Evaluating Health Interventions Over Time: Empirical Tests of the Validity of the Single Interrupted Time Series Design

Evaluating Health Interventions Over Time: Empirical Tests of the Validity of the Single Interrupted Time Series Design

A dissertation presented

by

Theodore Svoronos

to

The Harvard Committee on Higher Degrees in Health Policy

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Health Policy (Evaluative Sciences and Statistics)

Harvard University

Cambridge, Massachusetts

August 2016

Dissertation Advisors:
Jessica Cohen (Chair)
Katherine Baicker
Dan Levy

Author:
Theodore Svoronos

Evaluating Health Interventions Over Time:
Empirical Tests of the Validity of the Single Interrupted Time Series Design

Abstract

Single interrupted time series (ITS) is a quasi-experimental evaluation design used frequently in the health policy literature. This manuscript investigates the validity of single ITS through two within-study comparisons (WSCs), comparing the results of a randomized controlled trial (RCT) with the results that would have been obtained had a single ITS design been employed.

In Part 1, I discuss the theory underlying both within-study comparisons and single ITS. I propose an assessment framework to determine whether results from a given design should be deemed "concordant" with an RCT for a given intervention. This framework aims to unify metrics for concordance used in the existing literature, and considers both practical and statistical significance. After summarizing best practices of single ITS analysis, I propose two falsification tests to determine whether the single ITS design is well suited for the trend stability of a particular dataset. These tests draws from literature on determining structural breaks in time series data, as well as work on the optimal binning of data in the regression discontinuity design.

In Part 2, I conduct two within-study comparisons for single ITS. The first study evaluates a behavior change campaign in Uganda aimed at increasing uptake of rapid diagnostic tests for malaria. The WSC finds that single ITS estimates are highly concordant with that of the RCT, producing almost identical results in both point estimate and standard error. This result is robust to multiple specifications. The second study evaluates the effect of the expansion of Medicaid on emergency department use in Oregon. In this case, the single ITS estimates are so discordant

with the RCT as to produce statistically significant results in the wrong direction. This result is also robust to multiple specification decisions.

In comparing these differing results, I note important differences between the two datasets. The Uganda data passed the falsification tests for trend stability proposed in Part 1, while the Oregon data failed. Additionally, the Oregon sample is likely subject to a manifestation of self-selection known as "Ashenfelter's dip," whereas the Uganda sample is not. The implication of this shift in outcomes just before the intervention's introduction is especially damaging to single ITS, in comparison to traditionally "weaker" pre-post designs.

In Part 3, I attempt to generate hypotheses as to when single ITS should and should not be used. First, samples defined by self-selection are particularly problematic for single ITS analysis. Second, the advantages of relying on time trends must be weighed against the additional strong assumptions that the single ITS design carries with it. Third, trend stability in the pre period is a crucial factor in getting reliable estimates from single ITS. Fourth, the robustness of results in both WSCs suggests that whether to evaluate a given program with a single ITS design is a more important decision than how to implement ITS.

Contents

List of Tables

List of Figures

Acknowledgments

First and foremost, I want to thank my dissertation committee for their thoughtfulness, patience, and encouragement. Jessica: your work on program evaluation is what made me want to come to Harvard, and I'm grateful to have grown from your mentorship these past five years. Kate: your ability to get at the heart of complex problems has had an enormous impact on my thinking; I feel very lucky to have you on my committee. Dan: I hope to one day have a fraction of the wisdom and energy you bring to both research and teaching. To all three of you: thank you.

I am also grateful to Tom Cook, Atle Fretheim, Steve Soumerai, Matthew Sweeney, and Fang Zhang for their expertise on interrupted time series and within-study comparisons. Joe Newhouse, Kathy Swartz, Alan Zaslavsky, and Debbie Whitney were major sources of support throughout my PhD, and especially during my dissertation.

I would have been unable to finish this PhD had it not been for the intelligence and friendship of Hannah Neprash, the constant support of Spencer Robins, and the insights of Ben Lockwood. To Liana Rosenkrantz Woskie: I have learned so much from you already, and learn more every day.

Finally, thank you to my family: Alex Reynolds, Dean Reynolds, Paris Svoronos, and Soraya Svoronos. I love you very much, and owe you in more ways than I can count.

Part I

Theory and Motivation

Chapter 1

Introduction

The randomized controlled trial (RCT) is widely regarded as the gold standard for assessing the efficacy of health interventions. Randomization continues to provide researchers with the closest approximation to a true counterfactual, and experimental designs have been given an increasingly public role in discourse on health [28, 29]. Yet randomized designs can be complex, expensive, unethical, or simply impossible to implement in many contexts [25, 32]. This is especially true for large-scale social welfare programs, such as those aimed at increasing access to health services, where opportunities to randomize are few and far between. This presents a significant challenge to policymakers interested in evidence-based decision making.

Quasi-experimental designs present an alternative [2, 49, 92]. These designs promise flexibility in the design of evaluations, making them appealing to practitioners. By loosening requirements on control groups and randomization schemes, these designs allow for evaluations of programs that would otherwise be infeasible with a strict RCT design. Yet these designs are only as useful as their ability to produce reliable estimates of program impacts that are free of bias. These designs rely heavily on untestable assumptions and, as a result, may be especially sensitive to even minor violations of them [26, 46, 61, 86].

This manuscript tests the validity and robustness of the single interrupted time series (ITS) design, currently regarded as an especially strong study design [14, 76, 84]. I first outline the assumptions required for a single ITS design to provide unbiased estimates of program impact.

Then, in Part II, I subject these assumptions to empirical tests via two within-study comparisons (WSCs) of large scale randomized trials in the health sector. Finally, in Part III, I extract lessons from the results of the two WSCs to provide guidance to those considering the use of ITS in their research.

Chapter 2

Within-study comparisons

2.1 Background

2.1.1 Rationale

Within-study comparisons (WSCs) emerged out of an explosion of program evaluations funded by governments and non-profit organizations starting in the 1970s [15]. The vast array of programs, contexts, and populations involved were often not amenable to randomized designs, leading to an interest in the internal validity of alternative strategies. Debates over which quasi-experimental study designs (if any) could reliably produce unbiased estimates of program impact resulted in a proliferation of studies wherein evaluation designs themselves were the subject of inquiry.

While statisticians and econometricians have outlined the conditions under which a given design performs well [84], WSCs aim to empirically test whether or not these requirements hold in practice. By quantifying the magnitude of bias introduced by a particular threat to internal validity, WSCs illustrate what would have happened if a program had been evaluated with an alternative design. This allows researchers to determine whether the degree of bias introduced would have led to a different conclusion along a policy-relevant margin [44].

2.1.2 Structure

Figure 2.1 outlines the structure of a typical WSC. In contrast to meta-analyses and "between-study comparisons" [46], which compare the results of different studies, WSCs consist of conducting two separate analyses of the same data. This has traditionally involved the following steps (note that these steps are not necessarily performed by the same individual):

1. Conduct a randomized trial;
2. Take only the "Treatment" group of the RCT, leaving the randomized "Control" group aside;
3. Generate a "Comparison" group using a quasi-experimental technique;
4. Estimate the impact on a given outcome using the RCT data (Treatment versus Control) and using the quasi-experimental data (Treatment versus Comparison);
5. Compare the impacts using some metric for concordance between designs.

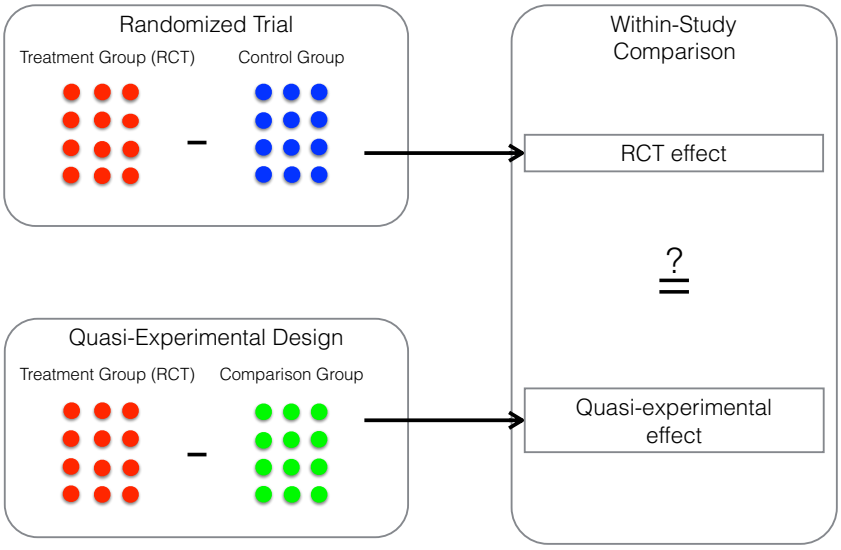

Figure 2.1: *Structure of a within-study comparison*

While this description captures the essence of WSCs in broad strokes, there is wide variation in the details of each step in the literature. For example, the nature of the "Comparison" group in step three depends largely on the quasi-experimental approach being used. Also, each analysis in step four is ideally conducted by separate groups without knowledge of the other's results [23], but this is complicated somewhat by the timing of each analysis and available resources. Some studies do away with the treatment group altogether, and simply compare the outcomes of a randomized control group with a quasi-experimentally determined comparison group [15]. The comparison in step five can also take many forms. Single interrupted time series, for example, relies on an extrapolation of pre intervention trends to construct a counterfactual rather than relying on a separate sample for its comparison group. This will be discussed in more detail in Chapter 3.

2.2 Within-study comparisons in the social science literature

This section summarizes the WSC literature to date. After outlining the evaluation designs that have been studied and how well they have fared, I extract lessons that will inform the WSCs conducted in this manuscript.

2.2.1 History

The first, and perhaps most influential, instance of a WSC is LaLonde's 1986 study of the National Supported Work Demonstration, an experiment that measured the impact of a randomly allocated training program on earnings [61]. In it, LaLonde took the study's experimental treatment group and compared it to two non-experimental control groups drawn from national surveys. LaLonde controlled for age, education, and race, and found large differences in estimated effects. These differences were mitigated somewhat by controlling for pre-treatment earnings, by limiting the sample to female participants, and by using two-step Heckman selection models. The general conclusion of the paper, however, was that econometric methods to control for bias were insufficient substitutes for RCTs.

LaLonde's study was not without its critics. Its detractors pointed out that the comparison

group was drawn from national surveys, not from the same locality of the treatment group [23]. Subsequent studies reanalyzed LaLonde's results using updated methodology. Dehejia and Wahba (1999) used propensity score matching (PSM) to generate a matched comparison group, which brought results much closer to experimental estimates [26]. However, a later study by Smith and Todd found that these results were very sensitive to researcher decisions, including the set of covariates used to generate the propensity score and the sample from which the matched sample was drawn [86], a finding consistent with a simulation of misspecified propensity score models [27].

Many WSCs have been conducted since LaLonde's original study, focusing on various quasi-experimental techniques. Matching methods, and propensity scores in particular, have been the focus of a large number of these studies. Some WSCs explicitly compare matched groups from increasingly dissimilar populations [15], while others focus on the relative effectiveness of different types of matching methods [17, 37, 38]. Some researchers have gone so far as to prospectively randomize a sample between experimental and non-experimental studies, in order to control for other design elements that could confound the comparison of study designs [79, 85]. While these studies came to differing conclusions regarding the utility of these methods [12, 39, 47, 75], a meta-analysis by Glazerman, Levy, and Myers concluded that several quasi-experimental methods were unreliable due to unpredictability in both magnitude and direction of bias.

In contrast to the performance of matching methods, WSCs on the regression discontinuity (RD) design have found it to be a promising alternative to traditional randomized designs. Cook, Shadish, and Wang (2008) discuss three WSCs involving RD designs [24]. One, conducted by Aiken et al. (1998), compared a randomized trial and RD design in the context of a remedial writing program at a large state university [3]. The authors found that randomized and RD results were quite close to one another, in terms of both the direction and size of most effects. However, they point out that the study had several elements that encouraged this correlation. For example, the writing program was already a requirement for all students who fell below a certain threshold, thus eliminating a large portion of selection issues. Additionally, the program intrinsically collected a large amount of pre-study data, as the university registrar kept

the students' performance and prior coursework. As such, the authors were able to control for numerous confounding variables from the outset.

A second WSC of the RD design focused on an evaluation of PROGRESA, a conditional cash transfer program in Mexico that offered incentives to ensure student attendance in school, may be more instructive [19]. This study had some issues in comparability between designs. For example, the RD component included data collected at different times from the randomized component. Though this had the potential to bias its results, the effect sizes measured in both experimental and RD contexts were quite similar. These findings were robust, as the authors collected data along multiple strata, phases of program rollout, and outcome variables. With a few exceptions, Buddelmeyer et al. (2004) were able to reproduce experimental results using an RD design.

A third study by Black and Galdo (2005) compared experimental and RD outcomes in a job training program and, similar to the above studies, found that RD results were quite similar to experimental data [13]. The results are generally favorable to advocates of RD as a worthwhile alternative to randomized trials, but with one caveat: the results of the RD method contained large standard errors compared to the randomized method.

2.2.2 Interrupted time series

To date, there has been one published set of WSCs interrogating the interrupted time series design [40, 41]. Fretheim et al. (2014) reanalyzed cluster randomized control trials of nine interventions taking place at the health facility level across various health systems. These interventions included a clinical alerting system, computer-based decision support tools, and educational/outreach materials for patients and providers. The authors used overlapping 95% confidence intervals as their metric for whether the designs produced similar results. In all but one case, Fretheim and colleagues found that single interrupted time series yielded results that were largely concordant with randomized trial results. In the one discordant case, the addition of a non-random comparison group brought the results in line with that of the RCT. Otherwise, the single interrupted time series design was found to be reliable for the selected interventions.

This initial analysis suggests that the single interrupted time series design has the potential to replicate the results of an RCT. However, a few issues are worth noting. First, several of the ITS estimates produced confidence intervals that were much wider than the original RCT, which made it easier to meet the authors' metric for concordance. The RCT confidence interval of one dataset, for example, was (-5.5%, 6.1%), while its ITS confidence interval was (-39%, 28%). Second, the interventions used in these comparisons were from local-level interventions, which generally provided flat pre intervention trends and were not subject to the types of concurrent events discussed in Chapter 3 that have the potential to bias the design's results. In contrast, the studies in this manuscript are large-scale interventions affecting populations outside the confines of health facilities. As a result, they are likely subject to more sources of bias than these interventions.

2.3 Ideal characteristics of a within-study comparison

2.3.1 Cook's criteria

Given the variation in methods and rigor across WSCs, Cook, Shadish, and Wong (2008) proposed a set of seven criteria with which the quality of a WSC should be judged [24]:

1. **Two or more counterfactual groups.** This basic criteria stipulates that a randomized control group must be compared to some other quasi-experimentally determined comparison group.

2. **Each study estimates the same causal quantity.** This requirement speaks to the sub populations that various study designs implicitly target. For example, as part of their identification strategies, many research designs produce estimates that are applicable only to "compliers," those who receive the treatment who would not have in the absence of the intervention. Estimates of the effect of an intervention on compliers is known as the Local Average Treatment Effect (LATE) [48], and refers only to this subpopulation. If one design estimates the LATE while another estimates the treatment effect for the entire population,

a comparison is not valid. When comparing two studies, it is essential that both are estimating the same parameter for comparable populations. In addition, the two studies must be estimating an effect for the same time period, which is particularly important for the comparisons in this manuscript.

3. **Differences between control and comparison groups unrelated to outcome.** In order to isolate the effect of a different study design unconfounded by extraneous differences, it is important that the control and comparison data are measured at the same time and in the same way, and the experiences of control and comparison participants should mirror one another as closely as possible. For example, any sort of "onboarding" process for participants of an RCT has the potential to bias results if the non-experimental group did not experience a comparable process. Similarly, data collection mechanisms such as repeated surveys should be the same across the RCT and quasi-experimental design.

4. **Blind analyses between studies.** Specification decisions by researchers can have a major impact on the estimated effects of an intervention, as the debate over propensity scores illustrates. Ensuring that each analysis is conducted blindly prevents the possibility of analytic decisions being influenced by the desired outcome.

5. **Experiment meets technical adequacy criteria.** WSCs implicitly assume that the RCT result represents an unbiased estimate of the true effect of the intervention. As such, the RCT should meet all the standard requirements of a well-conducted randomized trial.

6. **Observational study meets technical adequacy criteria.** Similar to criteria 5, the quasi-experimental study should adhere closely to the "state of the art" for that study design.

7. **Comparison of estimates is sufficiently rigorous and relevant.** A sufficiently rigorous WSC uses a set of criteria to determine whether the quasi-experimental estimate is sufficiently different from the experimental one. This can come in many forms, as will be discussed in the next section.

2.3.2 Metric for concordance

Conceptual framework

Of all the variation in the literature on within-study comparisons, perhaps the most pronounced is in the criteria used to compare the designs. Most published WSCs use some kind of qualitative determination of whether a quasi-experimental design's estimates are "close enough" to experimental ones [3, 15, 17, 26, 47, 79]. Fewer studies measure mean differences in bias between quasi-experimental and experimental estimates, and compare this to a threshold considered meaningful by policymakers [27, 44, 96]. Fewer still attempt to measure the statistical significance of differences in effect sizes between designs [13, 37, 38].

Each of these metrics captures a different dimension of what it means for effect estimates to be discordant. Are the effect estimates in the same direction and magnitude? If their magnitudes differ, is the difference in estimates statistically distinguishable from zero (statistical significance)? Are the two estimates different along a "policy relevant margin" (practical significance)?

Figure 2.2 consolidates these issues into a framework to interpret WSC findings.

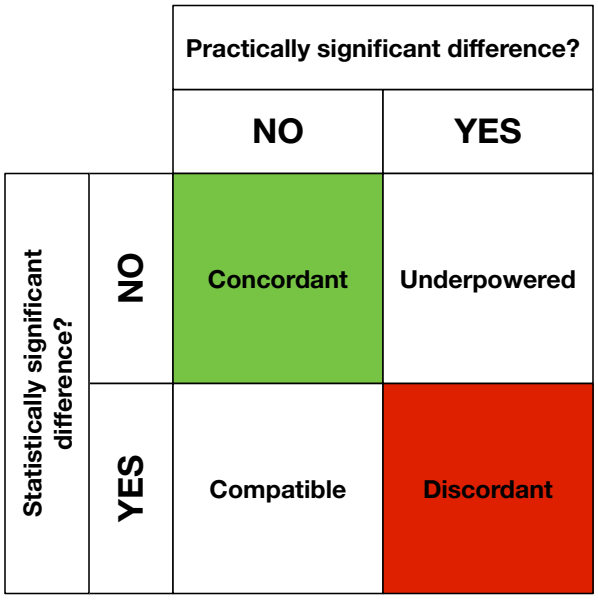

Figure 2.2: *Assessing concordance of RCT and quasi-experimental estimates*

At the extremes, interpretation is straightforward. If the difference between the RCT estimate and the quasi-experimental estimate is *neither* statistically significant nor practically significant, the conclusion of the WSC is that the two designs are concordant. That is to say, the quasi-experimental design was able to replicate the RCT result to the point that it can be considered a reliable substitute for this particular context. Note that this is not necessarily an endorsement of the quasi-experimental method more generally, but rather a data point in support of the design for a given intervention. The result should also be weighed against its sensitivity to various specifications in order to assess robustness.

Conversely, if the difference in estimates is *both* statistically and practically significant, the WSC concludes that the studies are discordant. In addition to a statistically significant difference, in this scenario the quasi-experimental design would have produced misleading results for the true impact of the program along a margin that can be considered "policy relevant". This term is difficult to determine and necessarily context-specific. In addition to differences across interventions and environments, even different policymakers might disagree over what is policy relevant for a given context. Attempts have been made to create standards against which WSC results can be compared, such as using the convention of whether a policymaker would alter support for a program [44, 96]. This convention is not ideal however, as it places too much importance on the particular resource constraints of a given policy environment.

As an alternative, I propose using the same metric that is used to weigh the importance of individual program impacts: if a program produced a change in the outcome that was the size of the difference between the estimates of the two designs, would the program be considered a success? If so, the difference between the studies is practically significant. This maps most closely to the notion of the quasi-experimental design as the "intervention" of interest, and the difference between the two as the "impact" of the intervention. If using an alternative design produces a difference that meets this criteria, it should be considered discordant.

