Series Name: WPS
Paper No.: 104
Issue Date: 8 Oct 2019

Inference without randomization or ignorability: A stability-controlled quasi-experiment on the prevention of tuberculosis

Chad Hazlett, Werner Maokola, and David Ami Wulf

CEGA

Center for Effective Global Action

Working Paper Series

Center for Effective Global Action
University of California

eScholarship
University of California

The CEGA Working Paper Series showcases ongoing and completed research by faculty affiliates of the Center. CEGA Working Papers employ rigorous evaluation techniques to measure the impact of large-scale social and eco-nomic development programs, and are intended to encourage discussion and feedback from the global development community.

Recommended Citation:

Hazlett, Chad; Maokola, Werner; Wulf, David Ami. (2019). Inference without randomization or ignorability: A stability-controlled quasi-experiment on the prevention of tuberculosis. Working Paper Series No. WPS-104. Center for Effective Global Action. University of California, Berkeley.

Inference without randomization or ignorability: A stability-controlled quasi-experiment on the prevention of tuberculosis

Chad Hazlett*, Werner Maokola†, David Ami Wulf‡

Abstract

When determining the effectiveness of a new treatment, randomized trials are not always possible or desirable. The stability-controlled quasi-experiment (SCQE) (Hazlett, 2019) is an observational approach that replaces the usual "no-unobserved confounding" assumption with one on the change in non-treatment outcome between successive cohorts, or the "baseline trend." We extend this method to allow variance estimation and inference, and apply it for he first time by examining whether isoniazid preventive therapy (IPT) reduced tuberculosis (TB) incidence among 26,715 HIV patients in Tanzania. After IPT became available in the clinics we studied, a non-random 25% of patients received it. Within a year, fewer than 1% of patients on IPT developed TB, compared to 16% of the untreated. Regression adjustment using available covariates produces an equally large and highly significant estimate of -15 percentage point (pp) [95% CI: -16.6, -13.7]. While those estimates may generate confidence in IPT's effectiveness, they cannot eliminate confounding. By contrast, SCQE reveals that the average treatment effect on the treated must be small and indistinguishable from zero, if we assume the baseline trend was flat over the study period. Rather, to argue that IPT was beneficial requires claiming that the (non-treatment) incidence rate rose by at least 0.5 pp per year. This is plausible, but far from certain. The SCQE approach has broad applicability and will sometimes lead to definitive claims of effectiveness. In this case, it usefully aids in protecting against over-confidence in claims that IPT was effective.

Keywords: Real world evidence, observational studies, causal inference, randomized trials, isoniazid preventative therapy, epidemiology, tuberculosis.

*Corresponding author. UCLA Statistics and Political Science; chazlett@ucla.edu. We thank Marcella Alsan, Mark Handcock, Patricia Kipnis, Frederick Sävje, and Teppei Yamamoto for valuable comments and discussion, together with participants of the 2019 Political Methodology Meeting and 2019 CEGA EASST Summit. Thanks to Patricia Valdezco for helpful edits. We are grateful to the California Center for Population Research at UCLA (CCPR) for support. CCPR receives population research infrastructure funding (P2C-HD041022) from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD). UCLA IRB approval #19-001447. The data that support the findings of this study are openly available from the Dataverse, at http://doi.org/[TBD], reference number [TBD].

†Research Coordinator, National AIDS Control Program (NACP); Department of Epidemiology, Kilimanjaro Christian Medical University College, Moshi, Tanzania

‡UCLA Statistics; amiwulf@ucla.edu

1 Introduction

Suppose a treatment becomes newly available or an effort is made to greatly increase access or uptake in a given population. Conventional statistical wisdom holds that only a randomized trial can reliably reveal the beneficial or harmful effects of that treatment. However, randomized trials may not be possible or suitable. Randomization may be unethical in some contexts, particularly if an alternative research design that does not require withholding treatment becomes available, as proposed here. Further, randomized control trials (RCTs) often study very particular populations. In the social sciences, the population drawn into a randomized experiment may be limited by feasibility or other concerns. In the case of medical trials, there may be strict eligibility criteria, often limiting study populations to those in better health and low comoborbidity, and to those willing to consent to randomization. Randomized trials also cannot answer questions about the "real world effect" of treatments, policies, or events that already took place.

