WORKING PAPER · NO. 2019-97

Testing the Validity of the Single Interrupted Time Series Design

Katherine Baicker and Theodore Svoronos

JULY 2019

Becker Friedman Institute | Health Economics Initiative

5757 S. University Ave.
Chicago, IL 60637
Main: 773.702.5599
bfi.uchicago.edu

TESTING THE VALIDITY OF THE SINGLE INTERRUPTED TIME SERIES DESIGN

Katherine Baicker
Theodore Svoronos

July 2019

We are grateful to Jessica Cohen and Dan Levy for their guidance and feedback throughout the creation of this paper. Theodore Svoronos received funding from Mathematica Policy Research during the preparation of this paper. Katherine Baicker is co-Principal Investigator of the original Oregon Health Insurance Experiment, supported by grants from the Assistant Secretary for Planning and Evaluation in the Department of Health and Human Services, the California HealthCare Foundation, the John D. and Catherine T. MacArthur Foundation, the National Institute on Aging (P30AG012810, RC2AGO36631 and R01AG0345151), the Robert Wood Johnson Foundation, the Sloan Foundation, the Smith Richardson Foundation, and the U.S. Social Security Administration (through grant 5 RRC 08098400-03-00 to the National Bureau of Economic Research as part of the SSA Retirement Research Consortium). We also gratefully acknowledge Centers for Medicare and Medicaid Services' matching funds for this evaluation. Baicker is a director of Eli Lilly and HMS Holdings, and serves on the U.S. Congressional Budget Office's Panel of Health Advisers. The findings and conclusions expressed are solely those of the author(s) and do not represent the views of SSA, the National Institute on Aging, the National Institutes of Health, any agency of the Federal Government, any of our funders, or the National Bureau of Economic Research.

Testing the Validity of the Single Interrupted Time Series Design
Katherine Baicker and Theodore Svoronos
July 2019
JEL No. C1,I1,I13

ABSTRACT

Given the complex relationships between patients' demographics, underlying health needs, and outcomes, establishing the causal effects of health policy and delivery interventions on health outcomes is often empirically challenging. The single interrupted time series (SITS) design has become a popular evaluation method in contexts where a randomized controlled trial is not feasible. In this paper, we formalize the structure and assumptions underlying the single ITS design and show that it is significantly more vulnerable to confounding than is often acknowledged and, as a result, can produce misleading results. We illustrate this empirically using the Oregon Health Insurance Experiment, showing that an evaluation using a single interrupted time series design instead of the randomized controlled trial would have produced large and statistically significant results of the wrong sign. We discuss the pitfalls of the SITS design, and suggest circumstances in which it is and is not likely to be reliable.

Katherine Baicker
Harris Public Policy
1155 E. 60th Street
Chicago, IL 60637
and Harvard Chan School
and also NBER
kbaicker@uchicago.edu

Theodore Svoronos
Harvard Kennedy School of Government
79 John F. Kennedy Street
Cambridge, MA 02138
theodore_svoronos@hks.harvard.edu

A randomized controlled trials registry entry is available at AEARCTR-0000028

1 Introduction

Given the complex relationships between patients' demographics, underlying health needs, access to care, and health outcomes, establishing the causal effects of health policy and delivery interventions on health outcomes is often empirically challenging. While randomized controlled trials (RCTs) remain the gold standard for assessing the unbiased causal effect of an intervention on health outcomes, there are many situations in which an RCT is deemed too complex, expensive, unethical, or simply infeasible to implement. There are a wide range of quasi-experimental techniques available that aim to address the challenges of causal inference in the absence of a randomized control group. The single interrupted time series (SITS) approach uses longitudinal data from before an intervention to construct a counterfactual (what the trend would look like in the absence of an intervention) that is compared to the actual trend seen afterwards to estimate the intervention's effect. SITS has been promoted as a plausible quasi-experimental design because of this use of longitudinal data and pre-intervention trends [11, 50, 60]. In addition, SITS is one of few quasi-experimental study designs that does not require an explicit comparison group [22, 41, 44, 57, 65, 72, 76].

The strength of this approach hinges on its ability to produce reliable estimates of program impacts, and SITS relies on a set of assumptions that must be met in order to produced unbiased results [67]. While these assumptions have been well documented, there has been little research on whether they hold in practice and on the empirical consequences of their violation. This paper provides an empirical test of the SITS design by using data from the treatment group of an RCT to construct a SITS estimate, comparing that SITS estimate to the results from the RCT itself, and using that comparison to shed light on the strengths – and hazards – of the approach.

We use data from the Oregon Health Insurance Experiment [27], a large-scale RCT begun in 2008. This study measured the effect of receiving Medicaid (a means-tested health insurance program in the United States) on a variety of outcomes. One of the central findings of this study was that receiving Medicaid substantially increased utilization of emergency department (ED) services in the treatment group relative to the randomly assigned control group [73]. In this paper, we show that, had a SITS design been used to evaluate the effect of Medicaid instead of the RCT design,

we would instead have found a large reduction in ED utilization. Not only do the SITS estimates have the wrong sign (relative to the "gold standard" RCT estimates) – they are often substantial in magnitude. Our SITS estimates are robust to multiple approaches intended to account for bias that are suggested in the SITS literature. We conclude with a discussion of the potential causes of this divergence and the implications for use of SITS by health researchers and policy analysts.