In the top-right scenario, where there is a practically significant difference between the two designs but the difference is not statistically significant, the WSC should be considered underpowered. That is to say, the variance in outcomes between the two designs was large enough that the WSC could not detect the measured impact as significant. Traditionally, underpowered is

used to describe an instance of Type II error; the design was unable to detect an effect that is real. Applying this terminology in this way places the concept of practical significance in the privileged position of determining whether a difference is in fact "real". Given that WSCs aim to determine whether a given design would have produced misleading results, this characterization seems appropriate.

Finally, if the WSC finds a statistically significant difference that is not practically significant, it does not provide us with much information regarding the quasi-experimental design. On one hand, the fact that the two designs yielded estimates close enough to result in the same policy outcome suggests that the quasi-experimental design was successful at approximating the RCT. On the other hand, the fact that the WSC was able to detect a statistically significant difference might seem to call the reliability of the quasi-experimental design into question. However, the statistical significance may also be the result of using a design that produces highly precise estimates, in which case it would be penalized for having a characteristic considered desirable. This framework uses the term "compatible" to describe a WSC producing a statistically significant difference that is not practically significant. This result presents evidence, albeit not as conclusive as a fully concordant scenario, that the quasi-experimental design is able to replicate the results of an RCT to some extent.

Computing standard errors

In order to determine whether differences between two designs are statistically significant, I will use a bootstrapping method used in some of the more rigorous WSCs to date [13, 37, 38, 86]. The procedure for generating a standard error estimate for the difference in estimated impacts of a WSC is as follows:

1. Draw a sample from the RCT dataset (with replacement) of equal size to the original RCT dataset;
2. Estimate the treatment effect of the bootstrapped RCT dataset;
3. Draw a sample from the quasi-experimental dataset (with replacement) of equal size to the original quasi-experimental dataset;

4. Estimate the treatment effect of the bootstrapped quasi-experimental dataset;
5. Take the difference of these two estimates;
6. Repeat steps 1-5 1,000 times;
7. Take the standard deviation of the 1,000 differences, to serve as the standard error of the difference in estimated effects of the original analyses.

This bootstrapped error is then used to conduct a t-test of the difference in impacts measured by the two designs.

Chapter 3

Interrupted Time Series

3.1 Theory

The focus of this manuscript is the interrupted time series (ITS) design. ITS is a widely used quasi-experimental design that relies on trends before and after the introduction of a discrete intervention to assess impact [84]. It is one of the few credible evaluation designs that is often implemented without a comparison group.

This manuscript focuses on short, single interrupted time series. The "short" qualifier means that my analysis will not rely on autoregressive integrated moving average (ARIMA) techniques to model time trends [45, 52]. An ITS design requires 100 or more time points to effectively leverage ARIMA techniques [30], which neither of these datasets have. Instead, I will focus on ordinary least squares estimation with autocorrelated errors [64], as will be described in Section 3.1.1. The "single" qualifier means that my analyses do not include a non-experimental control group [84]. Instead, counterfactuals are constructed via a projection of the trend before the intervention was introduced ("pre period") into the time period after it was introduced ("post period"). While this work can be extended to multiple ITS comparisons, the prevalence of single ITS in the health literature is the primary motivator of this analysis. Note that, for the remainder of this manuscript, ITS will be used synonymously with single ITS.

As discussed by Bloom (2003), the design is premised on two claims. First, absent some systemic

change, past experience is the best predictor of future experience. Second, using multiple observations from the past to establish a trend is a more reliable predictor of the future than a single observation [14].

3.1.1 Structure

In an ITS analysis, the unit of observation is some equally spaced time unit such as days or weeks. Data are usually collapsed to the time level, as opposed to a dataset with observations at the person-time level.[1] The standard model for an ITS design is as follows [64, 93]:

$$Y_t = \beta_0 + \beta_1 time_t + \beta_2 post_t + \beta_3 timepost_t + u_t \tag{3.1}$$

$$u_t = \rho u_{t-k} + z_t \tag{3.2}$$

where:

- *time* is a variable which equals one at the first time point t and is incremented by one for each subsequent time point;
- *post* is a dummy variable which equals one at the time immediately following the introduction of the intervention of interest (p) and for every time point thereafter;
- *timepost* is a variable which equals zero until time $p + 1$, and is incremented by one for each subsequent time point;

and, by extension:

- β_0 is the starting level of outcome Y;
- β_1 is the pre period slope;

[1] There is some work on constructing ITS estimates at the person-time level using Generalized Estimating Equations. While this has been described at a theoretical level [99], there have been almost no examples of it used in practice. This manuscript adheres to the traditional convention of time-level data.

- β_2 is the change in level at time p;
- β_3 is the change in slope in the post period.

To account for autocorrelation, the error term in Equation 3.2 is a Newey-West standard error with lag k [70].

A defining characteristic of an ITS analysis is that there is not a single coefficient that represents program impact. In Equation 3.1, β_2 and β_3 represent the immediate and subsequent effects of the intervention, respectively. This is seen by many as a strength of the design, since it allows researchers to disaggregate short term effects from longer term effects [76]. For the purpose of a WSC, however, care must be taken to generate effect estimates that are comparable to the single impact estimate of a comparison of means in a traditional RCT.

3.1.2 Assumptions

The interrupted time series design's validity rests on the following assumptions:

- **Assumption 1:** The expectation of the pre intervention level and trend would be the same irrespective of whether the sample received the treatment;
- **Assumption 2:** In the absence of the intervention, the post intervention trendline would have been equivalent in expectation to an extrapolated pre intervention trend.
- **Assumption 3:** The time trends in the pre and post periods can be expressed as a linear combination of parameters.

Let us illustrate this more formally, using the potential outcomes framework [57]. Assume Y_{t1} denotes the potential outcome for some group at time t if they receive the treatment, while Y_{t0} denotes the potential outcome for the group at time t if they do not receive the treatment. Then we can specify the following two equations using an ITS model:

$$Y_{t1} = \alpha_0 + \alpha_1 time_t + \alpha_2 post_t + \alpha_3 timepost_t + \epsilon_t \tag{3.3}$$

$$Y_{t0} = \gamma_0 + \gamma_1 time_t + \gamma_2 post_t + \gamma_3 timepost_t + \nu_t \tag{3.4}$$

In a single ITS context, we only have data to estimate Equation 3.3. We are, however, making the following implicit assumptions regarding Equation 3.4:

- **Assumption 1a:** $\alpha_0 = \gamma_0$
- **Assumption 1b:** $\alpha_1 = \gamma_1$
- **Assumption 2a:** $\gamma_2 = 0$
- **Assumption 2b:** $\gamma_3 = 0$

If the components of **Assumption 1** are met, then we have an unbiased estimate of the pre period trendline.

If the components of **Assumption 2** are met, then an extrapolation of the pre period trendline provides an unbiased estimate of post period outcomes in the absence of the intervention.

If **Assumption 3** is added, then **Assumptions 1** and **2** apply to pre period trends and extrapolations that are linear.

Taken together, **Assumptions 1 - 3** imply that *a linear extrapolation of the pre period trendline into the post period provides an unbiased representation of the counterfactual for a treated sample.*

3.2 Threats to internal validity

In discussing the threats to the internal validity of the ITS design, Shadish, Cook, and Campbell point to a number of potential threats [84]. These are discussed in the context of changes taking place at the time of an intervention's introduction, though pre period events that linked to the treatment pose a risk as well. In addition, the ITS design is particularly vulnerable to misspecification issues, known broadly as misspecifications of functional form.

3.2.1 Concurrent changes

The primary threat to an ITS design is the existence of changes that affect the outcome at the same time as the intervention's introduction p [84]. Since a single ITS design lacks a control group, any shifts in level or trend at the time of the intervention's introduction is fully attributed to the intervention itself [81]. Thus, any changes at time p other than the intervention which are related to the outcome of interest will be incorrectly attributed to the intervention.

In practice, "concurrent changes" can come in a number of forms [84]:

- **History threat**: Changes *external to the sample* such as other programs, policies, or economic changes. For example, the measured effect of a job training program on employment with single ITS would be biased if the program took place just as an economic recession began.
- **Selection threat**: Changes in the *composition of the sample* at the time of the intervention's introduction. For example, the introduction of a tax on firms may cause firms to relocate.
- **Instrumentation threat**: Changes in *measurement of the outcome* at the time of the intervention's introduction. The adoption of an electronic medical record system, for example, may require that health outcomes be recorded electronically rather than on paper. If this change makes it easier or harder for a physician to note a given condition, the ITS design may detect an effect at the intervention's introduction unrelated to the intervention's actual efficacy.

While each of these threats comes from a different source, they affect the validity of the design in the same way: by introducing a change in the data at the time of an intervention's introduction, these threats make it difficult to disentangle the true program impact from the impact of these other events.

Using the framework of Section 3.1.2, the threat of concurrent changes can be seen as a violation of **Assumption 2**: a concurrent event at the time of the intervention's introduction p implies that, even without the intervention, there would be a shift of the outcome variable in level ($\gamma_3 \neq 0$), slope ($\gamma_4 \neq 0$), or both after time p.

3.2.2 Differential pre period changes

The threat of concurrent changes at the time of an intervention's introduction is the primary focus of most ITS analyses. However, an equally important threat lies in the violation of **Assumption 1**. For example, knowledge that a cigarette tax will soon come into effect may lead to a sharp increase in cigarette sales leading up to the tax's introduction. In this scenario, a change related to the intervention taking place *during the pre period* leads to a trendline that is a poor approximation of the outcome variable's trend in the absence of the intervention.

Consider Figure 3.1. The diamonds in this figure represent data for the sample had it not received the intervention (Y_{t0}), while the squares represent data for the sample if it had received the intervention (Y_{t1}). The diamonds within squares represent points that are identical in either potential outcome.

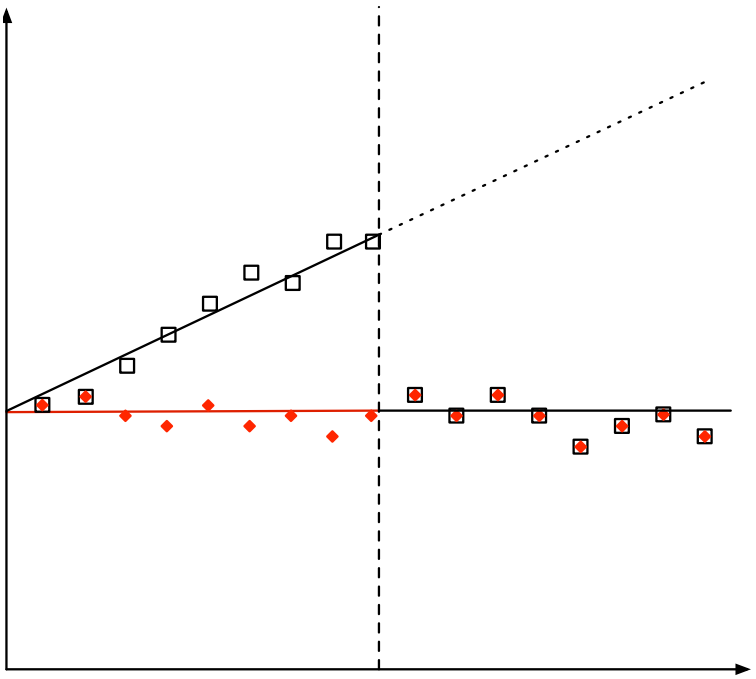

Figure 3.1: *Violation of Assumption 1*

Figure 3.1 illustrates that outcomes in the post period remain the same irrespective of whether the sample received the treatment. Additionally, **Assumption 2** holds in that the post period outcome for the sample had it not received the intervention is identical in level and slope to the

sample's pre period ($\gamma_3 = 0, \gamma_4 = 0$). However, something about the intervention induced a change in outcomes during the pre period so that, in this case, the pre period slope of the sample had it received the intervention in the post period is different from the pre period slope of the sample had it not.

The implications of this violation are clear: a pre period trend that is differential between the potential outcomes of units will lead to incorrect estimates of the change in level and slope at time p. In the case of Figure 3.1, the change in slope and level should both be zero (referring to Equation 3.3: $\alpha_2 = 0, \alpha_3 = 0$). Instead, this analysis would find a decrease in both level and slope.

A frequent cause of this phenomenon is referred to as "anticipation effects," wherein knowledge of the impending intervention leads to a change in behavior different from what would otherwise have occurred [67]. Similar issues can arise through history, selection, and instrumentation. Note that, for any of these threats to lead to bias in impact estimates, the pre period change must be somehow tied to the intervention itself. An event that does not affect potential outcomes differentially would not violate **Assumption 1**.

Ashenfelter's dip

One particular violation of **Assumption 1**, called "Ashenfelter's dip," merits further discussion. Ashenfelter's dip refers to a scenario wherein individuals self-select into a program on the basis of the outcome variable that the program aims to address [54]. This phenomenon was originally documented in the context of job training programs, wherein participant earnings appeared to decrease in the time leading up to a program's introduction [8]. This decrease in earnings was being driven by the fact that those choosing to enroll in the program were recently unemployed. As a result, those being selected into the sample were individuals who had decreasing earnings. We can illustrate the bias caused by Ashenfelter's dip in the absence of a randomized control group using an exercise adapted from Heckman and Smith (1999) [54]. Let the following represent the experimental estimate obtaining by taking the difference of a randomized treatment and control group in the post period:

$$E(Y_{1post}|D=1) - E(Y_{0post}|D=0) \tag{3.5}$$

Where $D = 1$ for individuals in the randomized treatment group, and $D = 0$ for individuals in the randomized control group. Under the assumption of random assignment, Equation 3.5 estimates the effect of the treatment on the treated. If instead we use a simple pre-post comparison

$$E(Y_{1post}|D=1) - E(Y_{1pre}|D=1) \tag{3.6}$$

which, in the presence of randomization, is equivalent to

$$E(Y_{1post}|D=1) - E(Y_{0pre}|D=0) \tag{3.7}$$

then the bias of the pre-post estimator is the difference between Equations 3.5 and 3.7

$$E(Y_{0post}|D=0) - E(Y_{0pre}|D=0) \tag{3.8}$$

In words, the bias of the pre-post estimator is the change in earnings of the control group before and after the intervention. In the event that the pre-treatment dip in earnings was transient, and strictly the product of self-selection, then this bias term would be positive. Using a simple pre-post estimator would thus lead to an overstatement of the true effect.

The risk of Ashenfelter's dip is especially strong in the context of a single interrupted time series design. Whereas a simple pre-post estimator is biased by artificially low outcomes in the pre period, a single ITS estimator is biased by an extrapolation of an artificially low trend. If the pre period values are stable and the pre period trend is zero, this would produce the same bias as the pre-post estimator. However, if the dip is especially transient, and only occurs in a few time points leading up to the intervention, the single ITS counterfactual will extrapolate the decreasing trend, thus exacerbating the bias by a large amount.[2]

Figure 3.2 illustrates this issue. Assume both scenarios represent a dip in outcomes from the

[2]A positive trend would also produce bias, but it would consist of an extrapolation that is too high, not too low.

trend prior to the start of data collection. Scenario A (circles) has a dip that is constant in the pre period, whereas Scenario B (diamonds) has a decreasing dip. Both scenarios have the same pre-treatment mean, and thus would produce the same degree of bias in a simple pre-post comparison. However, in the context of a single ITS design, Scenario B would produce much greater bias.

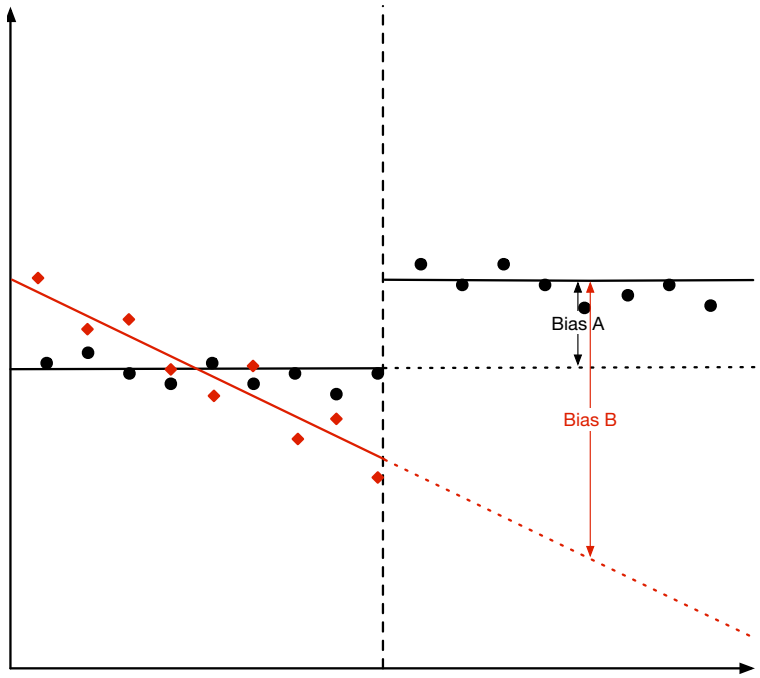

Figure 3.2: *Ashenfelter's Dip in a single interrupted time series design*

3.2.3 Misspecification of functional form

The final significant threat to the validity of ITS estimates is related to statistical specification. Since the "control" group of a simple ITS is represented by the extrapolation of pre period trends, the design relies more heavily on assumptions related to the timing of the intervention, the nature of its diffusion, and the presence of autocorrelation in the data [84]. For example, failing to account for a phased rollout of an intervention may lead to an underestimate of its effect, since untreated units at time t could be mischaracterized as treated, and vice versa. Similarly, a mischaracterization of the timing of an intervention's impact could lead to an overstatement or understatement of effect. On the other hand, failing to account for autocorrelation in the data

may lead to artificially low standard errors, increasing the likelihood of Type I error. While these risks are present in other study designs, the granularity of time in the data make ITS especially vulnerable.

Note that these issues do not necessarily violate **Assumption 3**. For example, the true relationship between the outcome variable and time can contain quadratic or interaction terms and still be linear in parameters. However, if an ITS model does not account for these realities in the data, it will lead to biased estimates.

The number and importance of specification decisions in the ITS design provide a great deal of discretion to the researcher. As such, it is important that any within-study comparison of ITS scrutinizes these decisions and the extent to which results are sensitive to them. A failure to do so may overstate its robustness as a quasi-experimental design [26, 86].

3.3 Single interrupted time series in practice

3.3.1 Use of single interrupted time series in health policy literature

The previous sections outlined the general structure, characteristics, and risks of the single ITS design. This section focuses on its varied use in the health policy literature, both in terms of intervention types and statistical specifications. I then propose a set of characteristics for ITS that represents the "best practice" in the literature, which will be the approach used in the subsequent within-study comparisons.

Broadly, the interventions studied using the ITS design tend to fall into one of two groups. The first, hereafter referred to as "local-level" interventions, involve introducing a change in the management of a health facility or group of facilities, often aimed at improving healthcare quality. Example of facility interventions include interventions to alter prescribing behavior [18, 33, 91], reduce antimicrobial resistance [7, 31, 66], improve patient adherence to medication [16], reduce readmissions [62], and improve referral behavior [51].

The second type of intervention, hereafter referred to as "population-level" interventions, aim to assess the impact of large-scale policy change on a given population. These interventions include

an effort to reduce perinatal mortality at a state level [43], a national change in pharmaceutical reimbursement schedules [4, 87], a statewide excise tax imposed on cigarette sales [65], and a national pay for performance program [83]. Single ITS designs are particularly attractive to evaluate population-level interventions, since a reasonable control group is often not possible. However, these interventions may be more problematic for ITS than local-level ones. As noted in Section 2.2.2, localpolic-level interventions may be less subject to the kind of concurrent events that the single ITS design is especially vulnerable to. In scenarios where there are additional events affecting the outcome throughout the study period, it is perhaps easier for a researcher to identify them and account for them in the analysis. Population-level interventions, on the other hand, take place in a much less controlled environment, and concurrent influential events may be more difficult to identify and account for.

