Standard Errors and 95% confidence intervals calculated by bootstrap resampling matched village pairs with 1000 iterations.

4.4.6 Hygiene and handwashing

The hygiene and handwashing information component of the intervention has not led to detectable improvements in discrete spot check observations of hygienic conditions or self-reported handwashing practices. Based on spot check observations collected by interviewers, intervention households fare the same or worse across a large number of indicators (Table 4.9). For example, over 12 monthly measurements intervention and control households are equally likely to have a dedicated handwashing station with water, but intervention households are slightly less likely to have a handwashing station with water and either soap, detergent or ash (61% vs. 64%; risk difference [RD] = −0.03, 95% CI: $-0.05, -0.01$). Intervention households are also less likely than controls to have soap available for handwashing at their toilet (45% vs. 59%; RD = −0.142, 95% CI: $-0.224, -0.035$), and are more likely to have feces observed in the living area (34% vs. 26%; RD = 0.079, 95% CI: $0.035, 0.125$). There are no differences between intervention and in observations of cleanliness for children < 5.

The count of critical times that primary caregivers reported washing hands with water alone and with soap are very similar between groups (Figure 4.13). Caregivers report

washing with soap with less frequency than with water alone: the median count of critical times reported washing with water alone is 5 in both intervention and control, and the median count of critical times reported washing with water and soap is 0 in both groups.

Self-reported handwashing with water alone is most common around eating, followed by feeding children, defecation and cooking (Table 4.10). Overall, reported handwashing with soap is rare: in 24.3% of caregiver interviews the woman reported washing her hands after defecation (the most common time, Table 4.10). Caregivers report handwashing with soap primarily around contact with feces (defecation, changing the baby, cleaning the house or cattle shed). Although caregivers in intervention households are slightly more likely to report washing their hands after defecation or handling their baby's feces, the differences are small ($\leq 2\%$) and not statistically significant (Table 4.11)

Table 4.9: Summary of hygiene spot check observations. Unless noted, data were collected during 12 monthly visits in 456 control households (648 children < 5) and 444 intervention households (637 children < 5). Numbers reported (N) are the number of positive instances. Total N varies slightly by indicator.

Outcome	Control		Intervention		Risk Difference
	N	(%)	N	(%)	(95% CI)*
Handwashing station spot check					
Station with water	5297	(72.0)	5130	(70.0)	−0.020 (−0.053, 0.014)
Station with water & soap/detergent/ash	5297	(63.6)	5130	(61.0)	−0.026 (−0.047, −0.004)
Station with basin/sink	5297	(13.3)	5130	(11.2)	−0.021 (−0.032, −0.008)
Latrine spot check					
Hole is covered	1048	(4.3)	2291	(4.1)	−0.001 (−0.020, 0.019)
Water available for handwashing †	97	(93.8)	222	(92.3)	−0.015 (−0.058, 0.029)
Soap available for handwashing	1048	(58.7)	2290	(44.5)	−0.142 (−0.222, −0.035)
Toilet paper available	951	(83.8)	2069	(84.2)	0.004 (−0.063, 0.083)
Feces on ground (not in hole)	951	(0.9)	2069	(0.6)	−0.004 (−0.010, 0.004)
Animals observed in the living area					
Cows/buffalo/oxen ‡	1346	(23.4)	1294	(34.3)	0.109 (0.047, 0.167)
Goats/sheep ‡	1346	(22.7)	1305	(25.4)	0.028 (−0.019, 0.074)
Chickens ‡	1346	(14.9)	1305	(19.0)	0.041 (0.001, 0.086)
Dogs/cats ‡	1346	(19.3)	1306	(19.1)	−0.003 (−0.034, 0.041)
Feces observed in living area	5293	(25.8)	5120	(33.7)	0.079 (0.035, 0.125)
Staff could smell feces during interview	5297	(10.6)	5130	(14.8)	0.042 (0.016, 0.072)
Kitchen spot check †					

Table 4.9 – continued on next page

Table 4.9 – continued from previous page

Outcome	Control		Intervention		Risk Difference (95% CI)*	
	N	(%)	N	(%)		
Food is covered	456	(97.8)	444	(98.2)	0.004	(−0.018, 0.029)
Garbage present inside home	454	(8.1)	436	(5.7)	−0.024	(−0.054, 0.010)
Flies present inside home	454	(15.2)	436	(15.4)	0.002	(−0.058, 0.059)
Can produce a bar of soap	456	(86.4)	444	(84.7)	−0.017	(−0.059, 0.021)
Soap is in plain view	454	(26.2)	436	(19.5)	−0.067	(−0.104, −0.028)
Children < 5 spot check						
Hands dirty	4903	(8.5)	4865	(8.6)	0.001	(−0.011, 0.016)
Dirt/mud in fingernails	4903	(20.7)	4864	(20.3)	−0.004	(−0.027, 0.021)
Face dirty	4904	(16.0)	4863	(16.2)	0.002	(−0.021, 0.028)
Clothes dirty	4208	(18.0)	4185	(19.1)	0.011	(−0.014, 0.039)
No clothes	4900	(14.4)	4864	(14.1)	−0.003	(−0.026, 0.020)
Shoes	4903	(0.3)	4865	(0.6)	0.003	(0.000, 0.007)

* Standard Errors and 95% confidence intervals calculated by bootstrap resampling matched village pairs with 1000 iterations.

† Measurement in survey round 1 only.

‡ Measurement in survey rounds 1-3 only.

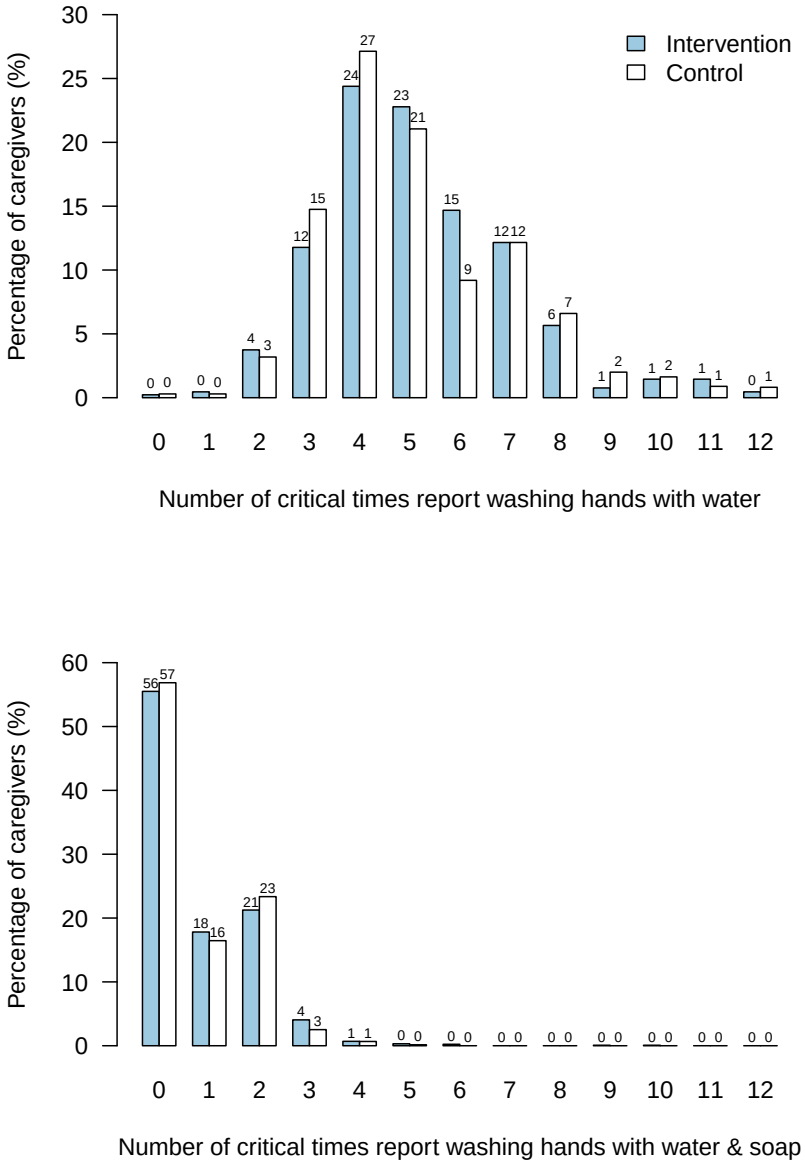


Figure 4.13: Counts of self-reported handwashing with water alone (top plot) and water plus soap (bottom plot) during 12 critical times. Counts are sums over coded responses to an open ended question to caregivers about handwashing in the previous 24 hours. N=1,349 control, N=1,308 intervention.

Table 4.10: Proportion of caregivers reporting washing their hands with water alone or water with soap during 12 critical times. Coded responses from an open-ended question: *When and how did you wash your hands in the last 24 hours (since this time yesterday)? ... Any other times?* N = 2,657 caregiver interviews.

	Critical time	Water Alone	Water & Soap
1	Before preparing food or cooking	39.9	0.2
2	After preparing food or cooking	45.8	0.3
3	Before eating	90.5	0.8
4	After eating	87.5	1.2
5	Before serving food	20.5	0.3
6	After serving food	18.9	0.3
7	Before feeding children	59.4	1.7
8	After changing baby / handling baby's feces	36.7	15.7
9	After defecation	60.9	24.3
10	After attending to cattle	9.9	3.5
11	After cleaning house / cattle shed	19.1	19.8
12	After returning from work / outside visit	24.8	8.7

Table 4.11: Caregiver self-reported handwashing with soap after four critical times with potential for contact with human or animal feces. Coded responses from an open-ended question. N=1,349 control and N=1,308 intervention interviews.

Report washing hands	Control		Intervention		Risk Difference	
with soap after:	N	(%)	N	(%)	(95% CI)*	
Changing baby/ handling baby's feces	201	(14.9)	216	(16.5)	0.016	(−0.009, 0.037)
Defecation	321	(23.8)	324	(24.8)	0.010	(−0.029, 0.047)
Attending to cattle	38	(2.8)	54	(4.1)	0.013	(−0.002, 0.029)
Cleaning house/ cattle shed	283	(21.0)	243	(18.6)	−0.024	(−0.058, 0.015)

* Standard Errors and 95% confidence intervals calculated by bootstrap resampling matched village pairs with 1000 iterations.

4.4.7 Sustainability of sanitation and hygiene behaviors

The oldest private toilets reported by participants were more than 20 years old (N=11), though we focus on latrines five years or newer due to the potential for measurement error over longer recall periods. Private toilets appear to be highly sustainable over at least a five year window. The proportion of toilets in use did not vary greatly by age, but was lowest among latrines less than or equal to 1 year old (Table 4.12).

Table 4.12: Summary of the proportion of private toilets in use by age of the toilet (reported by household members).

Age (years)	N	In Use (%)
1	58	75.9
2	77	94.8
3	65	94.8
4	54	98.1
5	30	100.0
>5	57	100.0
Unknown	33	100.0
Total	374	94.4

Figure 4.14 summarizes the village mean prevalence of open defecation by time since intervention completion at the initial survey. Ten of the 12 villages have similar open defecation prevalence, regardless of time since intervention completion. Two villages that were completed 17 months prior to our survey (Melanaduvalur and Kanganipatti) have substantially lower prevalence of reported open defecation in both adults and children under age 5. This is, in part, because they have the highest coverage of private toilets, which is negatively associated with open defecation (Figure 4.15).

Figure 4.16 summarizes mean village prevalence of four spot check hygiene indicators by time since intervention completion. Unlike open defecation, we measured hygiene indicators in every visit, so we have more village-level measurements (each village contributes 12 measurements). There is no clear increase or decrease in these hygiene indicators with time since intervention completion.

4.4.8 Child growth

Children in the study population are very small for height and weight by international standards. A Z-score of 0 is average by WHO international standards, and Z-scores below -2 indicate stunting (height), underweight (weight), or wasting/malnutrition (height-for-weight and mid-upper arm circumference). By these measures, 57% of the children are

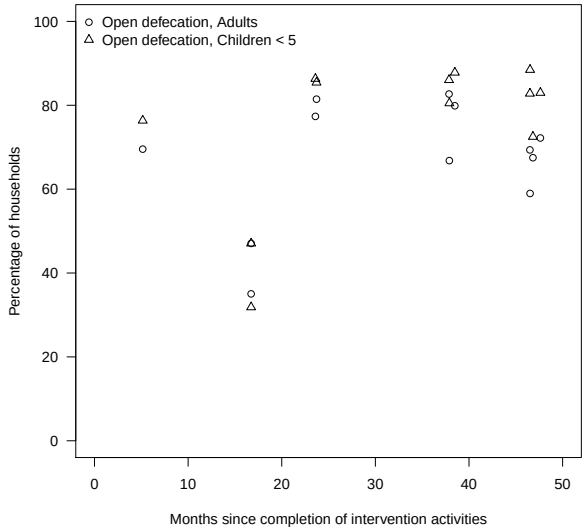


Figure 4.14: Village mean proportion of households that report open defecation by time since completion of the intervention. Data include 444 intervention households in 12 villages.

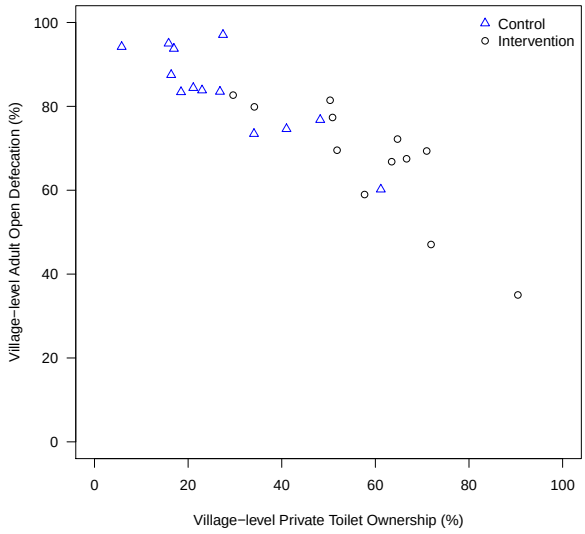


Figure 4.15: Village-level adult open defecation prevalence versus village-level private toilet ownership. Data include 900 households measured in 25 villages.

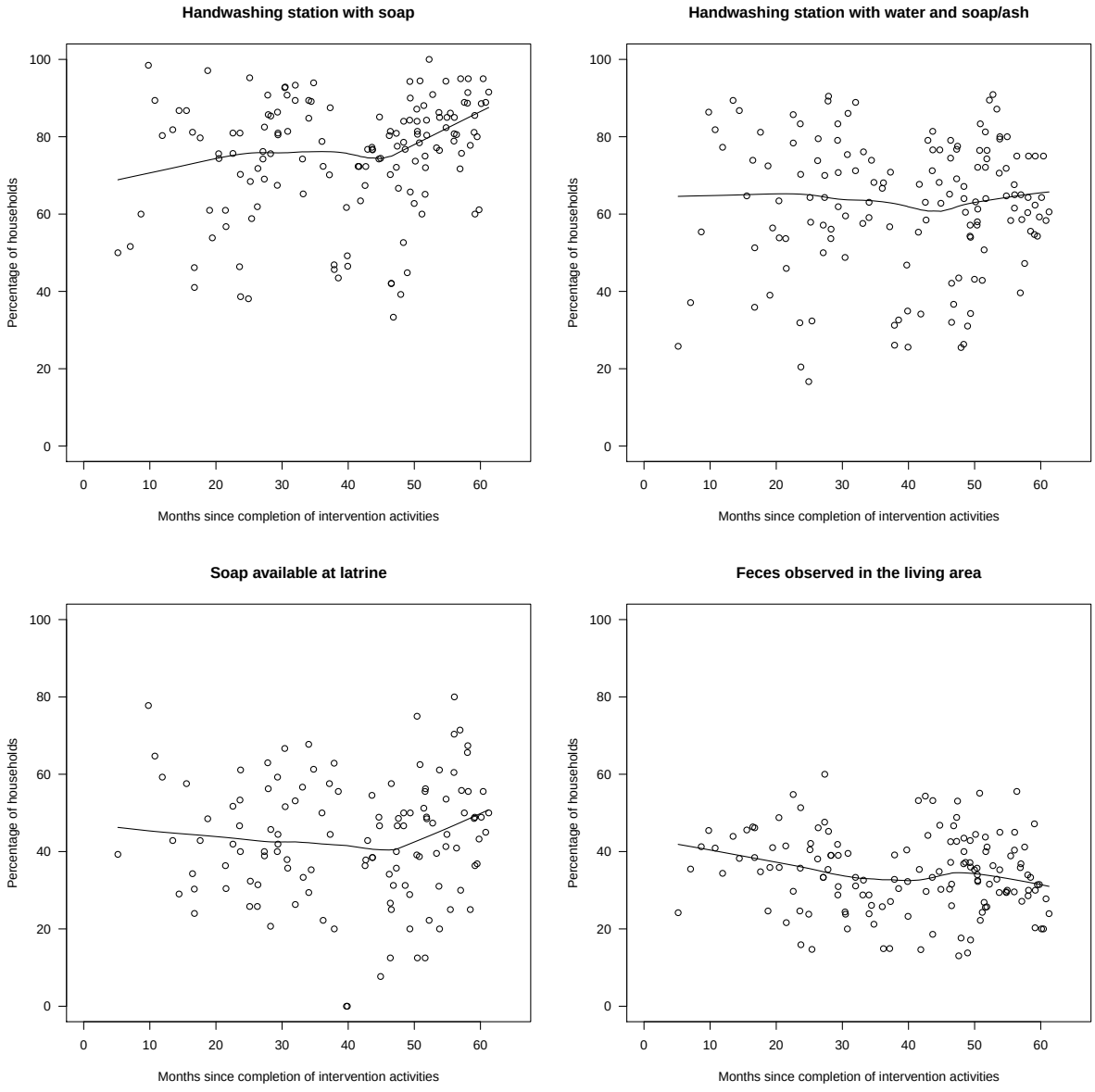


Figure 4.16: Village mean proportion of households with four hygiene indicators by time since intervention completion. Data include 444 intervention households in 12 villages measured 12 times over one year. Solid lines are locally weighted regression fits.

stunted, 53% are underweight, and between 26% (based on weight-for-height) and 44% (based on upper arm circumference) are malnourished. In addition, 42% of the children are both stunted and wasted. These prevalence estimates are based on the minimum of at most two measures for each child over the follow-up period. Most of the growth faltering occurs during the first 24 months, particularly in height (Figure 4.17). Z-scores for height correlate with all other anthropometry measures, but they most strongly correlate with weight Z-scores (Pearson's $R = 0.67$, Figure 4.18).

Child growth does not differ between intervention and control villages. Figures 4.19 includes box plots of Z-scores for each growth measurement, and the two populations are virtually indistinguishable. Figure 4.20, plots the same data but illustrates the continuous distributions of the two populations. Ideally, the Z-scores would be centered on zero, but instead the distributions are shifted to the left and a large proportion of children fall into poor growth categories (as described in the previous paragraph).

In unadjusted analyses children in intervention villages fall on average between 0.04 and 0.12 standard deviations below children in control villages (Table 4.13). Adjusted analyses with the G-computation and targeted MLE estimators led to slightly different point estimates, but do not modify our conclusions: differences between groups are small and fall at or below 0.07 standard deviations (Table 4.14).

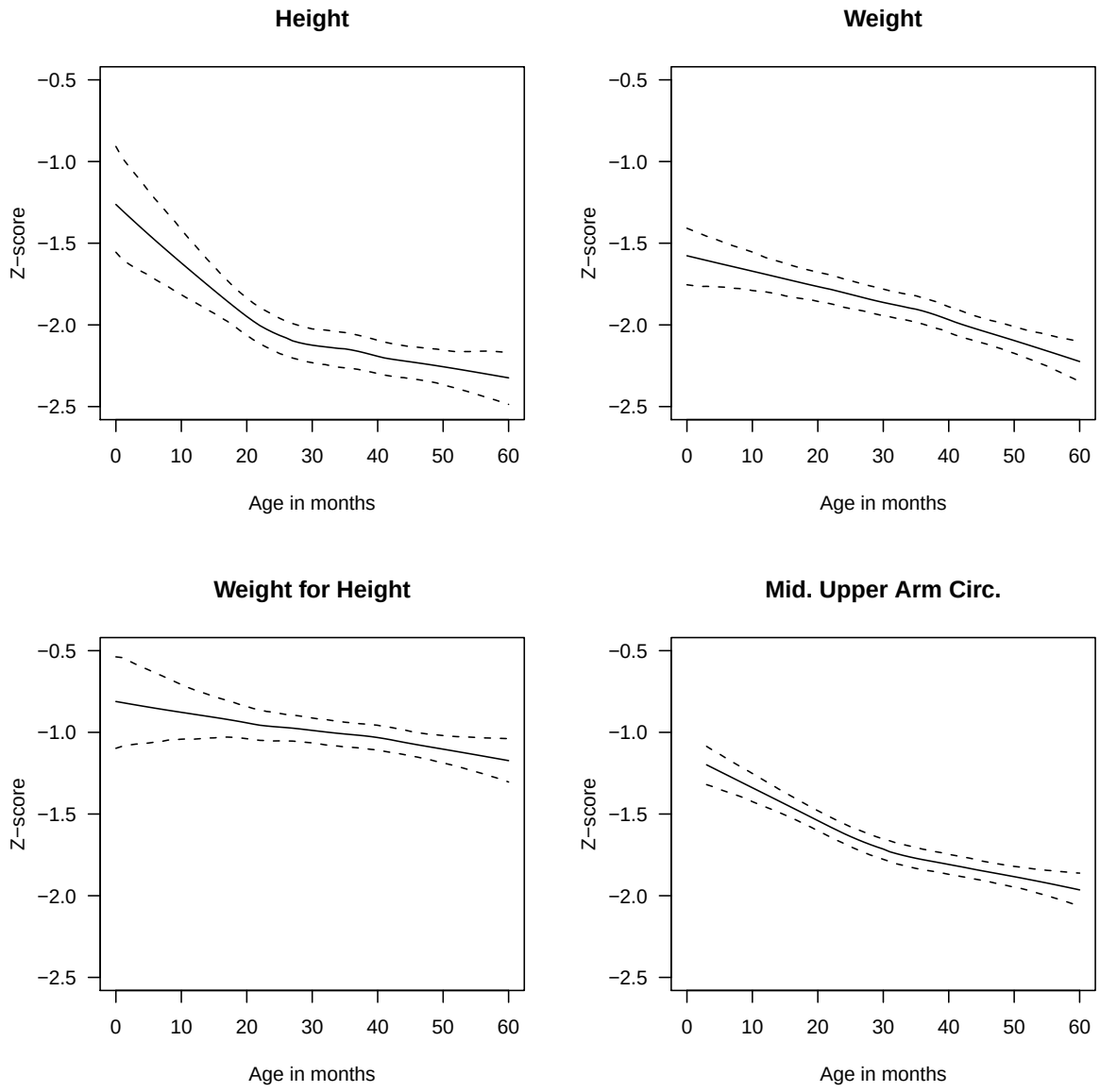


Figure 4.17: Anthropometric Z-scores by age in children under age 5. Solid lines represent a locally weighted regression line (lowess). Dashed lines are bootstrapped 95% confidence intervals for the lowess curves. Data were collected in the first and last survey rounds.