Existing research designs that seek to address these concerns fall almost entirely into one of two categories. The first is a variety of designs that allow for partial self-selection, including "comprehensive cohort studies" (Olschewski and Scheurlen, 1985) and "patient preference trials" (Brewin and Bradley, 1989). These include designs in which patients' preferences may be elicited, some individuals are randomized, and some receive a treatment of their choosing. In a recently proposed patient-preference design (Knox et al., 2019), treatment preferences are elicited from all individuals, who are then randomized into two groups: one that will have their treatment assigned at random, and one that can choose their own treatment. These designs seek to solve the representational shortcomings of RCTs: The population that would consent to be randomized likely differs from those who would choose to receive the treatment. They do not, however, sidestep the need for randomized treatment and control groups, and thus do not solve the feasibility or ethical problems with experiments nor speak to the effects of treatments already made available without randomization.

Second, and more commonly, observational studies are frequently used when randomization is not an option. Under these designs, investigators as asked to assume that there are "no unobserved confounders." This is universally the assumption called upon when investigators employ covariate-adjustment approaches, regardless of the conditioning technology used (e.g. regres-

sion, weighting, matching, or sub-classification). Such approaches are, of course, susceptible to debilitating biases when unobserved confounders are present, violating this assumption.

This article employs a new alternative, the stability-controlled quasi-experiment (SCQE; Hazlett, 2019), which requires neither randomization nor the absence of unobserved confounding. It instead requires an assumed "baseline trend" in the outcome. More precisely, we assume a value or range of values for how the average *non-treatment* outcome (i.e. the outcome individuals would have, had they not taken the treatment) would change from one cohort to the next. Such an assumption alone allows us to identify the average treatment effect on the treated (ATT), *regardless of unobserved confounding.* Because the ATT tells us the average treatment effect over a group who chose or was chosen to receive the treatment, this is a very desirable quantity, especially for retrospectively assessing real world effects of a treatment on those who actually received it.

In some cases the combination of a large effect size and the ability to support a narrow assumption on the baseline trend will lead to a firm conclusion regarding the ATT estimate. For example, if a disease has long had a stable fatality rate and we see no reason for this (non-treatment) rate to change over the time period studied, a sharp estimate would be possible and credible. In cases where results are not as decisive, the approach nevertheless reveals what can and cannot reliably be concluded, or what assumptions about the baseline trend must be believed to support a particular ATT estimate. For example, if the ATT estimate changes in sign over a range of baseline trend assumptions that cannot credibly be ruled out, we learn that we cannot justify a conclusion as to the sign of the estimate without a smaller assumed range of trends. Further, the reader who wishes to argue for a certain range of effects (say, beneficial ones) must also argue convincingly for the corresponding range of baseline trends. By contrast, comparisons based on covariate adjustment typically report results as if the assumption of zero confounding holds precisely and with full confidence, risking over-confidence in the result.

This article provides the first use of the SCQE approach. In doing so, it provides a practical test of this approach, and explores its difference and equivalences to a number of other approaches that may be familiar to methodologists and practitioners. To allow for inference and hypothesis testing, not developed in Hazlett (2019), we also provide standard error estimators for a variety of scenarios. We use this approach to estimate the effectiveness of isoniazid preventive therapy (IPT) on preventing tuberculosis (TB) among people living with HIV who visit

health clinics in Tanzania. As is often the case, there is little reason to believe that observed covariates are sufficient to rule out unobserved confounding, particularly since few covariates are available to us. We find a stark contrast between what policymakers are likely to conclude based on standard approaches, and what can be said after applying SCQE. For example, a naive comparison tells us that once IPT was made available, 16% of untreated patients developed TB compared to fewer than 1% of those on IPT. Similarly, regression-based covariate adjustment produces an estimate of -15 percentage points (pp) with a narrow confidence interval (-16.4, -13.7) and with a t-statistic of greater than 22. The superficially impressive scale and significance of these results exacerbates the risk that policymakers believe these estimates to be reliable evidence of a causal effect, despite any warnings we attempt to give regarding the severe threat of unobserved confounding. Applying SCQE, we find treatment effects that are not as compelling. In fact, both positive and negative treatment effect estimates can be supported under reasonable assumptions. For example, under the assumption of a baseline trend equal to zero, the ATT estimate is -2 pp, but with a wide confidence interval including zero (-10, 5). We further show results under a wider range of choices for δ, informed both by expert opinion and existing trends in the data. This ability to recast a dramatic but potentially misleading naive or covariate-adjusted estimate of -15 pp into a range of ATT estimates that do not exclude zero is a powerful tool, both to avoid over-confidence in results built on infeasible assumptions and to correctly characterize what we can and cannot validly claim about a causal effect.