In addition to these differences in scope, the health literature utilizing the ITS design lacks a consensus as to what elements are required for a strong ITS analysis. A systematic review by Jandoc et al. [59] that focused exclusively on drug utilization research is instructive. The authors compared 220 drug utilization studies employing ITS against a common rubric of characteristics that define ITS applications. They find commonalities along basic metrics, such as clearly defining a time point (84.5%) and using graphical figures to display results (83.6%). But these commonalities cease when going deeper than these superficial characteristics. For example, only 66.4% of studies attempted to account for autocorrelation, a fundamental issue in ITS designs as described above. There was also wide variation in the use of lag periods (27.7%), seasonality (30.6%), and sensitivity analyses (20.5%).

3.3.2 Current best practice in ITS analysis

Given the inconsistent way the ITS design is used in practice, it is necessary to define a standard against which other study designs can be compared. The characteristics presented in Figure 3.3 represent my assessment of best practices of short, single ITS currently found in the health policy literature. Note that the "short" and "single" descriptors exclude ARIMA modeling and a comparison group; this standard therefore does not represent the most robust version of ITS given unlimited data. That said, the vast majority of ITS designs found in the literature are of

this abbreviated form.

Sections 1-3 of Figure 3.3 present characteristics of many high quality short single ITS studies currently in the literature [59, 93]. Note that many of these requirements represent one of many suggested best practices. For example, the only other WSCs on the interrupted time series design use the requirement of six data points in each period, as opposed to twelve [41]. In these cases, I list the requirement most frequently referenced in the literature.

Figure 3.3: *Best practice of the short, single interrupted time series design*

1. **Data** [6, 93]
    - $\geq$ 12 time points in each period ("pre" and "post")
    - $\geq$ 100 observations/units per period
    - Data collapsed to the time level
2. **Statistical Analysis** [30, 64, 81]
    - Model using segmented regression analysis
    - Test for and account for autocorrelation in error term
    - Account for seasonality via dummy variables and/or lag in error term
    - Control for time-varying covariates that could potentially affect outcome
3. **Sensitivity analyses** [84, 93]
    - Allow for lag/transition period if intervention or context requires it
    - Assess sensitivity of results to changes in intervention point, changes in functional form, and the addition of covariates
    - Consider adding nonlinear terms

In addition to these characteristics found in the ITS literature, the next section suggests two falsification tests not currently employed in ITS analyses. This suggestion draws from broader

work on time series analysis, as well as a technique from the regression discontinuity design, which shares many structural similarities to ITS [55, 56].

3.3.3 Proposed falsification tests

The ITS design involves fitting a rigid structure onto time data, wherein the regression is permitted to diverge from a straight line only at a specific point and in a specific way. The risk of "forcing" the data into this structure is therefore high; it is possible that the researcher will ignore other potential break points in the data, or impose an artificial break point where there is none. To address these risks, I propose the following two procedures, drawing from techniques in both time series analysis and the regression discontinuity literature.

The first falsification test involves conducting a search for "data-driven" structural breaks in the data using a test for an unknown break point [5]. This involves taking the maximum value of the test statistic obtained from a series of Wald tests over a range of potential break points in the data [78]. The test can be represented formally as follows [88]:

$$\text{supremum } S_T = \sup_{b_1 \leq b \leq b_2} S_T(b) \tag{3.9}$$

Where $S_T(b)$ is the Wald test statistic testing the hypothesis $H_0 : \delta = 0$ at potential break point b:

$$y_t = x_t + (b \leq t)x_t\delta + \epsilon_t \tag{3.10}$$

I conduct this test using the `estat sbsingle` command in Stata 14 [88]. The purpose of this test is to determine whether there is a sufficient break in the data to be detected and, if there is, whether it corresponds to the theorized break point in the ITS design.

The second falsification test attempts to characterize the amount of variability across time points in the dataset. It is implemented as follows:

1. Generate a set of bins using an optimal bin width algorithm from the regression discontinuity literature [58]. Since these bins are generated to smooth the plot of data against a

running variable in order to better discern break points, the edges of these bins represent potential candidates for structural breaks in the data. I determine these bins using the `rdplot` command in Stata [20]. The purpose of this step is to create a set of theorized break points to test in the subsequent step.

2. Test for the presence of a structural break at the meeting point between adjacent bins using a Chow test, a variant of a Wald test and a technique common in the time series literature [21]. Briefly, a Chow test tests for a known break point by fitting a regression with a dummy variable equalling one for every time point after the theorized break. I conduct this test using the `estat sbknown` command in Stata 14 [88]. The frequency of statistically significant p-values across the potential break points provides a picture of the underlying variability in the data.

While the results of these two procedures are not conclusive, they provide insight into the underlying data, the intervention, and the appropriateness of the single ITS design. If the first test for a data-driven structural break is unable to identify a break in the data, it suggests that the intervention of interest did not lead to enough of a break in trend to be detectable by ITS. Thus, a failure to reject the null of no break point provides evidence that the intervention had no effect.

If instead the first test detects a structural break at a point other than the intervention point, it suggests that some outcome-influencing event took place during the study period, and that the impact of this event exceeds that of the intervention of interest. This could mean several things: perhaps the other event had an especially large effect on the outcome, or perhaps the impact of the intervention of interest is quite small. Regardless of the cause, there are two potential implications for detecting a break other than the intervention point. First, it would be prudent to allow for a second break at this point using multiple segments, in order to account for its influence. However, this may not be sufficient to account for it, particularly if it violates ITS **Assumption 1**. In this case the presence of a major event - more significant than the intervention itself - suggests that ITS may not be the most appropriate study design for these data.

Concerns about the appropriateness of ITS would be further confirmed by the presence of multiple, statistically significant break points detected in the second falsification test. Detecting multi-

ple break points suggests that the underlying variation in the data - driven by external events or simply the result of "noise" - may simply be too great for a single ITS design to perform reliably. This is especially concerning in the pre period, the trend of which is the sole determinant of the modeled counterfactual. Since the internal validity of the single ITS design relies so heavily on the nature of the time trend, instability in the underlying trend calls the projected counterfactual into question.

3.3.4 Framing this manuscript's studies

Figure 3.3 is the standard of ITS analysis that will be used in the within-study comparisons detailed in this manuscript. Adhering to these requirements ensures that the following analyses represent a "state of the art" version of ITS to be compared against the "gold standard" of RCTs. Both ITS studies meet the data requirements in section 1, and the analyses I conduct use all the tools outlined in section 2. Adherence to sections 3 varies somewhat depending on the nature of each intervention and the characteristics of the data in each study. For example, the intervention described in Chapter 5 lends itself to testing for a transition period, whereas the intervention in Chapter 4 does not. That said, I conduct the falsification test proposed in Section 3.3.3 for both studies.

One shortcoming of the following analyses is that I will not be blind to the RCT results, which has the potential to influence my specification decisions. Still, I have attempted to conduct a set of analyses analogous to what a researcher conducting an ITS analysis would reasonably do. Each WSC in this manuscript has a section justifying the methodological decisions made in constructing the ITS models. While this is not nearly as reliable as having the researcher be blind to the original study results, my hope is that making my decisions transparent will help to address the possibility of researcher bias.

Part II

Empirical Tests of the Interrupted Time Series Design

Chapter 4

Effect of a Behavior Change Campaign on Uptake of Rapid Diagnostic Tests for Malaria: Uganda

4.1 Background

The first within-study comparison of the interrupted time series design focuses on a behavior change campaign (BCC) in Uganda aimed at increasing the use of rapid diagnostic tests (RDTs) for malaria. RDTs aim to improve management of febrile illnesses in malaria-endemic countries by allowing individuals with basic training to effectively diagnose malaria cases [22]. RDTs are a simple and inexpensive alternative to microscopy [95], a technique which requires substantially more equipment and training to administer correctly [50]. Quick and accurate diagnosis of malaria helps to ensure that cases of febrile illness receive appropriate treatment [77]. A frequent response to febrile illness in malaria-endemic areas is to presume that the cause of fever is malaria and prescribe antimalarial medication, which is ineffective against other causes of fever. Reducing the cost of testing is an important step in preventing the presumptive treatment of malaria-like symptoms with antimalarials, which can lead to worse health outcomes, waste of scarce resources, and increased risk of parasital resistance to antimalarials [60, 68, 72, 89].

The RCT in question consists of two independently randomized interventions that were introduced six months apart from one another. The first intervention involved training vendors from licensed drugs shops to test patients with RDTs. They were also given the option to purchase RDTs at a subsidized price from wholesale vendors [22]. The aim of this intervention was to increase the availability of RDTs in the study area, while also ensuring that drug shops were adequately trained in their use. The second intervention, a behavior change campaign (BCC), involved community dialog meetings aimed at sensitizing communities to the benefits of RDTs over presumptive treatment of malaria.

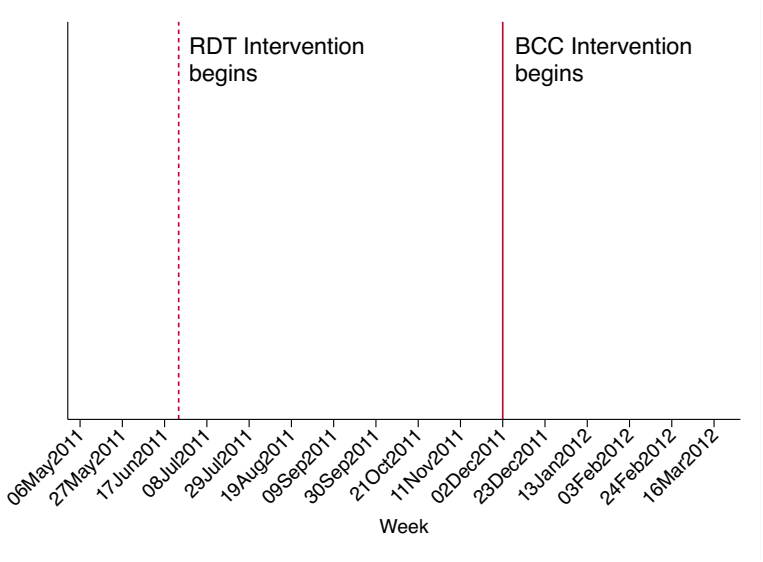

Figure 4.1: *Timeline of studies in Uganda*

This within-study comparison focuses on replicating the results of the *second* randomized trial - the BCC intervention to promote uptake of RDTs for febrile illness episodes. The existence of the *first* RCT, the intervention for which was introduced six months prior to the BCC intervention, provides sufficient pre period time points to conduct an ITS analysis of the BCC intervention (Figure 4.1). For the remainder of this chapter, "the study period" refers to the time period that includes both RCTs, and "the intervention" refers to the randomized trial of the BCC intervention.

4.1.1 Context

The study took place between March 2011 and April 2012 in the mid-Eastern region of Uganda, a country where malaria accounts for 30-50% of outpatient visits and 9-14% of inpatient deaths [80]. Presumptive treatment of malaria based on symptoms is common despite Ugandan Ministry of Health guidelines that emphasize parasitological confirmation of suspected cases [10]. This is especially true outside of higher level public health facilities, and many Ugandans seek malaria treatment in private sector drug shops [69, 71, 74].

The study was designed as a cluster-randomized controlled trial involving 92 villages across six districts: Budaka, Kibuku, Pallisa, Kumi, Ngora, and Bukedea. In June and early July 2011, the RDT intervention was introduced [22]. At this time, 67 of the 92 villages were randomly selected into the treatment group. All the drug shops in these villages received training in administering RDTs and were given access to subsidized rapid diagnostic tests (RDTs) sold through local wholesale providers. 12 villages were randomized into the RDT control group, while 13 villages did not have an eligible drug shop. The behavior change campaign (BCC) intervention - the focus of this within-study comparison - was introduced in early December 2011. The number of villages in each arm is presented in Table 4.1.

Table 4.1: *Villages in each arm of 2x2 design*

		RDT Intervention (June 2011)			
		Treatment	Control	Ineligible	Total
BCC Intervention	Treatment	34	5	7	46
(December 2011)	Control	33	7	6	46
	Total	67	12	13	92

Of the twelve villages in the control group of the RDT intervention, five villages were also randomized into the treatment group of the BCC intervention. For these villages, it should be noted that malaria testing was still available at public facilities and many private providers. While access was somewhat higher in the RDT intervention group than in the control group, the BCC intervention still had the potential to affect testing behavior in both the treatment and the control arms of the RDT trial.

4.1.2 Intervention

Between December 1 and December 9, 2011, two community dialogue meetings took place in each treated village, one for men and one for women. The meetings lasted approximately two hours, and targeted the primary female household member in charge of health care decisions (for the women's meeting), and the male household head and local leaders (for the men's meeting). Meetings were coordinated and conducted by the Uganda Health Marketing Group (UHMG).

The objectives of the meetings were: a) to sensitize the community members to the benefits of getting proper testing with RDTs; and b) to encourage community members to seek early malaria testing before treatment. A total of 16 facilitators were trained to lead meetings by addressing 14 talking points related to these objectives. Meetings began with a discussion of common causes of illness in the community, with a particular focus on malaria. Other diseases with similar symptoms to malaria were then discussed, to segue into a discussion on the risks of presumptive treatment. Facilitators then presented revised guidelines by the Ugandan Ministry of Health and the World Health Organization. Meeting attendees were then introduced to the concept of RDTs, their advantages, and the procedure wherein individuals can get tested before treatment. This was reinforced using visual aids and a live demonstration of RDT testing.

4.2 Randomized controlled trial

4.2.1 Data

For each village in the study, 30 households were randomly selected and monitored throughout the study period. A baseline survey on demographic information was administered to all sampled households, followed by monthly follow up visits and an endline survey to those same households. If the household reported any health problems, a treatment-seeking module collected information on the nature of the health problem, type of health service used to address it, whether blood-test-based diagnosis was conducted, and what medications were taken.

The sample was limited to cases of febrile illness, in accordance with existing literature on malaria testing and treatment. The dataset consists of 25,358 cases of febrile illness, encompassing 10,445

individuals across 2,347 households.

The primary outcome of interest is the likelihood that a given case of febrile illness resulted in a malaria test. Secondary outcomes include, for a given case of febrile illness, where treatment (if any) was sought, and what medication (if any) was ultimately taken.[1]

4.2.2 Methods

I estimate the effect of the BCC campaign using the following linear probability model:

$$Pr(tested_{ijt} = 1) = \beta_0 + \beta_1 post_{it} + \beta_2 treatment * post_{jt} + \alpha_t + \gamma_j + \epsilon_{ijt} \tag{4.1}$$

where

- $tested_{ijt}$ is a dummy variable which equals one if the febrile illness episode i which began at time t in village j resulted in the individual receiving a malaria test;
- $post_{it}$ is a dummy variable which equals one if the start of febrile illness episode i at time t is after the introduction of the BCC intervention;
- $treatment * post_{jt}$ is a dummy variable which equals one if village j has received the BCC intervention by time t;
- α_t are fixed effects for ten survey rounds;
- γ_j are village fixed effects, which subsume a dummy variable for being in the treatment group (hence its exclusion from the model).

β_2 captures the effect of the BCC intervention in the post period. I use a linear probability model (as opposed to logit or probit models) to simplify interpretation and facilitate comparison to the ITS design in Section 4.3.

[1] Results on these outcomes can be found in Appendix A.

4.2.3 Results

Balance

Baseline values for demographics and behavior related to treatment seeking, testing behavior, and medication taking are presented in Table 4.2. The table shows statistically insignificant baseline differences along all variables between the treatment and control arms of the BCC intervention, which is consistent with a successful randomization process. The only exception to this baseline equivalence is the percent of cases involving taking an antimalarial wherein the globally recommended treatment, Artemisinin-based Combination Therapy (ACT), was taken. This difference, though statistically significant, is not a practically significant margin, and is the only example of a statistically significant difference in the baseline measures.

Table 4.2: *Balance table - febrile illness episodes*

	Treatment	Control	Treatment vs. Control (p-value)
Demographics			
Under 5 years old	0.339	0.349	0.535
	(0.015)	(0.011)	
Female	0.550	0.557	0.492
	(0.011)	(0.008)	
Treatment seeking			
Ever visited public facility (%)	0.270	0.258	0.671
	(0.029)	(0.020)	
Ever visited private facility (%)	0.155	0.177	0.411
	(0.026)	(0.019)	
Ever visited any drugshop or pharmacy (%)	0.440	0.417	0.533
	(0.037)	(0.028)	
Ever visited program drug shop (%)	0.179	0.169	0.777
	(0.033)	(0.023)	
Ever sought any care (%)	0.769	0.743	0.152
	(0.018)	(0.014)	
Malaria test received if visited . . . [a]			
Public facility (%)	0.512	0.560	0.402
	(0.057)	(0.041)	
Private facility (%)	0.342	0.438	0.183
	(0.072)	(0.055)	
Any drugshop or pharmacy (%)	0.096	0.074	0.349
	(0.024)	(0.016)	
Program drugshop (%)	0.127	0.109	0.579
	(0.033)	(0.023)	
Medication taking			
Took ACT (%)	0.298	0.319	0.396
	(0.024)	(0.018)	
Took ACT (% Among those taking an antimalarial)	0.538	0.595	0.046**
	(0.028)	(0.020)	
Took any antimalarial (%)	0.554	0.536	0.583
	(0.033)	(0.024)	
Took any antibiotic (%)	0.279	0.258	0.191
	(0.016)	(0.010)	
Number of households	1,172	1,152	
Number of observations	10,015	9,460	

Standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1
All regressions include village and survey period fixed effects.
[a] Denominator represents the number of people that visited the facility.

Impact Estimates

Table 4.3 presents impact estimates for the BCC intervention, expressed as the change in percentage points of the probability that a given febrile illness episode resulted in a malaria test.

Table 4.3: *RCT impact estimates (Uganda)*

	(1)	(2)
VARIABLES	Full Sample	Under 5
Post	-0.0157	-0.0119
	(0.0169)	(0.0282)
Treatment x Post	0.0200	0.0197
	(0.0147)	(0.0216)
Constant	0.257***	0.322***
	(0.0116)	(0.0186)
Observations	25,207	8,990
Control Mean	0.230	0.246

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Outcome variable is the likelihood of receiving a malaria test
Regression includes fixed effects for villages and survey rounds
Standard errors are clustered at the village level

The RCT is unable to detect a statistically significant increase in the likelihood of receiving a malaria test, both for the full sample and for children under five. The absolute change in likelihood of getting tested for the treatment group as compared to the control group is approximately two percentage points. This is not a practically significant change, particularly when compared to the control mean of approximately 25 percent.

The following ITS analysis, then, will be compared against a statistically insignificant impact of two percentage points.

4.3 Interrupted time series

4.3.1 Data

For the interrupted time series specification, data was collapsed to the week level, in contrast to the RCT specification which used "pre" and "post" periods and survey round fixed effects.

Figure 4.2 details the number of episodes recorded per week.

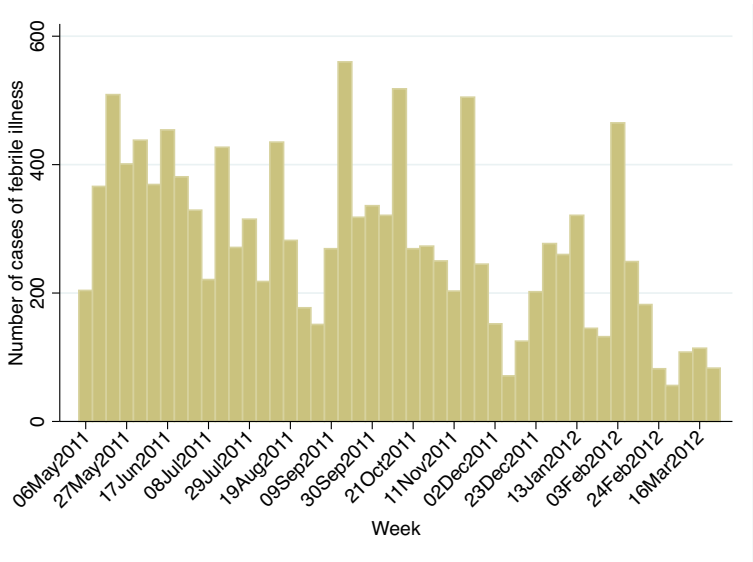

Figure 4.2: *Weekly febrile illness cases for BCC treatment group over study period*

All outcome measures in this analysis are of conditional probabilities: the probability of x event occurring given that a febrile illness episode took place. As a result, each measure is standardized by its shifting denominator.