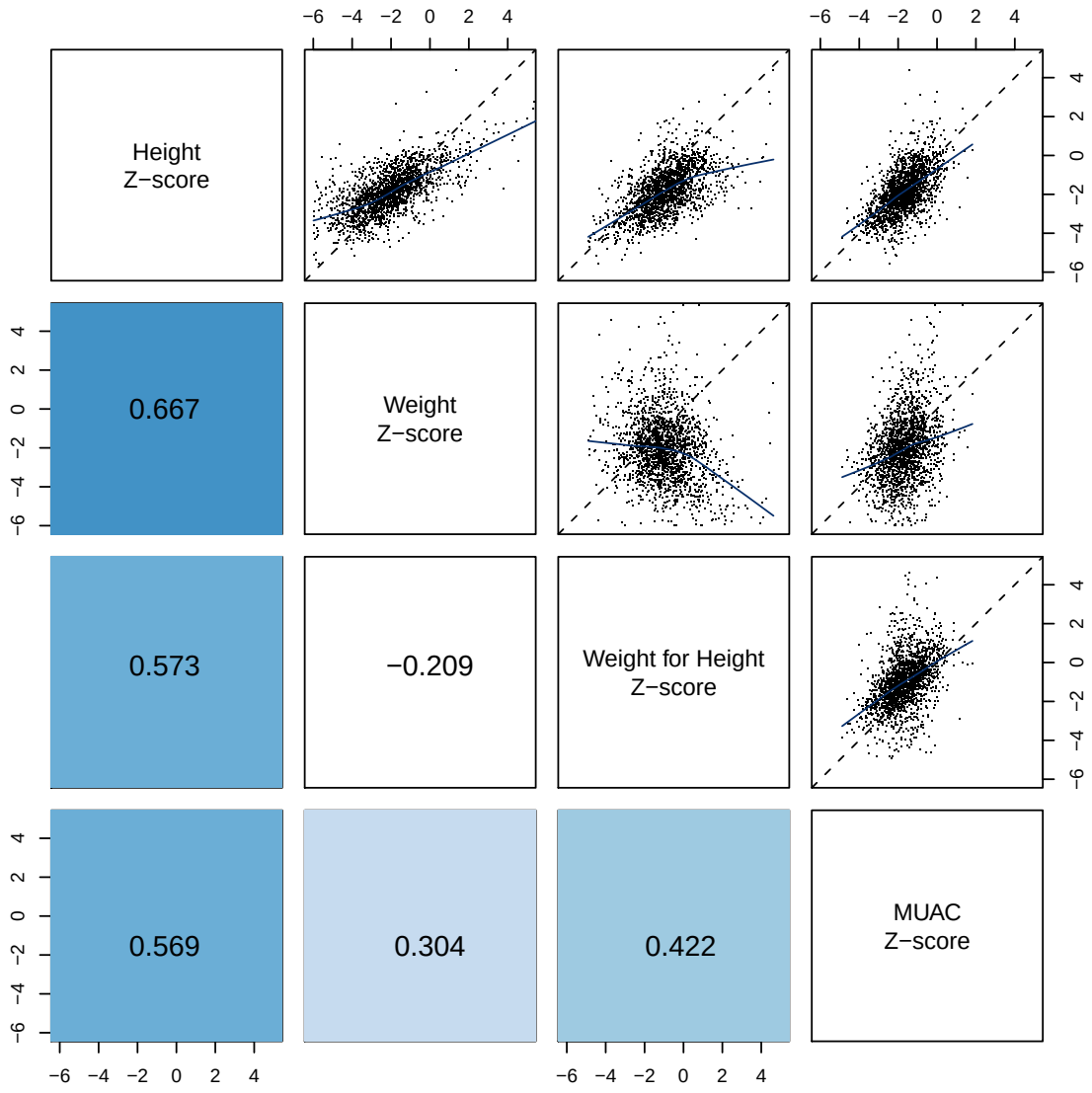


Figure 4.18: Pairs plot of anthropometric Z-scores. The lower panel includes Pearson correlation coefficients (colored by strength). The upper panel includes bivariate scatter plots; the dashed lines are 1:1 and the solid lines are locally weighted regression fits (lowess). Data were collected in the first and last survey rounds.

Table 4.13: Anthropometric Z-scores in children under age 5. Data were collected in the first and last survey rounds.

Z-score	Control			Intervention			Difference	
	N	Mean	SD	N	Mean	SD		(95% CI)*
Height	994	-1.96	1.69	974	-2.00	1.69	−0.037	(−0.337, 0.209)
Weight	1006	-1.86	1.16	983	-1.90	1.19	−0.044	(−0.262, 0.123)
Weight-for-height	990	-0.92	1.31	962	-1.02	1.34	−0.098	(−0.302, 0.099)
Upper Arm Circ.	1000	-1.63	0.90	977	-1.75	0.97	−0.123	(−0.308, 0.019)

* Standard Errors and 95% confidence intervals calculated by bootstrap resampling matched village pairs with 1000 iterations.

Table 4.14: Treatment effect estimates of intervention minus control for anthropometry Z-score outcomes in children under age five. Table 4.13 includes sample size information.

Estimator	Z-score, Height			Z-score, Weight		
	Difference	SE*	95% CI	Difference	SE*	95% CI
Unadjusted	-0.0370	0.1417	$(-0.345, 0.215)$	-0.0554	0.0998	$(-0.272, 0.116)$
G-comp	0.0321	0.1092	$(-0.181, 0.231)$	0.0155	0.0869	$(-0.173, 0.171)$
T-MLE	0.0527	0.1245	$(-0.186, 0.275)$	0.0317	0.0950	$(-0.176, 0.191)$

Estimator	Z-score, Weight-for-Height			Z-score, UAC		
	Difference	SE*	95% CI	Difference	SE*	95% CI
Unadjusted	-0.1021	0.1025	$(-0.306, 0.092)$	-0.1264	0.0850	$(-0.311, 0.016)$
G-comp	-0.0590	0.0880	$(-0.260, 0.091)$	-0.0714	0.0712	$(-0.222, 0.035)$
T-MLE	-0.0706	0.1104	$(-0.297, 0.125)$	-0.0713	0.0862	$(-0.248, 0.071)$

* Standard Errors and 95% confidence intervals calculated by bootstrap resampling matched village pairs with 1000 iterations.

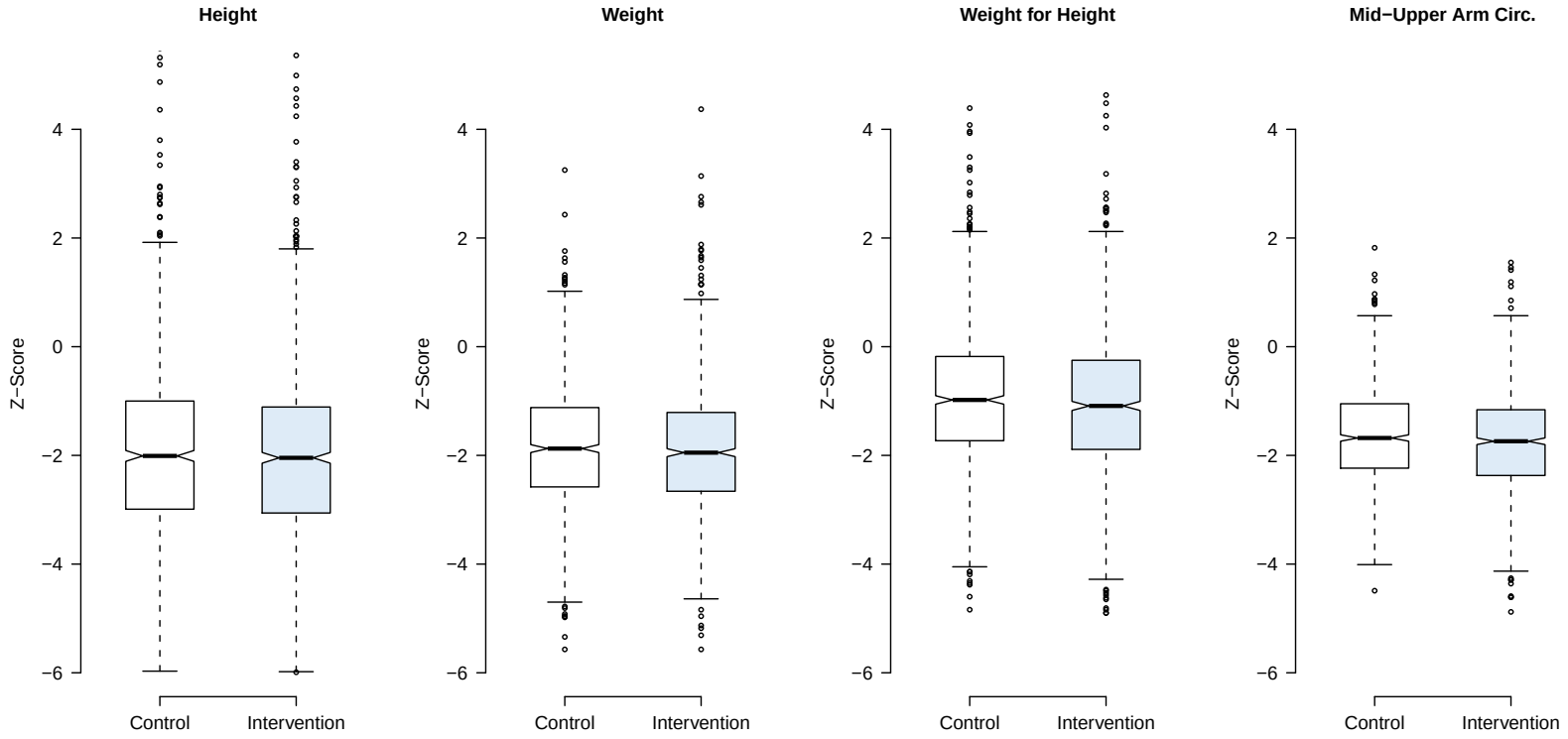


Figure 4.19: Box plots of anthropometric Z-scores for children in control and intervention villages. The heavy lines and wedges indicate median values, the ends of the boxes are 25th and 75th percentiles, and the whiskers are 1.5 times the interquartile range. Outliers are indicated with dots.

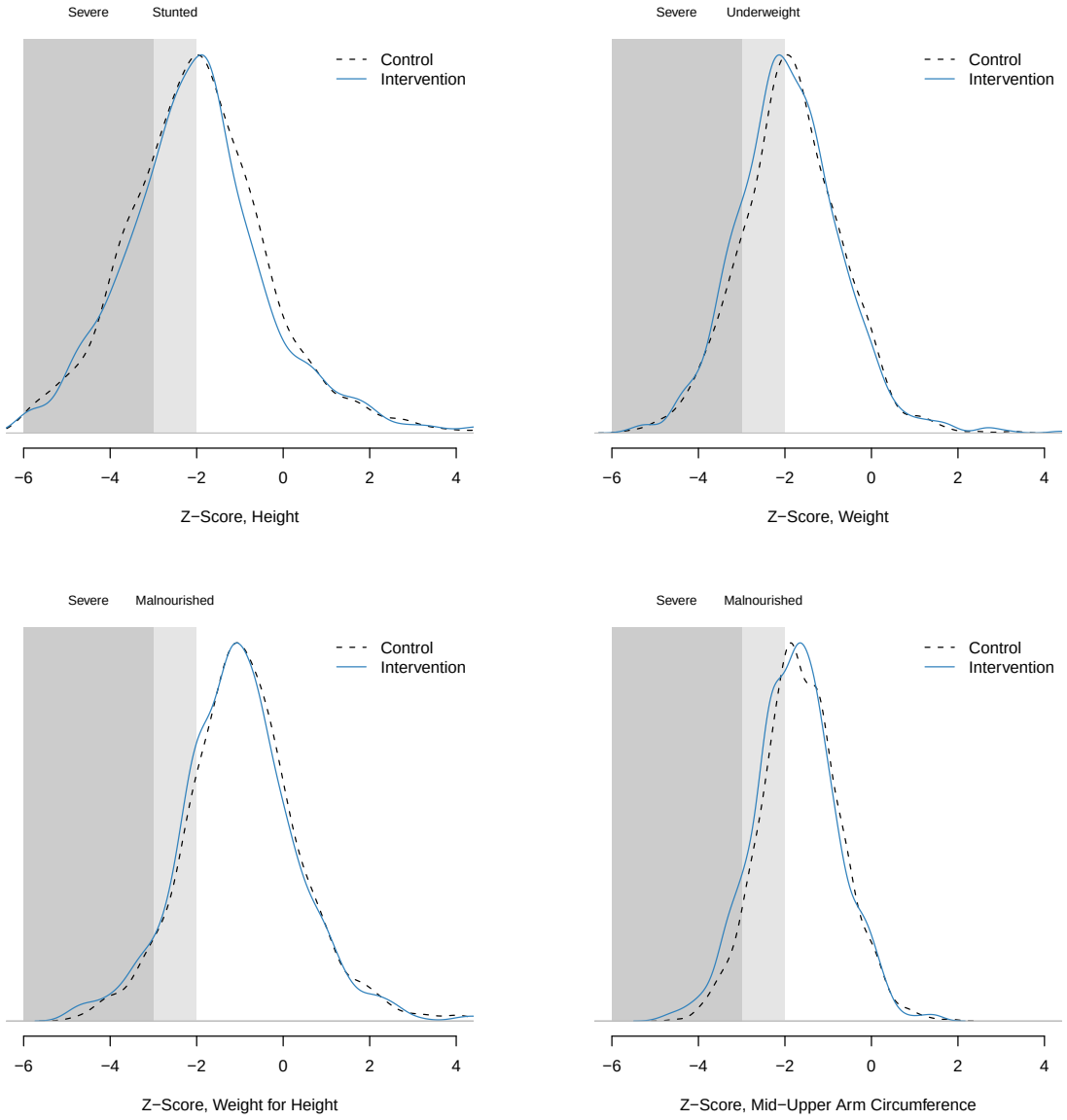


Figure 4.20: Smoothed kernal density plots of anthropometric Z-scores for children in control and intervention villages.

4.4.9 Diarrhea and gastrointestinal illness

Overall, the prevalence of gastrointestinal illness in children under five years old is very low in the study population relative to other developing country populations. The mean

prevalence of diarrhea over 14,259 child-weeks of follow-up was 1.8%, and the mean prevalence of HCGI over the same period was 2.6%. The very low prevalence of diarrhea makes it extremely difficult to detect differences between groups: we powered the study under an assumed baseline prevalence of 10%, and there is relatively little room for improvement below 2% prevalence.

The prevalence of gastrointestinal illness in children under five years old varied over the year. We observed a seasonal trend with seasonally higher prevalence of diarrhea and HCGI (3% – 5%) during the warm, dry summer months (June – September, Figure 4.21). Diarrhea is highly variable, with small localized outbreaks occurring in villages throughout the year (Figure 4.22). There was no village in the study with consistently high or low diarrhea.

The mean prevalence of diarrhea is slightly higher in intervention villages than in control villages (1.97% vs 1.62%), and the two groups differed primarily during the summer months (Figure 4.23). In unadjusted analyses, we did not observe differences in diarrhea between children in intervention and control villages (longitudinal prevalence difference = 0.0035, 95% CI: $-0.0012, 0.0083$, Table 4.15). Adjusted estimates from the G-computation and targeted MLE analysis, which accounted for a large set of potentially confounding characteristics, did not modify our conclusions (Table 4.16).

Like diarrhea, intervention villages have higher mean prevalence of HCGI than control villages (2.86% vs. 2.28%, longitudinal prevalence difference = 0.0058, 95% CI: 0.0018, 0.0093). Adjusted analyses led to similar estimates (Table 4.16).

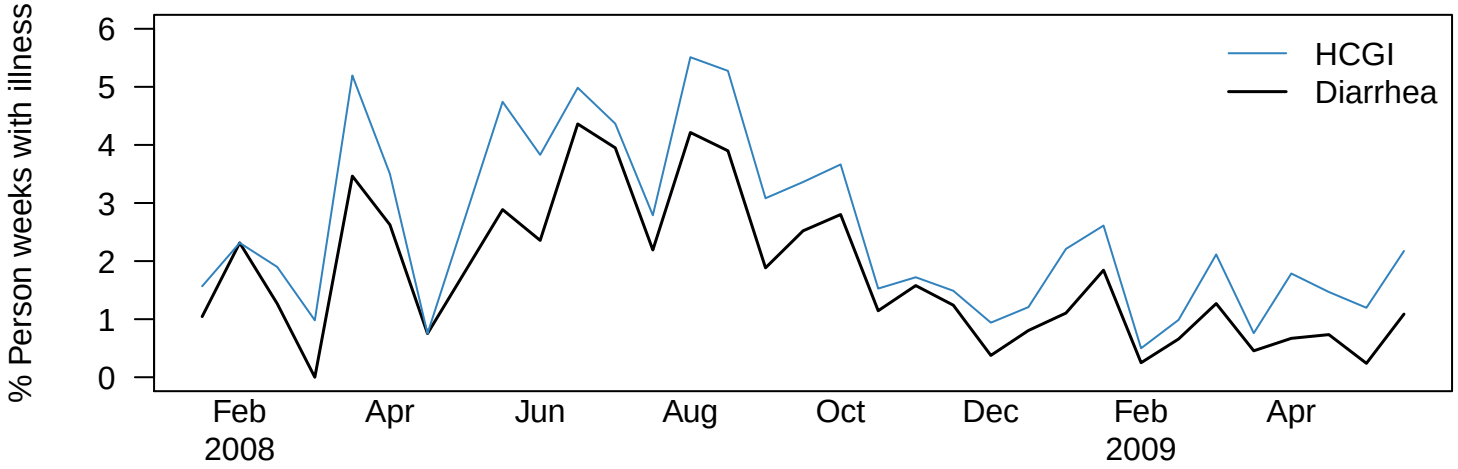


Figure 4.21: Longitudinal prevalence of diarrhea and highly credible gastrointestinal illness (HCGI) among children < 5 years over 65 weeks of follow-up (aggregated into 2-week periods). In addition to diarrhea, HCGI includes vomiting, soft stool and stomach cramps, and nausea and stomach cramps. Data include 1,285 children and 14,259 child-weeks of observation.