In what follows, Section 2 describes the proposed method and the inferential extensions developed here. Section 3 describes the application in greater detail and gives results. Section 4 discusses, compares the approach to other identification strategies, and concludes.

2 Stability-controlled quasi-experiments

This section provides methodological details of the approach. Though not given a name in Hazlett (2019), we propose to call this method the stability-controlled trial (SCT) when it is deployed by design, and the stability-controlled quasi-experiment (SCQE) when it is applied retrospectively. This application is of the latter type.

2.1 Setup

We use the potential outcomes framework (Neyman, 1923), in which, for each individual indexed by i from 1 to N, we can write the outcome the individual would have had under treatment ($Y_i(1)$) and the outcome they would have had under non-treatment ($Y_i(0)$), regardless of their actual realized treatment status ($D_i \in \{0, 1\}$). We consider two time periods: $T_i = 0$ for those individuals observed before the treatment becomes available, and $T_i = 1$ for those observed afterwards. Note that the cohorts observed in this framework are assumed to be separate groups, not repeated measures as in a panel. The sample (including potential outcomes) $\{Y_i(1), Y_i(0), D_i, T_i\}_{i=1}^N$ is assumed to be drawn independently from common joint density $p(Y(1), Y(0), D, T)$. For notational ease we often suppress the index i when referring to the common distribution, e.g. $\mathbb{E}[Y(0)|T = 1]$. Finally, we denote the proportion of individuals taking the treatment at time $T = 1$ as $\pi_1 \equiv Pr(D = 1|T = 1)$ and the proportion taking it at time $T = 0$ as $\pi_0 \equiv Pr(D = 1|T = 0)$. For this application, we can limit attention to the simpler case in which the treatment is newly introduced, and thus nobody in the first cohort receives it ($\pi_0 = 0$).[1]

The key assumption required is a postulated value or range of values for the shift in the expected non-treatment potential outcome between the pre-treatment and post-treatment cohorts, which we call δ,

$$\delta \equiv \mathbb{E}[Y(0)|T = 1] - \mathbb{E}[Y(0)|T = 0]. \tag{1}$$

While various sources of information may be useful to inform beliefs about δ, as explored here, it is fundamentally unknowable. If the outcome historically followed a stable and consistent trend, and subject matter experts agree that nothing else able to influence outcomes changed over this time (besides the treatment introduction in question), then a δ representing a continuation of that trend may be a reasonable assumption. A choice of $\delta = 0$ states that the average outcome would not be expected to change at all, in the absence of the new treatment.

With an assumed δ, the ATT is identifiable as follows. We have an unbiased estimate for the non-treatment outcome among the whole group in time period one, using the mean observed (non-treatment) outcome in time period zero, shifted by δ. This group average is, in turn, a

[1] Hazlett (2019) also generalizes this method to cases where $\pi_0 > 0$.

weighted combination of two other averages: the average non-treatment outcome among the untreated, which we observe, and the average non-treatment outcome among the treated, for which we can solve by applying the law of iterated expectations. That is,

$$\begin{aligned}\mathbb{E}[Y(0)|T=0] &= \mathbb{E}[Y(0)|T=1] - \delta \\ &= \mathbb{E}[Y(0)|D=1,T=1]\pi_1 \\ &\quad + \mathbb{E}[Y(0)|D=0,T=1](1-\pi_1) - \delta,\end{aligned}$$

which we can re-arrange to identify the important counterfactual quantity, $\mathbb{E}[Y(0)|D=1,T=1]$ in terms of observables,

$$\begin{aligned}\mathbb{E}[Y(0)|D=1,T=1] &= \frac{\mathbb{E}[Y(0)|T=0] - \mathbb{E}[Y(0)|D=0,T=1](1-\pi_1) + \delta}{\pi_1} \\ &= \frac{\mathbb{E}[Y|T=0] - \mathbb{E}[Y|D=0,T=1](1-\pi_1) + \delta}{\pi_1}.\end{aligned} \tag{2}$$