4.3.2 Methods

Naive specification

The main specification for a single ITS design for outcome y_t is as follows:

$$y_t = \phi_0 + \phi_1 time_t + \phi_2 post_t + \phi_3 timepost_t + \kappa_t \tag{4.2}$$

$$\kappa_t = \rho \kappa_{t-k} + \varrho_t \tag{4.3}$$

where *time* denotes the number of weeks since data collection began, *post* denotes the introduction of the intervention, and *timepost* denotes the number of weeks since the introduction of the intervention. The error term τ_t in Equation 4.3 is a Newey-West standard error with lag k [70]. The value of k is determined by a Cumby-Huizinga general test for autocorrelation in time series data [82], implemented using the `actest` command in Stata [11]. I conduct the test for lags 1 to 10, and use the lag with the smallest p-value as k. If none of the lags are significant at the 5% level, then no lag is used.

For comparisons with RCT results, I use the midpoint of the post period and compare the value of the extrapolated pre period trend with the value of the fitted post period trend. I use Stata's `lincom` command to produce estimates for a linear combination of the *post* coefficient and the *timepost* coefficient at the midpoint of the post period.

Concurrent changes

As discussed in Section 3.2.1, any event taking place at the same time as program rollout that could affect the outcome poses a risk for ITS analyses. In the context of this intervention, any event that would affect the likelihood of receiving a malaria test would be an issue. Two time-dependent variables arise as potential confounders: Uganda's rainy seasons and the influx of antimalarial drugs into Uganda during this time. Both of these phenomena have the potential to influence testing behavior. Rainfall increases the overall prevalence of malaria, which could influence consumer confidence that a given febrile illness is caused by malaria. The increased availability (and subsequent lower cost) of antimalarial drugs into the health sector could influence the cost-benefit calculation of using malaria tests, both from the consumer and provider perspectives.

Rainfall Given that the spread of malaria is highly correlated with rainfall [73], one may expect testing behavior to shift with the two rainy seasons that Uganda experiences. The first rainy season begins in February, peaks in May and ends in June. The second begins in September, peaks in November, and ends in December [42, 73]. While the rainy season fluctuates throughout the study period, the main source of bias lies in the decrease in rainfall around December, the time of the intervention's introduction. Additionally, it is likely that the rainy season might affect the timing and immediacy of the effect of the intervention, since use of RDTs might vary with the perceived likelihood of contracting malaria given the time of year.

However, it is not clear *a priori* which direction testing behavior would move, and whether the effect would be gradual or immediate, as a result of the rainy season. More rain implies both more cases of malaria (increasing testing behavior) and more of a presumption that a given fever episode is malaria (decreasing testing behavior). Which of these two outweighs the other is not obvious.

To answer this question, I rerun Equation 4.2 with a control variable for rainfall, a continuous variable for average rainfall for that month from 1990-2012 [97]. This rainfall data are shown in Figure 4.3.

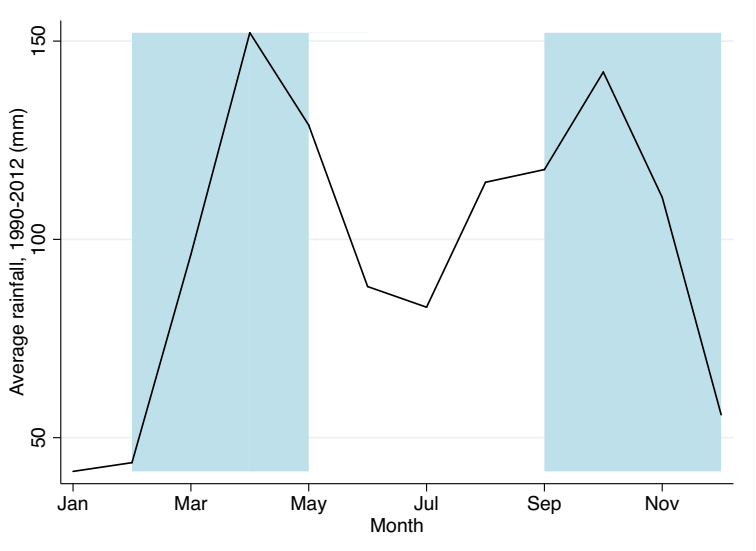

Figure 4.3: *Average rainfall in Uganda, 1990-2012 (shaded rainy seasons) [97]*

Drug stocks Throughout the study period, an influx of artemisinin combination therapies (ACTs) for malaria was taking place in Uganda thanks to the Affordable Medicines Facility - malaria (AMFm) [34]. Launched in 2009 as a pilot in seven countries including Uganda, the AMFm is a Global Fund program aiming to increase the availability and affordability of ACTs [1]. By reducing the cost of ACTs to first-line buyers by approximately 95% [63], the AMFm increased the market share of ACTs over inferior monotherapies and older antimalarials [34]. Note that this reduction in cost relates to first-line buyers, who did not necessarily transfer these savings to consumers.

Since ACT stocks were increasing throughout the study period, there are a number of mechanisms through which the AMFm could influence testing behavior that would confound the results of a single ITS. On the consumer side, increased availability of ACTs in local drug shops and in the private sector could reduce reliance on the public sector, where individuals are more likely to get a malaria test. Alternatively, a lower price for ACTs might reduce the relative value of a malaria test, since one of the benefits of testing is preventing spending on unnecessary antimalarial drugs. This could therefore reduce the willingness of individuals to pay for a test. On the provider side, the influx of ACTs might correspond to an increase in training of health workers with respect to malaria, leading to more appropriate use of malaria tests and treatment. Alternatively, health workers may perceive malaria tests as a way to ration the use of ACTs when there is high demand and low supply. The increased availability of ACTs might reduce the perceived need of health workers to ration their use via RDTs, leading to a decrease in RDT use. Note that, despite the fact that this influx took place throughout the study period, any fluctuations in the rate of drug stock replenishment occurring in the pre period or coinciding with the BCC intervention's introduction would constitute a concurrent event and potentially bias results.

To account for this influx of ACTs, I include national ACT stocks of public *and* private facilities in my ITS analyses (see Figure 4.4, with vertical line representing intervention introduction). Controlling for these stocks should isolate the impact of drug stocks on testing behavior over the course of the study period.

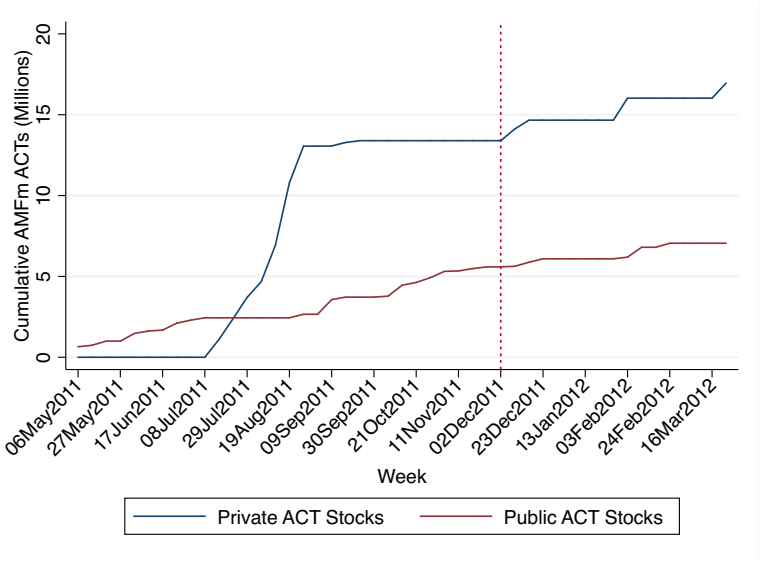

Figure 4.4: *AMFm ACTs arriving in public and private facilities nationally*

4.3.3 Results

Table 4.4 presents the ITS estimates for each specification described above.

Table 4.4: *ITS impact estimates (Uganda)*

VARIABLES	(1) Naive	(2) Rainfall	(3) Drugs	(4) Combined
Week	-0.00170**	-0.00170*	-0.00295	-0.00296
	(0.000831)	(0.000904)	(0.00831)	(0.00846)
Post	0.0386	0.0388	0.0404	0.0400
	(0.0243)	(0.0404)	(0.0283)	(0.0427)
Week*Post	-0.00228	-0.00228	-0.00165	-0.00163
	(0.00259)	(0.00273)	(0.00520)	(0.00555)
Rainfall (mm)		2.24e-06		-6.31e-06
		(0.000364)		(0.000415)
Public stocks			0.00515	0.00513
			(0.0339)	(0.0342)
Private stocks			0.000628	0.000654
			(0.00445)	(0.00475)
Constant	0.274***	0.274***	0.274***	0.274***
	(0.0177)	(0.0386)	(0.0201)	(0.0481)
Observations	47	47	47	47
Lag	0	0	0	0

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Outcome variable is the likelihood of receiving a malaria test
Data collapsed to the week level for 47 weeks

What follows is a discussion of the results for each model.

Naive results

The results of the naive regression are presented in column 1 of Table 4.4. The coefficients suggest a statistically significant decline in testing over the study period of 0.2 percentage points per week ("*Week*"). After the BCC intervention, there is an increase in testing of 3.8 percentage points ("*Post*"), followed by a steeper decline in testing over time of 0.40 percentage points ("$Week * Post$"). However, neither of these shifts is found to be statistically significant.

Figure 4.5 presents these results visually:

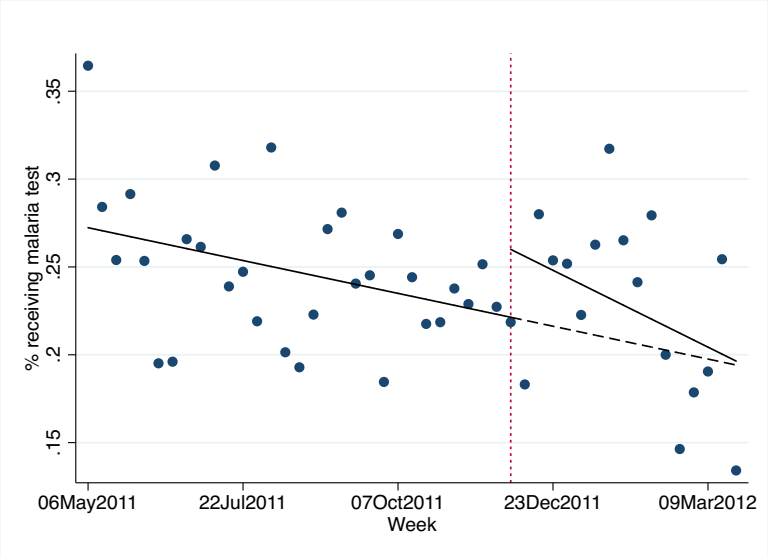

Figure 4.5: *Visual representation of naive ITS estimates (Uganda)*

The extrapolated counterfactual in Figure 4.5 provides some insight regarding the insignificant effect of the BCC intervention. If the counterfactual is to be believed, the sample ended up with almost exactly the same likelihood of receiving a test as it would have had if the BCC intervention had never been introduced. However, the shape of the impact over time suggests that there was a short term "bump" that gradually wore off throughout the post period. The notion that a behavior change intervention would affect behaviors in the very short term but not after is certainly plausible, and is an insight that a comparison of means would be unable to illustrate.

Concurrent changes

Columns 2 through 4 of Table 4.4 present the effect on ITS estimates of controlling for rainfall, national ACT drug stocks, and both, respectively.

Rainfall Column 2 suggests that controlling for monthly rainfall levels has essentially no effect on ITS estimates. The negligible effect of controlling for rainfall is further illustrated in the almost identical visuals of Figures 4.5 and 4.6:

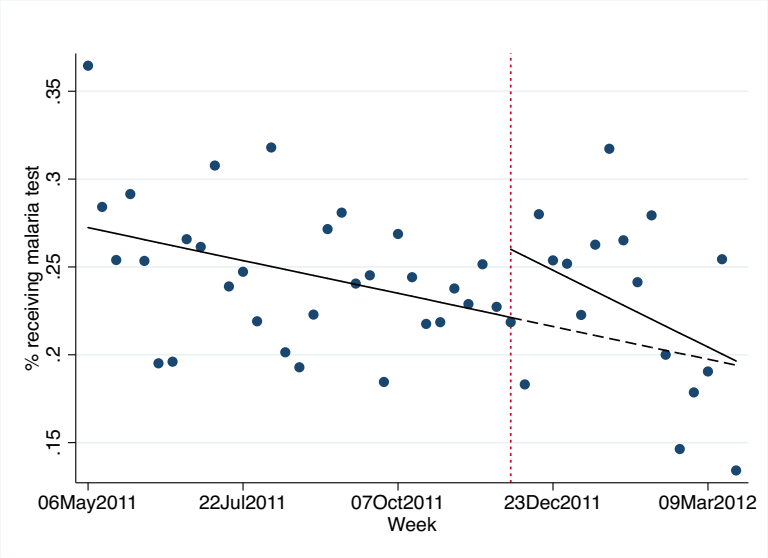

Figure 4.6: *Visual representation of ITS estimates controlling for rainfall (Uganda)*

These results suggest that the ITS specification is robust to a variable tightly related to seasonality. Additionally, the Cumby-Huizinga tests in all specifications found no significant evidence for autocorrelation, further assuaging concerns about time-related factors biasing results.

Drug stocks Column 3 of Table 4.4 shows the effect of controlling for public and private ACT stocks on the ITS estimates. The primary change between columns 1 and 3 is that the time trend shifts from being statistically significant to insignificant, despite increasing in magnitude. This is likely a consequence of the high correlation between public and private drug stocks with time, as shown in Figure 4.4.

This change, however, is not along a practically significant margin. Figure 4.7 illustrates just how similar the ITS estimates are as compared to the naive regression illustrated in Figure 4.5:

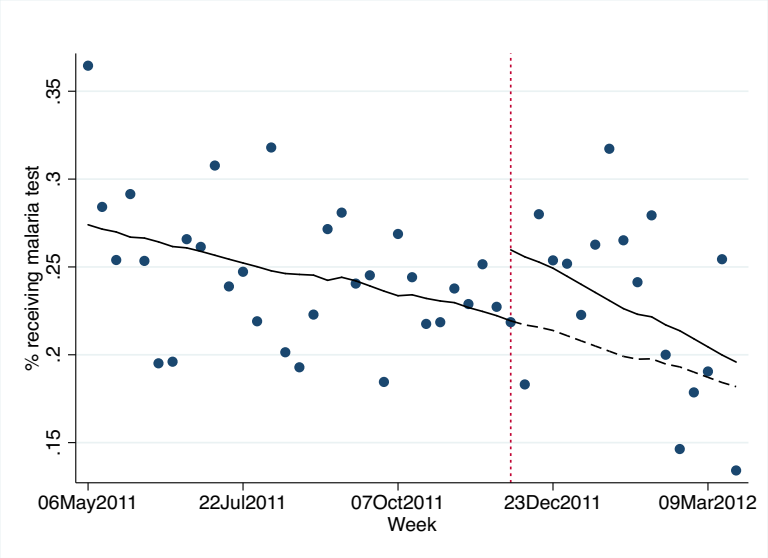

Figure 4.7: *Visual representation of ITS estimates controlling for national drug stocks (Uganda)*

This visual illustrates the slight fluctuations in predicted likelihood of a malaria test attributable to the presence of drug stocks. It is still possible that the influx of ACTs throughout the entire study period is causing the secular decrease in RDT use shown in the "week" variable and in all figures. However, the similarity of results in the naive and drug stock regressions suggests that any shifts in drug stocks during the study did not lead to a change in testing behavior.

Combined The results of the previous sections suggests that drug stocks and rainfall provide little benefit in predicting the effect of the BCC intervention on testing behavior. Nonetheless, it represents a reasonable approximation of what a researcher without access to control group data would propose. Column 4 of Table 4.4 and Figure 4.8 present the estimates of this fully specified model.

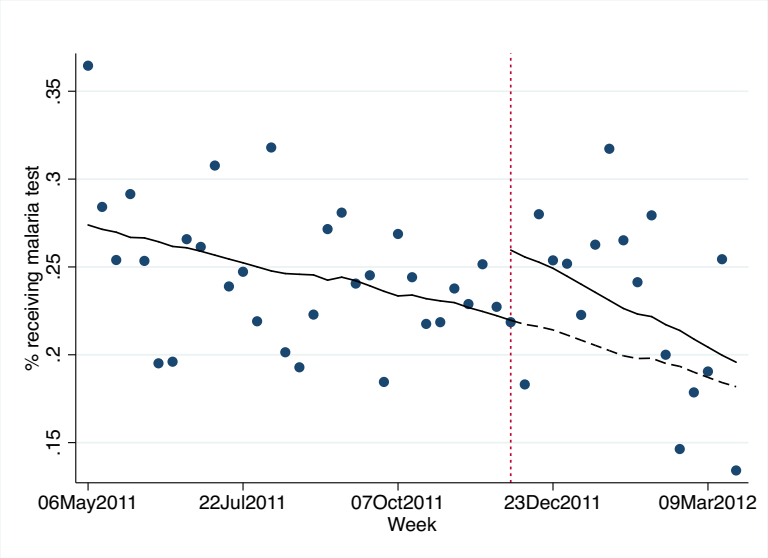

Figure 4.8: *Visual representation of ITS estimates in fully specified model (Uganda)*

The implication of these exercises is that the ITS specification is highly robust to potential confounders discussed in Section 4.3. This lends credibility to the notion that the estimates in Table 4.4 represent a best attempt to use the single ITS design to its fullest. The question, then, is how well this ITS "best guess" fares with respect to the original RCT.

RCT comparison

Table 4.5 compares the ITS estimates at the midpoint of the post period to the RCT estimates from Section 4.2.3. The last column of the table shows the difference between each ITS estimates with that of the RCT, including bootstrapped standard errors produced using the method described in Section 2.3.2.

Table 4.5: *ITS vs RCT Estimates (Uganda)*

	Impact Estimate	Difference
Randomized Controlled Trial	0.020	
	(0.015)	
Naive	0.020	0.000
	(0.020)	(0.020)
Rainfall	0.021	0.001
	(0.033)	(0.035)
Drugs	0.027	0.007
	(0.053)	(0.034)
Full	0.027	0.007
	(0.057)	(0.036)

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Outcome variable is the likelihood of receiving a malaria test
Data collapsed to the week level for 47 weeks

The ITS estimates are remarkably close to that of the RCT. The estimates never differ by more than 0.7 percentage points, and the differences are never statistically significant. An interesting note is that the model that comes closest to the RCT is the naive specification. This result is identical and magnitude and almost identical in standard error. While some secondary outcomes do in fact differ from the RCT results (see Appendix A), the impact of the BCC intervention on the primary outcome is measured as effectively with a single ITS design, using time trends and no control group, as it is with a cluster RCT design.

4.4 Discussion

Returning to the framework outlined in Figure 2.2, the differences in results between the ITS specifications and RCT are neither practically nor statistically significant. In this case, the single ITS design is concordant with the RCT result.

Understanding why the single ITS performs so well is difficult to determine with only one study's results. That said, the result of the falsification test described in Section 3.3.3 is instructive. First, the data-driven test for a structural break was unable to reject the null hypothesis of no structural

break in the data. In addition to aligning with the RCT result of no detectable effect of the intervention, this result also suggests that the outcome of interest was not subject to any large breaks at other points.

The binning method described in the latter part of the falsification test found only two points of possible breaks in the post period. Figure 4.9 shows dotted lines corresponding to these potential breaks, and a solid line at the intervention's introduction (which was not found to be statistically significant). Thus the underlying data - particularly in the pre-period and at the intervention's introduction - can be thought of as relatively stable.