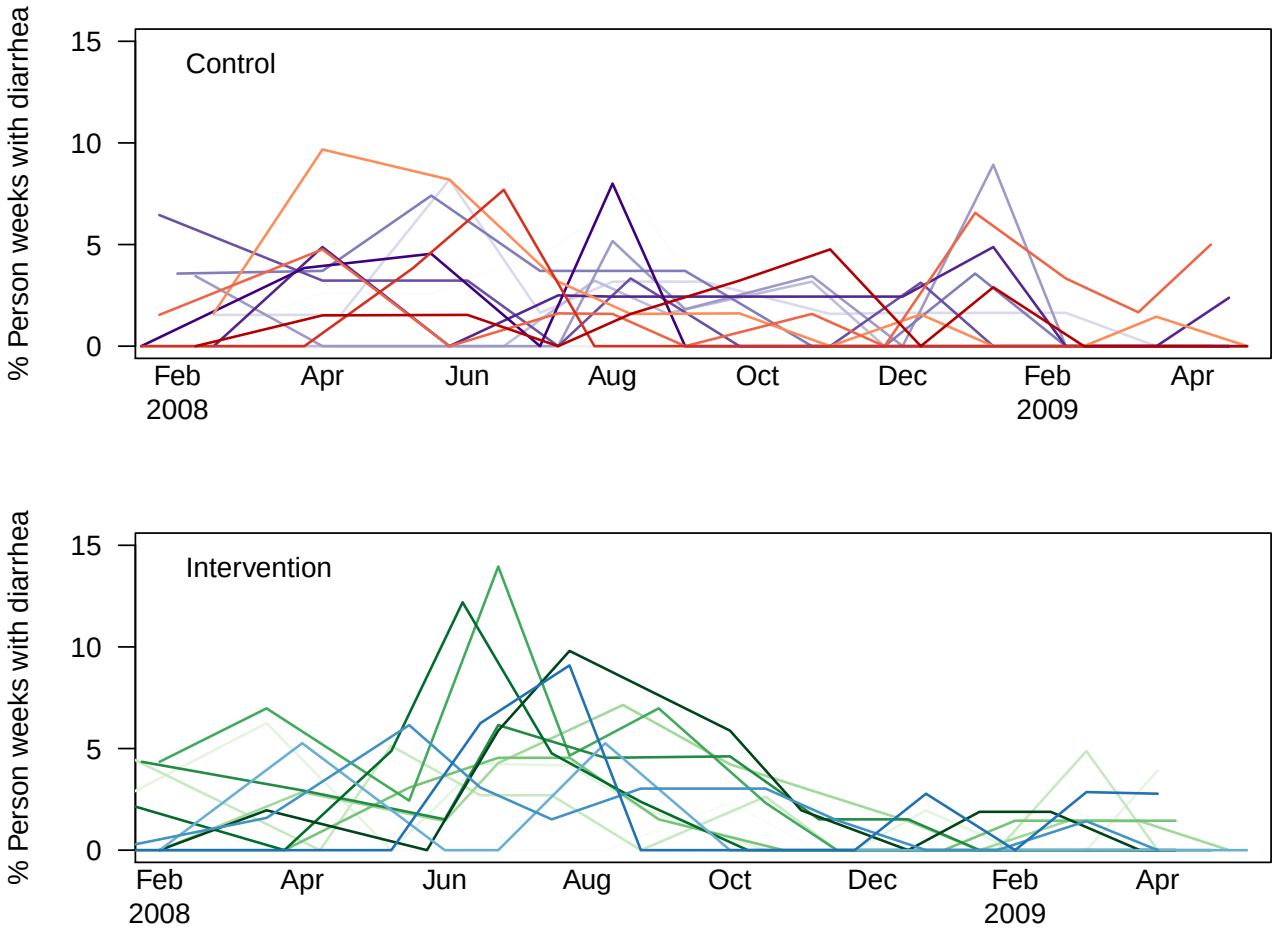


Figure 4.22: Longitudinal prevalence of diarrhea among children < 5 years by village over 65 weeks of follow-up (aggregated into 2-week periods). Line colors and weights are arbitrary and for visualization only. The village-level data illustrate the hyper-variability of diarrhea in this population.

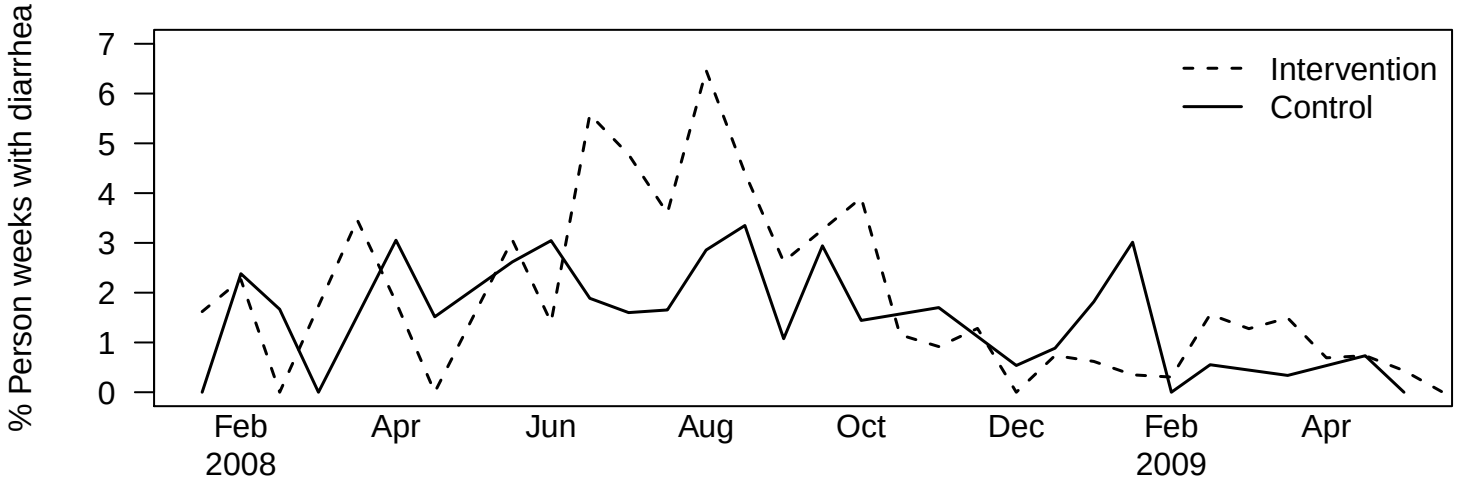


Figure 4.23: Longitudinal prevalence of diarrhea among children < 5 years by intervention group over 65 weeks of follow-up (aggregated into 2-week periods). Data include 7,076 child-weeks of observation in intervention villages and 7,183 child-weeks of observation in control villages.

Table 4.15: Weeks of illness and weekly longitudinal prevalence (%) of diarrhea, highly credible gastrointestinal illness (HCGI) and related symptoms in children under age 5. There were a 7,183 child weeks of observation in the control group, and 7,076 child weeks of observation in the intervention group. Data were collected February 2008 – April 2009.

Outcome	Total		Control		Intervention	
	N	(%)	N	(%)	N	(%)
Diarrhea	257	1.80	117	1.63	140	1.98
HCGI	367	2.57	164	2.28	203	2.87
Vomiting	149	1.04	63	0.88	86	1.22
Stomach cramps	17	0.12	10	0.14	7	0.10
Nausea	8	0.06	4	0.06	4	0.06
Blood or mucus in stool	17	0.12	9	0.13	8	0.11

Table 4.16: Estimates of the difference in weekly longitudinal prevalence for diarrhea and highly credible gastrointestinal illness (HCGI) in children under age 5. There were a 7,183 child weeks of observation in the control group, and 7,076 child weeks of observation in the intervention group. Data were collected February 2008 – April 2009.

Estimator	Diarrhea			HCGI		
	Difference	SE*	95% CI	Difference	SE*	95% CI
Unadjusted	0.0035	0.0024	$(-0.001, 0.008)$	0.0058	0.0019	$(\ 0.002, 0.009)$
G-comp	0.0028	0.0031	$(-0.003, 0.009)$	0.0062	0.0026	$(\ 0.001, 0.011)$
T-MLE	0.0000	0.0067	$(-0.013, 0.013)$	0.0001	0.0080	$(-0.015, 0.016)$

* Standard Errors and 95% confidence intervals calculated by bootstrap resampling matched village pairs with 1000 iterations.

4.4.10 Subgroup analyses

The quantiles of the principal components-based wealth index score usefully categorized households into different wealth categories based on household characteristics. For all household characteristics and assets used to create the index there is a gradation across the wealth quintiles in the expected direction (Table 4.17). For example, mobile phone ownership is 4%, 16%, 21%, 53% and 69% from the poorest to the richest wealth quintile. Based on this categorization of households, intervention households make up 57% of the poorest quintile and 45% of richest quintile (Table 4.17). Although these imbalances are relatively small (and the wealth index score distributions are overall quite similar: Figure 4.24), it further reinforces the importance of adjusted analyses to control for potential confounding by wealth.

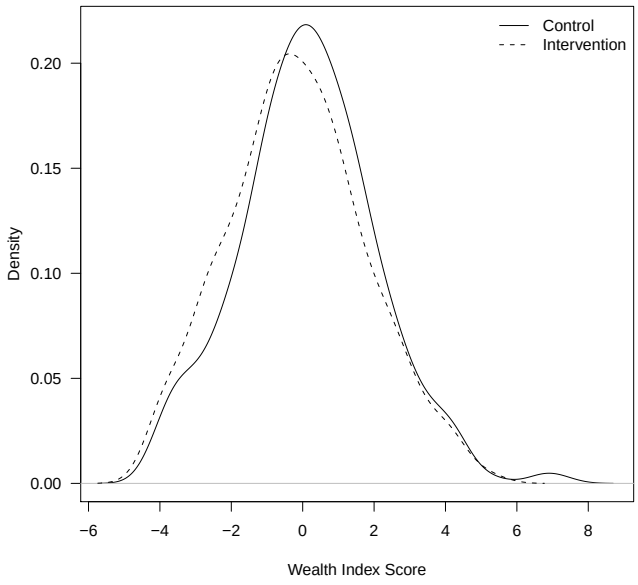


Figure 4.24: Smoothed kernel density distributions of the wealth index score by treatment group. N=456 control, N=444 intervention households.

After stratifying by wealth quintile, it is clear that the intervention expanded private toilet construction to the poorest segments of the population. In the poorest wealth quintile, just 1.3% of control households built private toilets between 2003 and 2008, while 29% of similar intervention households built private toilets over the same period (risk difference = 0.28, 95% CI: 0.15, 0.45). Differences between control and intervention groups in private toilet construction were greatest among households in the second lowest

wealth category, and were smallest in the richest category (Table 4.18). Similarly, the intervention expanded private toilet ownership disproportionately among the scheduled caste population: 11.3% of scheduled caste households in control villages built new private toilets, while 76.5% of scheduled caste households in intervention villages built new private toilets (risk difference = 0.65, 95% CI: 0.31, 0.85) (Table 4.18).

The differences by wealth quintile were less dramatic for new private tap construction. The intervention increased private tap construction by between 5% and 7% beyond the control in the first four wealth quintiles (Table 4.19). None of the differences are statistically significant at the 95% confidence level. However, the intervention greatly expanded private tap access among scheduled caste households: in control villages just 1.8% of scheduled caste households installed new private taps, versus 22.5% of scheduled caste households in intervention villages (risk difference = 0.21, 95% CI: 0.07, 0.30) (Table 4.19).

The net effect of these increases in private toilet and tap construction has led to different overall levels of toilet and tap ownership by wealth quintile in intervention versus control villages, with differences greatest among the poorest households (Figure 4.25). Scheduled caste households in intervention villages also have dramatically higher ownership of private toilets and private taps compared to scheduled caste households in control villages (Figure 4.26).

Table 4.17: Summary of household characteristic means by wealth index quintile. Household characteristics were used in the principal components analysis to derive the wealth index. The factor loading is the eigenvector from the first principal component.

Mean	Wealth Index Quintile					Factor
	1	2	3	4	5	Loading
Work in agriculture	0.87	0.77	0.74	0.64	0.50	−0.14
Women in the home works	0.76	0.78	0.71	0.62	0.60	−0.07
Participates in a committee/group	0.41	0.43	0.49	0.51	0.49	0.04
Women participate in self help group	0.29	0.32	0.38	0.38	0.36	0.03
Soil floor	0.89	0.50	0.13	0.03	0.02	−0.34
Thatched roof	0.82	0.29	0.07	0.02	0.01	−0.32
Total persons in the household	4.62	4.66	4.59	4.82	5.17	0.09
Total rooms in the house	1.38	1.86	2.47	3.04	4.72	0.40
Sleeping rooms in the house	1.13	1.30	1.56	1.84	3.05	0.35
Has electricity	0.56	0.97	0.98	1.00	0.99	0.25
Owns their home	0.91	0.92	0.94	0.92	0.94	0.02
Owns their land	0.92	0.93	0.96	0.95	0.97	0.04
Uses banking services	0.01	0.08	0.12	0.29	0.58	0.25
Refrigerator	0.00	0.01	0.00	0.01	0.12	0.16
Radio	0.29	0.46	0.61	0.64	0.79	0.19
Television	0.21	0.57	0.69	0.83	0.97	0.28
Mobile phone	0.04	0.16	0.21	0.53	0.69	0.26
Motorcycle/scooter	0.03	0.04	0.13	0.32	0.74	0.29
Bicycle	0.60	0.73	0.79	0.81	0.89	0.12
Mosquito net	0.02	0.07	0.08	0.17	0.33	0.17
Household in an intervention village	0.57	0.53	0.46	0.46	0.45	
Number of households	180	180	180	180	180	

Table 4.18: Subgroup analysis of the proportion of households that installed new private toilets wealth index quintile and scheduled caste status. Calculations include 810 households that did not have a latrine at baseline (in 2003).

Subgroup	Control		Intervention		Risk Difference (95% CI)*	
	N	(%)	N	(%)		
Wealth quintile						
1 (poorest)	78	1.3	99	29.3	0.280	(0.154, 0.454)
2	83	2.4	89	50.6	0.482	(0.369, 0.598)
3	96	14.6	77	55.8	0.413	(0.293, 0.524)
4	82	24.4	71	66.2	0.418	(0.294, 0.524)
5 (richest)	67	47.8	68	75.0	0.272	(0.078, 0.475)
Caste status						
Scheduled Caste	62	11.3	51	76.5	0.652	(0.313, 0.848)
Non-Scheduled Caste	344	18.0	353	49.9	0.318	(0.258, 0.376)

* Standard Errors and 95% confidence intervals calculated by bootstrap resampling matched village pairs with 1000 iterations.

Table 4.19: Subgroup analysis of the proportion of households that installed new private taps wealth index quintile and scheduled caste status. Calculations include 733 households that did not have a private tap at baseline (in 2003).

Subgroup	Control		Intervention		Risk Difference (95% CI)*	
	N	(%)	N	(%)		
Wealth quintile						
1 (poorest)	77	1.3	95	7.4	0.061	(−0.004, 0.152)
2	74	4.1	87	10.3	0.063	(−0.026, 0.145)
3	85	12.9	66	18.2	0.052	(−0.078, 0.175)
4	79	15.2	63	22.2	0.070	(−0.106, 0.213)
5 (richest)	54	20.4	53	22.6	0.023	(−0.170, 0.183)
Caste status						
Scheduled Caste	56	1.8	40	22.5	0.207	(0.074, 0.304)
Non-Scheduled Caste	313	11.8	324	13.9	0.021	(−0.075, 0.123)

* Standard Errors and 95% confidence intervals calculated by bootstrap resampling matched village pairs with 1000 iterations.

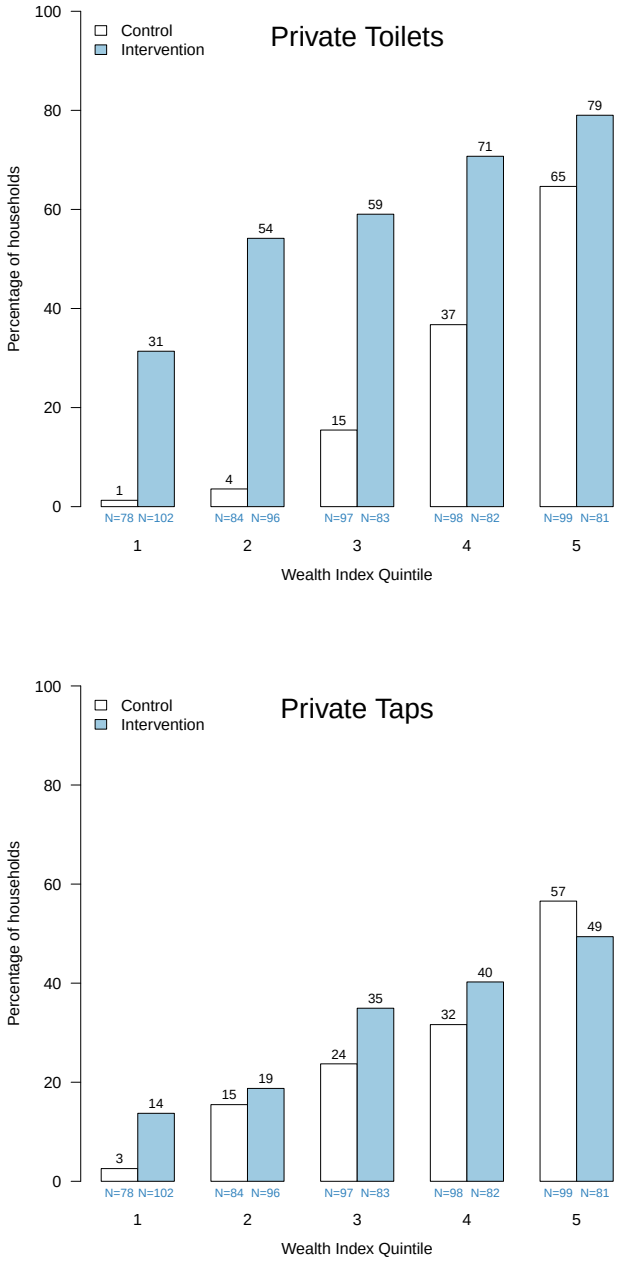


Figure 4.25: Proportion of households that own a private toilet (top plot) or private water tap (bottom plot) by intervention group and wealth index quintile. N=456 control, N=444 intervention households.

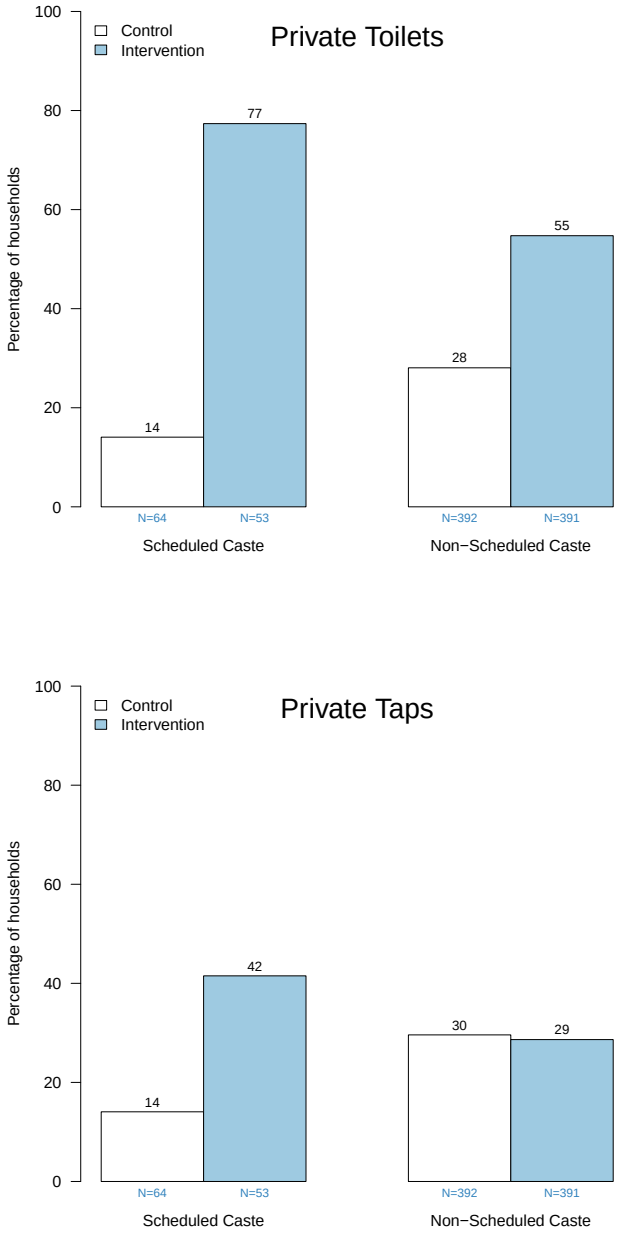


Figure 4.26: Proportion of households that own a private toilet (top plot) or private water tap (bottom plot) by intervention group and scheduled caste status. N=456 control, N=444 intervention households.

Despite the clear pattern of disproportionate improvements in the mid and lower quintiles of the wealth distribution, there is no clear pattern of diarrhea prevalence or intervention health impacts in the different wealth quintiles (Table 4.20). The subgroup analyses of height-for-age Z-scores is more mixed. Mean height Z-scores improve with nearly each wealth quintile in the control group, with children in the richest quintile 0.6 standard deviations above those in the poorest quintile ($-1.74 vs. -2.34$) (Table 4.21). In intervention villages there is no clear association between wealth and child height Z-scores. The highest average height Z-score is in the middle wealth category, though in both groups children in the poorest wealth quintile show the greatest growth faltering. Scheduled caste children in intervention villages fare 0.46 standard deviations better than scheduled caste children in control villages (-1.70 vs -2.15, but the difference is not statistically significant at the 95% confidence level (95% CI: $-0.076, 0.824$).

Table 4.20: Subgroup analysis of the longitudinal prevalence (%) of diarrhea by wealth index quintile and scheduled caste status. N indicates the child weeks of observation in each subgroup.