2 Within-study comparisons

2.1 Rationale

Within-study comparisons (WSCs) emerged out of an explosion of program evaluations funded by governments and non-profit organizations starting in the 1970s [12]. 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 formal derivations have outlined the conditions under which a given design performs well [67], 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 [35].

2.2 Structure

Figure 1 outlines the structure of a typical WSC. In contrast to meta-analyses and "between-study comparisons" [37], 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.

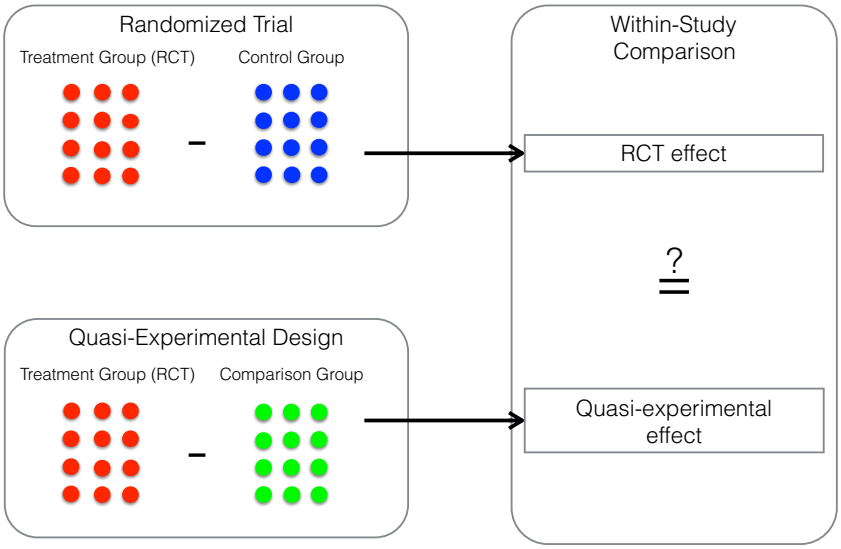

Figure 1: *Structure of a within-study comparison*

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 [19], 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 [12]. 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.

2.3 Within-study comparisons in the social science literature

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 [51]. 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 [19]. 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 [21]. 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 [69], a finding consistent with a simulation of misspecified propensity score models [23].

Many WSCs have been conducted since LaLonde's original study, focusing on various quasi-experimental techniques. These have most commonly focused on matching methods [9, 12, 14, 29–31, 35, 38, 59, 62, 68] and the regression discontinuity (RD) design [1, 10, 16, 20].

Of note to this paper, one published set of WSCs focused on the interrupted time series design [32, 33]. 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 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 interrupted time series design was found to be reliable for the selected interventions.

2.4 Metric for concordance

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 [1, 12, 14, 21, 38, 62]. Fewer studies measure mean differences in bias between quasi-experimental and experimental estimates, and compare this to a threshold considered meaningful by policymakers [23, 35, 78]. Fewer still attempt to measure the statistical significance of differences in effect sizes between designs [10, 29, 30].

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 consolidates these issues into a framework to interpret WSC findings.

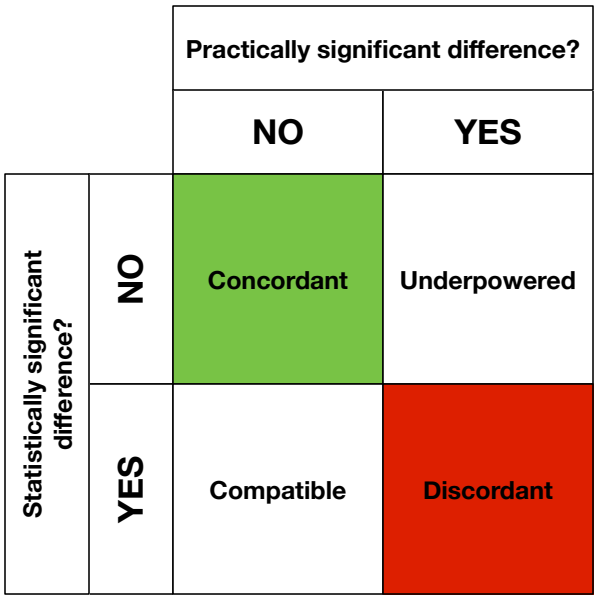

Figure 2: *Assessing concordance of RCT and quasi-experimental estimates*

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 [35, 78]. 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, we 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 conclu-

sive as a fully concordant scenario, that the quasi-experimental design is able to replicate the results of an RCT to some extent.

2.4.1 Computing standard errors

In order to determine whether differences between two designs are statistically significant, we use a bootstrapping method used in some of the more rigorous WSCs to date [10, 29, 30, 69]. 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.

3 Single Interrupted Time Series (SITS)

The focus of this paper is the single interrupted time series (SITS) design. SITS is a widely used quasi-experimental design that relies on trends before and after the introduction of a discrete

intervention to assess impact [67]. It is one of the few credible evaluation designs that is often implemented without a comparison group.

This paper 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 [36, 40]. An ITS design requires 100 or more time points to effectively leverage ARIMA techniques [24], which neither of these datasets have. Instead, we focus on ordinary least squares estimation with autocorrelated errors [53], as will be described in Section 3.1. The "single" qualifier means that my analyses do not include a non-experimental control group [67]. 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.

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 [11].