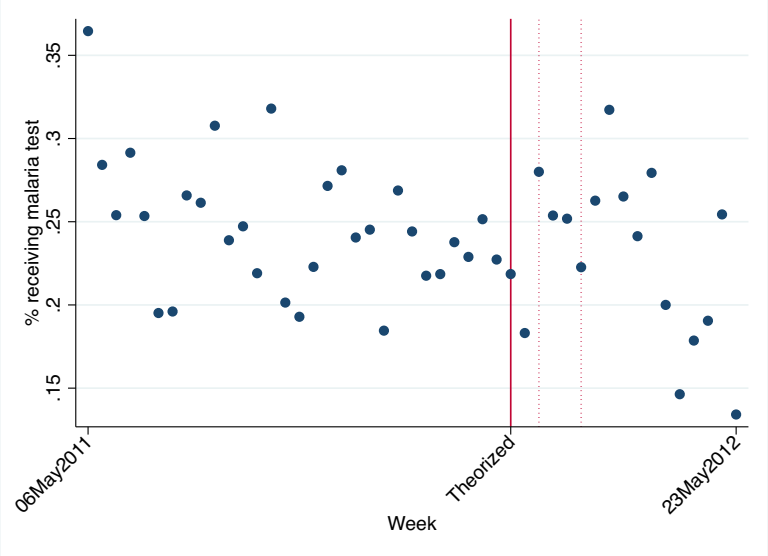

Figure 4.9: *Detected structural breaks (Uganda)*

In addition to having a stable trend, it is worth noting that the spread of the data appears to adhere to the linearity implied by **Assumption 2**. This is especially important during the pre period, since it ensures that the projected counterfactual provides a stable baseline against which the post period data can be compared. Moreover, the shape of the data looks plausible; the trend looks relatively stable other than seasonal fluctuations, and the influx of antimalarial drugs into the country throughout the study period helps explain the slightly negative trend.

The robustness of the results to specifications that include potential confounders lends some credibility to the main specification. This is further exemplified by Figure 4.10, which illustrates

the spread of difference between the single ITS and RCT designs for the primary outcome of being tested in addition to the secondary outcomes listed in Appendix A.

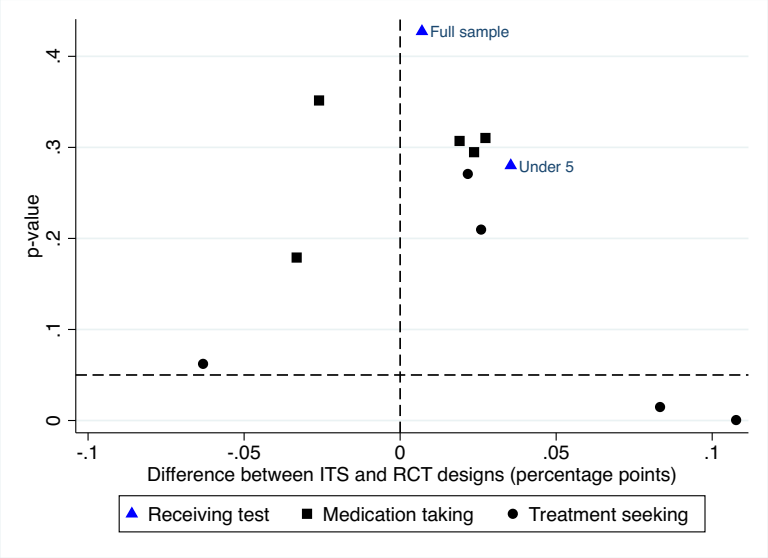

Figure 4.10: *Distribution of WSC results for primary and secondary outcomes*

The primary outcome of receiving a malaria test fairs best, while secondary outcomes are more likely to produce discordant estimates (though not by much). This is especially true of outcomes related to treatment seeking behavior, arguably the set of variables least directly tied to the BCC intervention. These outcomes are more discordant with the RCT result, and are more likely to have a significant difference with the RCT result. One could argue that the dynamics underlying an individual's treatment seeking behavior are subject to a more complex set of covariates than the determinants of whether an individual receives a malaria test for a given episode, such as changes in the distribution of public and private facilities, as well as an individual's trust in public and private providers. While further research is required to test this theory, the results are compatible with the argument that an outcome influenced by fewer external variables is more likely to provide a reliable ITS estimate.

Chapter 5

Effect of Health Insurance on Emergency-Department Use in Oregon

5.1 Background

5.1.1 Intervention

The second WSC in this manuscript tests the single interrupted time series design in the context of an expansion of health insurance in the United States. In 2008, the state of Oregon expanded its Medicaid program to a group of previously uninsured adults via a lottery [35]. 30,000 names were drawn from a waiting list of 90,000 people. These individuals were given the opportunity to apply for Medicaid and, if they met requirements for eligibility, enroll [9]. The randomized nature of the expansion allowed for a large-scale randomized trial to study the effects of health insurance provision on self-reported general health [35], measured physical health [9], and emergency-department usage [90]. This analysis will focus only on the study assessing the impact of insurance provision on usage of emergency departments [90], since it relied primarily on administrative data for which multiple pre period time points were available.

5.2 Randomized controlled trial

5.2.1 Data

Data consists of all emergency-department visits to 12 Portland area hospitals from 2007 to 2009. Though these data are not comprehensive of all ED visits in Oregon, it comprises almost all visits in Portland and about half of all hospital admissions in Oregon [90]. The dataset includes emergency-department records and, for those that were admitted to the same hospital, inpatient records.

Of the 90,000 names in the lottery, approximately 75,000 remained in the Oregon Health Insurance Experiment after excluding ineligible entries. Of these individuals, 24,646 lived in a zip code at the time of the lottery where residents used one of the twelve study hospitals almost exclusively ($\geq$ 98% of admissions). Within this sample, 9,626 were assigned to the treatment while 15,020 were controls [90]. Emergency-department data were probabilistically matched to the individuals in the experiment on the basis of name, date of birth, and gender [9, 35, 90]. Since randomization was at an individual level, larger households were more likely to receive the treatment than smaller ones. To account for this, the RCT specification below controls for household size.

5.2.2 Methods

The RCT results are estimated using both intent to treat (effect of lottery selection) and local average treatment effect (effect of Medicaid coverage) specifications. The intent to treat (ITT) effect was estimated using the following equation:

$$y_{ih} = \beta_0 + \beta_1 LOTTERY_h + \beta_2 hhsize_h + \beta_3 preoutcome_i + \beta_4 preoutcome\_missing_i + \epsilon_{ih} \quad (5.1)$$

where

- y_{ih} is the total number of ED visits for person i in household h between March 9, 2008 and

September 30, 2009;

- $LOTTERY$ is a dummy variable for the selection of household h into treatment;
- *hhsize* is the number of individuals in household h, a variable which was correlated with treatment selection;
- *preoutcome* is the pre-randomization value of person i for the outcome (before March 9, 2008);
- *preoutcome_missing* is an indicator for an observation lacking a pre-randomization value for outcome y (the *preoutcome* value for these observations is the mean for non-missing observations). Of the 24,646 individuals in the dataset, 12 were missing pre-randomization values.

The effect of Medicaid coverage is estimated using an instrumental variable (IV) approach, wherein the variable $LOTTERY$ is used as an instrument for Medicaid coverage. A two-stage least squares (2SLS) regression is modeled using the following equation:

$$y_{ih} = \pi_0 + \pi_1 MEDICAID_{ih} + \pi_2 hhsize_h + \pi_3 preoutcome_i + \pi_4 preoutcome\_missing_i + \nu_{ih} \quad (5.2)$$

where π_1 is estimated using the first stage equation:

$$MEDICAID_{ih} = \delta_0 + \delta_1 LOTTERY_{ih} + \delta_2 hhsize_h + \delta_3 preoutcome_i + \delta_4 + preoutcome\_missing_i + \mu_{ih} \quad (5.3)$$

Using the IV approach, π_1 represents the impact of receiving Medicaid on the compliers, i.e., those who received Medicaid via the lottery who would not have done so have without it.

5.2.3 Results

Results of the ITT and IV estimates for the outcome variable "total number of ED visits in the post period" are presented in Table 5.1.[1] Column (1) displays ITT results (Equation 5.1). The effect of selection in the lottery is an increase in ED visits of .101 ($p < 0.01$), indicating a 10% increase in the number of ED visits for the treatment group as compared to the control group mean of 1.022. Column (2) displays IV results (Equation 5.2). The effect of enrollment in Medicaid for compliers is an increase of .408 visits ($p < 0.01$), a 40% increase as compared to the control group. While these two estimates differ in terms of magnitude (by construction), they are consistent in positing a positive effect of insurance allocation on ED use that is both practically and statistically significant.

Table 5.1: *RCT impact estimates (Oregon)*

	(1)	(2)
VARIABLES	ITT	IV
Selected in the lottery	0.101***	
	(0.0287)	
Enrolled in Medicaid		0.408***
		(0.116)
No. of ED visits by 3/9/08	0.762***	0.755***
	(0.0252)	(0.0253)
Missing no. of ED visits by 3/9/08	19.50***	19.42***
	(4.654)	(4.645)
Constant	0.438***	0.381***
	(0.0200)	(0.0305)
Observations	24,622	24,622

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Outcome variable is the number of ED visits per person
Dummy variables for number of individuals in household not shown
Control group mean is 1.022

For the purposes of comparison, all subsequent analyses will use the ITT estimates as the RCT

[1] These results are from a reanalysis of the original Oregon data, which reproduce the published results in Taubman et al. [90].

results, in adherence to WSC criteria two discussed in Section 2.3.1. In order to ensure that the population of the subsequent ITS analysis is equivalent to that of the RCT, it is important to include all individuals allocated to treatment rather than just the compliers.

5.3 Interrupted time series

5.3.1 Data

I include only the 9,612 individuals from the treatment group in the ITS analysis. Data are collapsed to the biweekly level instead of "pre" and "post" periods. I use biweeks as the unit of time to allow for time trends while avoiding the noise introduced by a more granular measure [2] Data are available from January 1, 2007 to September 30, 2009. At the biweek level, this produces 72 time points, 31 of which were pre-intervention.

Notification of acceptance into Medicaid began on March 3, 2008, and continued until September 11, 2008. A new round of notifications took place every two to three weeks, with approximately 1,000 new individuals notified during each round (see Table 5.2).

Table 5.2: *Notification of insurance provision by date*

Notification Date	*N*
3/10/2008	1,010
4/7/2008	1,004
4/16/2008	1,014
5/9/2008	1,004
6/11/2008	932
7/14/2008	1,849
8/12/2008	1,885
9/11/2008	914
Total	9,612

The outcome variable of interest for the RCT was total number of ED visits per person in the post period. The analogous measure for ITS was total number of ED visits in the sample per person,

[2]The RCT regressions in the previous section were run with data collapsed to this level. The results were robust to either specification.

per biweek.

5.3.2 Methods

Naive specification

Using the specification for a single ITS design, I estimate the following OLS regression:

$$y_t = \zeta_0 + \zeta_1 biweek_t + \zeta_2 post_t + \zeta_3 biweekpost_t + \tau_t \tag{5.4}$$

$$\tau_t = \rho\tau_{t-k} + \psi_t \tag{5.5}$$

where τ_t is a Newey-West standard error with lag k [70]. The value for k is determined using the Cumby-Huizinga general test for autocorrelation in time series [82], implemented using `actest` in Stata [11]. I conduct the test for lags 1 to 10, and use the lag with the smallest p-value as k. If none of the lags had a p-value less than 0.05, no lag is used.

The variable *post* equals one for all times after March 3, 2008, the beginning of insurance rollout.

In addition to this standard single ITS specification, I attempt to address each of the potential threats to validity outlined in Section 3.1.2.

Concurrent changes

Given that the outcome of interest is emergency-department visits and the intervention was introduced in early March, there is potential for the flu season to generate an increase in ED usage around the time of program rollout. To address this concern, I gather data on cases of the flu in Oregon during the 2006-07 season, 2007-08 season, and 2008-09 season, in order to assess the degree of correlation between ED usage and flu cases, as well as explicitly controlling for it in the ITS regression [36].

Differential pre period changes

Events taking place in the pre period that are related to treatment provision can also introduce bias, as outlined in Section 3.2.2. One such event is the signup period for entry into the lottery, which began at the start of 2008. There are several mechanisms through which this could plausibly introduce bias.

First is the possibility of differential selection. If we assume that hospitals prefer insured patients to uninsured ones, they would have an incentive to encourage uninsured ED patients to sign up for the lottery once it was announced. This would lead to a sample that is defined by patients with an especially high number of ED visits in the months leading up to insurance provision, either in the form of higher levels or an increasing slope. This would be a violation of **Assumption 1**, since the sample receiving the treatment is fundamentally different from the sample that would have existed without an announced signup period.

A second mechanism is differential history. If the existence of the signup period led to an environment for the treatment group that is different from what otherwise would have occurred, this would cause bias. Though less plausible than the previous example, one might imagine a more "hostile" admissions environment. Hospitals may have an incentive to delay providing care to to uninsured patients for routine procedures if there is some chance of them obtaining insurance in the near future. Doing so would allow hospitals to avoid incurring the risk of non-payment from uninsured patients, as well as the cost of seeking out individuals with delinquent payments. Since hospitals were made aware of the upcoming lottery, they may have had artificially low ED admissions for the months leading up to provision.

To address this, I model the signup period explicitly in a multiple segmented regression:

$$y_t = \theta_0 + \theta_1 biweek_t + \theta_2 signup_t + \theta_3 biweeksignup_t + \theta_4 post_t + \theta_5 biweekpost_t + g_t \tag{5.6}$$

Where $signup$ is a dummy variable which equals one at the start of the signup period and after, and $biweeksignup$ is a variable which is zero until the signup period, and incremented by one for every subsequent biweek. g_t maintains the autocorrelation structure from Equation 5.5.

Misspecification of functional form

Finally, misspecifying the timing and dynamic of an intervention's introduction can threaten the internal validity of an ITS design [84]. In the context of the Oregon Medicaid rollout, the simplifying assumption that the program began on March 3, 2008 may introduce bias or noise. To address this, I estimate a respecified ITS model which accounts for the eight different notification dates occurring between March and September. Specifically, I use the following model:

$$y_t = \lambda_0 + \lambda_1 biweek_t + \lambda_2 post_{tg} + \lambda_3 biweekpost_{tg} + h_t \tag{5.7}$$

Where *post* is a dummy variable which equals one at the period that group g was notified of Medicaid enrollment and all subsequent time periods, and *biweekpost* equals zero until the the period group g was notified of Medicaid enrollment, and incremented by one for every subsequent period h_t maintains the autocorrelation structure from Equation 5.5.

5.3.3 Results

The results of each specification are presented in Table 5.3.

Table 5.3: *ITS impact estimates (Oregon)*

VARIABLES	(1) Naive	(2) Flu season	(3) Signup period	(4) Recentered
Biweek	0.000241***	0.000249***	0.000118***	0.000174***
	(5.08e-05)	(6.02e-05)	(4.10e-05)	(3.07e-05)
Post	-0.000828	-0.00105	-0.00222***	-0.000203
	(0.00132)	(0.00163)	(0.000649)	(0.000983)
Biweek*Post	-0.000300***	-0.000302***	2.48e-05	-0.000255***
	(5.75e-05)	(6.16e-05)	(9.85e-05)	(4.38e-05)
Signup			0.00512***	
			(0.000785)	
Biweek*Signup			-0.000202*	
			(0.000103)	
Flu rate per 100,000		7.44e-05		
		(7.68e-05)		
Constant	0.0210***	0.0207***	0.0222***	0.0208***
	(0.000666)	(0.000843)	(0.000536)	(0.000795)
Observations	72	72	72	85
Lag	1	4	0	0

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Outcome variable is the number of ED visits per person per biweek
Data collapsed to the biweek level for 72 biweeks

Naive results

Column 1 of Table 5.3 presents the results of a naive single ITS analysis. The results show a positive, statistically significant increase in ED visits per person during the pre period. At the time of insurance provision, there is a non-significant drop in level, followed by a significant, sharply negative change in slope. The magnitude of the slope change actually reverses the trend in ED utilization from a positive trend to a negative trend.

Figure 5.1 shows these results visually:

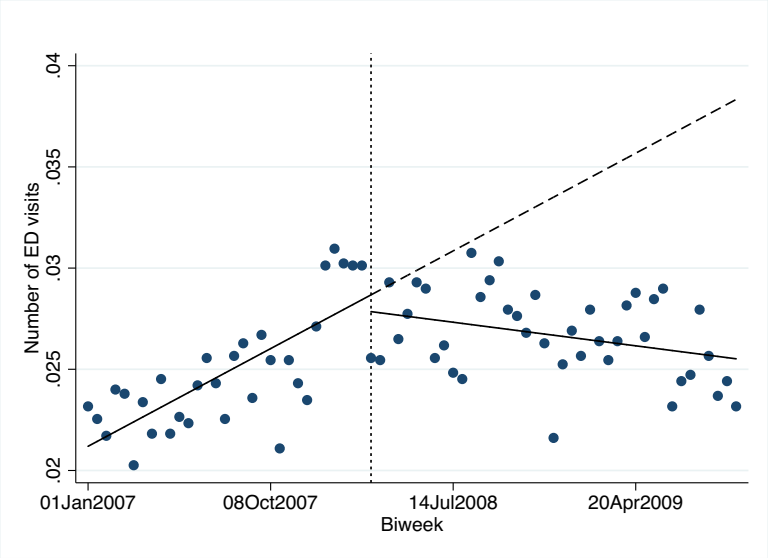

Figure 5.1: *Visual representation of naive ITS estimates (Oregon)*

The naive specification thus implies that provision of Medicaid in Oregon in mid-2008 reversed an increasing trend of ED use.

The pre period data has two notable characteristics. First is the cluster of ED visits in the five time points immediately preceding the intervention's introduction. This period corresponds to the time after which the lottery was announced, when hospitals were likely encouraging ED patients to apply and patients themselves were considering their need for insurance. It is thus plausible that patients who happened to come in during this period were more likely to be included in the lottery sample.

The second issue is the general positive trend in ED visits throughout the entire period. It is unclear whether this is a secular trend, the result of seasonal variation, or a manifestation of Ashenfelter's dip described in Section 3.2.2.

Each of these potential explanations for the time trend in Figure 5.1 will now be interrogated.

Flu season

Figure 5.2 overlays average number of ED visits with data on flu. While it does look as though the 2007-08 season was especially acute, there is no evidence for a statistically significant difference

in means across the three season ($F_{2,88} = 0.13, p = 0.874$).

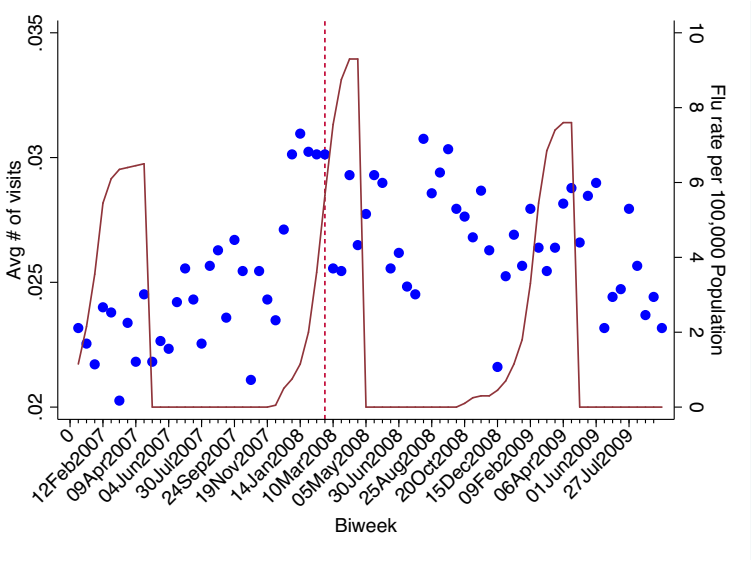

Figure 5.2: *ED usage and flu seasons, 2007-09*

Additionally, flu rates are not a significant predictor of ED visits, as shown in Column 2 of Table 5.3. Augmenting the naive regression from Equation 5.4 with a continuous variable for weekly flu rate had essentially no effect on estimates. Taken together, these results suggest that the positive pre period trend is not driven by seasonality related to the flu.

Signup period

Column 3 of Table 5.3 shows that modeling the signup period using the specification from Equation 5.6 does change the estimates from the naive specification, though not by much. In addition to lowering the slope of the pre period, accounting for the signup period made the change in slope insignificant, while making the drop in level significant. These changes are to be expected, as they are in contrast to the slope and level of the signup period as opposed to a pre period which included the signup period. Figure 5.3 illustrate these changes.