Subgroup	Control		Intervention		Difference	
	N	%	N	%		(95% CI)*
Wealth quintile						
1 (poorest)	1225	1.47	1498	2.87	0.014	(−0.003, 0.029)
2	1251	1.84	1456	2.34	0.005	(−0.009, 0.016)
3	1541	1.30	1348	0.96	-0.003	(−0.013, 0.006)
4	1604	2.18	1393	2.08	-0.001	(−0.014, 0.015)
5 (richest)	1562	1.34	1381	1.52	0.002	(−0.006, 0.011)
Caste status						
Scheduled Caste	1106	1.08	866	2.31	0.012	(−0.008, 0.027)
Non-Scheduled Caste	6077	1.73	6210	1.93	0.002	(−0.003, 0.008)

* Standard Errors and 95% confidence intervals calculated by bootstrap resampling matched village pairs with 1000 iterations.

4.5 Discussion

Results in context

We measured a large set of outcomes in intervention and control villages up to four years after the completion of combined intervention activities that included sanitation

Table 4.21: Subgroup analysis of the height-for-age Z-scores by wealth index quintile and scheduled caste status.

Subgroup	Control		Intervention		Difference	
	N	Mean	N	Mean	(95% CI)*	
Wealth quintile						
1 (poorest)	164	-2.34	218	-2.36	-0.020	(−0.505, 0.485)
2	183	-2.15	193	-2.01	0.139	(−0.396, 0.629)
3	205	-2.00	182	-1.72	0.279	(−0.023, 0.624)
4	227	-1.71	186	-1.83	-0.119	(−0.511, 0.231)
5 (richest)	215	-1.74	195	-2.00	-0.263	(−0.569, 0.050)
Caste status						
Scheduled Caste	145	-2.15	115	-1.70	0.455	(−0.076, 0.824)
Non-Scheduled Caste	849	-1.93	859	-2.04	-0.110	(−0.403, 0.186)

* Standard Errors and 95% confidence intervals calculated by bootstrap resampling matched village pairs with 1000 iterations.

and water supply improvements and hygiene education. We observed large differences between intervention and control villages in private toilet access, but no difference in water source access, hygiene behaviors or child health measured by acute gastrointestinal illness and anthropometric growth. Although the intervention program improved access to private toilets and private taps for all households, improvements were largest relative to control villages in middle and lower income households, and among scheduled caste members.

Intervention and control villages are highly similar in their current water supply (Figure 4.8), household drinking water quality (Table 4.8) and hygiene conditions (Table 4.11), which implies that the primary difference between intervention and control villages during our measurement period is private toilet access. The sanitation improvements took place against a backdrop of moderate water and hygiene conditions. For example, 100% of the population has access to "improved" water sources by the JMP definition (Table 1.1), the geometric mean E. coli concentration was 1.7 per 100 ml (22.5% of 2,510 household water samples had detectable *E. coli*). In 10,427 spot-checks, 62% of the time a study household had a dedicated handwashing station stocked with both water and soap or ash (Table 4.9). Child caretaker self-reported handwashing practices were generally poor (just 24% report washing their hands with soap after defecation, Table 4.10).

Despite low coverage of private toilets (26% of households) and widespread open defecation (88% of households) in control villages, we observe low diarrhea prevalence (1.6%)

and HCGI prevalence (2.3%) in children under five years old. With such low disease prevalence, if differences exist between intervention and control villages they would have to be extremely small. Given our sample size, the prevalence of diarrhea and HCGI in intervention villages was not statistically different from control villages during the study, differing at most by one half of one percent (Table 4.16).

Child gastrointestinal illness has been reduced to very low levels in this population without improving sanitation, which suggests that large and costly sanitation improvements such as those implemented in this intervention have not provided additional health benefits. Based on the disease experience of children living in control villages over one year of measurement, improved water supply and moderate hygiene conditions are sufficient to reduce gastrointestinal illness to very low levels. This study provides evidence that in at least some populations, it is not necessary to improve all of water quality, hygiene conditions and sanitation to achieve very low levels of gastrointestinal illness. We infer (though have not tested) that in this population the primary transmission pathways for gastrointestinal illness among young children do not include human fecal deposition in the environment. In the only randomized controlled trial specifically designed and powered to evaluate combined interventions, the two interventions evaluated were point-of-use water treatment and handwashing promotion with soap. Individually, both interventions reduced child diarrhea (51 and 64% reduction),but there was no additional reduction in diarrhea with the combined intervention (55% reduction) [37]. These results are consistent with results of meta- analyses of all published interventions to improve water quality, sanitation and hygiene, which found that studies of combined interventions had no greater reduction in diarrheal disease than single intervention studies [38].

Given that the intervention program was motivated in large part by the reduction of gastrointestinal illness in young children, our findings suggest that the program should have been located in a different population (gastrointestinal illness is not a significant health burden in this population). Including health measurement and perceptions of health problems in basic needs assessment surveys prior to program implementation will help ensure that organizations deploy resources to populations with the greatest health burdens. If sanitation, water and hygiene interventions are deployed to populations with high prevalences of gastrointestinal disease, the overall benefits from the interventions would likely be greater because we expect that non-health benefits such as improved privacy of women with private toilets would accrue independent of child disease burden.

Consistent with no difference in gastrointestinal illness, we also observed no difference in child anthropometric growth (Table 4.13). Few studies have attempted to measure the impacts of water and sanitation improvements on child health, but previous studies (all non-randomized) have reported large gains in child height. Esrey used Demographic and Health Survey data from eight countries to estimate the effect of various combinations of water and sanitation improvements on child growth [39]. Among children under five, the adjusted effects of having a latrine (relative to no sanitation) increased the average height-for-age Z-score by 0.26. Optimal sanitation (flush toilets or water-sealed latrines

– similar to sanitation improvements in our cohort) compared to no sanitation increased the height-for-age Z-score by 0.49. Optimal water supply (tap in the yard or inside the house) compared to unimproved surface water did not increase the mean height-for-age Z-score. Both optimal sanitation and water increased mean height-for-age Z-scores by 0.56. A second observational study by Checkley *et al.* followed a birth cohort for two years in peri-urban Peru [40]. At 24 months, they estimate that having a sewer connection increased linear growth by 0.9 cm. For children with and without tap water in their yard or household, the difference in linear growth at 24 months was 0.6 cm.

Chronic diarrhea in the first two years of life leads to growth faltering [41–43]. Given the low prevalence of diarrhea we observed, the growth faltering in this cohort likely results primarily from nutritional deficiencies. However, it remains possible that despite improvements in water and sanitation, some of the growth faltering could result from bacterial exposure that is not sufficient to cause symptomatic illness such as diarrhea, but is sufficient to lead to enteropathy in the digestive tracts of young children[44]. This enteropathy in turn leads to poor nutrient absorption and reduced growth [44]. We did not measure intestinal permeability or blood antibody levels in our cohort to evaluate enteropathy, but future studies of the association between environmental interventions to reduce bacterial exposure and child growth should consider including biometric measurements to measure this causal pathway.

Despite the lack of health impacts identified by this study, we have documented important non-health benefits that follow from improving water supply and sanitation. In this population where nearly every household has access to nearby public taps (median walking time to a public tap is 10 minutes), the median time savings per day of installing a private tap in the home is 25 minutes (50 vs. 75 minutes, Figure 4.9). This time savings estimate is lower than the assumption of 90 minutes of time saved used in global cost-effectiveness analyses of in-home water taps [45]). Yet, our estimate of total time spent gathering water per day (median = 60 minutes) is consistent with a recent study of 5,000 households in six states in India, which reports that rural households spend a mean of 56 minutes per day fetching water [46]. Our estimates suggest that for rural villages in southern India that tend to have tightly clustered households and multiple public taps, the time savings from in-home water supply are substantial, but lower than for more dispersed populations that rely on surface water [45].

The combined CLTS and subsidized loan sanitation campaign led to large gains in new private toilet construction: 48% of households in intervention villages constructed a toilet since 2003 versus 15% in control villages (difference = 33%). All of the private toilets constructed were still in use by the time of our study (up to four years after initial construction). This increase in toilet access has consequently reduced adult open defecation by 15 percentage points compared to control, and has increased perceived privacy and safety for women and girls by 13 percentage points compared to control (Table 4.6). Although we observed large increases in private toilet coverage, the intervention fell short of 100% coverage (by 2008, 57% of intervention households had a private toilet). Our

findings with respect to sanitation are broadly consistent with other published evaluations of CLTS programs.

Pattanayak *et al.* evaluated a similar marketing and subsidy CLTS campaign in the state of Orissa, India, and found broadly consistent impacts on private toilet ownership: they observed an increase of 29% (95% CI: 15%, 43%) relative to control villages [2]. Unlike Pattanayak *et al.* , we observed sanitation improvements in control villages independent of the intervention: private toilet ownership increased from 11% in 2003 to 26% in 2008 among control households. The observed overall improvements are larger in our study compared to Pattanayak *et al.* , but the treatment effects are similar, which is consistent with a higher background level of private toilet construction independent of the intervention.

A case-control study conducted within a social marketing latrine promotion program in Ghana found that at the time of their survey only 60% latrines were functional and in use – much lower than our estimate of 94% in-use [47]. The authors also suggest that the Ghana campaign failed to reach the most marginal households: households that built latrines during the promotion campaign were more educated and wealthier than households that did not build latrines [47]. Their results are difficult to interpret, however, because the study did not include a control group and so it is impossible subtract out latrine construction that would have happened in the absence of the intervention. Although we observed the largest increases in private toilet construction among the wealthiest households, the difference in improvements relative to control villages were greatest among the middle income households, and among the scheduled caste population (Tables 4.18 and 4.19). In the lowest wealth quintile, the proportion of households without a private toilet in 2003 that installed a new private toilet was just 1.3% in control villages and 29.3% in intervention villages. Although this improvement was large (28%), it was smaller than the improvement in the second wealth quintile (48%), which suggests that the poorest households still face resource constraints that may be alleviated in part by hardware subsidies (as in the Orissa study [2]).

We are aware of two additional evaluations of CLTS programs. In Zimbabwe, a pilot study of the Participatory Hygiene And Sanitation Transformation (PHAST) campaign that combines social marketing with subsidized hardware increased latrine ownership to 43%, up from 2% coverage in (subjectively) matched control villages [48]. An evaluation of the CLTS promotion campaign in Ethiopia that constructed over 89,000 latrines in 2004 found that 87% of 160 randomly selected participants had completed latrines and that 90% of these latrines were in use (the study provides no information about latrine coverage before and after the campaign) [1].

To our knowledge this is the first published study to report the impact of a CLTS program on the perceived privacy and safety of women and girls and on open defecation practices. Over 81% of households with private toilets report that women and girls have privacy and feel safe while defecating, versus 53% in households without a private toilet (difference = 28%, 95%CI: 18%, 37%). This is an important finding that reinforces

the importance of considering non-health benefits in addition to health benefits when contemplating investments in sanitation infrastructure.

Our measurements of open defecation practices are less optimistic. Village-level estimates suggest that open defecation behavior is reduced greatly by increasing private toilet ownership (Figure 4.15), and that time since the conclusion of intervention activities is not relevant (Figure 4.14). Yet, nearly 40% of households with a private toilet report that adults practice daily open defecation, and 52% of the same households report that children under 5 years old practice daily open defecation. These figures underscore the difficulty of both defecation behavior change, and the technical difficulties of properly disposing child feces despite in-home hardware improvements. Our results also highlight the nuanced and complicated relationship between toilet construction and actual defecation practice.

Although this study cannot provide detailed explanations for why households continue to practice open defecation after installing a private toilet, 50% of households report that they have "no choice" but to practice open defecation and 25% report that it is "convenient." These responses are consistent with the inherently challenging conditions of properly disposing baby and toddler feces (usually without the aid of diapers), and with inadequate toilet facilities at work (among households with private toilets, the presence of an adult who works in agriculture increases the probability of adult defecation by 15 percentage points: 44.4% vs. 29.5%). Future research that focuses on the motivations and barriers among toilet owners who still practice open defecation will make important contributions to the field. Simple interventions that facilitate the proper disposal of baby and toddler feces, such as inexpensive child potties (currently unstudied), may help mitigate this persistent problem that households face even after installing a private toilet.

Taken together, the data suggest that hardware improvements (taps and toilets) have been highly sustainable over the five year period since the conclusion of the intervention. More than 94% of private toilets were in use up to five years after installation. In contrast, the behavioral components that relate to defecation and hygiene practices have been less successful. Although there is an unequivocal decline in adult open defecation in intervention villages (Figure 4.6), it is not proportional to the expansion of private toilets. For example, open defecation is 19% lower among adult women in intervention villages compared to control, yet private toilet ownership is 31% higher. Hygiene indicators measured repeatedly in households over the year suggest they are relatively stable with time since the end of intervention activities (Figure 4.16). Our study contributes additional evidence that CLTS can result in adoption of improved sanitation technologies across all socio-economic classes, though empirically the campaigns fall short of the goal of "total" sanitation. In this specific application, subsidized hardware provision to the poorest households could further increase the private toilet coverage. Even among private toilet owners open defecation remains prevalent, particularly among children under age five, and behavior change has not kept pace with hardware improvements.

Comments on methodology

This study demonstrates the usefullness of using pre-intervention secondary data to construct a control group in the design stage of evaluating a pre-existing intervention. It was possible to select a group of villages that were well balanced across a broad range of baseline characteristics using multivariate matching based on a propensity score. The quasi-experimental study design was feasible because we had detailed records from WPI and Gramalaya on the details of the interventions, pre-intervention census data were available, and the intervention was relatively homogeneous across villages. Unlike the analysis in Chapter 3, the method of using pre-intervention data to select control villages had greater limitations in this analysis. The data used to select villages was collected two years before the intervention started and seven years before we collected outcome measurements. Although the longer time period allowed us to evaluate the sustainability of the interventions, we found that in the intervening years control villages made water source improvements independent of the intervention. Thus, while the intent of the study was to evaluate a combined sanitation, water and hygiene intervention, by the time of our measurement the intervention villages differed from control villages only in their access to sanitation.

The implications of finding no difference in key exposures between intervention and control groups depends on whether the control group has improved on its own or whether the intervention group either failed to improve or has regressed from an improved condition back to its original state. In this study we have examples of both. Although intervention villages had more water source improvements than control villages in the previous five years (Figure 4.4), overall the two groups had similar water supplies in 2008 (Figure 4.8). It is not possible for us to evaluate the full impact of the program on child health because we do not observe a counterfactual population without water improvements. Control and intervention villages also have similar hygiene exposures, but both have poor hygiene based on objective indicators (Table 4.9) and self-reported handwashing (Table 4.11). In contrast to the impact of the program through water sources, we conclude that the program did not measurably improve hygiene practices in this population and thus there are likely no positive health benefits from that component of the intervention.

As we have demonstrated here, this study design has potential to create intervention and control groups that do not differ in key intermediate exposures of interest when applied in populations undergoing rapid development. We were fortunate that there were large differences between groups in sanitation exposure, but that difference was not guaranteed. In general, we expect that it will be increasingly difficult to maintain a pure control group with increased time from the initiation of a pre-existing intervention. This limitation does not apply to the matching method, but instead to the specific intervention and population of interest that will vary in each application. Future evaluations of pre-existing interventions should consider this important context-specific design limitation and determine whether they expect non-intervention groups to have made substantive improvements independent of the intervention.

In our analysis we could measure the construction of private taps and latrines retrospectively over the intervention period. This retrospective measurement allowed us to estimate a differences-in-differences (DID) measure of the change in these outcomes. The advantage of this approach is that controls for potential baseline differences in the outcome and eliminates time-invariant unmeasured confounding [20]. We expect that there is some measurement error in the retrospective data, but we would expect this error to be non-differential with respect to intervention status. As long as the measurement error is non-differential, it will lead to conservative bias toward a null finding [49]. Intervention and control villages were highly comparable in their private toilet and private tap ownership in 2003 (Figure 4.5), which reinforces the baseline comparability of the two groups, and lends additional credibility to post-intervention-only comparisons of child health in the villages.

In adjusted analyses of gastrointestinal illness we observed very little difference between the unadjusted treatment effect and the treatment effect estimated using G-computation (Table 4.16). However, after including the one-step targeted update in the targeted MLE estimator, the point estimate shifted to the null, and the standard errors increased by 3 to 4 times. This increased variability in the targeted MLE estimator followed from controlling residual confounding by modeling the treatment mechanism ($h(A, W)$ in equation 4.6). The update coefficients $\hat{\epsilon}_n$ were often non-zero (Figure 4.27) which indicates that in most bootstrap samples there was residual bias in the unadjusted and G-computation estimators. In this specific application there is a clear bias-variance tradeoff between unadjusted and G-computation estimator, and the targeted MLE estimator – a bias-variance tradeoff that we do not see in the child growth analyses (Table 4.14). Brookhart and van der Laan [28] and Brookhart *et al.* [29] demonstrated through simulation that if an estimator relies on a treatment model, then the variability of the estimator will increase dramatically if covariates are included in the treatment model that are associated only with treatment, but not the outcome of interest. Following this work, we restricted potential covariates in the model selection to only those that strong univariate associations with the outcomes of interest. In future work we plan to identify the covariates that most strongly predict treatment, and then serially delete these covariates from the treatment model selection routines to see if the variability of the targeted MLE estimator is reduced without losing its bias reduction.

Given that our adjusted analyses suggest that there a small amount of residual confounding between intervention and control villages, it suggests that the matching in the design stage was imperfect. This is unsurprising for three reasons. First, many of the characteristics used to match were from the Panchayat rather than village level, and that our rapid assessment suggested that there were some inaccuracies in the census data themselves. Second, we selected the propensity score matching model iteratively, by hand. After our village selection was complete and the study was underway, we become aware of the `GenMatch` routine in R, which uses a genetic algorithm to optimally match treatment and control units based on a loss function based on the minimum p-value for baseline

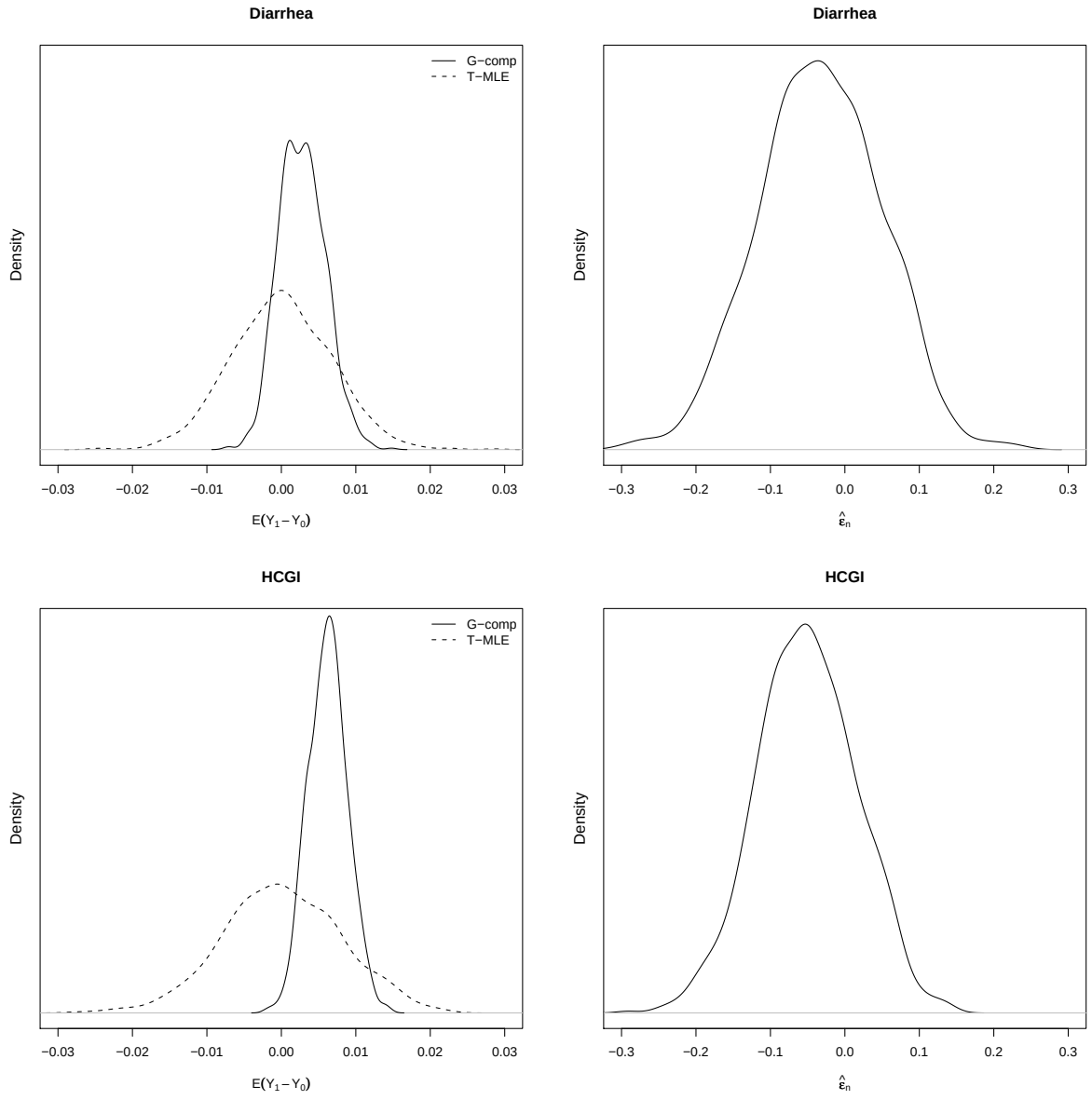


Figure 4.27: The left plots summarize smoothed kernel densities of the bootstrap distributions of G-computation and targeted MLE (T-MLE) estimators for diarrhea and HCGI. The right plots include smoothed kernal distributions of the targeted MLE update coefficient ($\hat{\epsilon}_n$) for the two outcomes.

covariate differences between groups [36]. We recommend that future studies that implement this design use this machine learning tool that exhaustively searches for optimal balance. Finally, there was a relatively small number of intervention villages in our sample, and better matches are obtained with larger pools of both intervention and control units [50].