Identification of the average non-treatment outcome among the treated in Equation (2) can be of direct interest in many contexts, such as analysis of the real world use of a treatment, because it tells us "who" is receiving treatment in terms of how they would have done in the absence of the treatment. It also leads directly to the ATT, $\mathbb{E}[Y(1) - Y(0)|D=1,T=1]$, which is identifiable and given by

$$\begin{aligned}ATT &= \mathbb{E}[Y(1)|D=1,T=1] - \mathbb{E}[Y(0)|D=1,T=1] \\ &= \mathbb{E}[Y|D=1,T=1] - \left(\frac{\mathbb{E}[Y|T=0] - \mathbb{E}[Y|D=0,T=1](1-\pi_1) + \delta}{\pi_1}\right).\end{aligned} \tag{3}$$

2.2 Instrumental variables and variance estimation

An equivalence that will prove useful momentarily in deriving standard errors is that this estimator is equal to a modification of the Wald estimator with an adjustment due to a given choice

of δ,

$$\begin{aligned} ATT &= \mathbb{E}[Y|D=1,T=1] - \left(\frac{\mathbb{E}[Y|T=0] - \mathbb{E}[Y|D=0,T=1](1-\pi_1) - \delta}{\pi_1}\right) \\ &= \frac{1}{\pi_1}\left(\pi_1\mathbb{E}[Y|D=1,T=1] + (1-\pi_1)\mathbb{E}[Y|D=0,T=1] - \mathbb{E}[Y|T=0] - \delta\right) \\ &= \frac{\mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0] - \delta}{\pi_1}. \end{aligned} \tag{4}$$

This formulation suggests that time can be thought of as an instrumental variable, or more intuitively an "encouragement" to receive the treatment, because the cohort at time $T=1$ has a higher probability of taking treatment than the cohort at time $T=0$.[2] A conventional instrumental variable must satisfy two assumptions. The first is the exclusion restriction: the instrument (time) does not cause the outcome to change except through treatment. Second is that the relationship between the instrument (time) and the outcome is not confounded by unobserved common causes. Violating either has the consequence that $\mathbb{E}[Y(0)|T=1] \neq \mathbb{E}[Y(0)|T=0]$. It is this relationship that is relaxed by incorporating δ, instead allowing $\mathbb{E}[Y(0)|T=1] = \mathbb{E}[Y(0)|T=0] + \delta$. That is, δ makes up for any differences between the two cohorts in terms of the expected non-treatment outcomes, whether that difference arises through confounding of the outcome with time or a direct effect of time on the outcome.

A brief intuition for the Wald formulation is as follows. Consider a pseudo-outcome $\tilde{Y}$ that subtracts δ from the outcome of units in the second cohort, i.e. $\tilde{Y}_i = Y_i - \delta T_i$, with potential outcomes $\tilde{Y}_i(0)$ and $\tilde{Y}_i(1)$. While the expectation of $Y(0)$ can differ by δ conditionally on the time, it is the case that $\tilde{Y}_i(0) \perp\!\!\!\perp T_i$. Any difference seen in the expected $\tilde{Y}_i$ between cohorts cannot then be due to differences in their non-treatment outcome, but rather is generated by the treatment effect. This "reduced form" effect (difference in δ-adjusted average outcomes between $T=1$ and $T=0$) is the numerator in Equation (4). Since, under the exclusion restriction, all of this difference must be caused only by the subset of units that were treated, dividing by the proportion treated (π_1) recovers "how large the treatment effect must be for each of the treated on average," i.e. the ATT.[3]

[2] We refer readers to Baiocchi, Cheng and Small (2014) for a recent tutorial on instrumental variables in medicine.

[3] More generally instrumental variable approaches identify the treatment effect among "compliers," i.e. those who would have received the treatment if they appeared in the second cohort but would not have if they appeared in the first. However, in this case this is simply the ATT: the assumption that the treatment is newly available ($\pi_0 = 0$) implies "one-sided non-compliance," under which the treated are representative of the compliers.