3.1 Structure

In a SITS 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 a SITS design is as follows [53, 75]:

$$Y_t = \beta_0 + \beta_1 time_t + \beta_2 post_t + \beta_3 timepost_t + u_t \tag{1}$$

$$u_t = \rho u_{t-k} + z_t \tag{2}$$

[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 [80], there have been almost no examples of it used in practice. This paper adheres to the traditional convention of time-level data.

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;
- 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 2 is a Newey-West standard error with lag k [58].

A defining characteristic of a SITS analysis is that there is not a single coefficient that represents program impact. In Equation 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 [60]. 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.

3.2 Assumptions

The single 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 [47]. 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 a SITS model:

$$Y_{t1} = \alpha_0 + \alpha_1 time_t + \alpha_2 post_t + \alpha_3 timepost_t + \epsilon_t \tag{3}$$

$$Y_{t0} = \gamma_0 + \gamma_1 time_t + \gamma_2 post_t + \gamma_3 timepost_t + \nu_t \tag{4}$$

In a single ITS context, we only have data to estimate Equation 3. We are, however, making the following implicit assumptions regarding Equation 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.*

3.3 Threats to internal validity

In discussing the threats to the internal validity of the SITS design, Shadish, Cook, and Campbell point to a number of potential threats [67]. 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 SITS design is particularly vulnerable to misspecification issues, known broadly as misspecifications of functional form.

3.3.1 Concurrent changes

The primary threat to a SITS design is the existence of changes that affect the outcome at the same time as the intervention's introduction p [67]. 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 [63]. 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 [67]:

- **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 SITS 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 3.2, 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.

3.3.2 Differential pre period changes

The threat of concurrent changes at the time of an intervention's introduction is the primary focus of most SITS 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. 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.

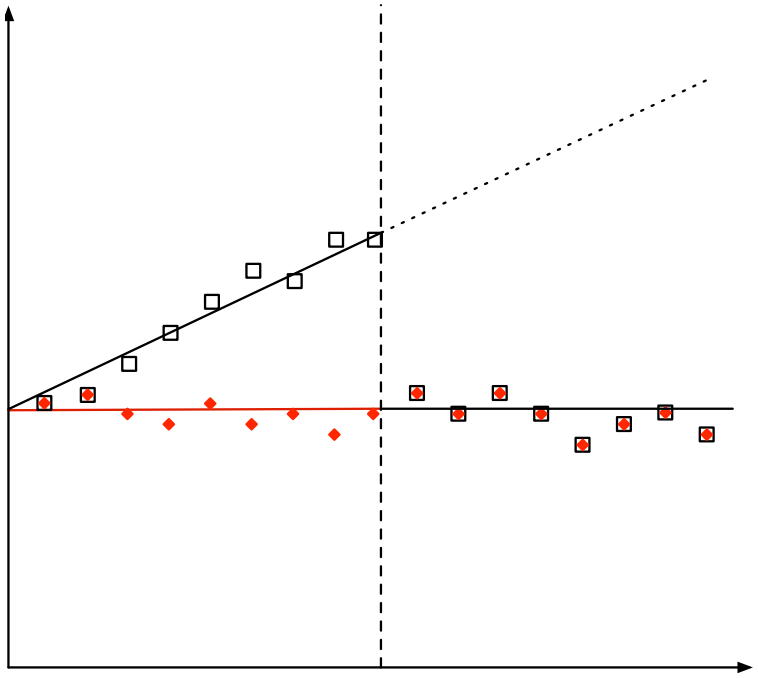

Figure 3: *Violation of Assumption 1*

Figure 3 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, the change in slope and level should both be zero (referring to Equation 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 [56]. 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**.

One particular violation of **Assumption 1**, called "Ashenfelter's dip," merits further discussion. Ashenfelter's dip refers to a scenario wherein individuals are included in a program on the basis of the outcome variable that the program aims to address [43]. 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 [6]. 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) [43]. Let the following represent the experimental estimate obtaining by taking the difference of a randomized treatment and control group in the post period:

$$E(Y_{1post}|D=1) - E(Y_{0post}|D=0) \tag{5}$$

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 5 estimates the effect of the treatment on the treated. If instead we use a simple pre-post comparison

$$E(Y_{1post}|D=1) - E(Y_{1pre}|D=1) \tag{6}$$

which, in the presence of randomization, is equivalent to

$$E(Y_{1post}|D=1) - E(Y_{0pre}|D=0) \tag{7}$$

then the bias of the pre-post estimator is the difference between Equations 5 and 7

$$E(Y_{0post}|D=0) - E(Y_{0pre}|D=0) \tag{8}$$

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 (defined as selection based on pre-treatment outcomes), 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 4 illustrates this issue. Assume both scenarios represent a dip in outcomes from the 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.