Study limitations

This study does not include baseline outcome measurement so it remains possible that intervention villages were in worse health than control villages prior to the intervention, and that their health improved to control levels by our post-intervention measurement in 2008. This scenario is unlikely given the comparability between the two groups across a large set of characteristics that were strongly associated with the outcomes at follow-up (Table 4.4).

The prevalence of diarrhea and HCGI was lower than we expected and consequently we have little power to detect differences between groups if differences exist. Based on our adjusted targeted MLE estimates, we have power to detect differences of approximately 1.5 percentage points in the prevalence of diarrhea and HCGI (Table 4.16). Although our best estimate is that there is no difference between groups, if differences exist that are smaller than our level of detection, an intervention would need to be extremely inexpensive to be cost-effective based on health outcomes alone.

It is possible that this study took place during a year with unusually low prevalence of gastrointestinal illness and that our findings are not representative of the population's experience over many years. For example, the prevalence of diarrhea in children under five in a surveillance area of Karachi, Pakistan varied 8 fold over six years of follow-up (e.g., the prevalence in 2000 was 1.3%, and the prevalence in 2003 was 10.6%; Stephen P. Luby, personal communication based on an unpublished manuscript). Although we observe similar hypervariability in our diarrhea and HCGI outcomes within our follow-up period, we suspect that our mean estimates (averaged over 12 months and 25 villages) are generally representative of this population for two reasons. First, when asked an open-ended question about what they believed to be the most important disease in the community, just 2% (17/900) of caregivers reported diarrhea – the most common responses were fever (70%) and cough, cold, pneumonia (13%). We asked this same question in our Guatemala cohort (Chapter 3), where diarrhea prevalence in children under five was 11.9%, and 36% of caregivers reported that diarrhea was the most important disease in the community. Second, surveillance data collected on hospitalized cases from the administrative blocks of our study also suggest that the disease experience in the year of our study was comparable to the previous three years.

Although it is possible that unmeasured confounding exists that is masking intervention effects, our adjusted child health treatment effects that control for a large range of demographic and socio-economic characteristics are consistent with the unadjusted esti-

mates (Tables 4.14, 4.16). Matching in the design stage to select control villages led to more comparable groups than a pure random sample of controls, but we observe residual differences between the groups in potentially important confounding characteristics. The intervention villages are generally more agricultural than the control villages and slightly poorer, but the two groups have good support across the joint distribution of covariates that are most strongly associated with child health outcomes in this population (Figure 4.3). Given that the two groups have good support in the covariates, our adjusted estimates should adequately control for observed differences. If unobserved residual confounding exists, it would need to be large to mask child health treatment effects.

An important limitation of our analysis is that we have ignored heterogeneity in the intervention program across villages. For example, four of the 12 intervention villages did not include subsidized loan programs through local self-help groups for hardware improvements, and private household taps were not emphasized in these villages (Table 4.1). Our analyses have used the complete intervention program package as our treatment of interest, and they do not estimate the individual impacts from intervention components such as installing a private toilet or dedicated handwashing station with water and soap. We plan to look at the individual impact of intervention components in future analyses.

Conclusion

The combined CLTS marketing, water supply improvement, hygiene education led to large increases private toilet ownership in this population, and relative to control villages the increases were greatest among middle income and scheduled caste households. Nearly all private toilets were in use up to five years from the installation date. Although the intervention led to improvements in both public and private water taps, control villages had nearly commensurate improvements through channels independent of the intervention. An exception was among scheduled caste households in intervention villages, who installed private taps with much greater frequency than scheduled caste households in control villages (23% vs. 2%). The water supply improvements have improved village source water quality, but household drinking water quality is similar in the two groups due to contamination during storage. Hygiene was uniformly poor in both intervention and control villages, and open defecation remained common, even among households with private toilets. Sanitation and hygiene behavior change has not kept pace with hardware improvements in this population.

All villages in our study had low diarrhea and gastrointestinal illness prevalence. Given that diarrhea prevalence was extremely low in control villages, which had persistently poor sanitation conditions characterized by frequent open defecation and low private toilet coverage, our data indicate that (i) low diarrhea prevalence can be achieved without sanitation improvements, and (ii) sanitation improvements provide no marginal improvement in diarrhea prevalence given the already low background levels. These findings suggest that defecating in the environment is not an important source of pathogen transmission

in this setting, and that the primary benefits of private toilet and tap installation in this population include increased privacy and time savings for women.

This quasi-experimental design that selects a matched control set using pre-intervention secondary data improved the comparability of intervention and control groups across a wide range of characteristics. For older pre-existing interventions, this quasi-experimental design does not guarantee differences in the primary exposures of interest, particularly in populations that undergo rapid development between the pre-intervention period and the follow-up survey. Future applications of this design for evaluating pre-existing interventions should consider this important design limitation, which may or may not apply depending on the specific application.

Bibliography

[1] O'Loughlin R, Fentie G, Flannery B, Emerson PM. Follow-up of a low cost latrine promotion programme in one district of Amhara, Ethiopia: characteristics of early adopters and non-adopters. Trop Med Int Health. 2006;11(9):1406–15.

[2] Pattanayak SK, Yang JC, Dickinson KL, Poulos C, Patil SR, Mallick RK, et al. Shame or subsidy revisited: social mobilization for sanitation in Orissa, India. Bull World Health Organ. 2009;8:580 – 587.

[3] WSP. Measuring the Cost-Effectiveness of Sanitation Interventions: Experience from TSSM-Indonesia. World Bank; 2009. Available from: `http://www.wsp.org/UserFiles/file/Cost-effectiveness_Indonesia_1__2.pdf`.

[4] Kar K. Subsidy or self-respect? Participatory total community sanitation in Bangladesh. Institute of Development Studies, Working paper 184; 2003.

[5] Arney H, Damodaran S, Meckel M, Barenberg A, White G. Creating Access to Credit for Water and Sanitation: Women's Self-Help Groups in India. IRC Symposium: Sanitation for the Urban Poor. Partnerships and Governance.; 2008. Available from: `www.irc.nl/page/44897`.

[6] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55.

[7] Luby SP, Halder AK. Associations among handwashing indicators, wealth, and symptoms of childhood respiratory illness in urban Bangladesh. Trop Med Int Health. 2008;13(6):835–44. Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. England Tm & ih.

[8] Goldman N, Vaughan B, Pebley AR. The use of calendars to measure child illness in health interview surveys. Int J Epidemiol. 1998;27(3):505–12.

[9] Baqui AH, Black RE, Yunus M, Hoque AR, Chowdhury HR, Sack RB. Methodological issues in diarrhoeal diseases epidemiology: definition of diarrhoeal episodes. Int J Epidemiol. 1991;20(4):1057–63.

[10] Colford J J M, Wade TJ, Sandhu SK, Wright CC, Lee S, Shaw S, et al. A randomized, controlled trial of in-home drinking water intervention to reduce gastrointestinal illness. Am J Epidemiol. 2005;161(5):472–82.

[11] ORC-Macro. Demographic and Health Survey Interviewer's Manual. MEASURE DHS Basic Documentation No. 2. ORC Macro; 2006.

[12] Diggle PJ, Heagerty P, Liang KY, Zeger SL. Analysis of Longitudinal Data (2nd ed.). Oxford: Oxford University Press; 2002.

[13] Leon AC. Sample-size requirements for comparisons of two groups on repeated observations of a binary outcome. Eval Health Prof. 2004;27(1):34–44.

[14] Reller ME, Mendoza CE, Lopez MB, Alvarez M, Hoekstra RM, Olson CA, et al. A randomized controlled trial of household-based flocculant-disinfectant drinking water treatment for diarrhea prevention in rural Guatemala. Am J Trop Med Hyg. 2003;69(4):411–9.

[15] Katz J, Carey VJ, Zeger SL, Sommer A. Estimation of design effects and diarrhea clustering within households and villages. Am J Epidemiol. 1993;138(11):994–1006.

[16] Rahmathullah L, Underwood BA, Thulasiraj RD, Milton RC. Diarrhea, respiratory infections, and growth are not affected by a weekly low-dose vitamin A supplement: a masked, controlled field trial in children in southern India. Am J Clin Nutr. 1991;54(3):568–77.

[17] Schmidt WP, Luby SP, Genser B, Barreto ML, Clasen T. Estimating the longitudinal prevalence of diarrhea and other episodic diseases: continuous versus intermittent surveillance. Epidemiology. 2007;18(5):537–43.

[18] Morris SS, Cousens SN, Kirkwood BR, Arthur P, Ross DA. Is prevalence of diarrhea a better predictor of subsequent mortality and weight gain than diarrhea incidence? Am J Epidemiol. 1996;144(6):582–8.

[19] WHO. WHO Anthro Software. Geneva: WHO; 2008. Available from: `http://www.who.int/childgrowth/software/en/`.

[20] Meyer BD. Natural and quasi-experiments in economics. Journal of Business & Economic Statistics. 1995;13(2):151–161.

[21] Freedman DA. Statistical Models: Theory and Application. New York: Cambridge University Press; 2005.

[22] van der Laan M, Rubin D. Targeted Maximum Likelihood Learning. Int J Biostatistics. 2006;2(1):1–38.

[23] van der Laan MJ. The Construction and Analysis of Adaptive Group Sequential Designs; 2008. Available from: `http://www.bepress.com/ucbbiostat/paper232`.

[24] Robins J. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. Journal of Chronic Diseases. 1987;40:S139–S161.

[25] van der Laan MJ, Polley EC, Hubbard AE. Super learner. Stat Appl Genet Mol Biol. 2007;6:Article25.

[26] Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Stanford University; 2009. Available from: `www-stat.stanford.edu/~hastie/Papers/glmnet.pdf`.

[27] Hastie T, Tibshirani R. Generalized Additive Models. London: Chapman and Hall; 1990.

[28] Brookhart MA, van der Laan MJ. A semiparametric model selection criterion with applications to the marginal structural model. Computational Statistics & Data Analysis. 2006;50(2):475–498. 0167-9473 doi: DOI: 10.1016/j.csda.2004.08.013.

[29] Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable Selection for Propensity Score Models. Am J Epidemiol. 2006;163(12):1149–1156.

[30] Houweling TA, Kunst AE, Mackenbach JP. Measuring health inequality among children in developing countries: does the choice of the indicator of economic status matter? Int J Equity Health. 2003;2(1):8.

[31] Vyas S, Kumaranayake L. Constructing socio-economic status indices: how to use principal components analysis. Health Policy Plan. 2006;21(6):459–68.

[32] Luby SP, Mendoza C, Keswick BH, Chiller TM, Hoekstra RM. Difficulties in bringing point-of-use water treatment to scale in rural Guatemala. Am J Trop Med Hyg. 2008;78(3):382–7.

[33] Stauber CE, Ortiz GM, Loomis DP, Sobsey MD. A randomized controlled trial of the concrete biosand filter and its impact on diarrheal disease in Bonao, Dominican Republic. Am J Trop Med Hyg. 2009;80(2):286–93.

[34] Jalan J, Somanathan E. The importance of being informed: Experimental evidence on demand for environmental quality. Journal of Development Economics. 2008;87(1):14–28.

[35] Abadie A. Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models. Journal of the American Statistical Association. 2002;97(457):284–292.

[36] Sekhon J. Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching package for R. Journal of Statistical Software. Forthcoming;.

[37] Luby SP, Agboatwalla M, Painter J, Altaf A, Billhimer W, Keswick B, et al. Combining drinking water treatment and hand washing for diarrhoea prevention, a cluster randomised controlled trial. Trop Med Int Health. 2006;11(4):479–89.

[38] Fewtrell L, Kaufmann RB, Kay D, Enanoria W, Haller L, Colford J J M. Water, sanitation, and hygiene interventions to reduce diarrhoea in less developed countries: a systematic review and meta-analysis. Lancet Infect Dis. 2005;5(1):42–52.

[39] Esrey SA. Water, waste, and well-being: a multicountry study. Am J Epidemiol. 1996;143(6):608–23.

[40] Checkley W, Gilman RH, Black RE, Epstein LD, Cabrera L, Sterling CR, et al. Effect of water and sanitation on childhood health in a poor Peruvian peri-urban community. Lancet. 2004;363(9403):112–118.

[41] Guerrant RL, Schorling JB, McAuliffe JF, de Souza MA. Diarrhea as a cause and an effect of malnutrition: diarrhea prevents catch-up growth and malnutrition increases diarrhea frequency and duration. Am J Trop Med Hyg. 1992;47(1 Pt 2):28–35.

[42] Checkley W, Gilman RH, Epstein LD, Suarez M, Diaz JF, Cabrera L, et al. Asymptomatic and Symptomatic Cryptosporidiosis: Their Acute Effect on Weight Gain in Peruvian Children. Am J Epidemiol. 1997;145(2):156–163.

[43] Checkley W, Buckley G, Gilman RH, Assis AM, Guerrant RL, Morris SS, et al. Multicountry analysis of the effects of diarrhoea on childhood stunting. Int J Epidemiol. 2008;37(4):816–30.

[44] Campbell DI, Elia M, Lunn PG. Growth Faltering in Rural Gambian Infants Is Associated with Impaired Small Intestinal Barrier Function, Leading to Endotoxemia and Systemic Inflammation. J Nutr. 2003;133(5):1332–1338.

[45] Hutton G, Haller L, Bartram J. Global cost-benefit analysis of water supply and sanitation interventions. J Water Health. 2007;5(4):481–502.

[46] Barnes D, Sen M. The impact of energy on women's lives in rural India. IBRD/World Bank; 2003. Available from: `http://www.esmap.org/filez/pubs/finalindiaforweb.pdf`.

[47] Rodgers AF, Ajono LA, Gyapong JO, Hagan M, Emerson PM. Characteristics of latrine promotion participants and non-participants; inspection of latrines; and perceptions of household latrines in Northern Ghana. Trop Med Int Health. 2007;12(6):772–82.

[48] Waterkeyn J, Cairncross S. Creating demand for sanitation and hygiene through Community Health Clubs: a cost-effective intervention in two districts in Zimbabwe. Soc Sci Med. 2005;61(9):1958–70.

[49] Rothman K, Greenland S. Modern Epidemiology. 2nd ed. Philadelphia: Lippincott-Raven Publishers; 1998.

[50] Rubin DB. Multivariate Matching Methods that are Equal Percent Bias Reducing, II: Maximums on Bias Reduction for Fixed Sample Sizes. Biometrics. 1976;32(1):121–132. Biometrics.

Chapter 5

Integrated discussion

5.1 Compendium of scientific results

Below I summarize the key findings from Chapters 2, 3 and 4 using narrative descriptions and tables that list the goals and related learning points. In the section that follows, I draw on results from all chapters to make broad conclusions about evaluating non-randomized, pre-existing interventions based on my experiences in the water, sanitation and hygiene sector.

Chapter 2
A quasi-experimental design to evaluate non-randomized, pre-existing community interventions

In Chapter 2 I outline a conceptual framework for the design and analysis of non-randomized, pre-existing interventions. The design arises naturally from the Neyman-Holland-Rubin causal model that conceptualizes treatment effects in terms of potential outcomes. Multiple assumptions are required to interpret findings from non-randomized, pre-existing interventions as causal. Studies of such interventions must address many threats to validity not encountered in randomized, prospective studies. In addition, since the interventions are, by definition, not under the control of investigators, I outline six minimum criteria that are necessary for an intervention to be suitable to include in a study. Table 5.1 summarizes the main goals and learning points from Chapter 2, along with the most relevant page references. The design can contribute information that is either not possible to obtain with prospective studies, or it obtains similar information at much lower financial cost. As with all non-randomized, retrospective studies, the design requires strong assumptions and has limitations that investigators must address to establish a study's validity.

Table 5.1: Summary of goals and key learning points from Chapter 2.

Goals	Learning Points	Page References
Identify the minimum necessary conditions that a pre-existing community intervention program must meet to be studied.	I summarize six conditions that an intervention must meet to make it amenable to a quasi-experimental design.	34

Continued on Next Page...

Table 5.1 Summary of goals and key learning points from Chapter 2 – Continued

Goals	Learning Points	Page References
Describe a quasi-experimental design that investigators can use to evaluate non-randomized, pre-existing community interventions.	The design uses secondary data collected prior to the commencement of intervention activities to select control communities using propensity score matching, and collects outcomes from the sample using a field team.	36
Identify the main threats to the validity of the design.	The main threats to validity include unmeasured confounding, informative censoring, measurement error and sampling bias.	41, 57
Outline analysis strategies for the design.	Matching in the design imposes no constraints on subsequent statistical analyses. The design naturally estimates the average treatment effect among the treated (ATT) parameter.	43, 45
Summarize the main advantages of the design.	Matching improves the chances that intervention and control communities will have overlap in covariates used to match, which leverages the design to remove bias and relies less on statistical adjustment.	67
	Studies of pre-existing community interventions avoid the problems of including scientific measurement in the intervention itself.	67
	Studies of pre-existing community interventions can obtain information about medium or longer term impacts quickly without costly prospective follow-up.	67
Summarize the main limitations of the design.	Like all observational studies, investigators must assess very carefully the assumption of no unmeasured confounding.	26, 41, 67, 57
	It is possible that intervention communities will be so different from the potential control communities that no close matches exist.	67
	The design does not guarantee contrasts on key post-intervention exposures. This is a context-specific problem, but will likely be a larger threat to studies in highly dynamic populations or studies with long periods between baseline and follow-up.	67

Chapter 3
Evaluation of a 3-year, pre-existing household water treatment and handwashing intervention in rural Guatemala

Table 5.2 summarizes the goals and learning points from Chapter 3. Six months after the end of a three-year intervention in rural Guatemala the study observed minimal sustained water treatment and handwashing behavior of the types promoted by the intervention. The study also found no difference between children living in control and intervention villages with respect to acute gastrointestinal, respiratory, or anthropometric measures. The lack of health impacts is internally consistent with only small differences in behavioral outcomes, and the findings highlight the difficulty of achieving sustained new behavior adoption in the context of non-research intervention campaigns. The study is a successful empirical application of the design that I propose in Chapter 2. Its application in this context highlights the difficulty of interpreting a null finding in the context of sustainability: an intervention with large initial impact but poor sustainability will be indistinguishable from an intervention with poor initial impact (and nothing to sustain) without outcome measurement at the immediate conclusion of intervention activities.

Table 5.2: Summary of goals and key learning points from Chapter 3 (Guatemala study).

Goals	Learning Points	Page References
Use the design developed in Chapter 2 to assemble a matched cohort of intervention and control study villages that are balanced at baseline across a large set of potentially confounding covariates	It is possible to find good matches and identify a balanced sample using PSM, even with relatively small sample sizes.	91, 93
	In practice, it is important to conduct rapid assessments prior to field activities to "ground truth" the secondary data and potentially refine inclusion and exclusion criteria.	80
	It is possible to conduct cross-sectional studies of pre-existing interventions rapidly and at relatively low cost.	
	Without outcome measurement at the immediate conclusion of intervention activities, and again during the post-intervention period, it is impossible to distinguish between an intervention with large impact but low sustainability, and an intervention with low impact overall for outcomes that cannot be measured retrospectively (e.g., child health).	69, 115
Evaluate the impact of a 3-year household water treatment and handwashing campaign on behavior following the conclusion of intervention activities.	The intervention increased household water treatment behavior in intervention villages by 5.3% relative to control villages (8.7% vs. 3.3%).	96, 97
	There are no differences between groups in self-reported handwashing behavior, or spot-check observations of hygiene conditions.	100

Chapter 4
Evaluation of a pre-existing, combined sanitation, water and hygiene intervention in rural Tamil Nadu, India

Table 5.3 summarizes the goals and learning points from Chapter 4. The combined community-led total sanitation (CLTS) marketing, water supply improvement, hygiene education led to large increases private toilet ownership in this population, and relative to control villages the increases were greatest among middle income and scheduled caste households. Over 94% of private toilets were in use up to five years from the installation date, suggesting high sustainability of this hardware intervention. Private toilets have not led to a 1:1 reduction in open defecation, and nearly 40% of households with private toilets report that adults still practice daily open defecation. Nonetheless, village level open defecation rates are strongly, negatively correlated with private toilet ownership and not with time since intervention activities cease. Complete eradication of open defecation will require extending private toilets – such as those installed under this intervention – to additional poor households (those with the lowest adoption rates) and will require a thorough understanding of the cultural and behavioral barriers that contribute to the persistence of open defecation among toilet-owning households.

Although the intervention led to improvements in both public and private water taps, control villages had nearly commensurate improvements through channels independent of the intervention. An exception was among scheduled caste households in intervention villages, who installed private taps with much greater frequency than scheduled caste households in control villages (23% vs. 2%). The water supply improvements have improved village source water quality, but household drinking water quality is similar in the two groups due to contamination during storage. Hygiene is uniformly poor in both intervention and control villages, and hygienic practices are unrelated to time since intervention activities ceased. Taken together, the study concludes that sanitation and hygiene behavior change has not kept pace with hardware improvements in this population.