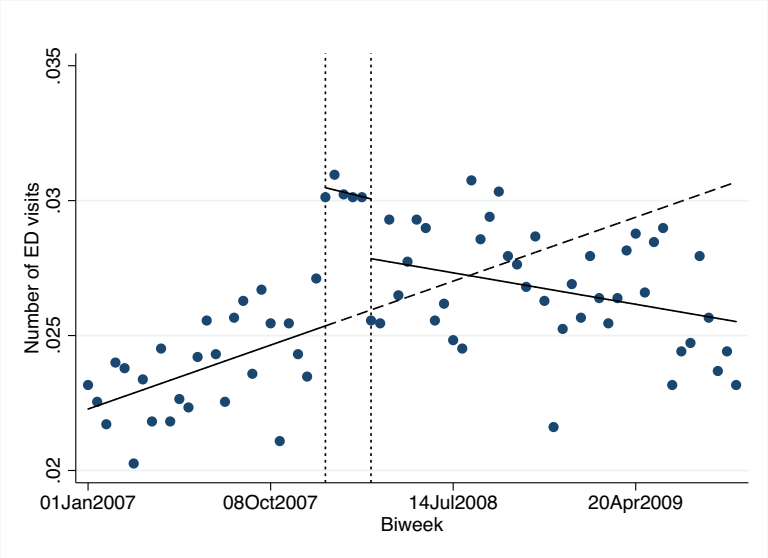

Figure 5.3: *Visual representation of ITS estimates with "washing out" signup period (Oregon)*

Accounting for the signup period leading to the spike in ED visits in the weeks preceding the lottery does change the estimated coefficients to a degree. However, these changes are not along a practically significant margin, and Figure 5.3 shows that the findings remain unchanged in broad terms. The data are still characterized by a positive pre period slope (albeit a more shallow one) followed by a negative slope (albeit an insignificant one) in the post period.

Recentered specification

Recentering the specification around notification does not significantly affect ITS estimates, as shown in Column 4 of Table 5.3. While Figure 5.4 shows a level change of zero in contrast to Figure 5.1, the naive regression's level change is not statistically significant either.

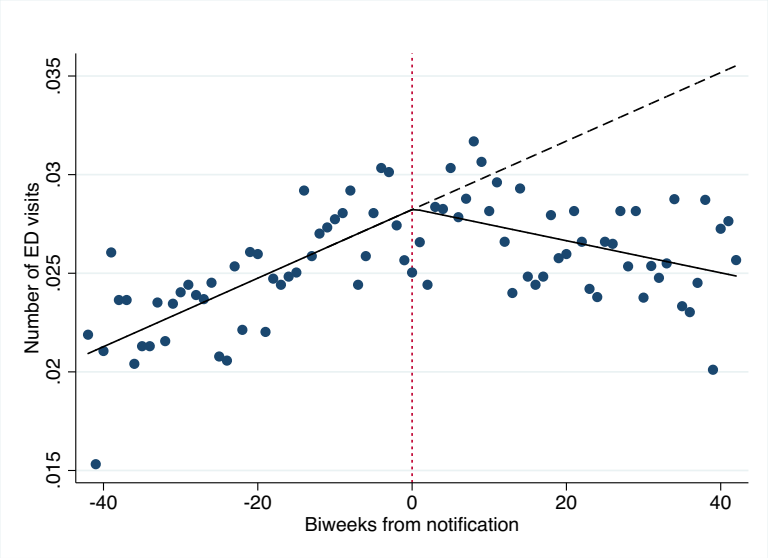

Figure 5.4: *Visual representation of ITS estimates with recentered specification (Oregon)*

The fact that the recentered specification has a negligible effect on the naive result is telling. By recentering the data around each group's notification date, this specification theoretically offsets biweek-specific drivers of the result. This includes the possibility of particular events taking place during the study period that would drive this upward trend. It does not, however, rule out Ashenfelter's dip. If individuals were self-selecting into the lottery on the basis of increased ED use in the run up to the intervention, this upward trend would be reflected in the recentered specification as well.

RCT comparison

For each of the specifications above, estimates were translated into an aggregated measure of the effect of the program on total number of ED visits, in order to make results comparable to those of the RCT. Table 5.4 presents these results. [3] Differences between the RCT estimate and each ITS specification's estimate are presented as well. Standard errors for differences are obtained via the bootstrapping method outlined in Section 2.3.2.

[3]Ien order to ensure that this analysis compared "apples to apples," the RCT was rerun using week-level data instead of individual observations. The estimate using these data was extremely close to the original RCT estimate.

Table 5.4: *ITS vs RCT estimates (Oregon)*

	Impact Estimate	Difference
Randomized Controlled Trial (ITT)	0.101***	
	(0.029)	
Naive	-0.280***	-0.380***
	(0.087)	(0.112)
Flu Season	-0.291***	-0.392***
	(0.104)	(0.112)
Signup Period	-0.069	-0.169
	(0.083)	(0.116)
Recentered	-0.228***	-0.329***
	(0.056)	(0.115)

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Control mean is 1.022

All four ITS specifications show a complete failure to replicate the RCT result. Each of them is in the opposite direction from the RCT finding by a wide margin, which easily falls into the category of "practically significant". The model that comes closest to the RCT result is the signup period (Column 3 of Table 5.3), in that it is statistically indistinguishable from zero. Using the framework described in Figure 2.2, the single ITS design is clearly discordant with the RCT result.

In order to further explore the possibility of Ashenfelter's dip driving this discordance, Table 5.5 compares the ITS estimates to the estimates of a simple pre-post comparison.

Table 5.5: *ITS vs Pre-Post estimates (Oregon)*

	Effect of Medicaid	
Randomized Controlled Trial (ITT)	.101***	
	(0.029)	
	ITS	**Pre Post**
Naive	-0.280***	0.037***
	(0.087)	(0.012)
Flu Season	-0.291***	0.037***
	(0.104)	(0.012)
Signup Period	-0.069	0.041***
	(0.083)	(0.009)
Recentered	-0.228***	0.039***
	(0.056)	(0.012)

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

In each specification, a pre-post estimator comes much closer to replicating the RCT result than the ITS estimator [4]. While all pre-post estimates understate the effect by several percentage points, they are in the correct direction and have similar statistical significance as the RCT impact. This once again lends credibility to the possibility of Ashenfelter's dip, which predicts that an ITS estimator will produce more bias than a simple pre-post comparison (see Section 3.2.2).

5.4 Discussion

The results of this analysis paint a discouraging picture for the single ITS design. Using the framework described in Section 2.3.2, the presence of statistically and practically significant differences between designs suggests that the results of the single ITS design are discordant with those of the RCT. Put concretely, if the state of Oregon had chosen to analyze the effect of its Medicaid expansion using ITS, the measured impact would have been statistically significant and in the wrong direction. To make things worse, this incorrect result is robust to alternative

[4]For completeness, two additional models were run: one incorporating quadratic time terms, and an individual level model estimated using Generalized Estimating Equations. The latter had a negligible effect on results, while the former produced estimates that diverged even further from the RCT due to a positive quadratic term in the pre period.

specifications, which would only further mislead policymakers with respect to the validity of these estimates.

The question of *why* this analysis was unable to reproduce the RCT result is difficult to answer definitively. However, some issues are worth pointing out.

The falsification test described in Section 3.3.3 provides some useful insight. The data-driven test for a structural break detected a highly significant break at December 31, 2007 ($p < 0.001$), corresponding to the start of the signup period. This is illustrated by a thick dashed line in Figure 5.5. The fact that this break taking place in the pre period was found to be more significant than the intervention introduction calls the validity of the counterfactual into question. Additionally, the binning method and subsequent tests for structural breaks found 11 statistically significant breaks in the pre period and six in the post period (illustrated with thin dotted lines in Figure 5.5), as well as the intervention point itself.

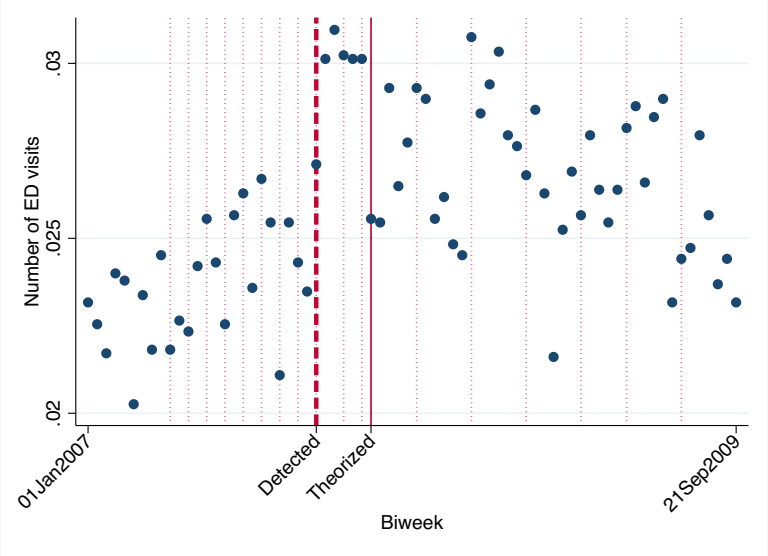

Figure 5.5: *Detected structural breaks (Oregon)*

The sheer number of structural breaks detected in the data implies that this dataset has far too many fluctuations to provide a credible estimate of the intervention's effect.

The result from this falsification test on the sample data is further confirmed by emergency-department data more generally. Figure 5.6 illustrates annual ED visit data for Oregon and two

neighboring states (California and Washington) since 2001. In this ten year period, there are a large number of fluctuations in ED admissions for each state. Depending on the intervention point and state chosen, there are many points where an ITS analysis would detect a significant effect (e.g., 2007 in California, 2008 in Oregon, or 2009 in Washington).

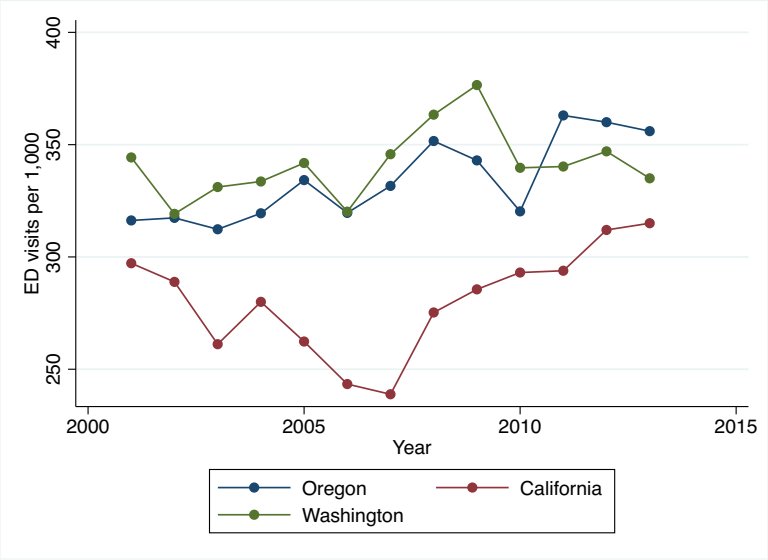

Figure 5.6: *Emergency department visits in Oregon, California, and Washington, 2001-2013*

The most immediate implication for an ITS analysis is that, for these data, *any extrapolated linear trend will be misleading*. The discordance of the ITS estimates with the RCT (and even pre-post) estimates appears largely attributable to this poor counterfactual.

In addition, the results of an ITS analysis for a given state at a given time produces conflicting results depending on the time horizon used. Table 5.6 presents an example. In this table I run an ITS specification for Oregon using 2008 as the intervention point (the year the lottery took place). The only difference between columns 1 and 2 is the inclusion of three more data points in column 2 (years 2011-2013). Yet the estimated impacts go from highly significant to non-existent, an intervention that reduces ED use to having no effect at all.

Table 5.6: *Sensitivity of ITS results for ED data in Oregon by timeframe*

	(1)	(2)
VARIABLES	2001-2010	2001-2013
Year	2.589***	2.589***
	(0.675)	(0.629)
Post	22.05***	8.805
	(4.637)	(9.829)
Year Post	-18.26***	0.718
	(2.131)	(2.524)
Constant	311.2***	311.2***
	(3.424)	(3.188)
Observations	10	13
Lag	1	1

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Outcome variable is number of ED visits per 1,000 population
Data derived from annual state-level reports of ED visits

Finally, the bias potentially introduced by Ashenfelter's dip appears to manifest itself in this analysis. The fact that individuals self-selected into the lottery produces a pre period trend that is a poor counterfactual. Building off the literature on this phenomenon, even a non-experimental control group would only be able to address this issue if it was characterized by the same dip as the treated sample [54]

Inherent noisiness, sensitivity to timeframe, and issues of self-selection are properties of the data itself which drive the discordance of ITS results with RCT estimates. Each of these qualities implies that the noisiness of ED visit data makes the series itself a poor candidate for an ITS analysis.

Part III

Conclusion

Chapter 6

Lessons

6.1 Introduction

The following sections represent a set of hypotheses aimed at making generalizable statements about the interrupted time series design. These lessons should be taken with a grain of salt, as they are the product of only two analyses, though I make every attempt to link these findings with existing literature on the topic.

6.2 Samples defined by self-selection may be problematic for single ITS

The Oregon Health Insurance Experiment is an example of a "randomized encouragement design." In this design a population is offered the opportunity to participate in some program, and the sample is then selected based on who volunteers. The sample is then randomly allocated between treatment and control arms, the former of which is given the opportunity to take up treatment [94]. The results are then analyzed using an instrumental variable approach to account for the probability that someone offered the intervention actually took it up. This is a common design in the literature on evaluating social programs, as it allows for the "messiness" of the political and social elements of program recruitment while still preserving the advantages of a

randomized controlled trial.

Unfortunately, the analyses in this manuscript suggest that this approach - or any in which a sample self-selects from a population of interest - may be highly damaging to the internal validity of the single ITS design. The issues underlying Ashenfelter's dip explained in Section 3.2.2 are readily apparent in the Oregon WSC. By allowing individuals to voluntarily select into the lottery, the pre period data was defined by an increasing trend, with an especially pronounced spike immediately leading up to the intervention. This pre period behavior introduced bias into a simple pre-post comparison, but introduced substantially *more* bias into a single ITS specification, as shown in Table 5.5. While this WSC represents a single data point, it adheres closely to the theory and empirical work done on Ashenfelter's dip in literature on labor markets [53, 54].

The sample in Uganda, by contrast, was not subject to self-selection. It was comprised of individuals, living in villages randomly selected to receive the BCC intervention, who experienced a febrile illness during the study period. Thus the only selection mechanism was whether or not an individual had an illness episode involving fever, which can be plausibly argued is largely orthogonal to the conditional probability of receiving a test given a febrile illness episode. One can imagine a scenario where the Uganda experiment would be subject to issues of self-selection. For example, if instead individuals experience a febrile illness in the pre period were offered to participate in a BCC information campaign, the sample would be subject to the same issues of self-selection as Oregon.

These results suggest that single ITS should be avoided when evaluating any program wherein participants self-select on the basis of the outcome that the program aims to change. While more rigorous evaluative methods are able to account for this self-selection in some way, the single ITS design is not.

6.3 A trend is not always superior to a mean

Short interrupted time series is premised on the notion that a trend of an outcome over time is preferable to a single point [14]. While it is true that several data points are always preferable to a single one, relying on a time trend instead of a mean requires the researcher to make parametric

assumptions related to level and slope that may not hold in practice. In particular, the linear extrapolation implied by **Assumption 2** may be an especially strong assumption that cannot be easily validated in the absence of a control group. This is most clearly evident in the Oregon WSC, where a simple pre-post comparison was actually superior to a parametrized single ITS model.

In many econometric strategies, adding a covariate such as a quadratic term or exogenous control variable is seen as potentially helpful and rarely harmful; these covariates have little effect if they do not contribute to model fit. By contrast, single ITS not only allows the level and slope to vary; it stakes the strength of the counterfactual on these parameters regardless of their level of significance. In addition, it only allows the level and slope to vary *at a single point* in time, which assumes a great deal about the nonexistence of other breaks in the dataset. The first part of the falsification test proposed in Section 3.3.3 aims to test the strength of this assumption. Specifically, the test makes no assumption about the location of the biggest break in the data, and makes a "best guess" at where this break could be. If the test determines that the largest break is not at the intervention point, and if this break is statistically significant, this assumption may be too strong for the data. This appeared to be the case in Oregon, where the largest break in the data occurred when the lottery was announced, not when insurance was introduced. This suggests a clear violation of the strong assumption of a break point at (and only at) a pre-specified time. In contrast to Oregon, the data in Uganda did not reveal *any* statistically significant breaks in the data, which is concordant with an outcome that was influenced by neither the intervention of interest nor other external forces.

In summary, the preceding analyses illustrate that the benefits of relying on trends for inference must be weighed against the strong assumptions that accompany their use.

6.4 Trend stability is crucial, especially in the pre period

Much of the documented guidance regarding the appropriateness of ITS has focused on having a sufficient number of time points to allow for stable trends and the ability to model seasonality [52, 93, 98]. Yet this may be only part of the story, particularly when dealing with "short"

interrupted time series designs that do not have statistical requirements for a minimum number of data points, since ARIMA modeling is not feasible [14, 64]. In these contexts, some measure of trend "stability" - particularly in the pre period - would be more appropriate.

The second part of the falsification test proposed in Section 3.3.3 is a useful starting point. In contrast to the first part, which tests whether the intervention point is the most significant break point, the second part tests for overall variability between adjacent times at various "bin points" in the data. If many of these points have statistically significant breaks, the data may not be stable enough for a single ITS analysis. Again, a contrast between Oregon and Uganda is instructive. Only two potential break points were detected in Uganda, while 17 were detected in Oregon (see Figures 4.9 and 5.5). Moreover, the two points in Uganda were both in the post period, while 11 of the 17 points in Oregon were in the pre period. As discussed in Section 3.2.2, the trend in the pre period is especially important in establishing a credible counterfactual, thus making the poor performance of the Oregon data even more concerning.

Similarly, the time frame that is deemed appropriate for an ITS analysis should be a function of the presence of trends and fluctuations in the pre period data. Whenever possible, historical data for the outcome of interest should be obtained to develop a strong prior for underlying trends in the data. These data need not be from the actual sample, provided that the population is at least somewhat comparable to the sampling frame.

6.5 Whether to implement an ITS design is more important than how to implement it

The two WSCs in this manuscript have clear results in opposite directions using the framework in Figure 2.2. While the ITS analysis of the BCC intervention in Uganda was entirely concordant with the RCT results, the ITS model of the Oregon Health Insurance Experiment was fully discordant with the RCT. In both cases, the result of the WSC was robust to every specification attempted (see Tables 4.5 and 5.4). In both studies, the naive ITS model was as good (or as bad) as a fully specified one, accounting for various threats to validity outlined in Chapter 3.

In addition to the robustness of these results, it should be noted that both studies fulfilled the criteria for a "best practice" ITS design enumerated in Figure 3.3. In addition to meeting the minimum criteria for number of time points and observations per time point, the ITS models accounted for autocorrelation and seasonality, controlled for potential confounders, and were robust to multiple sensitivity analyses.

The fact that results across these two studies were equally robust is troubling. It suggests that the robustness of an ITS model provides little information about the validity of its results. Granted, an especially sensitive model may imply that the model is poor. However, the fact that a model provides consistent results across multiple specifications does not even guarantee that results will be in the right direction, as the case in Oregon showed.

Thus, the results of the analyses in this manuscript suggest that the underlying properties of the data have far more impact on the validity of single ITS results than modeling decisions. This is an interesting contrast to WSCs using other quasi-experimental techniques. For example, the literature on propensity score matching has found that the validity of inferences based on matching is highly sensitive to analytic discretion [44, 79, 84, 86]. For single ITS, however, emphasis should be placed on the choice of whether or not to employ a given design, while how to best implement it should be a secondary concern. To understand if and when single ITS should be used, further research could employ multiple within-study comparisons of the same study. For example, a randomized trial with multiple time points could be analyzed via interrupted time series, a traditional difference-in-differences, and a difference-in-differences using various matching techniques.