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

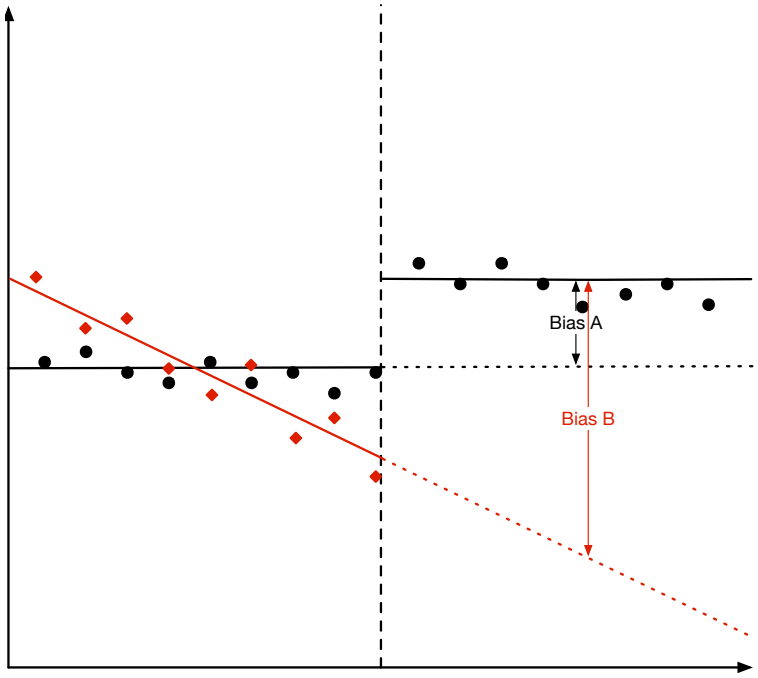

Figure 4: *Ashenfelter's Dip in a single interrupted time series design*

3.3.3 Misspecification of functional form

The final significant threat to the validity of SITS estimates is related to statistical specification. Since the "control" group of a SITS 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 [67]. 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 SITS 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 a SITS 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 SITS design provide a great deal of discretion to the researcher. As such, it is important that any within-study comparison of SITS 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 [21, 69].

3.4 Single interrupted time series in practice

3.4.1 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. We then propose a set of characteristics for SITS 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 SITS design tend to fall into one of two groups. The first are "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 [15, 26, 74], reduce antimicrobial resistance [5, 25, 55], improve patient adherence to medication [13], reduce readmissions [52], and improve referral behavior [39].

The second type of intervention are "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 [34], a national change in pharmaceutical reimbursement schedules [2, 70], a statewide excise tax imposed on cigarette sales [54], and a national pay for performance program [66]. 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 SITS than local-level ones, since local-level interventions may be less subject to the kind of concurrent events that the single ITS design is especially vulne-

rable 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 SITS design lacks a consensus as to what elements are required for a strong SITS analysis. A systematic review by Jandoc et al. [49] that focused exclusively on drug utilization research is instructive. The authors compared 220 drug utilization studies employing SITS against a common rubric of characteristics that define SITS 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 SITS 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%).

3.4.2 Current best practice in SITS analysis

Given the inconsistent way the SITS 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 5 represent my assessment of best practices of short, single SITS 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 SITS given unlimited data. That said, the vast majority of SITS designs found in the literature are of this abbreviated form.

Sections 1-3 of Figure 5 present characteristics of many high quality short single ITS studies currently in the literature [49, 75]. 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 [33]. In these cases, we list the requirement most frequently referenced in the literature.

Figure 5: *Best practice of the short, single interrupted time series design*

1. **Data** [4, 75]
    - $\geq$ 12 time points in each period ("pre" and "post")
    - $\geq$ 100 observations/units per period
    - Data collapsed to the time level
2. **Statistical Analysis** [24, 53, 63]
    - 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** [67, 75]
    - 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 SITS literature, the next section suggests two falsification tests not currently employed in SITS 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 SITS [45, 46].

3.5 Proposed falsification tests

The SITS 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, we 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 [3]. 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 [61]. The test can be represented formally as follows [71]:

$$\text{supremum } S_T = \sup_{b_1 \leq b \leq b_2} S_T(b) \tag{9}$$

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

$$y_t = x_t + (b \leq t)x_t\delta + \epsilon_t \tag{10}$$

We conduct this test using the `estat sbsingle` command in Stata 14 [71]. 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 SITS 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 [48]. 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. We determine these bins using the `rdplot` command in Stata [17]. 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

[18]. 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. We conduct this test using the `estat sbknown` command in Stata 14 [71]. 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 SITS. 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 SITS **Assumption 1**. In this case the presence of a major event - more significant than the intervention itself - suggests that SITS may not be the most appropriate study design for these data.

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

4 Effect of Health Insurance on Emergency-Department Use in Oregon

4.1 Intervention

In 2008, the state of Oregon expanded its Medicaid program to a group of previously uninsured adults via a lottery [27]. 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 [7]. 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 [27], measured physical health [7], and emergency-department usage [73].

The effect of winning the Medicaid lottery on ED use provides an ideal context in which to compare RCT and SITS approaches, since we have sufficiently granular observations of ED use over time for periods both before and after the Medicaid expansion. Administrative data were gathered from the 12 Portland-area EDs for 2007-2009 to measure the number of ED visits for the 24,646 lottery list members (9,626 selected in the lottery, 15,020 not selected) living in the area. These data show the universe of visits to those EDs, including the date of the visit.

4.2 Randomized controlled trial

4.2.1 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 [73]. 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 [73]. Emergency-department data were probabilistically matched to the individuals in the experiment on the basis of name, date of birth, and gender [7, 27, 73]. 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.

4.2.2 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:

$$y_{ih} = \beta_0 + \beta_1 LOTTERY_h + \beta_2 hhsize_h + \beta_3 preoutcome_i + \beta_4 preoutcome\_missing_i + \epsilon_{ih} \qquad (11)$$

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:

$$y_{ih} = \pi_0 + \pi_1 MEDICAID_{ih} + \pi_2 hhsize_h + \pi_3 preoutcome_i + \pi_4 preoutcome\_missing_i + \nu_{ih} \quad (12)$$

where π_1 is estimated using the first stage equation:

$$MEDICAID_{ih} = \delta_0 + \delta_1 LOTTERY_{ih} + \delta_2 hhsize_h + \delta_3 preoutcome_i + \delta_4 + preoutcome\_missing_i + \mu_{ih} \quad (13)$$

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.