Intervention and control villages differ primarily in their access to private toilets, and the differences are large: 57% of intervention households versus 26% of control households have a private toilet. Yet, we observe no differences between groups in child diarrhea, and diarrhea is very low (1.8%) in the study population. This population provides an example where it was not necessary to intervene on all three water, sanitation and hygiene pathways to reduce diarrhea to very low levels (indeed, sanitation and hygiene conditions are poor in control villages). Thus, the primary benefits in this intervention are non-health benefits, that include increased perception of privacy and safety among women during defecation due to private toilets (previously unstudied), and time savings for women following the installation of private water taps.

The quasi-experimental design successfully created a cohort of similar intervention and control villages using pre-intervention census data to identify matched controls. Its application in this study highlights one of the limitations of the design in the context of sustainability research: when the study includes a long period between baseline and follow-

up there is potential for control villages to make improvements similar to the intervention on their own (independent from the intervention) and so differences in key exposures that result from the intervention program may not exist in the follow-up survey. Studies of any pre-existing intervention (randomized or not) should consider this potential difficulty, which will be context specific.

Table 5.3: Summary of goals and key learning points from Chapter 4 (India study).

Goals	Learning Points	Page References
Use the design developed in Chapter 2 to assemble a matched cohort of intervention and control study villages that are balanced at baseline across a large set of potentially confounding covariates.	It is possible to find good matches and identify a balanced sample using PSM, even with a small number of intervention villages (N=12).	129, 143
	Matching in the design stage balanced most characteristics, but intervention villages are slightly more agricultural than control. Residual bias is likely and further adjustment with statistical analysis changed the point estimates (though not conclusions) slightly.	147
	A rapid assessment of potential study villages indicated that the census data were inaccurate in some cases. The "ground truth" exercise was necessary to validate and refine the initial control village selection.	129
	The design does not guarantee differences between intervention and control groups in post-treatment exposures. In this case, control and intervention villages differed in sanitation conditions, but had highly similar water supply and hygiene conditions.	199

Continued on Next Page...

Table 5.3 Summary of goals and key learning points from Chapter 4 (India study) – Continued

Goals	Learning Points	Page References
Evaluate the impact of a combined sanitation, water supply and hygiene education intervention on sanitation infrastructure and open defecation behavior.	The intervention greatly expanded private toilet construction relative to control villages (48% vs. 15%).	151, 155
	Relative to control villages, gains in private toilet construction are largest among middle-to-lower income and scheduled caste (SC) households. For example, among households with no toilet at baseline, 76.5% of SC households in intervention villages constructed a toilet between 2003 and 2008 versus 11.3% of SC households in control villages.	189, 191, 192
	Providing private toilets reduces open defecation, but the reduction is not 1:1. Behavioral changes lag hardware improvements. Nearly 40% of households with a private toilet report that adults practice open defecation daily, and 52% of these same household report that children < 5 practice open defecation daily.	171, 154
	Over 94% of private toilets are still in use, with most units installed in the last 5 to 10 years.	170
	Village open defecation prevalence is unrelated to the time since intervention activities ceased, but is strongly related to the strength of the intervention (measured by the proportion of households that have private toilets).	171, 171

Continued on Next Page...

Table 5.3 Summary of goals and key learning points from Chapter 4 (India study) – Continued

Goals	Learning Points	Page References
Evaluate the impact of a combined sanitation, water supply and hygiene education intervention on water supply infrastructure and water quality.	The intervention led to more improvements in public and private water supplies than control villages (26% vs. 18%), but overall water supply conditions are similar in 2008.	150, 158
	The gains in private tap construction are largest among the scheduled caste (SC) households. Among households without private taps at baseline, over 22.5% of SC households in intervention villages constructed a new tap between 2003 and 2008, versus just 1.8% of SC households in control villages.	190, 191, 192
	Recent improvements in intervention village water sources have improved village source water quality relative to control villages, but household drinking water quality is not statistically different between the two groups. The lack of difference at the household level is likely due to recontamination during storage.	161, 162, 163
	Over 96% of households that have a private water tap use it as their primary water source. I infer that the taps are sustainable over at least a five-year horizon.	157
Evaluate the impact of a combined sanitation, water supply and hygiene education intervention on hygiene practices.	Hygienic conditions and self-reported handwashing are similarly pooer in intervention and control villages.	165, 168, 167, 169
	Hygiene indicators are unrelated to time since intervention completion. Similarity between control and intervention households suggests that the intervention had minimal impact on these indicators.	172

Continued on Next Page...

Table 5.3 Summary of goals and key learning points from Chapter 4 (India study) – Continued

Goals	Learning Points	Page References
Evaluate the impact of a combined sanitation, water supply and hygiene education intervention on child health measured by growth and acute gastrointestinal illness.	Children in this population suffer from substantial growth faltering relative to international standards. There are no differences in child growth between children in intervention villages and children in control villages.	174, 178, 177
	Given the low prevalence of gastrointestinal illness in this population, the growth faltering likely results from poor nutrition and from asymptomatic infections that cause tropical enteropathy (both mechanisms untested in this study).	193
	Diarrhea and highly credible gastrointestinal illness (HCGI) are rare in children under five. The mean prevalence of diarrhea was 1.8% over the year, and there are no differences between groups.	184, 185
	Diarrhea prevalence is very low in control villages despite poor sanitation conditions. This population experiences very low child diarrhea prevalence with good water quality, marginal hygiene practices, and poor sanitation.	193
	Future intervention programs could target populations with large illness burdens by conducting pre-intervention rapid assessments to measure child health prior to selection.	193
Quantify some of the non-health benefits that follow from sanitation and water hardware improvements.	Installing a private toilet increases the perception of safety and privacy during defecation for women and girls by 28% (81% versus 53%).	155
	Installing a private tap saves a median of 25 minutes per day (50 min versus 75 min) for water collection activities.	159

5.2 Conclusions

Study designs for sustainability research

In this dissertation I have focused on the sustainability of water, sanitation and hygiene interventions in developing countries. By definition, measuring the sustainability of an intervention requires measuring outcomes long after the conclusion of major intervention activities. An ideal design would measure outcomes at baseline, randomize communities to treatment and control, prospectively follow communities over time, collect outcomes at the conclusion of intervention activities and collect them again in the period following the intervention (Table 2.8, Figure 2.10). The time required to evaluate intervention sustainability using such a design leads to costly research efforts that are still not protected from some threats to validity such as informative censoring.

Studies of non-randomized, pre-existing interventions can evaluate intervention sustainability more quickly at a lower financial cost, but they require more work by investigators to justify their validity. The quasi-experimental approach that I develop in Chapter 2 imposes many constraints on the intervention (Table 2.1) and requires additional assumptions for validity beyond ignorable censoring. The strongest assumption, which is common to all observational designs, is that intervention treatment is randomized conditional on observed characteristics (no unmeasured confounding). There are no easy solutions to the problem of unmeasured confounding in observational studies, but the approach I have used in this dissertation leverages the design to remove bias and help guarantee overlap in observed covariates between control and intervention communities. This, in turn, places less emphasis on statistical models and the additional assumptions that they require. There are many situations when randomization is impossible or undesirable, and there is much we can learn from carefully conducted observational studies. However, observational designs require stronger assumptions than randomized designs, and so they require a more complicated, thoughtful, theory-dependent process of argument construction than inference from randomized experiments [1, 2].

The design requires further nuance when applied in sustainability research. First, it does not guarantee clear interpretations for intervention sustainability in the case of a null finding (Table 5.2). Second, the design does not guarantee a "pure" control group in highly dynamic populations or in studies with long periods between baseline and follow-up surveys (Table 5.2). This second issue applies to any sustainability evaluation, even those that are randomized with prospective outcome measurement. The central issue is the practical (and ethical) dilemma created by measuring intervention sustainability using outcomes that require a counterfactual where it becomes increasingly difficult to expect the control group to remain intervention-free over long periods of time. A pragmatic approach would establish the health efficacy of the intervention during an initial trial, and then measure compliance to the intervention exposure (e.g., handwashing) in the post-intervention period. Sustainability inference would then rely on measures of compliance and the assumption that health benefits follow high compliance. This is clearly an area

for future methodologic development and empirical research.

Contributions to sustainability research and extensions

In both the Guatemala and India interventions, I found that behavioral components to the interventions had a limited impact on behavior change – particularly in the areas of handwashing and hygiene. This is perhaps not a surprise. Behavior change has been a stumbling block in many key public health interventions including HIV prevention [3], cardiovascular disease reduction [4] and obesity reduction [5] (the complete list of failures would be much, much longer). My findings are also broadly consistent with other disappointing findings from recent sustainability studies of household water treatment and handwashing interventions [6–8].

There is considerable evidence across the social sciences that convenience is a key factor in promoting behavior change. One important lesson from the behavior change literature is that making something easy can be more effective at inducing change than education or promotional messaging [9–11]. Additionally, there is evidence – at least in a western context – that people are much more likely to wash hands after using a toilet if they believe they may be observed [12]. Improved environmental modifications such as in-home water supply and a dedicated, centrally-located handwashing station with available water and soap may complement handwashing behavior change messages [13] (our group at Berkeley is currently initiating research on this specific intervention).

In the context of the India intervention, eliminating open defecation requires a combination of environmental modification (installing private toilets) and behavioral modification (actually using them). The intervention program expanded private toilet access to all segments of the population. The largest gains relative to control villages were among the middle income and scheduled caste households. Households in the lowest wealth quintile installed private toilets at a much higher rate than similar control households (29% vs 1%), but the majority of the poorest households still have no private toilet (Figure 4.25). The poorest households apparently still face economic barriers that the subsidized loans provided by the intervention did not overcome. Expanding toilet access to the most marginalized populations may require additional hardware subsidies, such as those implemented in addition to a CLTS intervention in Orissa, India [14]. The balance between marketing and subsidies in toilet provision interventions is an area of debate and ongoing research.

Expanding access to private toilets reduces open defecation (Figure 4.15), but the relationship is not 1:1. Behavioral modification has not kept pace with environmental modification. Among households that own private toilets, the primary reasons given for continuing to practice open defecation are: no choice (50%), privacy (26%), convenience (25%) and safety (9%). Hypothesized causes for lack of behavior change in this context include the persistent difficulty of safely disposing feces from babies and young children ("no choice"), no available facilities at places where people work ("no choice" and "con-

venience"), and lingering cultural perceptions and customs ("privacy").[1] Additional formative research that focuses on residual barriers to behavior change among toilet owners will provide valuable information to future community led total sanitation campaigns.

Potential for additional analyses using data collected in this dissertation

All of the analyses that I have presented in this dissertation have focused on community-level impacts of the combined intervention programs. I have defined the treatment of interest as living in an intervention village and evaluated the impact of that treatment on outcomes. A related question is whether there are child health impacts from individual components of the intervention programs, such as owning a private latrine or a private tap. Examples of hardware improvements of immediate interest include private toilets, private taps, and fully stocked handwashing stations (all in the Tamil Nadu dataset).

Unlike the main intervention program, the design did not carefully select balanced treatment and control groups for these treatments, so it is possible that there is not support in the data to compare children from households with and without these hardware improvements. For example, if all households who have a private tap are wealthy, then we can not remove the confounding effect of wealth from the relationship between private taps and child health. However, since the WPI/Gramalaya intervention program expanded private toilet and private tap access to households that may have not had them otherwise, this dataset may provide a possibility for reasonably good observational inference for these household-level treatments. As a segue into future analyses, I have estimated the predicted probability of owning a private toilet, a private tap, and having a fully-stocked handwashing station conditional on household covariates listed in Table 4.3) using the Super Learner (see methods in Chapter 4). Figure 5.1 summarizes the distributions. Given the good overlap in the distributions for private tap ownership and fully-stocked handwashing stations future observational analyses may be useful. The distributions of predicted probabilities of toilet ownership is fairly good, but there is a large component of the toilet owner distribution without mutual support among non-owners. This indicates that for many toilet owners there are not comparable families in the dataset that do not own toilets, and observational analyses will rely on extrapolation beyond the empirical dataset without restricting the population of inference [15].

These preliminary figures suggest that data collected in the context of a matched design where intervention programs expand access of hardware can create an opportunity to identify the impact of the hardware in a larger part of the population than would otherwise be possible. Without the intervention program and matched design, treatment effects would likely not be identifiable because there would probably not be overlap on key confounding covariates needed for adjusted estimates. For example, the intervention

[1] Anecdotally, for example, it is common among men living in rural India to meet their friends in the morning for a walk to the fields to defecate. Men may decide that the social benefits of this behavior outweigh the costs.

in India has "artificially" increased toilet ownership among poor households, allowing for experimentation in toilet ownership among poor households that would otherwise not occur.

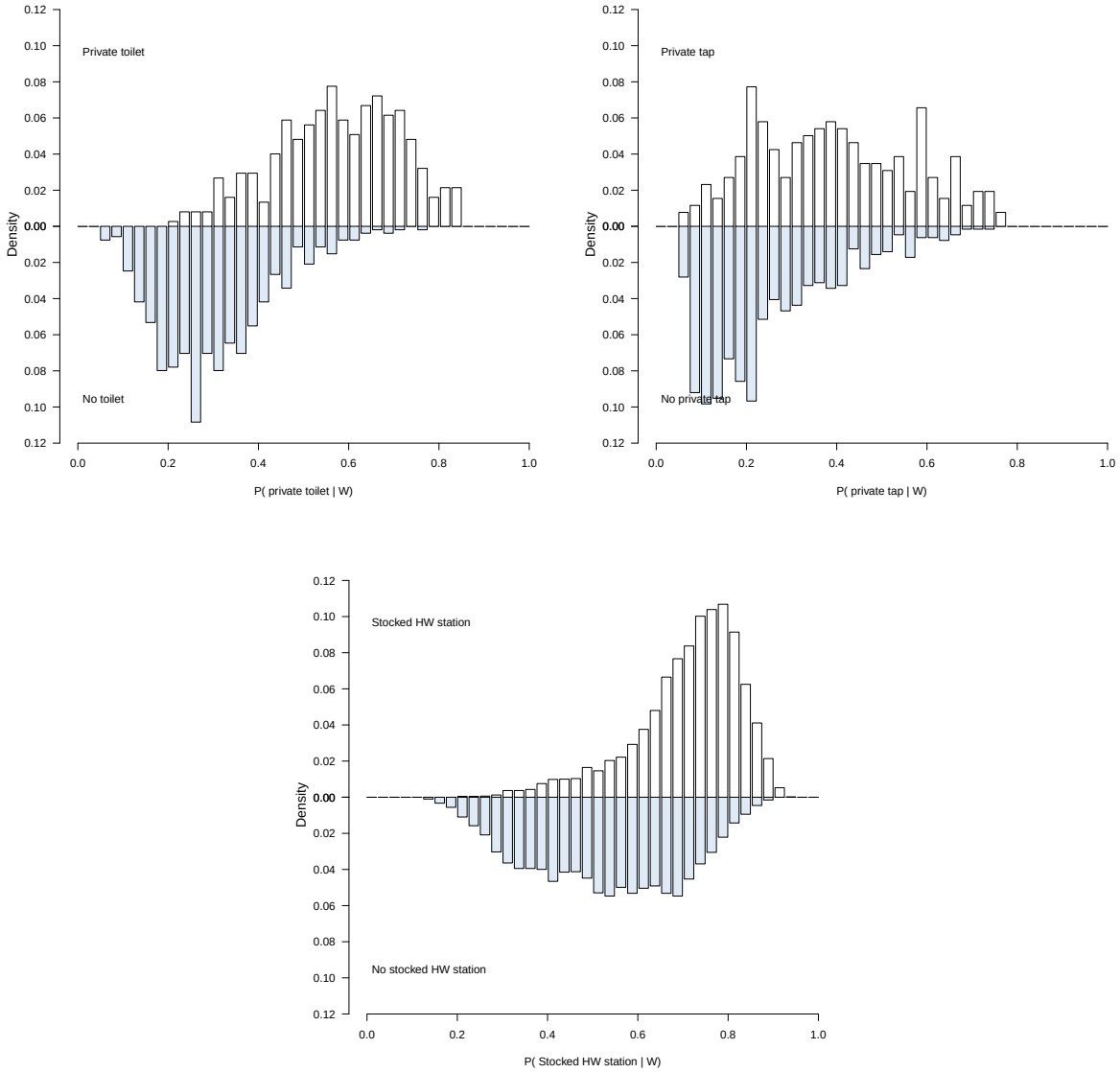


Figure 5.1: Distribution of predicted probabilities of private toilets, private taps and stocked handwashing (HW) stations (water + soap/ash present) conditional on adult and household level covariates listed in Table 4.3. Distributions are stratified by households with and without each facility. Probabilities were predicted using the Super Learner (see methods in Chapter 4).

Final remarks

Under favorable conditions, non-randomized, pre-existing interventions provide a valuable source of information about intervention sustainability. The interventions reflect implementation conditions that are not influenced by the process of scientific research, and studies of such interventions can collect outcomes after the completion of implementation activities without years of prospective follow-up. Studies of non-randomized intervention programs require care in their design, analysis and interpretation. Causal inference in this context relies on strong assumptions. The quasi-experimental approach that I developed in the course of this work has been successful in two applied field studies, and the design is useful for evaluations outside of water, sanitation and hygiene interventions. There will always be an abundance of non-randomized, pre-existing intervention programs: the methods and experience herein provide tools and guidance for investigators to learn from them empirically.

Bibliography

[1] Cook TD, Shadish WR. Social Experiments: Some Developments over the Past Fifteen Years. Annual Review of Psychology. 1994 Jan;45(1):545–580.

[2] Shadish WR, Cook TD, Campbell DT. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston: Houghton Mifflin Company; 2002.

[3] d'Cruz Grote D. Prevention of HIV infection in developing countries. The Lancet. 1996 Oct;348(9034):1071–1074.

[4] Fortmann SP, Varady AN. Effects of a Community-wide Health Education Program on Cardiovascular Disease Morbidity and Mortality The Stanfort Five-City Project. Am J Epidemiol. 2000;152(4):316–323.

[5] Sahota P, Rudolf MCJ, Dixey R, Hill AJ, Barth JH, Cade J. Randomised controlled trial of primary school based intervention to reduce risk factors for obesity. BMJ. 2001;323(7320):1029–.

[6] Luby SP, Mendoza C, Keswick BH, Chiller TM, Hoekstra RM. Difficulties in bringing point-of-use water treatment to scale in rural Guatemala. Am J Trop Med Hyg. 2008;78(3):382–7.

[7] Brown J, Proum S, Sobsey MD. Sustained use of a household-scale water filtration device in rural Cambodia. J Water Health. 2009;7(3):404–12.

[8] Luby SP, Agboatwalla M, Bowen A, Kenah E, Sharker Y, Hoekstra RM. Difficulties in maintaining improved handwashing behavior, Karachi, Pakistan. Am J Trop Med Hyg. 2009 Jul;81(1):140–145.

[9] Kaplan LM, McGuckin M. Increasing handwashing compliance with more accessible sinks. Infect Control. 1986;7(8):408–10.

[10] Sallis JF, Owen N, Fisher EB. Ecological Models of Health Behavior. 4th ed. Glanz K, Rimer BK, Viswanath K, editors. San Francisco: Josey-Bass; 2008.

[11] Kremer M, Miguel E, Mullainathan S, Null C, Zwane AP. Making water safe: Price, persuasion, peers, promoters, or product design? Working Paper. 2009;.

[12] Pedersen DM, Keithly S, Brady K. Effects of an observer on conformity to handwashing norm. Percept Mot Skills. 1986;62(1):169–70. Journal Article United states.

[13] Luby SP, Halder AK, Tronchet C, Akhter S, Bhuiya A, Johnston RB. Household characteristics associated with handwashing with soap in rural Bangladesh. Am J Trop Med Hyg. 2009 Nov;81(5):882–887.

[14] Pattanayak SK, Yang JC, Dickinson KL, Poulos C, Patil SR, Mallick RK, et al. Shame or subsidy revisited: social mobilization for sanitation in Orissa, India. Bull World Health Organ. 2009;8:580 – 587.

[15] Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. Biometrika. 2009;96(1):187–199.

Appendix A

Additional background: current evidence for intervention efficacy

In 1985 and 1991, Esrey *et al.* reviewed the effects of improved water supply and sanitation in developing countries and summarized results from 144 studies[1, 2]. In 2005 Fewtrell *et al.* updated and expanded the original review from Esrey *et al.* and reported meta-analysis estimates for water, sanitation and hygiene interventions in developing countries [3]. In 2007 Arnold and Colford [4] and Clasen *et al.* [5, 6] updated the water quality component of Fewtrell *et al.*'s systematic review, and the entire review was updated (using non-sytematic methods) by Zwane and Kramer [7].[1] Three recent systematic reviews summarize efficacy trials of handwashing interventions [9–11]. I will draw on these systematic reviews and more recent published work to summarize the current state of knowledge in this sector. In this section I focus on the interventions' impacts on diarrhea and respiratory outcomes, which are most relevant to this dissertation.

The summary below shows that there is a surprising scarcity of information about key interventions in the sector. There are virtually no rigorous studies to date on health impacts from water supply improvements, source water treatment and sanitation improvements. There have been a large number of efficacy studies on behavioral interventions that focus on household water treatment and handwashing with soap, but almost no information on their long-term adoption or effectiveness under non-trial conditions.