6.6 Conclusion

So when *should* single ITS be used? In this manuscript I identify a number of characteristics that should signal to researchers and implementers that they should be particularly wary of using a single ITS design. [1] Identifying conditions especially conducive to the design is a much more

[1] A note outside of the scope of this analysis: the presence of a comparable control group in a multiple ITS design may account for many of the shortcomings in single ITS identified in this manuscript [84]. For example,

difficult (and data intensive) task than identifying reasons that the design may fail. Still, the fact that the two WSCs in this manuscript generated such different conclusions for the single ITS design suggests that the following hypotheses be further tested:

1. The two falsification tests identified in Section 3.3.3 provides a useful metric for determining the adequacy of single ITS to detect an unbiased effect in a given scenario. If the first test fails to reject the null hypothesis of an alternate break point, and the second test finds few potential breaks in the pre period, single ITS may be a viable candidate for a study design.

2. The data should not have any kind of "dip" or "spike" in the outcome for the time points leading up to an intervention's introduction. The presence of such a shift may be a red flag that disqualifies single ITS as a possible design.

3. Samples derived from any sort of self-selection mechanism may be poor candidates for single ITS. In contrast, interventions that are distributed across a population, where a study sample can be drawn randomly, may be more desirable.

Again, the above list represents a set of hypotheses to be further explored in research scrutinizing single ITS. In the meantime, the results of this manuscript suggest that caution should be exercised before adopting this popular quasi-experimental study design.

---

a similarly selected control sample with the kind of spike identified in the Oregon study could potentially offset the bias introduced in the single ITS analysis. However, as noted earlier, the control sample would have to mirror the sampling of the treatment group quite closely for this to occur. Such a design is also less reliant on projected counterfactuals, though is subject to the issues outlined in the literature on control groups and matching techniques [61, 86].

References

[1] O Adeyi and R Atun. Universal access to malaria medicines: innovation in financing and delivery. *The Lancet*, 376:1869–1871, 2010.

[2] Roberto Agodini and Mark Dynarski. Are Experiments the Only Option? A Look at Dropout Prevention Programs. *The Review of Economics and Statistics*, 86(1):180–194, February 2004.

[3] LS Aiken, SG West, and DE Schwalm. Comparison of a Randomized and Two Quasi-Experimental Designs in a Single Outcome Evaluation . *Evaluation Review*, 22(207), 1998.

[4] Karolina Andersson, Max Gustav Petzold, Christian Sonesson, Knut Lönnroth, and Anders Carlsten. Do policy changes in the pharmaceutical reimbursement schedule affect drug expenditures? interrupted time series analysis of cost, volume and cost per volume trends in sweden 1986-2002. *Health Policy*, 79(2-3):231–43, Dec 2006.

[5] Donald W. K. Andrews. Tests for parameter instability and structural change with unknown change point. *Econometrica*, 61(4):821–856, 1993.

[6] Anonymous. Time series analysis. In *Pharmacoepidemiology: behavioral and cultural themes*. Newcastle: Center for Clinical Epidemiology and Biostatistics Australia, 2001.

[7] Faranak Ansari, Kirsteen Gray, Dilip Nathwani, Gabby Phillips, Simon Ogston, Craig Ramsay, and Peter Davey. Outcomes of an intervention to improve hospital antibiotic prescribing: interrupted time series with segmented regression analysis. *Journal of Antimicrobial Chemotherapy*, 52(5):842–848, 2003.

[8] Orley C Ashenfelter. Estimating the Effect of Training Programs on Earnings. *The Review of Economics and Statistics*, 60(1):47–57, February 1978.

[9] Katherine Baicker, Sarah L. Taubman, Heidi L. Allen, Mira Bernstein, Jonathan H. Gruber, Joseph P. Newhouse, Eric C. Schneider, Bill J. Wright, Alan M. Zaslavsky, and Amy N. Finkelstein. The oregon experiment — effects of medicaid on clinical outcomes. *New England Journal of Medicine*, 368(18):1713–1722, 2013. PMID: 23635051.

[10] Vincent Batwala, Pascal Magnussen, and Fred Nuwaha. Comparative feasibility of implementing rapid diagnostic test and microscopy for parasitological diagnosis of malaria in uganda. *Malaria journal*, 10:373, 2011.

[11] Christopher F Baum and Mark E Schaffer. ACTEST: Stata module to perform Cumby-Huizinga general test for autocorrelation in time series. Statistical Software Components, Boston College Department of Economics, July 2013.

[12] Stephen H. Bell, Larry l. Orr, John D. Blomquist, and Glen G. Cain. *Program Applicants as a Comparison Group in Evaluating Training Programs: Theory and a Test*. Number pacg in Books from Upjohn Press. W.E. Upjohn Institute for Employment Research, January 1995.

[13] D Black and J Galdo. Estimating the selection bias of the regression discontinuity design using a tie-breaking experiment . *Syracuse University working paper*, 2005.

[14] Howard S Bloom. Using "Short" Interrupted Time-Series Analysis To Measure The Impacts Of Whole-School Reforms: With Applications to a Study of Accelerated Schools. *Evaluation Review*, 27(1):3–49, February 2003.

[15] HS Bloom, C Michalopoulos, and CJ Hill. Can Nonexperimental Comparison Group Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs? *MDRC Working Papers on Research Methodology*, 2002.

[16] Patrick Boruett, Dorine Kagai, Susan Njogo, Peter Nguhiu, Christine Awuor, Lillian Gitau, John Chalker, Dennis Ross-Degnan, Rolf Wahlström, and INRUD –IAA. Facility-level intervention to improve attendance and adherence among patients on anti-retroviral treatment in Kenya-a quasi-experimental study using time series analysis. *BMC health services research*, 13:242, 2013.

[17] Espen Bratberg, Astrid Grasdal, and Alf Erling Risa. Evaluating social policy by experimental and nonexperimental methods. *Scandinavian Journal of Economics*, 104(1):147–171, 2002.

[18] J W Brufsky, D Ross-Degnan, D Calabrese, X Gao, and S B Soumerai. Shifting physician prescribing to a preferred histamine-2-receptor antagonist. Effects of a multifactorial intervention in a mixed-model health maintenance organization. *Medical care*, 36(3):321–332, March 1998.

[19] H Buddelmeyer and E Skoufias. An evaluation of the performance of regression discontinuity design on PROGRESA. World Bank Policy Research Working Paper No. 3386; IZA Discussion Paper No. 827, 2004.

[20] Sebastian Calonico, Matias D Cattaneo, and Rocio Titiunik. Optimal data-driven regression discontinuity plots. *Journal of the American Statistical Association*, 110(512):1753–1769, 2015.

[21] Gregory C. Chow. Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, 28(3):591–605, 1960.

[22] Jessica Cohen, Günther Fink, Kathleen Maloney, Katrina Berg, Matthew Jordan, Theodore Svoronos, Flavia Aber, and William Dickens. Introducing rapid diagnostic tests for malaria to drug shops in uganda: a cluster-randomized controlled trial. *Bulletin of the World Health Organization*, 2015.

[23] Thomas D Cook, William R Shadish, and Vivian C Wong. Within-Study Comparisons of Experiments and Non-Experiments: What the Findings Imply for the Validity of Different Kinds of Observational Study. December 2005.

[24] Thomas D Cook, William R Shadish, and Vivian C Wong. Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, 27(4):724–750, June 2008.

[25] Angus S. Deaton. Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development. Working Paper 14690, National Bureau of Economic Research, January 2009.

[26] Rajeev H. Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062, 1999.

[27] Christiana Drake. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, 49(4):1231–1236, 1993.

[28] E Duflo and R Glennerster. Using randomization in development economics research: A toolkit. *NBER Technical Working Paper 333*, 2007.

[29] E Duflo and M Kremer. *Use of Randomization in the Evaluation of Development Effectiveness*, chapter 10, pages 205–231. World Bank Series on Evaluation and Development. Transaction Publishers, 2005.

[30] Effective Practice and Organisation of Care (EPOC). *EPOC Resources for review authors: Interrupted time series analyses*. Norwegian Knowledge Centre for the Health Services, Oslo, http://epocoslo.cochrane.org/epoc-specific-resources-review-authors edition, January 2013.

[31] Marion Elligsen, Sandra A N Walker, Ruxandra Pinto, Andrew Simor, Samira Mubareka, Anita Rachlis, Vanessa Allen, and Nick Daneman. Audit and feedback to reduce broad-spectrum antibiotic use among intensive care unit patients: a controlled interrupted time series analysis. *Infect Control Hosp Epidemiol*, 33(4):354–61, Apr 2012.

[32] M English and J Schellenberg. Assessing health system interventions: key points when considering the value of randomization . *Bulletin of the World Health Organization*, 89(12):907–912, 2011.

[33] Adrianne C Feldstein, David H Smith, Nancy Perrin, Xiuhai Yang, Steven R Simon, Michael Krall, Dean F Sittig, Diane Ditmer, Richard Platt, and Stephen B Soumerai. Reducing warfarin medication interactions: an interrupted time series evaluation. *Archives of Internal Medicine*, 166(9):1009–1015, 2006.

[34] Günther Fink, William T. Dickens, Matthew Jordan, and Jessica L. Cohen. Access to subsidized act and malaria treatment—evidence from the first year of the amfm program in six districts in uganda. *Health Policy and Planning*, 2013.

[35] Amy Finkelstein, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi Allen, Katherine Baicker, and Oregon Health Study Group. The oregon health insurance experiment: Evidence from the first year*. *The Quarterly Journal of Economics*, 127(3):1057–1106, 2012.

[36] Centers for Disease Control and Prevention (U.S.). Fluview interactive, Jun 2015.

[37] Kenneth Fortson, Philip Gleason, Emma Kopa, and Natalya Verbitsky-Savitz. Horseshoes, hand grenades, and treatment effects? reassessing whether nonexperimental estimators are biased. *Economics of Education Review*, 44:100 – 113, 2015.

[38] Kenneth Fortson, Natalya Verbitsky-Savitz, Emma Kopa, and Philip Gleason. Using an experimental evaluation of charter schools to test whether nonexperimental comparison group methods can replicate experimental impact estimates. ncee 2012-4019. *National Center for Education Evaluation and Regional Assistance*, 2012.

[39] Thomas Fraker and Rebecca Maynard. The adequacy of comparison group designs for evaluations of employment-related programs. *The Journal of Human Resources*, 22(2):194–227, 1987.

[40] Atle Fretheim, Stephen B Soumerai, Fang Zhang, Andrew D Oxman, and Dennis Ross-Degnan. Interrupted time-series analysis yielded an effect estimate concordant with the cluster-randomized controlled trial result. *Journal of Clinical Epidemiology*, 66(8):883–887, August 2013.

[41] Atle Fretheim, Fang Zhang, Dennis Ross-Degnan, Andrew D. Oxman, Helen Cheyne, Robbie Foy, Steve Goodacre, Jeph Herrin, Ngaire Kerse, R. James McKinlay, Adam Wright, and Stephen B. Soumerai. A reanalysis of cluster randomized trials showed interrupted time-series studies were valuable in health system evaluation. *Journal of clinical epidemiology*, Dec 2014.

[42] Rowland J. Eilerts G. Funk, C. and L. White. A climate trend analysis of uganda. Report, U.S. Geological Survey Fact Sheet 2012-3062, 2012.

[43] D Gillings, D Makuc, and E Siegel. Analysis of interrupted time series mortality trends: an example to evaluate regionalized perinatal care. *American Journal of Public Health*, 1981.

[44] Steven Glazerman, Dan M. Levy, and David Myers. Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy of Political and Social Science*, 589(63), 2003.

[45] JM Gottman and GV Glass. Analysis of interrupted time-series experiments. In T.R. Kratochwill and J.R. Levin, editors, *Single-case Research Design and Analysis: New Development for Psychology and Education*. Lawrence Erlbaum Associates, 1992.

[46] David H Greenberg, Charles Michalopoulos, and Philip K Robin. Do experimental and nonexperimental evaluations give different answers about the effectiveness of government-funded training programs? *Journal of Policy Analysis and Management*, 25(3):523–552, 2006.

[47] R. Mark Gritz and Terry Johnson. National job corps study: Assessing program effects on earnings for students achieving key program milestones. Mathematica policy research reports, Mathematica Policy Research, 2001.

[48] Joshua D. Angrist Guido W. Imbens. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

[49] JP Habicht and CG Victora. Evaluation designs for adequacy, plausibility and probability of public health programme performance and impact. *International Journal of Epidemiology*, 28:10–18, 1999.

[50] Davidson H Hamer, Erin Twohig Brooks, Katherine Semrau, Portipher Pilingana, William B MacLeod, Kazungu Siazeele, Lora L Sabin, Donald M Thea, and Kojo Yeboah-Antwi. Quality and safety of integrated community case management of malaria using rapid diagnostic

tests and pneumonia by community health workers. *Pathogens and Global Health*, 106(1):32–39, 2012.

[51] Andria Hanbury, Katherine Farley, Carl Thompson, Paul M Wilson, Duncan Chambers, and Heather Holmes. Immediate versus sustained effects: interrupted time series analysis of a tailored intervention. *Implementation Science*, 8(1):130, 2013.

[52] D P Hartmann, J M Gottman, R R Jones, W Gardner, A E Kazdin, and R S Vaught. Interrupted time-series analysis and its application to behavioral data. *Journal of applied behavior analysis*, 13(4):543–559, 1980.

[53] James J. Heckman, Robert J. Lalonde, and Jeffrey A. Smith. Chapter 31 - the economics and econometrics of active labor market programs. volume 3, Part A of *Handbook of Labor Economics*, pages 1865 – 2097. Elsevier, 1999.

[54] James J. Heckman and Jeffrey A. Smith. The pre-programme earnings dip and the determinants of participation in a social programme. implications for simple programme evaluation strategies. *The Economic Journal*, 109(457):313–348, 1999.

[55] Guido Imbens and Karthik Kalyanaraman. Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, 2011.

[56] Guido W. Imbens and Thomas Lemieux. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615 – 635, 2008. The regression discontinuity design: Theory and applications.

[57] G.W. Imbens and D.B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press, 2015.

[58] Robin Tepper Jacob, Pei Zhu, Marie-Andrée Somers, and Howard S Bloom. *A practical guide to regression discontinuity*. Citeseer, 2012.

[59] Racquel Jandoc, Andrea M. Burden, Muhammad Mamdani, Linda E. LÃ©vesque, and Suzanne M. Cadarette. Interrupted time series analysis in drug utilization research is increasing: systematic review and recommendations. *Journal of Clinical Epidemiology*, 68(8):950 – 956, 2015.

[60] Joan N Kalyango, Tobias Alfven, Stefan Peterson, Kevin Mugenyi, Charles Karamagi, and Elizeus Rutebemberwa. Integrated community case management of malaria and pneumonia increases prompt and appropriate treatment for pneumonia symptoms in children under five years in eastern uganda. *Malar J*, 12:340, 2013.

[61] Robert J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4):604–620, 1986.

[62] Anthony A. Laverty, Sarah L. Elkin, Hilary C. Watt, Christopher Millett, Louise J. Restrick, Sian Williams, Derek Bell, and Nicholas S. Hopkinson. Impact of a copd discharge care bundle on readmissions following admission with acute exacerbation: Interrupted time series analysis. *PLoS ONE*, 10(2):e0116187, 02 2015.

[63] R. Laxminarayan and H. Gelband. A global subsidy: key to affordable drugs for malaria? *Health Aff (Millwood)*, 28(4):949–961, 2009.

[64] Ariel Linden et al. Conducting interrupted time-series analysis for single-and multiple-group comparisons. *Stata J*, 15(2):480–500, 2015.

[65] Zhen-qiang Ma, Lewis H. Kuller, Monica A. Fisher, and Stephen M. Ostroff. Use of interrupted time-series method to evaluate the impact of cigarette excise tax increases in pennsylvania, 2000-2009. *Prev Chronic Dis*, 10:E169, 2013.

[66] A Mahamat, F M MacKenzie, K Brooker, D L Monnet, J P Daures, and I M Gould. Impact of infection control interventions and antibiotic use on hospital mrsa: a multivariate interrupted time-series analysis. *Int J Antimicrob Agents*, 30(2):169–76, Aug 2007.

[67] Anup Malani and Julian Reif. Accounting for Anticipation Effects: An Application to Medical Malpractice Tort Reform. NBER Working Papers 16593, National Bureau of Economic Research, Inc, December 2010.

[68] Irene M Masanja, Majige Selemani, Baraka Amuri, Dan Kajungu, Rashid Khatib, S Patrick Kachur, and Jacek Skarbinski. Increased use of malaria rapid diagnostic tests improves targeting of anti-malarial treatment in rural tanzania: implications for nationwide rollout of malaria rapid diagnostic tests. *Malar J*, 11:221, 2012.

[69] A K Mbonye, I C Bygbjerg, and P Magnussen. Prevention and treatment practices and implications for malaria control in mukono district uganda. *Journal of biosocial science*, 40(2):283–296, 2008.

[70] Whitney Newey and Kenneth West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–08, 1987.

[71] Fred Nuwaha. People's perception of malaria in mbarara, uganda. *Tropical Medicine and International Health*, 7(5):462–470, 2002.

[72] Kathryn A O'Connell, Ghazaleh Samandari, Sochea Phok, Mean Phou, Lek Dysoley, Shunmay Yeung, Henrietta Allen, and Megan Littrell. "souls of the ancestor that knock us out" and other tales. a qualitative study to identify demand-side factors influencing malaria case management in cambodia. *Malar J*, 11:335, 2012.

[73] E Odongo-Aginya, G Ssegwanyi, P Kategere, and P C Vuzi. Relationship between malaria infection intensity and rainfall pattern in entebbe peninsula, uganda. *African health sciences*, 5(3):238–245, 09 2005.

[74] Uganda Bureau of Statistics. Uganda malaria indicator survey 2009-2010. Report, ICF Macro, 2010.

[75] Robert B. Olsen and Paul T. Decker. Testing different methods of estimating the impacts of worker profiling and reemployment services systems. Mathematica policy research reports, Mathematica Policy Research, 2001.

[76] Robert B. Penfold and Fang Zhang. Use of interrupted time series analysis in evaluating health care quality improvements. *Academic Pediatrics*, 13(6, Supplement):S38 – S44, 2013. Quality Improvement in Pediatric Health Care.

[77] Mark D Perkins and David R Bell. Working without a blindfold: the critical role of diagnostics in malaria control. *Malaria Journal*, 7(Suppl 1):S5–S5, 2008.

[78] Pierre Perron. Dealing with structural breaks. In TC Mills and K Patterson, editors, *Palgrave Handbook for Econometrics: Econometric Theory, Vol 1*, pages 278–352. Palgrave, Basingstoke, UK, 2006.

[79] Steffi Pohl, Peter M Steiner, Jens Eisermann, Renate Soellner, and Thomas D Cook. Unbiased Causal Inference from an Observational Study: Results of a Within-Study Comparison. *Educational Evaluation and Policy Analysis*, 31(4):463–479, December 2009.

[80] National Malaria Control Program. Uganda national malaria control policy. Report, Ministry of Health, 2010.

[81] Craig R Ramsay, Lloyd Matowe, Roberto Grilli, Jeremy M Grimshaw, and Ruth E Thomas. Interrupted time series designs in health technology assessment: Lessons from two systematic reviews of behavior change strategies. *International Journal of Technology Assessment in Health Care*, 19(04):613–623, April 2004.

[82] John Huizinga Robert E. Cumby. Testing the autocorrelation structure of disturbances in ordinary least squares and instrumental variables regressions. *Econometrica*, 60(1):185–195, 1992.

[83] Brian Serumaga, Dennis Ross-Degnan, Anthony J Avery, Rachel A Elliott, Sumit R Majumdar, Fang Zhang, and Stephen B Soumerai. Effect of pay for performance on the management and outcomes of hypertension in the united kingdom: interrupted time series study. *BMJ*, 342, 2011.

[84] William R Shadish, Thomas D Cook, and Donald Thomas Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin College Div, 2002.

[85] WR Shadish and MH Clark. Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments . *Journal of the American Statistical Association*, 103(484), 2008.

[86] JA Smith. Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125(2005):305–353, 2005.