4.2.3 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 1.[3] Column (1) displays ITT results (Equation 11). 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 12). 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.

[3]These results are from a reanalysis of the original Oregon data, which reproduce the published results in Taubman et al. [73].

Table 1: *RCT impact estimates*

VARIABLES	(1) ITT	(2) 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 resultsx. In order to ensure that the population of the subsequent SITS analysis is equivalent to that of the RCT, it is important to include all individuals allocated to treatment rather than just the compliers.

We treat these RCT estimates as the "truth" against which we gauge the SITS results. Of course, the RCT results are themselves subject to random noise, but there is no identifiable source of bias. When we compare the RCT results to the SITS results below, we gauge the degree to which observed differences are statistically significant.

4.3 Interrupted time series

4.3.1 Data

We include only the 9,612 individuals from the treatment group in the SITS analysis. Data are collapsed to the biweekly level instead of "pre" and "post" periods. We use biweeks as the unit of

time to allow for time trends while avoiding the noise introduced by a more granular measure.[4] 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 2).

Table 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 SITS was total number of ED visits in the sample per person, per biweek.

4.3.2 Methods

Simple specification

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

$$y_t = \zeta_0 + \zeta_1 biweek_t + \zeta_2 post_t + \zeta_3 biweekpost_t + \tau_t \tag{14}$$

$$\tau_t = \rho\tau_{t-k} + \psi_t \tag{15}$$

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

where τ_t is a Newey-West standard error with lag k [58]. The value for k is determined using the Cumby-Huizinga general test for autocorrelation in time series [64], implemented using `actest` in Stata [8]. We 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 simple SITS specification, we attempt to address each of the potential threats to validity outlined in Section 3.2.

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, we 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 SITS regression [28].

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 3.3.2. One such event is the signup period for entry into the lottery, which began at the start of 2008. The signup period for entry into the insurance lottery began in January 2008, in advance of the first lottery draw in March. During this period, hospitals may have influenced uninsured patients to sign up for the lottery, resulting in a sample that is disproportionately defined by an especially high number of ED visits in the months leading up to insurance provision.

To account for this, we treat the signup period as a separate "washout" period, and define the pre period to end in December 2007. We model the signup period explicitly in a multiple segmented regression:

$$y_t = \theta_0 + \theta_1 biweek_t + \theta_2 signup_t + \theta_3 biweeksignup_t + \theta_4 post_t + \theta_5 biweekpost_t + g_t \qquad (16)$$

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 15.

Misspecification of functional form

Finally, misspecifying the timing and dynamic of an intervention's introduction can threaten the internal validity of a single ITS design [67]. 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, we estimate a respecified SITS model which accounts for the eight different notification dates occurring between March and September. Specifically, we use the following model:

$$y_t = \lambda_0 + \lambda_1 biweek_t + \lambda_2 post_{tg} + \lambda_3 biweekpost_{tg} + h_t \tag{17}$$

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 15.

4.4 Results

The results of each specification are presented in Table 3.

Table 3: *SITS impact estimates*

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, 31 of which were pre-intervention.

4.4.1 Simple results

Column 1 of Table 3 presents the results of a simple 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 6 shows these results visually:

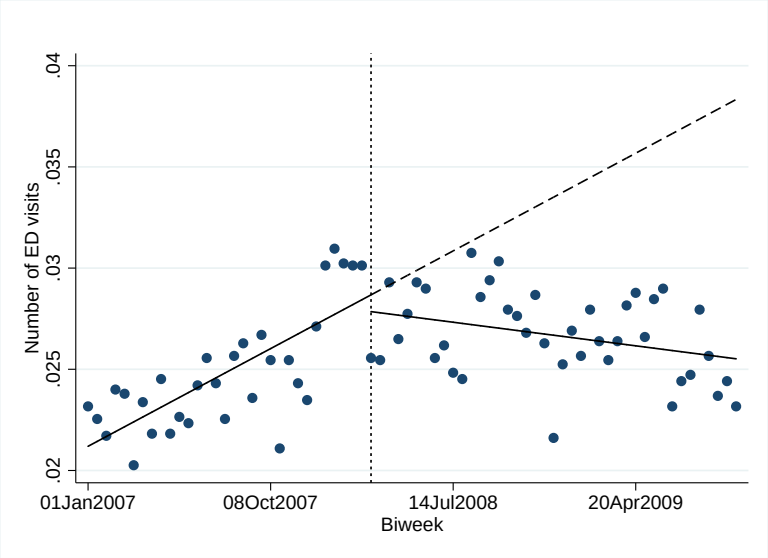

Figure 6: *Visual representation of simple SITS estimates*

The simple 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.

4.4.2 Flu season

Figure 7 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 ($F_{2,88} = 0.13, p = 0.874$).