A.1 Water supply improvements

Water supply improvements include interventions that make water collection more convenient or efficient for consumers. The most common example of water supply improvements

[1] As this dissesrtation is "going to press," an additional, highly comprehensive review was just published by Waddington *et al.* , which updates the meta-analyses through 2009 [8].

include community taps and private, household taps. Although water supply improvements may improve water quality by reducing the level of environmental contamination at the source and at the point of use, they differ from water quality improvements because they do not typically include a specific component that cleans the water, such as chlorination or filtration.

Esrey *et al.* (1991) identified 58 studies of health effects from water supply improvements. Of these, the authors identified 32 "rigorous" studies, from which they estimated median reductions of 17% (n = 2 studies of water quantity and quality improvements) and 20% (n=5 studies of water quantity improvement) in diarrhea morbidity [2]. By 2003, there were just six intervention studies that estimated the health effects of water supply improvements, and only two were deemed to be good quality based on basic quality measures suggested by Blum and Feachum (1983) [3, 12]. The pooled relative risk of water supply interventions on diarrhea (n = 4 studies reported by Fewtrell *et al.*) is 1.03 (95% CI 0.73 – 1.46, Figure A.1). This estimate only includes rural populations, and the most recent study was published in 1992.

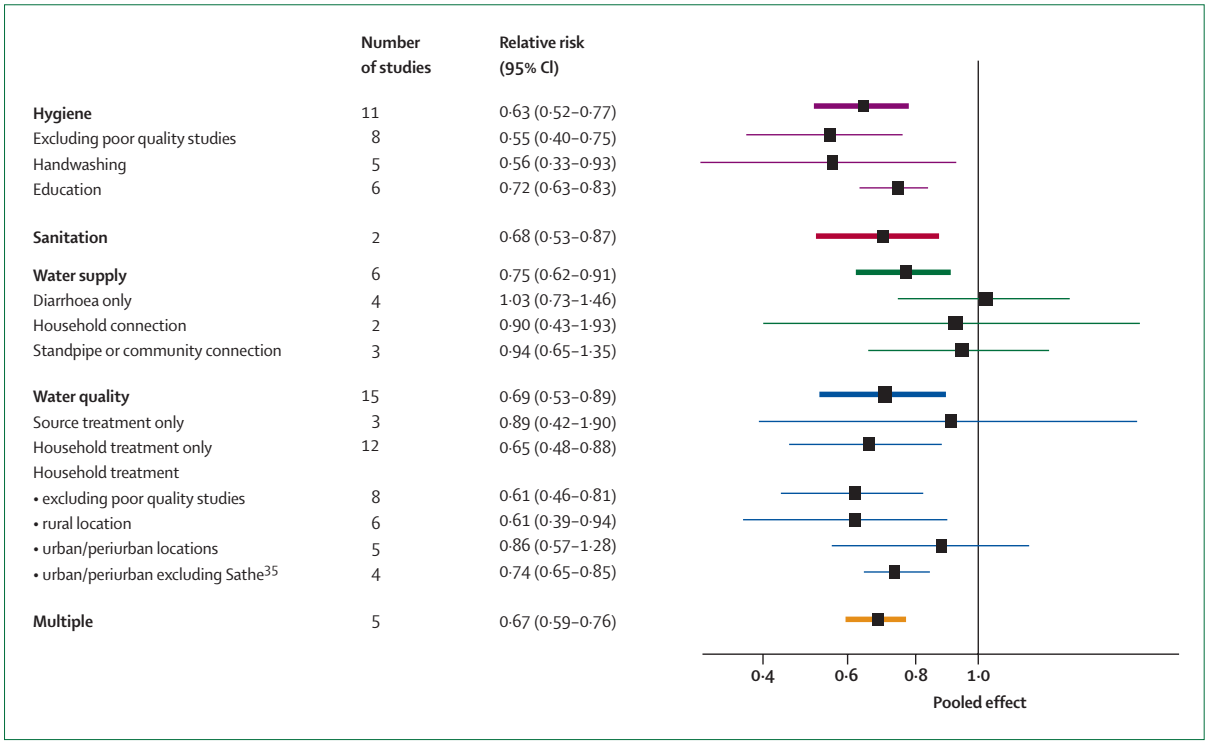


Figure A.1: Summary meta-analysis estimates of intervention efficacy from Fewtrell *et al.*, reproduced from the original publication [3].

In more recent work, Jalan and Ravallion use propensity score matching with cross-sectional data from 33,000 households in 1,765 rural Indian villages to estimate the impact

of private tap water on diarrhea among children under five [13]. Consistent with results from Esrey *et al.*, but in contrast to the current intervention literature, Jalan and Ravallion report a 21% reduction in diarrhea due to tap water improvements. They also report effect modification by income and maternal education: tap water provision led to no health improvements among the poorest two income quintiles unless mothers had more than a primary school education.

In sum, there is only a sparse literature on the health impacts of water supply improvements in developing countries and it focuses mainly on impacts in rural areas – there are no rigorous studies to date in urban areas. The results are mixed: current intervention evidence suggests no impact on diarrheal morbidity, but observational [13] and historic (pre-1986) studies [2] suggest a roughly 20% reduction in diarrheal morbidity following water supply improvements alone.

A.2 Water quality improvements

Water quality improvements include interventions with the specific aim of improving the microbiological quality of the water consumed. Water quality interventions can be centralized (e.g., municipal water treatment plants) or decentralized at the household level. In developing countries, the vast majority of research has focused on point of use water quality interventions [3, 4, 6]. Although centralized water distribution systems are in place in many urban and peri-urban areas in developing countries, the complexity and cost of centralized, source water treatment continues to make it relatively rare. Point of use water treatment allows households to clean their water using simple chemical additives (e.g., chorine or flocculant), filtration, boiling or UV disinfection. Unlike source water treatment, household treatment relies on individual-level behavior modification: users typically need to modify their water handling practices and incorporate water treatment into their daily routine. This behavioral component presents important challenges for compliance in trials and sustainability in practical applications (explored in Chapter 2).

Recent systematic reviews identified just three studies of source water quality improvements alone that included village level chlorination or sedimentation combined with ceramic filtration [3, 6]. A major limitation of all three studies is their small size: none included more than 2 villages (village is the unit of intervention). None identified reductions in diarrhea morbidity (pooled relative risk 0.89, 95% CI $0.42 - 1.90$).

In contrast, Clasen *et al.* identified numerous household water quality intervention trials that evaluated chlorination (16 trials), solar disinfection (three trials), filtration (eight trials) or combined flocculation and disinfection (PUR sachets, seven trials) [6]. With few exceptions, the household water quality intervention trials in developing countries have documented reductions in diarrhea in children under five: meta-analysis summary estimates range between $20 - 35\%$ reduction (Figure A.1) [3, 4, 6]. The finding of a near universal protective effect across numerous trials suggests that household water treatment methods effectively reduce diarrhea when used regularly and correctly.

A.3 Handwashing and hygiene promotion

In developing countries, handwashing and hygiene promotion interventions typically focus on mothers and caregivers of small children. Randomized trials provide soap to each participating family, and then motivate its use in group training exercises. In all trials to date, the studies have focused hygiene messages on critical times to wash hands (such as before cooking or eating, and after defecation or changing diapers), on proper disposal of human and animal feces, and on general health information related to washing hands [11]. Curtis [14, 15] recently proposed a change of course in hygiene promotion campaigns: rather than design campaigns as purely didactic, rational exercises, Curtis argues that they should exploit people's innate emotional disgust with dirty objects and disease, and highlight those conditions in their living environments (positioning handwashing as an natural antidote). This later intervention message has yet to be tested in a scientific study.

The 2008 systematic review by Ejemot *et al.* identified 14 handwashing intervention studies, and six of the trials took place in low or middle income countries [11]. Of these, the single institution-based intervention included a school-based intervention in rural China that promoted handwashing with soap [16]. The study did not have rigorous illness measurement, but did document a 40% relative reduction in school absenteeism (1.2 episodes versus 2.0 episodes of absence per 100 student-weeks). Illness-specific absence was too rare to estimate. The remaining five handwashing studies were cluster-randomized, community-based interventions [17–21]. The summary estimate across the five studies indicates a 31% reduction in diarrhea incidence (pooled rate ratio 0.69, 95% CI $0.55 - 0.87$).

Although fewer studies have measured respiratory health impacts of handwashing, a meta-analysis from Rabie and Curtis [10] estimates a 24% reduction in respiratory infections following from handwashing with soap interventions. A limitation of Rabie and Curtis' summary for this context is that all eight studies included in the meta-analysis were conducted in schools or daycares in industrialized countries, and their outcome included all acute respiratory infections (both upper and lower). To date, the only measurement of respiratory impacts from handwashing with soap in developing countries is a 2005 controlled trial by Luby *et al.* in Pakistan, which documented a 45 – 50% reduction in pneumonia incidence among children < 5 years in the handwashing intervention groups [22].

A.4 Sanitation improvements

Improved sanitation effectively separates human excreta from human contact and the environment, and the most common interventions in developing countries are various forms of private latrines. The logistical difficulties and time scale required to implement sanitation interventions make them difficult to study, and large scale interventions, such

as urban sewerage improvements, are essentially impossible to randomize. Nonetheless, one of the most surprising findings of Fewtrell *et al.*'s review is that by 2003 there had been just *two* pure sanitation intervention studies, neither randomized, and only one that the authors considered to be "good" quality [3].[2]

Azurin and Alvero evaluated the impact of communal latrines on cholera in an urban center in the Philippines, and found the intervention highly protective (relative risk 0.32, 95% CI 0.24 - 0.42) [24]. However, the study only included four communities and employed relatively poor surveillance methods. The second study by Daniels *et al.* evaluated the impact of private latrine installation in rural Lesotho using a hospital-based case-control design, and found that latrines reduced the odds of diarrhea in children under five by 24% (odds ratio 0.76, 95% CI 0.58 - 1.01) [25].

In 2007, Barreto *et al.* published a repeated sample, longitudinal study of child diarrhea in the city of Salvador, Brazil [26].[3] This is the only study that that I am aware of that documents the impact of a large urban sanitation project in a developing country. The initial 1997 cross-sectional sample was during the beginning of a massive sanitation effort in the study population. Sanitation improvements included laying more than 2000 km of sewer pipes, 86 pumping stations, and connection of more than 300,000 households to the sewerage network over eight years ($1996-2004$). The study team identified a second cross-sectional sample in 2003. Both samples included 8 months of follow-up to ensure sufficient diarrhea episodes. Over the six years between samples the study documented an increase of sewerage connection from 0% to more than 50% in the majority of study neighborhoods. After adjusting for potential confounders and baseline sewerage coverage, the sanitation improvements reduced diarrhea in children under age three by 22% (prevalence ratio 0.78, 95% CI $0.74 - 0.81$).

A forthcoming randomized trial evaluated the impact of a community-led total sanitation campaign in 40 rural villages in Orissa, India[27]. The village-level intervention used defecation mapping, "walks of shame" (a community walk to identify current conditions), and community education to motivate shifts in behavior from open defecation to using latrines. Latrine motivation and education promoted health and non-health benefits (e.g., dignity, time savings, and privacy for women). The intervention also provided technical assistance and materials for latrine construction. Over a one year period, private latrine ownership and usage increased from 6% to 32% in intervention villages and remained at 13% in control villages. There was a dramatic decline in diarrhea over the year in both intervention and control villages [28]. In intervention villages, the two-week period prevalence of diarrhea in children under age five dropped from 28% to 15% between 2005 and 2006; in control villages the same prevalence dropped from 23% to 15%. The authors do not have an explanation for the large secular decline in prevalence in the study population (the authors rule out handwashing, water sources, and water quality because these indica-

[2] A finding even more remarkable given the *BMJ* reader poll that voted sanitation as the most important medical advancement since 1840 [23].

[3] This paper was nominated for (but did not win) 2007 paper of the year in *Lancet*.

tors did not change much over the study period). The difference-in-differences estimate of the intervention effect was a 5% absolute prevalence reduction (17% relative reduction) in children under age five, and 10% absolute prevalence reduction (28% relative reduction) in children under age three (note: standard errors and confidence intervals are not provided, but the difference for children under age three is significant at p=0.1).

There have been additional observational studies of combined interventions that include sanitation (detailed in the Multiple Interventions section, below). Although the few existing studies of sanitation interventions alone demonstrate promising reductions in child diarrhea, the scientific evidence base is inadequate to guide the enormous investments needed in sanitation infrastructure. For example, a recent World Bank Water and Sanitation Program (WSP) assessment of 16 countries in Africa estimated that it will require investments of *$1.3 billion US dollars per year for 10 years* for the countries to meet the regional sanitation goal of 63% of the population with improved sanitation by 2015 [29].

A.5 Multiple interventions

The "F Diagram" in the Introduction (Figure 1.1, page 5) depicts the complexity of gastrointestinal disease transmission. It also implies that if all pathways are viable then no single intervention (water, sanitation, or hand washing) will be sufficient to block disease transmission. The relative importance and impact of each intervention will depend on the specific conditions endemic to particular settings and populations. For example, among breastfeeding infants we would expect very little direct transmission through water (fluids) and a majority through person-to-person transmission (fingers) or the environment (flies/floors). In this example, we would expect a water quality intervention to have a small impact on infant health relative to hand washing or sanitation.

As a matter of practice, it is generally difficult to assess the relative importance of different transmission pathways for a given population and setting. The consequence is that many implementing organizations make the rational decision to intervene on all three points (water, sanitation and hygiene). Eisenberg *et al.* have demonstrated with theoretical and simulation results that if each pathway alone is sufficient to maintain diarrheal disease, that single-pathway interventions will have minimal benefit [30]. However, if there is a single, dominant transmission pathway, then investing in a multiple intervention strategy over-allocates scarce resources that could be dispersed to a larger population if just one approach is sufficient to reduce or eliminate transmission. This remains a hotly debated point in the scientific and implementation fields. As I detail below, the empirical evidence remains sparse and conflicted on this important (perhaps central) point in designing water, sanitation and hygiene interventions to reduce diarrhea. Our results from Tamil Nadu, India (Chapter 4), will contribute an additional data point to this debate.

Nearly all of the evidence to date on multiple interventions arises from non-randomized (observational) studies. Three non-intervention, observational studies have documented

interactions between water quality and sanitation interventions in their effect on child diarrhea [31–33]. VanDerslice and Briscoe studied a cohort of 2,355 Filipino infants in a longitudinal study with bimonthly measurement during their first year of life [31]. Their model estimates suggest that water quality improvements would have no effect in neighborhoods with poor community sanitation, but reducing fecal coliform concentrations by two orders of magnitude in neighborhoods with good sanitation would reduce child diarrhea by 40%. They also estimated that providing private latrines would independently reduce diarrhea by 42%.

Esrey found a similar interaction between water supply and sanitation improvements in an analysis of Demographic and Health Surveys (DHS) from eight countries [32]. The analysis found that water supply improvements alone did not reduce diarrhea prevalence, regardless of the level of sanitation. Diarrhea reductions from improved sanitation varied depending on the water supply improvements: optimal sanitation led to a relative reduction in diarrhea prevalence of 44% in unimproved water conditions, but only 19% in the presence of optimal water conditions.

Esrey also analyzed the impacts of water and sanitation improvements on child growth and found interaction in those outcomes too. For example, improvements in water were associated with increases in child weight-for-age, but only when sanitation was improved and optimal water conditions were present: children with optimal water conditions compared to no water improvements had an average of 0.139 (95% CI 0.014 to 0.264) higher Z-scores. Esrey also identified synergy[4] between water and sanitation interventions in child growth outcomes. Using weight-for-age Z-scores as an example: intermediate improvements in only water did not improve weight-for-age Z-scores (difference = 0.017, 95% CI $-0.056 - 0.090$) nor did intermediate sanitation improvements alone (difference = 0.072, 95% CI $-0.003 - 0.147$). However, in the presence of both intermediate improvements in water and sanitation, weight-for-age Z-scores increased 0.115 (95% CI $0.040 - 0.189$).

Scott updated and expanded Esrey's original DHS diarrhea analysis using surveys from 27 African countries that spanned 1995 - 2003 [33]. Scott analyzed the data using point-treatment marginal structural models[35], and reports results consistent with Esrey, but more attenuated. Combined improvements[5] of both water and sanitation interventions would reduce diarrhea by 18% (95% CI $14\% - 24\%$), while improving water alone would reduce diarrhea by 4% (95% CI $-1\% - 13\%$) and improving sanitation alone would reduce diarrhea by 11% (95% CI $7\% - 17\%$). Scott's analysis identifies some synergy between water and sanitation interventions: their combined effect is larger than the sum of their independent effects.

More recently, Garrett *et al.* conducted a quasi-experimental evaluation of a combined chlorine-disinfection, safe water storage, latrine construction and improved water supplies (including shallow wells and rainwater harvesting) in rural Kenya [36]. The study included

[4]Synergy is a condition where the joint effect of two interventions is greater than the sum of their independent effects [34].

[5]Scott used the JMP definitions of improved water and sanitation conditions (Table 1.1).

12 intervention and 6 control villages (controls were selected by geographic proximity to the intervention villages). The authors measured water management, latrine coverage, and child diarrhea 1.5 years into the project. Promotion activities were ongoing throughout the study.

A quasi-experimental study from Pattanayak *et al.* evaluates a combined water supply, sanitation and hygiene intervention in rural Maharashtra [37, 38]. In contrast to observational studies, which estimate the treatment effects of water and sanitation improvements, the treatment in the Maharashtra study is a targeted, community-demand driven intervention that facilitates water and sanitation improvements and promotes hygiene, which is more consistent with a randomized trial. This large matched cohort study covers 242 villages (95 intervention, 147 control, >10,000 households). Similar to the design that I propose in Chapter 2, the study used a matched cohort design, with control village selection based on a propensity score match using baseline characteristics. Pattanayak *et al.* used a matched sample because the intervention was both targeted and community-demand-driven, so they needed to carefully construct the control set. The intervention was deployed between 2005 and 2006, and the team conducted two baseline surveys in 2005 (wet and dry seasons) and corresponding follow-up surveys in 2007.

Pattanayak *et al.* found that the intervention improved water supply and latrine coverage: difference-in-difference (DID) estimates of 6% for both. The rainy season diarrhea prevalence dropped dramatically in both the intervention and control villages, falling from 13% to 7.5% between 2005 and 2007, and the intervention effect was non-significant (4.2% absolute prevalence reduction in intervention villages versus 6.5% reduction in control villages, rainy season, p-value = 0.138).[6] The lack of observed health impacts raises questions about whether the health improvements identified in purely observational studies (above) are confounded and reflect differences in other characteristics such as income or access to health services. An additional question is whether the efficacy observed in smaller interventions can scale up to large interventions on a country or regional level.

In the only published randomized controlled trial specifically designed and powered to evaluate combined interventions, the two interventions evaluated were point of use water treatment and handwashing promotion with soap [21]. Individually, both interventions were associated with a marked reduction in diarrheal disease (51 and 64% reduction). There was no additional reduction in diarrhea with the combined intervention (55% reduction). These results are consistent with results of the Fewtrell *et al.* meta-analysis, which noted that studies of interventions that combined water quality, sanitation, and hygiene interventions had no greater reduction in diarrheal disease than published trials of single interventions (Figure A.1, page 226) [3].

[6]Note: diarrhea estimates are presented in the working paper but not in the published manuscript that describes the design.

A.6 Hypotheses for synergy and antagonism between interventions

A key question of multiple interventions is whether water, sanitation and hygiene interventions have additive or synergistic effects. A technical definition of synergy in this context is that the joint effect of two interventions is greater than the sum of their independent effects; antagonism is the opposite of synergy: the joint effect of two or more interventions is less than the sum of their independent effects [34]. Finally, additivity describes the condition when the joint effects equal the sum of the independent effects [34].[7] As detailed above, there is conflicting evidence across the range of different studies about this issue. Below, I outline briefly the hypotheses for synergy and antagonism.

A.6.1 Synergy

The observational studies of multiple interventions found synergy between water supply and sanitation interventions [31–33], which is supported by theoretical results [39]. The rationale for synergy follows the basic argument for multiple interventions. Consider the most simple hypothetical case of only two viable pathogen transmission pathways: person-to-person transmission and waterborne transmission.[8] A handwashing intervention reduces person-to-person transmission, but children may still be infected from contaminated water so the net effect on child diarrhea is a 15% reduction. Similarly, if water quality is improved it interrupts waterborne transmission, but will end person-to-person transmission, so its net effect on child diarrhea is a 20% reduction. Combined, however, the two interventions will prevent pathogen transmission entirely (100% reduction), which is greater than the additive effect of each intervention alone.

A.6.2 Antagonism

In contrast to theoretical results and observational findings, the recent meta-analysis by Fewtrell *et al.* and the single randomized trial of combined water quality and handwashing interventions suggest antagonism between multiple interventions [3, 21]. Antagonism is less-intuitive based on the theory of pathogen transmission alone, but there are at least four hypotheses for antagonism in the context of multiple interventions [21].[9]

First, when interventions are combined, the separate components may not be delivered as effectively by the implementing organization nor taken up as effectively by the

[7] For example, if independently water quality and sanitation interventions reduced diarrhea by 20% each, then their expected additive effect would be 40% = 20% + 20%. If the combined effect was 60% > (20% + 20%), then there would be evidence of synergy. In contrast, if the combined effect was 10% < (20% + 20%), then there would be evidence for antagonism.