[87] SB Soumerai, D Ross-Degnan, S Gortmaker, and J Avorn. Withdrawing payment for nonscientific drug therapy: Intended and unexpected effects of a large-scale natural experiment. *JAMA*, 263(6):831–839, 1990.

[88] StataCorp. *Stata 14 Base Reference Manual*. Stata Press, College Station, TX, 2015.

[89] Laura C Steinhardt, Jobiba Chinkhumba, Adam Wolkon, Madalitso Luka, Misheck Luhanga, John Sande, Jessica Oyugi, Doreen Ali, Don Mathanga, and Jacek Skarbinski. Quality of malaria case management in malawi: results from a nationally representative health facility survey. *PLoS One*, 9(2):e89050, 2014.

[90] Sarah L Taubman, Heidi L Allen, Bill J Wright, Katherine Baicker, and Amy N Finkelstein. Medicaid increases emergency-department use: evidence from oregon's health insurance experiment. *Science*, 343(6168):263–8, Jan 2014.

[91] Jasperien E. van Doormaal, Patricia M.L.A. van den Bemt, Rianne J. Zaal, Antoine C.G. Egberts, Bertil W. Lenderink, Jos G.W. Kosterink, Flora M. Haaijer-Ruskamp, and Peter G.M.

Mol. The influence that electronic prescribing has on medication errors and preventable adverse drug events: an interrupted time-series study. *Journal of the American Medical Informatics Association*, 16(6):816–825, 2009.

[92] CG Victora and JP Habicht. Evidence-based public health: moving beyond randomized trials. *American Journal of Public Health*, 94(3):400–405, March 2004.

[93] AK Wagner, SB Soumerai, and F Zhang. Segmented regression analysis of interrupted time series studies in medication use research. *Journal of clinical Pharmacy and Therapeutics*, 27:299–309, 2002.

[94] Stephen G. West, Naihua Duan, Willo Pequegnat, Paul Gaist, Don C. Des Jarlais, David Holtgrave, José Szapocznik, Martin Fishbein, Bruce Rapkin, Michael Clatts, and Patricia Dolan Mullen. Alternatives to the randomized controlled trial. *Am J Public Health*, 98(8):1359–1366, Aug 2008. 18556609[pmid].

[95] Nicholas J. White, Sasithon Pukrittayakamee, Tran Tinh Hien, M. Abul Faiz, Olugbenga A. Mokuolu, and Arjen M. Dondorp. Malaria. *The Lancet*, 383(9918):723–735, 2016/04/02 2013.

[96] Elizabeth Ty Wilde and Robinson Hollister. How close is close enough? evaluating propensity score matching using data from a class size reduction experiment. *Journal of Policy Analysis and Management*, 26(3):455–477, 2007.

[97] World Bank Group. Climate change knowledge portal, 2016.

[98] Fang Zhang, Anita K Wagner, and Dennis Ross-Degnan. Simulation-based power calculation for designing interrupted time series analyses of health policy interventions. *J Clin Epidemiol*, 64(11):1252–61, Nov 2011.

[99] Fang Zhang, Anita K Wagner, Stephen B Soumerai, and Dennis Ross-Degnan. Methods for estimating confidence intervals in interrupted time series analyses of health interventions. *J Clin Epidemiol*, 62(2):143–8, Feb 2009.

Appendix A

Uganda Secondary Outcomes

Table A.1: *RCT impact estimates for BCC intervention for secondary outcomes (Uganda)*

	Impact	Control Mean	N
Malaria Testing			
Received Any Malaria Test	0.020	0.236	25,207
	(0.015)		
Received Any Malaria Test (Under 5)	0.020	0.257	8,990
	(0.022)		
Treatment Seeking			
Ever Visited Public Hospital or Clinic	-0.002	0.240	25,358
	(0.019)		
Ever Visited Private Hospital or Clinic	0.033**	0.139	25,358
	(0.015)		
Ever Visited Any Drug Shop or Pharmacy	0.014	0.401	25,358
	(0.023)		
Ever Visited Trained Drug Shop	0.038**	0.154	25,358
	(0.015)		
Ever Sought Any Care	0.034*	0.713	25,358
	(0.020)		
Medication Taking			
Took ACT	0.025	0.328	25,358
	(0.018)		
Took ACT (Under 5)	0.022	0.338	9,037
	(0.028)		
Took ACT (of those taking Antimalarial)	0.016	0.631	13,681
	(0.021)		
Took Any Antimalarial	0.027	0.520	25,358
	(0.019)		
Took Any Antibiotic	0.010	0.251	25,358
	(0.017)		

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A.2: *Naive ITS impact estimates for secondary outcomes (Uganda)*

	Time	Post	Time Post	Constant
Malaria Testing				
Received Any Malaria Test	-0.002**	0.039	-0.002	0.274***
	(0.001)	(0.024)	(0.003)	(0.018)
Received Any Malaria Test (Under 5)	-0.004***	0.055	0.001	0.342***
	(0.001)	(0.039)	(0.003)	(0.021)
Treatment Seeking				
Ever Visited Public Hospital or Clinic	-0.003***	0.010	0.005***	0.313***
	(0.001)	(0.018)	(0.002)	(0.010)
Ever Visited Private Hospital or Clinic	-0.003**	0.034*	0.002	0.205***
	(0.001)	(0.020)	(0.003)	(0.025)
Ever Visited Any Drug Shop or Pharmacy	-0.002**	0.054***	-0.008***	0.471***
	(0.001)	(0.015)	(0.002)	(0.019)
Ever Visited Trained Drug Shop	-0.002***	0.024	0.000	0.204***
	(0.001)	(0.018)	(0.001)	(0.009)
Ever Sought Any Care	-0.003***	0.068***	-0.004*	0.819***
	(0.001)	(0.022)	(0.002)	(0.007)
Medication Taking				
Took ACT	0.002**	0.030**	-0.005***	0.277***
	(0.001)	(0.012)	(0.001)	(0.015)
Took ACT (Under 5)	0.001*	0.044**	-0.007***	0.326***
	(0.000)	(0.019)	(0.002)	(0.008)
Took ACT (of those taking Antimalarial)	0.005***	-0.027	0.001	0.471***
	(0.001)	(0.028)	(0.003)	(0.014)
Took Any Antimalarial	-0.002**	0.070***	-0.007**	0.583***
	(0.001)	(0.023)	(0.003)	(0.014)
Took Any Antibiotic	0.000	0.009	-0.005**	0.278***
	(0.001)	(0.018)	(0.002)	(0.011)

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Data collapsed to the week level for 47 weeks

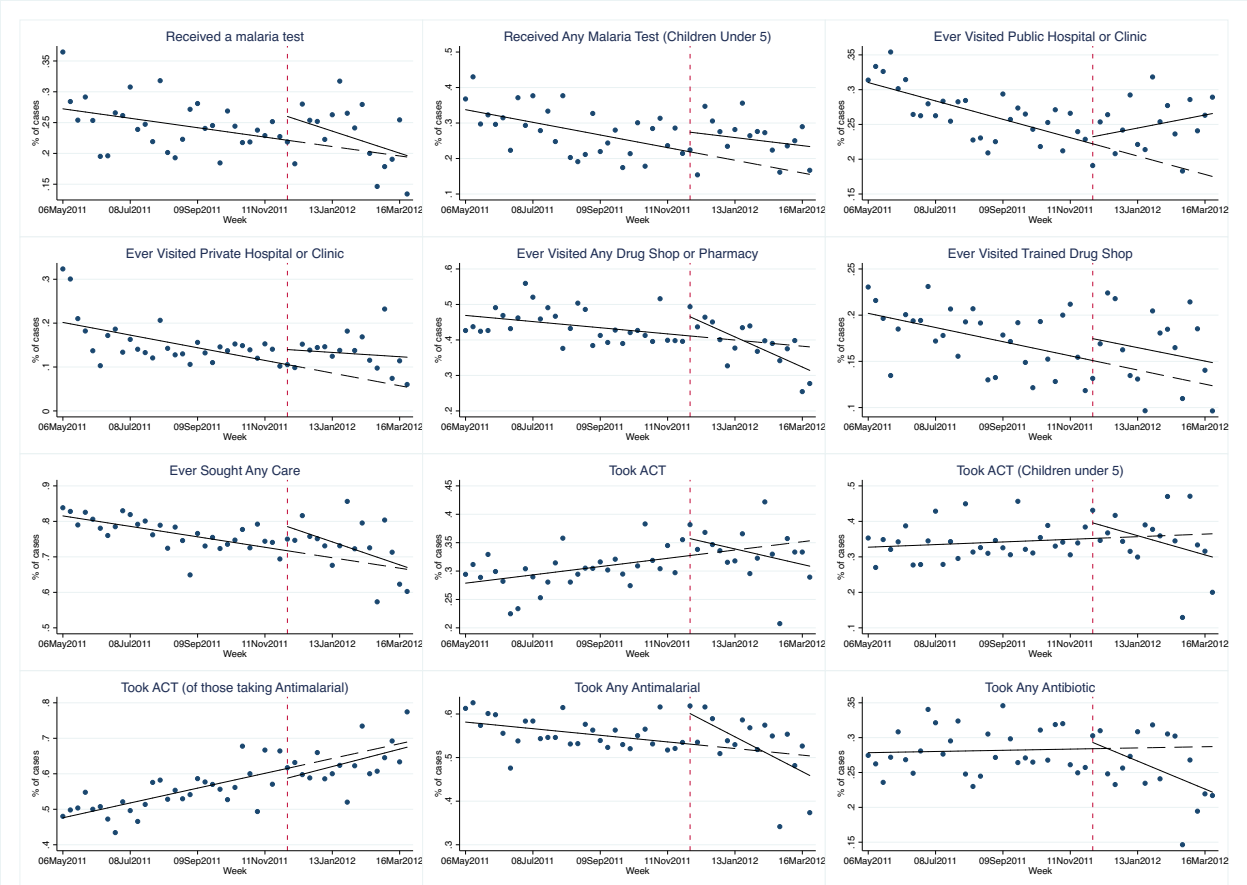

Figure A.1: *Visual representation of ITS estimates for secondary outcomes (Uganda)*

Table A.3: *Comparison of RCT and naive ITS results for secondary outcomes (Uganda)*

	ITS Impact	RCT Impact	Control Mean
Malaria Testing			
Received Any Malaria Test	0.020	0.020	0.236
	(0.020)	(0.015)	
Received Any Malaria Test (Under 5)	0.067**	0.020	0.257
	(0.031)	(0.022)	
Treatment Seeking			
Ever Visited Public Hospital or Clinic	0.050***	-0.002	0.240
	(0.016)	(0.019)	
Ever Visited Private Hospital or Clinic	0.051**	0.033**	0.139
	(0.025)	(0.015)	
Ever Visited Any Drug Shop or Pharmacy	-0.006	0.014	0.401
	(0.017)	(0.023)	
Ever Visited Trained Drug Shop	0.024	0.038**	0.154
	(0.015)	(0.015)	
Ever Sought Any Care	0.037*	0.034*	0.713
	(0.019)	(0.020)	
Medication Taking			
Took ACT	-0.007	0.025	0.328
	(0.015)	(0.018)	
Took ACT (Under 5)	-0.011	0.022	0.338
	(0.017)	(0.028)	
Took ACT (of those taking Antimalarial)	-0.021	0.016	0.631
	(0.031)	(0.021)	
Took Any Antimalarial	0.013	0.027	0.520
	(0.023)	(0.019)	
Took Any Antibiotic	-0.028	0.010	0.251
	(0.019)	(0.017)	

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Data collapsed to the week level for 47 weeks

Table A.4: *Comparison of RCT and ITS results controlling for rainfall and drug stocks for secondary outcomes (Uganda)*

	ITS				RCT
	Naive	Rainfall	Drugs	Combined	Impact
Malaria Testing					
Received Any Malaria Test	0.020	0.021	0.027	0.027	0.020
	(0.020)	(0.033)	(0.053)	(0.057)	(0.015)
Received Any Malaria Test (Under 5)	0.067**	0.042	0.056	0.055	0.020
	(0.031)	(0.046)	(0.047)	(0.054)	(0.022)
Treatment Seeking					
Ever Visited Public Hospital or Clinic	0.050***	0.074***	0.061	0.081**	-0.002
	(0.016)	(0.021)	(0.031)	(0.038)	(0.019)
Ever Visited Private Hospital or Clinic	0.051**	0.100**	0.109	0.141**	0.033**
	(0.025)	(0.045)	(0.060)	(0.063)	(0.015)
Ever Visited Any Drug Shop or Pharmacy	-0.006	-0.061*	-0.023	-0.050	0.014
	(0.017)	(0.032)	(0.042)	(0.043)	(0.023)
Ever Visited Trained Drug Shop	0.024	0.023	0.060	0.064*	0.038**
	(0.015)	(0.027)	(0.035)	(0.036)	(0.015)
Ever Sought Any Care	0.037*	0.034	0.045	0.056**	0.034*
	(0.019)	(0.027)	(0.025)	(0.027)	(0.020)
Medication Taking					
Took ACT	-0.007	0.037**	0.014	0.044**	0.025
	(0.015)	(0.016)	(0.039)	(0.021)	(0.018)
Took ACT (Under 5)	-0.011	0.008	-0.023	-0.004	0.022
	(0.017)	(0.043)	(0.040)	(0.040)	(0.028)
Took ACT (of those taking Antimalarial)	-0.021	0.033	0.007	0.043	0.016
	(0.031)	(0.038)	(0.040)	(0.036)	(0.021)
Took Any Antimalarial	0.013	0.045*	0.028	0.050*	0.027
	(0.023)	(0.025)	(0.033)	(0.029)	(0.019)
Took Any Antibiotic	-0.028	-0.057***	-0.005	-0.023	0.010
	(0.019)	(0.020)	(0.045)	(0.040)	(0.017)

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Data collapsed to the week level for 47 weeks

Table A.5: *ITS impact estimates for secondary outcomes, controlling for rainfall (Uganda)*

	Rainfall (mm)	Time	Post	Time Post	Constant
Malaria Testing					
Received Any Malaria Test	0.000	-0.002*	0.039	-0.002	0.274***
	(0.000)	(0.001)	(0.040)	(0.003)	(0.039)
Received Any Malaria Test (Under 5)	-0.000	-0.004***	0.025	0.002	0.377***
	(0.000)	(0.001)	(0.060)	(0.004)	(0.051)
Medication Taking					
Ever Visited Public Hospital or Clinic	0.000	-0.003***	0.039	0.004**	0.278***
	(0.000)	(0.001)	(0.028)	(0.002)	(0.023)
Ever Visited Private Hospital or Clinic	0.001*	-0.004***	0.094**	0.001	0.134***
	(0.000)	(0.001)	(0.043)	(0.003)	(0.034)
Ever Visited Any Drug Shop or Pharmacy	-0.001**	-0.001**	-0.013	-0.006***	0.550***
	(0.000)	(0.001)	(0.037)	(0.002)	(0.038)
Ever Visited Trained Drug Shop	-0.000	-0.002***	0.021	0.000	0.206***
	(0.000)	(0.001)	(0.030)	(0.001)	(0.025)
Ever Sought Any Care	-0.000	-0.003***	0.065*	-0.004*	0.823***
	(0.000)	(0.001)	(0.035)	(0.002)	(0.029)
Medication Taking					
Took ACT	0.001**	0.001**	0.084***	-0.006***	0.212***
	(0.000)	(0.001)	(0.023)	(0.002)	(0.034)
Took ACT (Under 5)	0.000	0.001	0.067	-0.007***	0.298***
	(0.001)	(0.001)	(0.049)	(0.002)	(0.059)
Took ACT (of those taking Antimalarial)	0.001**	0.004***	0.039	-0.001	0.392***
	(0.000)	(0.001)	(0.035)	(0.003)	(0.040)
Took Any Antimalarial	0.000*	-0.002***	0.109***	-0.008***	0.536***
	(0.000)	(0.001)	(0.034)	(0.002)	(0.027)
Took Any Antibiotic	-0.000**	0.000	-0.026	-0.004***	0.320***
	(0.000)	(0.001)	(0.024)	(0.001)	(0.021)

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Data collapsed to the week level for 47 weeks

Table A.6: *ITS impact estimates for secondary outcomes, controlling for private and public ACT stocks (Uganda)*

	Private	Public	Time	Post	Time Post	Constant
Malaria Testing						
Received Any Malaria Test	0.001	0.005	-0.003	0.040	-0.002	0.274***
	(0.004)	(0.034)	(0.008)	(0.028)	(0.005)	(0.020)
Received Any Malaria Test (Under 5)	-0.003	0.033	-0.007	0.041	0.002	0.317***
	(0.004)	(0.039)	(0.009)	(0.025)	(0.006)	(0.019)
Medication Taking						
Ever Visited Public Hospital or Clinic	0.001	0.015	-0.006	0.011	0.006*	0.308***
	(0.003)	(0.025)	(0.006)	(0.018)	(0.003)	(0.014)
Ever Visited Private Hospital or Clinic	0.006	0.039	-0.013	0.051*	0.007	0.204***
	(0.004)	(0.031)	(0.008)	(0.027)	(0.005)	(0.026)
Ever Visited Any Drug Shop or Pharmacy	-0.003	0.019	-0.003	0.041	-0.008***	0.453***
	(0.004)	(0.024)	(0.005)	(0.026)	(0.003)	(0.023)
Ever Visited Trained Drug Shop	0.002	0.045*	-0.011*	0.028**	0.004	0.190***
	(0.003)	(0.025)	(0.006)	(0.014)	(0.003)	(0.009)
Ever Sought Any Care	-0.001	0.039	-0.009	0.061***	-0.002	0.798***
	(0.003)	(0.028)	(0.006)	(0.015)	(0.004)	(0.010)
Medication Taking						
Took ACT	0.002	0.011	-0.002	0.037**	-0.003	0.279***
	(0.003)	(0.025)	(0.006)	(0.015)	(0.004)	(0.014)
Took ACT (Under 5)	-0.002	0.001	0.002	0.038	-0.008**	0.321***
	(0.003)	(0.027)	(0.006)	(0.023)	(0.004)	(0.011)
Took ACT (of those taking Antimalarial)	0.003	0.013	0.000	-0.018	0.003	0.474***
	(0.004)	(0.033)	(0.008)	(0.016)	(0.004)	(0.012)
Took Any Antimalarial	0.002	0.009	-0.004	0.075***	-0.006*	0.584***
	(0.003)	(0.025)	(0.006)	(0.016)	(0.003)	(0.014)
Took Any Antibiotic	0.002	0.027	-0.005	0.012	-0.002	0.271***
	(0.004)	(0.030)	(0.007)	(0.022)	(0.004)	(0.013)

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Data collapsed to the week level for 47 weeks

Table A.7: *Fully specified ITS model versus RCT estimates for secondary outcomes (Uganda)*

	ITS Impact	RCT Impact	Difference
Malaria Testing			
Received Any Malaria Test	0.027	0.020	0.007
	(0.057)	(0.015)	(0.038)
Received Any Malaria Test (Under 5)	0.055	0.020	0.035
	(0.054)	(0.022)	(0.061)
Treatment Seeking			
Ever Visited Public Hospital or Clinic	0.081**	-0.002	0.083**
	(0.038)	(0.019)	(0.038)
Ever Visited Private Hospital or Clinic	0.141**	0.033**	0.108***
	(0.063)	(0.015)	(0.031)
Ever Visited Any Drug Shop or Pharmacy	-0.050	0.014	-0.063
	(0.043)	(0.023)	(0.041)
Ever Visited Trained Drug Shop	0.064*	0.038**	0.026
	(0.036)	(0.015)	(0.032)
Ever Sought Any Care	0.056**	0.034*	0.022
	(0.027)	(0.020)	(0.035)
Medication Taking			
Took ACT	0.044**	0.025	0.019
	(0.021)	(0.018)	(0.038)
Took ACT (Under 5)	-0.004	0.022	-0.026
	(0.040)	(0.028)	(0.068)
Took ACT (of those taking Antimalarial)	0.043	0.016	0.027
	(0.036)	(0.021)	(0.055)
Took Any Antimalarial	0.050*	0.027	0.024
	(0.029)	(0.019)	(0.044)
Took Any Antibiotic	-0.023	0.010	-0.033
	(0.040)	(0.017)	(0.036)

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Data collapsed to the week level for 47 weeks