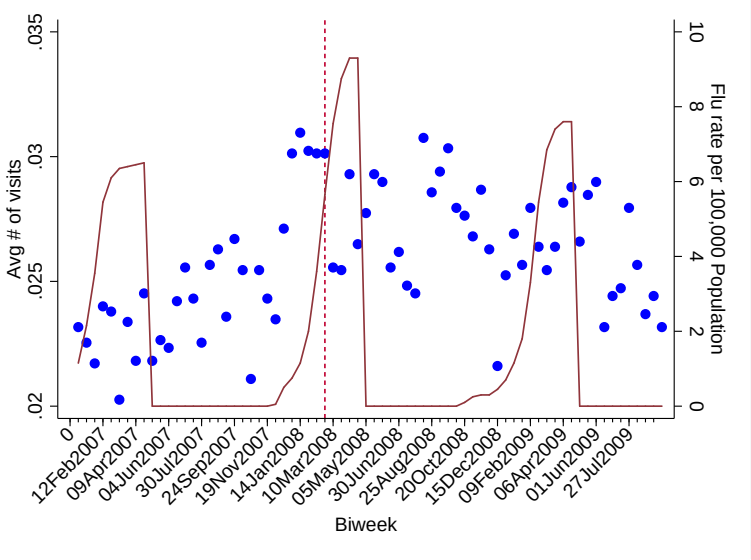

Figure 7: *ED usage and flu seasons, 2007-09*

Additionally, flu rates are not a significant predictor of ED visits, as shown in Column 2 of Table 3. Augmenting the simple regression from Equation 14 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.

4.4.3 Signup period

Column 3 of Table 3 shows that modeling the signup period using the specification from Equation 16 does change the estimates from the simple 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 8 illustrate these changes.

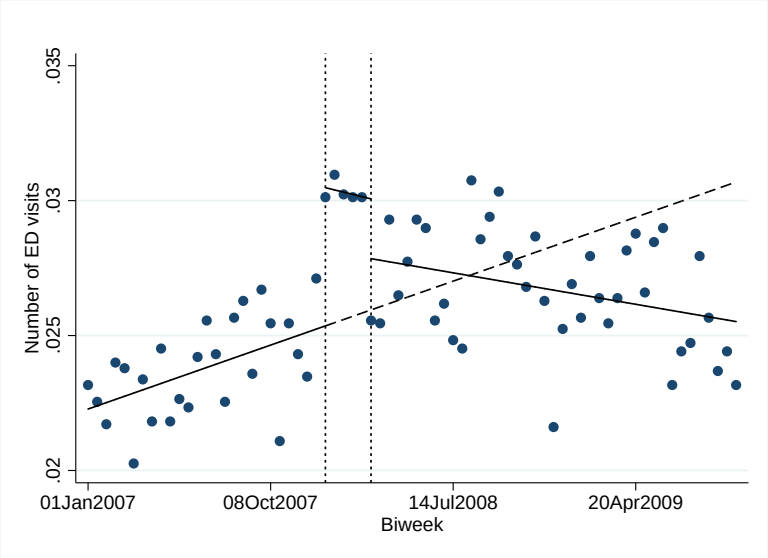

Figure 8: *Visual representation of SITS estimates with "washing out" signup period*

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 8 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.

4.4.4 Recentered specification

Recentering the specification around notification does not significantly affect SITS estimates, as shown in Column 4 of Table 3. While Figure 9 shows a level change of zero in contrast to Figure 6, the simple regression's level change is not statistically significant either.

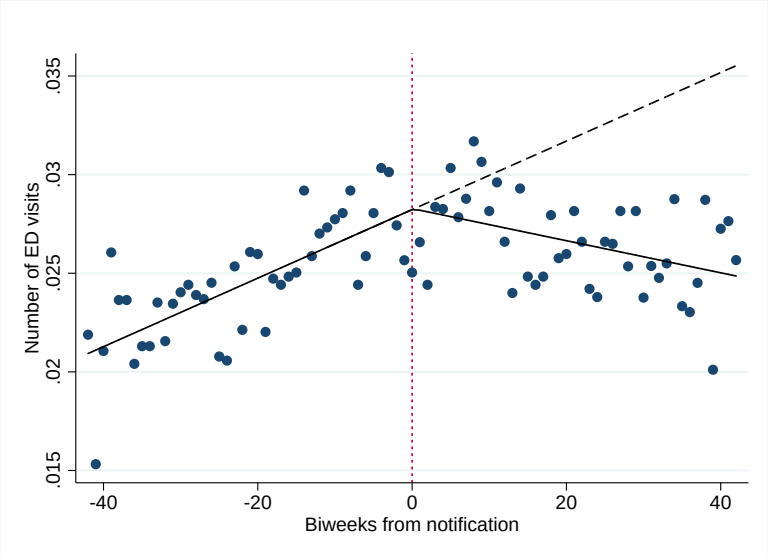

Figure 9: *Visual representation of SITS estimates with recentered specification*

The fact that the recentered specification has a negligible effect on the simple 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 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.

4.4.5 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 4 presents these results.[5] Differences between the RCT estimate and each SITS specification's estimate are presented as well. Standard errors for differences are obtained via the bootstrapping method outlined in Section 2.4.