[8] Briscoe makes a more rigorous and detailed argument [40]

[9] This set of hypotheses and discussion is based on material originally written by Steve Luby as part of a 2008 research proposal to the Bill and Melinda Gates Foundation.

target households compared to individual interventions which allow both the implementing organization as well as the recipients to focus on a single effective intervention. A comprehensive review of public health behavior change in developing countries concluded that simple focused messages were the most effective [41]. Multiple interventions mean multiple messages and multiple opportunities to lose interest or practice sub-optimal behaviors.

A second hypothesis for the failure to observe marginal reduction in diarrhea with additional inputs is that the diarrhea reductions measured in controlled trials are dominated by courtesy bias, which would not be additive with multiple interventions. The nature of the interventions make them impossible to blind participants and administrators. Households that receive an intervention to reduce diarrhea as part of a study generally understand that the study team expects to see less diarrhea, and the participants themselves often expect to see less diarrhea. The combination of no blinding with a clear connection made between the treatment and the outcome for participants could bias diarrhea reporting non-differentially. This differential misclassification would likely bias the treatment effects away from the null (with study participants or assessors under-reporting diarrhea in the intervention group), and the estimated effect sizes may be larger than their true values. In the context of multiple interventions, standard economic theory predicts that the marginal perceived value of additional inputs is less than the original input,[10] thus courtesy bias would not be expected to be proportional to the number of inputs given. If courtesy bias is a substantial contributor to the reported reduction in diarrhea, it may obscure more modest additive benefits of combined interventions.

Scott [33] argues that since blinding in this context is usually impossible, one remedy is to study interventions using carefully designed observational studies with statistical analysis to adjust for potential confounding variables. Another solution may be less frequent disease monitoring [42]. Alternatively, objective outcome measures such as anthropometrics or stool sampling could provide additional evidence to validate self-reported diarrhea. The problem with objective outcomes is that they are typically much more expensive to collect (but could be collected on a random subsample of the study population).

A third hypothesis for the lack of an additive effect on combined interventions may be that the relationship between pathogen dose and probability of diarrhea is non-linear. Many phenomena throughout biology and ecology including, for example, such wide ranging processes as glucose versus lactose utilization preference for Escherichia coli [43] and the reduction of insect pest populations with introduction of a natural enemy [44] are best described by a sigmoidal curve (Figure A.2). The central observation is that the relationship between dose and response throughout nearly all biological processes is quite different at the highest and lowest dose ranges, compared to the mid range. Applying this pattern to diarrhea transmission, initial reductions in pathogen load from an effective single intervention would yield marked reduction in the risk of diarrhea, but additional reductions in pathogen dose would have lower marginal benefits. The most relevant ques-

[10]For most individuals, a second piece of chocolate cake has less marginal utility than the first.

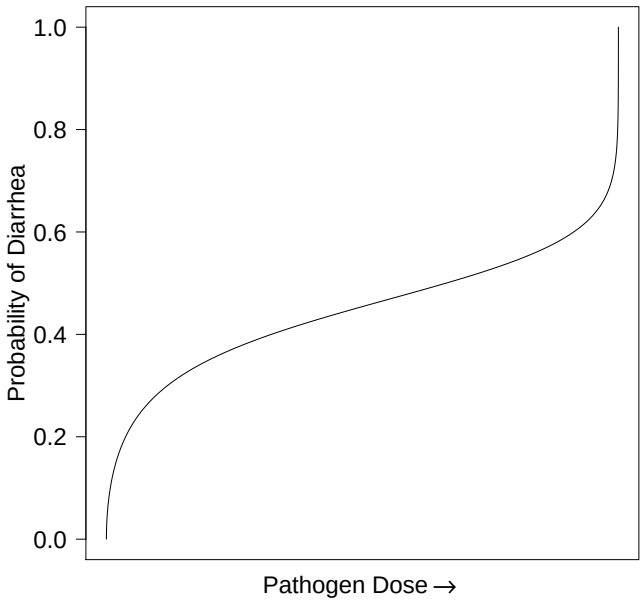


Figure A.2: Hypothetical dose-response curve between pathogen dose and diarrhea risk.

tion for the context of pathogen dose and diarrhea, is does a single intervention, in fact reduce dose enough to shift the household to a less steep part of the curve, and how steep is the slope in the middle of the range?

A fourth hypothesis for the lack of additive benefit may be that multiple interventions are interrupting the same chain of transmission. If enough of the pathways of the dominant pathogens are closely linked, then well-implemented single interventions could perform as well as combined interventions. For example, if a mother's hand contacts the environment and collects many fecal pathogens, and subsequently contacts and contaminates the household drinking water, this could transmit pathogens and cause diarrhea among others in the households. This chain of transmission could be interrupted by either keeping the water protected or by regular handwashing by the mother. If this was the dominant mode of pathogen transmission in the household then regular practice of either handwashing with soap or water treatment would interrupt transmission, and there would be no added benefit from both practices. Most likely some combination of these factors has contributed to the lack of observed combined benefits in diarrhea reduction seen with combined interventions.

Bibliography

[1] Esrey SA, Feachem RG, Hughes JM. Interventions for the control of diarrhoeal diseases among young children: improving water supplies and excreta disposal facilities. Bull World Health Organ. 1985;63(4):757–72.

[2] Esrey SA, Potash JB, Roberts L, Shiff C. Effects of improved water supply and sanitation on ascariasis, diarrhoea, dracunculiasis, hookworm infection, schistosomiasis, and trachoma. Bull World Health Organ. 1991;69(5):609–21.

[3] Fewtrell L, Kaufmann RB, Kay D, Enanoria W, Haller L, Colford J J M. Water, sanitation, and hygiene interventions to reduce diarrhoea in less developed countries: a systematic review and meta-analysis. Lancet Infect Dis. 2005;5(1):42–52.

[4] Arnold BF, Colford J J M. Treating water with chlorine at point-of-use to improve water quality and reduce child diarrhea in developing countries: a systematic review and meta-analysis. Am J Trop Med Hyg. 2007;76(2):354–64. The American journal of tropical medicine and hygiene.

[5] Clasen T, Roberts I, Rabie T, Schmidt W, Cairncross S. Interventions to improve water quality for preventing diarrhoea. Cochrane Database Syst Rev. 2006;3:CD004794. Cochrane database of systematic reviews (Online).

[6] Clasen T, Schmidt WP, Rabie T, Roberts I, Cairncross S. Interventions to improve water quality for preventing diarrhoea: systematic review and meta-analysis. Bmj. 2007;334(7597):782. BMJ (Clinical research ed.

[7] Zwane AP, Kremer M. What Works in Fighting Diarrheal Diseases in Developing Countries? A Critical Review. World Bank Research Observer. 2007;22(1):1–24.

[8] Waddington H, Snilstveit B, White H, Fewtrell L. Water, sanitation and hygiene interventions to combat childhood diarrhoea in developing countries. Int Initiative for Impact Eval. 2009 Aug;Synthetic Review 001.

[9] Curtis V, Cairncross S. Effect of washing hands with soap on diarrhoea risk in the community: a systematic review. Lancet Infect Dis. 2003;3(5):275–81.

[10] Rabie T, Curtis V. Handwashing and risk of respiratory infections: a quantitative systematic review. Trop Med Int Health. 2006;11(3):258–67.

[11] Ejemot RI, Ehiri JE, Meremikwu MM, Critchley JA. Hand washing for preventing diarrhoea. Cochrane Database Syst Rev. 2008;(1):CD004265.

[12] Blum D, Feachem RG. Measuring the impact of water supply and sanitation investments on diarrhoeal diseases: problems of methodology. Int J Epidemiol. 1983;12(3):357–65.

[13] Jalan J, Ravallion M. Does piped water reduce diarrhea for children in rural India? Journal of Econometrics. 2003;112(1):153–173.

[14] Curtis VA. A natural history of hygiene. Can J Infect Dis Med Microbiol. 2007;18(1):11–4.

[15] Curtis VA. Dirt, disgust and disease: a natural history of hygiene. J Epidemiol Community Health. 2007;61(8):660–4.

[16] Bowen A, Ma H, Ou J, Billhimer W, Long T, Mintz E, et al. A cluster-randomized controlled trial evaluating the effect of a handwashing-promotion program in Chinese primary schools. Am J Trop Med Hyg. 2007;76(6):1166–1173.

[17] Han AM, Hlaing T. Prevention of Diarrhea and Dysentery by Hand Washing. Transactions of the Royal Society of Tropical Medicine and Hygiene. 1989;83(1):128–131.

[18] Stanton BF, Clemens JD. An educational intervention for altering water-sanitation behaviors to reduce childhood diarrhea in urban Bangladesh. II. A randomized trial to assess the impact of the intervention on hygienic behaviors and rates of diarrhea. Am J Epidemiol. 1987;125(2):292–301.

[19] Haggerty PA, Muladi K, Kirkwood BR, Ashworth A, Manunebo M. Community-based hygiene education to reduce diarrhoeal disease in rural Zaire: impact of the intervention on diarrhoeal morbidity. Int J Epidemiol. 1994;23(5):1050–9.

[20] Luby SP, Agboatwalla M, Painter J, Altaf A, Billhimer WL, Hoekstra RM. Effect of intensive handwashing promotion on childhood diarrhea in high-risk communities in Pakistan: a randomized controlled trial. JAMA. 2004;291(21):2547–54.

[21] Luby SP, Agboatwalla M, Painter J, Altaf A, Billhimer W, Keswick B, et al. Combining drinking water treatment and hand washing for diarrhoea prevention, a cluster randomised controlled trial. Trop Med Int Health. 2006;11(4):479–89.

[22] Luby SP, Agboatwalla M, Feikin DR, Painter J, Billhimer W, Altaf A, et al. Effect of handwashing on child health: a randomised controlled trial. Lancet. 2005;366(9481):225–33.

[23] Ferriman A. BMJ readers choose the "sanitary revolution" as greatest medical advance since 1840. Bmj. 2007;334(7585):111–a–.

[24] Azurin JC, Alvero M. Field evaluation of environmental sanitation measures against cholera. Bull World Health Organ. 1974;51(1):19–26. Bulletin of the World Health Organization.

[25] Daniels DL, Cousens SN, Makoae LN, Feachem RG. A case-control study of the impact of improved sanitation on diarrhoea morbidity in Lesotho. Bull World Health Organ. 1990;68(4):455–63.

[26] Barreto ML, Genser B, Strina A, Teixeira MG, Assis AM, Rego RF, et al. Effect of city-wide sanitation programme on reduction in rate of childhood diarrhoea in northeast Brazil: assessment by two cohort studies. Lancet. 2007;370(9599):1622–8. Lancet.

[27] Pattanayak SK, Yang JC, Dickinson KL, Poulos C, Patil SR, Mallick RK, et al. Shame or subsidy revisited: social mobilization for sanitation in Orissa, India. Bull World Health Organ. 2009;8:580 – 587.

[28] Pattanayak S, Dickinson K, Yang J, Patil S, Praharaj P, Poulos C. Promoting Latrine Use: Midline Findings from a Randomized Evaluation of a Community Mobilization Campaign in Bhadrak, Orissa. RTI Working Paper 07-02; 2007.

[29] Getting Africa on Track to Meet the MDGs on Water and Sanitation. Washington, DC; 2006.

[30] Eisenberg JN, Scott JC, Porco T. Integrating disease control strategies: balancing water sanitation and hygiene interventions to reduce diarrheal disease burden. Am J Public Health. 2007;97(5):846–52.

[31] VanDerslice J, Briscoe J. Environmental interventions in developing countries: interactions and their implications. Am J Epidemiol. 1995;141(2):135–44.

[32] Esrey SA. Water, waste, and well-being: a multicountry study. Am J Epidemiol. 1996;143(6):608–23.

[33] Scott JC. Water Supply, Sanitation, and Gastrointestinal Illness: Estimating the risk of disease using cross-sectional data and marginal structural models. Saarbrucken: VDM Verlag; 2008.

[34] Rothman K, Greenland S. Modern Epidemiology. 2nd ed. Philadelphia: Lippincott-Raven Publishers; 1998.

[35] Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000;11(5):550–60.

[36] Garrett V, Ogutu P, Mabonga P, Ombeki S, Mwaki A, Aluoch G, et al. Diarrhoea prevention in a high-risk rural Kenyan population through point-of-use chlorination, safe water storage, sanitation, and rainwater harvesting. Epidemiol Infect. 2008;136(11):1463–71. Epidemiology and infection.

[37] Pattanayak SK, Poulos C, Yang JC, Patil SR, Wendland KJ. Of taps and toilets: quasi-experimental protocol for evaluating community-demand-driven projects. J Water Health. 2009;7(3):434–51. Journal Article England.

[38] Pattanayak S, Poulos C, Wendland K, Patil S, Yang J, Kwok R, et al. Informing the water and sanitation sector policy: Case study of an impact evaluation study of water supply, sanitation and hygiene interventions in rural Maharashtra, India. Working Paper 06-04; 2008.

[39] Briscoe ME. Research note: proxy responses in health surveys: a methodological issue. Sociology of Health and Illness. 1984;6(3):359–65.

[40] Briscoe J. Intervention studies and the definition of dominant transmission routes. Am J Epidemiol. 1984;120(3):449–55.

[41] Loevinsohn BP. Health education interventions in developing countries: a methodological review of published articles. Int J Epidemiol. 1990;19(4):788–94.

[42] Schmidt WP, Luby SP, Genser B, Barreto ML, Clasen T. Estimating the longitudinal prevalence of diarrhea and other episodic diseases: continuous versus intermittent surveillance. Epidemiology. 2007;18(5):537–43.

[43] Ozbudak EM, Thattai M, Lim HN, Shraiman BI, van Oudenaarden A. Multistability in the lactose utilization network of Escherichia coli. Nature. 2004;427(6976):737–740.

[44] Beddington JR, Free CA, Lawton JH. Characteristics of successful natural enemies in models of biological control of insect pests. Nature. 1978;273(5663):513–519.

Appendix B

Targeted maximum likelihood estimation for point-treatment studies

This appendix includes a short introduction to the main statistical methods used in Chapters 3 and 4. Targeted maximum likelihood estimation (MLE) is an estimation approach that attempts to reduce confounding bias by modeling both the treatment mechanism and the outcome mechanism [1, 2]. The approach is fundamentally tied to the Neyman-Holland-Rubin causal model, which conceptualizes causal inference in terms of potential outcomes under treatment and control, only one of which is observed ([3–5], see Chapter 2 for a detailed exposition and [6] for a review).

The targeted MLE approach is similar to standard maximum likelihood regression, but it targets the likelihood to a specific parameter of interest, for example, the risk difference [1]. Like the Double-Robust marginal structural model estimator [7], the targeted MLE estimator is considered "double robust" because it is consistent if the analyst correctly specifies the model for the outcome *or* the model for the treatment mechanism [1]. One advantage of targeted MLE over the Double-Robust marginal structural model estimator is that it is easier to implement using standard software. Targeted MLE also assumes the proper range of the parameter of interest (e.g., a risk difference will like between -1 and 1), which is not true of the estimating equation approach. This latter attribute can gain efficiency in relatively small samples [1].

Let Y be an outcome of interest, A be the intervention status equal to 1 for treatment units and 0 for control units, and W be a set of covariates that are potential confounders of the relationship between A and Y. It is possible to estimate the risk difference using the following steps (following notation from Bembom *et al.* [8]):

1. Estimate the conditional expectation of Y given A and W using a generalized linear model with maximum likelihood. I denote this initial estimate $Q^0(A, W)$. For binary

outcomes, I estimate this model using a logit link; for continuous outcomes I use an identity link.

2. Estimate the conditional probability of receiving the intervention, A, given W using a logit model. I denote this estimate $g^0(A|W)$.[1]

3. For each individual, calculate a covariate based on her observed values for A and W and using the estimate $g^0(A|W)$. I denote this covariate $h(A, W)$, where:

$$h(A, W) = \frac{I(A = 1)}{g^0(1|W)} - \frac{I(A = 0)}{g^0(0|W)} \tag{B.1}$$

4. Update the original regression by adding the covariate $h(A, W)$ and estimate the corresponding coefficient by maximum likelihood, holding the remaining coefficient estimates at their initial values. In practice, this is achieved by estimating a univariate regression of Y on $h(A, W)$ with $Q^0(A, W)$ as an offset with coefficient constrained to one. Let ϵ_n be the coefficient on $h(A, W)$. I denote this one-step updated regression $Q^1(A, W)$ where:

$$Q^1(A, W) = Q^0(A, W) + \epsilon_n h(A, W) \tag{B.2}$$

for the case of a continuous outcome, and:

$$Q^1(A, W) = \frac{1}{1 + \exp[-m^0(A, W) + \epsilon_n h(A, W)]} \tag{B.3}$$

for the case of a binary outcome (note that $m^0(A, W) = \text{logit } Q^0(A, W)$).

5. Evaluate the updated regression at $A = 1$ and $A = 0$ to get two predicted outcomes for each individual. Take the empirical mean of the difference across the population to obtain a targeted estimate of the difference:

$$\theta^{T-MLE} = \frac{1}{n} \sum_{i=1}^{n} Q^1(1, W_i) - Q^1(0, W_i) \tag{B.4}$$

The methods above outline the one-step update of targeted MLE estimator, which is the most simple case of targeted MLE. The estimator above only targets the marginal causal effect, but it is also possible to target the treatment mechanism [1]. If both quantities are targeted, it is necessary to iterate the algorithm k times to achieve convergence. Let $k = 0$, and let $g^0(A|W)$ and $Q^0(A, W)$ be given and defined as above. The $k-$step algorithm that targets the marginal causal effect, including the treatment mechanism is [1]:

[1]Note that $g^0(1|W)$ is the propensity score [9]

1. For each individual, define two constants h_1^k and h_2^k, where:

$$h_1^k = h_1(g^k, Q^k)(A, W) = \left(\frac{I(A=1)}{g^k(1|W)} - \frac{I(A=0)}{g^k(0|W)} \right) \sigma(Q^k)^2(A, W) \tag{B.5}$$

$$h_2^k = h_2(g^k, Q^k)(W) = \frac{Q^k(1|W)}{g^k(1|W)} - \frac{Q^k(0|W)}{g^k(0|W)} \tag{B.6}$$

2. Let $m^k(W) = \log(g^k(1|W)/g^k(0|W))$ so that $g^k(1|W) = 1/(1 + \exp(-m^k(W)))$.

3. Specify a univariate logistic regression model:

$$g^k(\epsilon_2)(1|W) = \frac{1}{1 + \exp(-m^k(W) - \epsilon_2(k) h_2^k(W))} \tag{B.7}$$

   where the term $-m^k(W)$ is entered as a constant, and $\epsilon_2(k)$ is a coefficient on $h_2^k(W)$ and $\hat{\epsilon}_2(k)$ is its maximum likelihood estimate.

4. Regress Y on h_1^k in a univariate regression using $Q^k(A, W)$ as an offset.
   Let $\epsilon_1(k)$ be the coefficient on h_1^k and let $\hat{\epsilon}_1(k)$ be its maximum likelihood estimate.

5. Update g^k and Q^k as follows:

$$\begin{aligned} Q^{k+1}(A, W) &= Q^k(A, W) + \hat{\epsilon}_1(k) h_1^k(A, W) && \text{(B.8)} \\ m^{k+1}(W) &= m^k(W) + \hat{\epsilon}_2(k) h_2^k(W) && \text{(B.9)} \\ g^{k+1}(A|W) &= \frac{1}{1 + \exp(-m^{k+1}(W))} && \text{(B.10)} \end{aligned}$$

6. Set $k = k + 1$ and iterate until $\hat{\epsilon}_1(k)$ and $\hat{\epsilon}_2(k)$ converge to zero.

After convergence, evaluate the updated regression at $A = 1$ and $A = 0$ to get two predicted outcomes for each individual. Take the empirical mean of the difference across the population to obtain a targeted estimate of the difference, as in equation (B.4), with Q^1 replaced by Q^k.

Bibliography

[1] van der Laan M, Rubin D. Targeted Maximum Likelihood Learning. Int J Biostatistics. 2006;2(1):1–38.

[2] van der Laan MJ. The Construction and Analysis of Adaptive Group Sequential Designs; 2008. Available from: `http://www.bepress.com/ucbbiostat/paper232`.

[3] Splawa-Neyman J, Dabrowska DM, Speed TP. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. Statistical Science. 1990;5(4):465–472.

[4] Holland PW. Statistics and Causal Inference. Journal of the American Statistical Association. 1986;81(396):945–960.

[5] Rubin DB. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. Journal of Educational Psychology. 1974;66(5):688–701.

[6] Freedman DA. Statistical models for causation - What inferential leverage do they provide? Evaluation Review. 2006;30(6):691–713.

[7] Neugebauer R, van der Laan M. Why prefer double robust estimators in causal inference? Journal of Statistical Planning and Inference. 2005;129(1-2):405–426.

[8] Bembom O, Petersen ML, Rhee SY, Fessel WJ, Sinisi SE, Shafer RW, et al. Biomarker discovery using targeted maximum-likelihood estimation: application to the treatment of antiretroviral-resistant HIV infection. Stat Med. 2009 Jan;28(1):152–172.

[9] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55.