[5]In 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 4: *SITS vs RCT estimates*

		Impact Estimate	Difference
	Randomized Controlled Trial (ITT)	0.101***	
		(0.029)	
SITS Estimates	Simple	-0.280***	-0.380***
		(0.087)	(0.112)
	Flu Season Control	-0.291***	-0.392***
		(0.104)	(0.112)
	Signup Period Wash-out	-0.069	-0.169
		(0.083)	(0.116)
	Individual-Specific Lottery Date	-0.228***	-0.329***
		(0.056)	(0.115)

Standard errors in parentheses. Standard errors of the difference in estimates was calculated by taking the standard deviation of 1,000 bootstrapped differences between the RCT and ITS results. Data taken from the Oregon Health Insurance Experiment. Observations are at the biweek level, from January 1, 2007 to September 30, 2009. This produces 72 time points, 31 of which were pre-intervention. ITS estimates of number of ED visits per person, per biweek were multiplied across the entire post period to produce estimates comparable to that of the randomized trial. Each cell in the "estimate" column is the result of a separate regression using the ITS specification described in the text. Control mean is 1.022

*** p<0.01, ** p<0.05, * p<0.1

All four SITS 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 3), in that it is statistically indistinguishable from zero. Using the framework described in Figure 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 compares the SITS estimates to the estimates of a simple pre-post comparison.

Table 5: *SITS vs Pre-Post estimates*

	Effect of Medicaid	
Randomized Controlled Trial (ITT)	.101***	
	(0.029)	
	SITS	**Pre Post**
Simple	-0.280***	0.037***
	(0.087)	(0.012)
Flu Season Control	-0.291***	0.037***
	(0.104)	(0.012)
Signup Period Wash-out	-0.069	0.041***
	(0.083)	(0.009)
Individual-Specific Lottery Date	-0.228***	0.039***
	(0.056)	(0.012)

Standard errors are in parentheses. Observations are at the biweek level, from January 1, 2007 to September 30, 2009. Each cell is the result of a separate regression using either ITS or before-after specifications. ITS estimates of number of ED visits per person, per biweek were scaled across the entire post period to produce estimates comparable to that of the randomized trial. Before-after estimates consist of biweek values collapsed to two periods and then subtracted from one another. For comparison, the randomized controlled trial estimate is .101*** (0.029).

*** 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 SITS estimator.[6] 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 a SITS estimator will produce more bias than a simple pre-post comparison.

5 Discussion

The results of this analysis paint a discouraging picture for the single ITS design. Using the framework described in Section 2.4, the presence of statistically and practically significant differences between designs suggests that the results of the single ITS design are discordant with

[6]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.

those of the RCT. Put concretely, if the state of Oregon had chosen to analyze the effect of its Medicaid expansion using SITS, the measured impact would have been statistically significant and in the wrong direction. To make things worse, this incorrect result is robust to alternative 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 3.5 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 10. 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 10), as well as the intervention point itself.

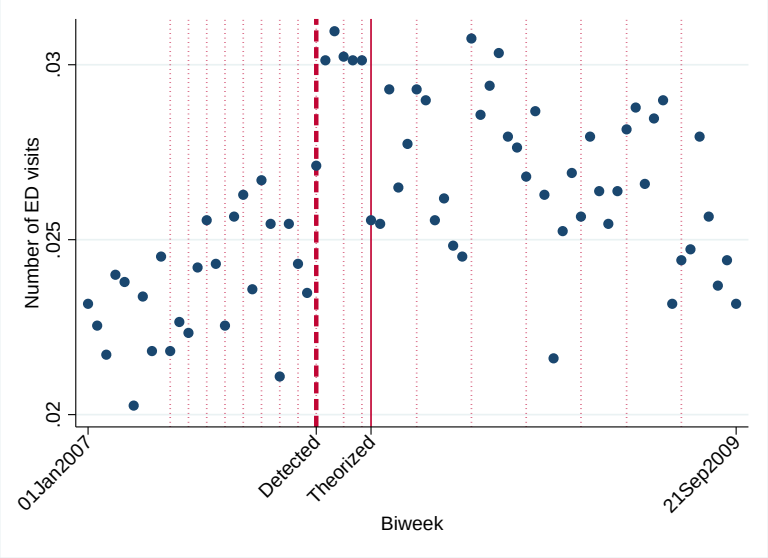

Figure 10: *Detected structural breaks*

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 11 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 a SITS analysis would detect a significant effect (e.g., 2007 in California, 2008 in Oregon, or 2009 in Washington).

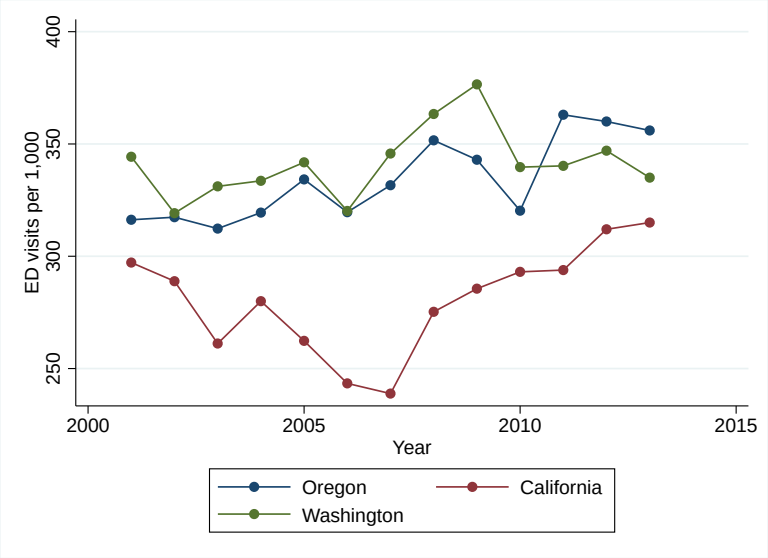

Figure 11: *Emergency department visits in Oregon, California, and Washington, 2001-2013*

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

In addition, the results of a SITS analysis for a given state at a given time produces conflicting results depending on the time horizon used. Table 6 presents an example. In this table we run a SITS 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 6: *Sensitivity of SITS 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 the individuals included in the lottery were defined by high levels of the outcome produced 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 [43]

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

5.1 Lessons for evaluators and policymakers interesting in using single ITS

5.1.1 Beware Ashenfelter's dip

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 [77]. 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 paper suggest that this approach - or any in which a sample is selected from a population of interest based on their prior outcomes - may be highly damaging to the internal validity of the single ITS design. The issues underlying Ashenfelter's dip explained earlier are readily apparent in this WSC. By allowing individuals to select voluntarily 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. 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 [42, 43].

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

5.1.2 Relying on prior trends requires additional strong assumptions

Short interrupted time series is premised on the notion that a trend of an outcome over time is preferable to a single point [11]. 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 this exercise, 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 3.5 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 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.

5.1.3 Trend stability is crucial, especially in the pre period

Much of the documented guidance regarding the appropriateness of SITS has focused on having a sufficient number of time points to allow for stable trends and the ability to model seasonality [40, 75, 79]. 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 [11, 53]. 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 3.5 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. While this is only one data point, this test suggested 11 potential break points in the pre period (see Figure 10). The trend in the pre period is especially

important in establishing a credible counterfactual, thus making the poor performance of the available data even more concerning.

Similarly, the time frame that is deemed appropriate for a SITS 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.

5.1.4 Whether to implement a SITS design is more important than how to implement it

The results of the WSC in this paper were robust to every specification attempted (see Table 4). The simple SITS model was as good (or as bad) as a fully specified one, accounting for various threats to validity outlined above. That is to say, the robustness of a SITS model to alternative specifications provided 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 this exercise shows.

Thus, the results of the analyses in this paper 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 [35, 62, 67, 69]. 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 Conclusion

So when *should* single ITS be used? In this paper we identify a number of characteristics that signal to researchers and implementers that they should be particularly wary of using a single ITS design.[7] Identifying conditions especially conducive to the design is a much more difficult (and data intensive) task than identifying reasons that the design may fail. Still, the fact that the two WSCs in this paper 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 3.5 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 for the validity of the single ITS as a possible design.

3. Samples selected based on their pre-treatment outcomes 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.

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

[7] 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 paper [67]. For example, a similarly selected control sample with the kind of spike identified in the treatment data 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 [51, 69].

References

[1] 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.

[2] 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.

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

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

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

[20] 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.

[21] 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.

[22] Jessica Dennis, Tim Ramsay, Alexis F Turgeon, and Ryan Zarychanski. Helmet legislation and admissions to hospital for cycling related head injuries in canadian provinces and territories: interrupted time series analysis. *BMJ : British Medical Journal*, 346, 2013.

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] Keith Hawton, Helen Bergen, Sue Simkin, Sue Dodd, Phil Pocock, William Bernal, David Gunnell, and Navneet Kapur. Long term effect of reduced pack sizes of paracetamol on poisoning deaths and liver transplant activity in england and wales: interrupted time series analyses. *BMJ : British Medical Journal*, 346, 2013.

[42] 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.

[43] 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.

[44] Sally Hopewell, Philippe Ravaud, Gabriel Baron, and Isabelle Boutron. Effect of editors' implementation of consort guidelines on the reporting of abstracts in high impact medical journals: interrupted time series analysis. *BMJ : British Medical Journal*, 344, 2012.

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

[46] 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.

[47] 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.

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

[49] 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.

[50] Evangelos Kontopantelis, Tim Doran, David A Springate, Iain Buchan, and David Reeves. Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. *BMJ : British Medical Journal*, 350, 2015.

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

[52] 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.

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

[54] 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.

[55] 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.

[56] 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.

[57] Anthony Matthews, Emily Herrett, Antonio Gasparrini, Tjeerd Van Staa, Ben Goldacre, Liam Smeeth, and Krishnan Bhaskaran. Impact of statin related media coverage on use of statins: interrupted time series analysis with uk primary care data. *BMJ*, 353, 2016.

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

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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: British Medical Journal*, 342, 2011.

[66] 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.

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

[68] 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.

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

[70] 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.

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

[72] Sheldon Paul Stone, Christopher Fuller, Joan Savage, Barry Cookson, Andrew Hayward, Ben Cooper, Georgia Duckworth, Susan Michie, Miranda Murray, Annette Jeanes, J Roberts, Louise Teare, and Andre Charlett. Evaluation of the national cleanyourhands campaign to reduce staphylococcus aureus bacteraemia and clostridium difficile infection in hospitals in england and wales by improved hand hygiene: four year, prospective, ecological, interrupted time series study. *BMJ : British Medical Journal*, 344, 2012.

[73] 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.

[74] 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.

[75] 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.

[76] Alexander Y Walley, Ziming Xuan, H Holly Hackman, Emily Quinn, Maya Doe-Simkins, Amy Sorensen-Alawad, Sarah Ruiz, and Al Ozonoff. Opioid overdose rates and implementation of overdose education and nasal naloxone distribution in massachusetts: interrupted time series analysis. *BMJ : British Medical Journal*, 346, 2013.

[77] 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].

[78] 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.

[79] 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.

[80] 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.