This re-expression of SCQE also suggests a natural approach to variance estimation. Consider the "instrumental variable regression" using our pseudo-outcome $\tilde{Y}_i$,

$$\tilde{Y}_i = \beta_0 + \beta_{IV} D_i + \mu_i, \tag{5}$$

in which β_{IV} is the (δ-adjusted) instrumental variable estimate in which time (T) was used as an instrument for treatment (D). The fitted residuals $\hat{\mu}$ are then obtained by $\hat{\mu}_i = \tilde{Y}_i - \hat{\beta}_0 - \hat{\beta}_{IV} D_i$. Assuming spherical errors, $Var(\mu) = \Sigma = \sigma^2 I_n$, an asymptotically valid variance estimator for $\hat{\beta}_{IV}$ is given by (Wooldridge, 2009),

$$\widehat{SE}(\hat{\beta}_{IV}) = \frac{\hat{\sigma}_\mu}{\sqrt{N}\,\hat{\rho}_{D,T}\,\hat{\sigma}_D}, \tag{6}$$

where $\hat{\sigma}_\mu = \sum_i \hat{\mu}_i^2/(n-3)$, $\hat{\rho}_{D,T}$ is the sample correlation of the treatment and time indicators, and $\hat{\sigma}_D$ is the sample standard deviation of the treatment indicator.[4] Further, the more general estimator for the standard error, without assuming $\Sigma = \sigma^2 I$, is given by the "sandwich estimator" form, $(\mathbf{Z}^\top \mathbf{D})^{-1} \mathbf{Z}^\top \hat{\mathbf{\Sigma}} \mathbf{Z} (\mathbf{Z}^\top \mathbf{D})^{-1}$, where $\mathbf{Z} = [1\ T]$ and $\mathbf{D} = [1\ D]$, and where $\hat{\mathbf{\Sigma}}$ is a consistent estimator for $Var(\mu)$, which can be constructed for cluster-robust, heteroskedasticity-consistent, or other specialized error covariance structures. In particular, in the clinic-level estimates given below we will employ the heteroskedasticity-robust form $\hat{\mathbf{\Sigma}} = diag(\hat{\mu}^2)$ (Cameron and Trivedi, 2005).

If our units are clustered into groups and we wish to avoid assumptions of independence within groups, we could apply a cluster-robust standard error under the IV approach. Alternatively, the bootstrap is valid for instrumental variables (Freedman et al., 1984). We use a block bootstrap to obtain confidence intervals for our estimates when we pool across clinics, below, wishing not to impose assumptions on within-cluster dependency. That is, we pool the (δ-adjusted) data across clinics before constructing the ATT estimate. We then construct standard errors by resampling (with replacement) at the clinic level, re-estimating the ATT on each iterate. We construct 95% confidence intervals from the 2.5^{th} and 97.5^{th} percentile of the bootstrap estimates.

[4] Note that Wooldridge (2009) proposes $n-2$ as the finite sample correction for $\hat{\sigma}_\mu$, but software such as the `ivreg` function in `AER` for `R` use $n-3$.

3 Application: Evaluating IPT for TB prevention

3.1 Isoniazid preventive therapy

Tanzania is experiencing a major crisis in TB prevalence and mortality. For those already immunocompromised by HIV, developing an active TB infection is both more likely and more fatal. Isoniazid, an antibiotic, has long been used in the treatment of active TB. The prophylactic use of isoniazid to prevent latent TB from developing into active TB is referred to as isoniazid preventative therapy (IPT). Randomized trials have shown the effectiveness of isoniazid in combination with other agents to treat active TB (see Fox, Ellard and Mitchison, 1999 for review), and more importantly here, the efficacy of IPT in preventing it (see Smieja et al., 1999; Akolo et al., 2010 for reviews). As a result, the World Health Organization strongly encourages the use of IPT to prevent active TB in those immunocompromised by HIV, even in settings where testing for latent TB cannot be provided (WHO, 2008). However, we are not aware of any evidence as to the actual effectiveness of national IPT-promoting policies in developing country settings.

Beginning in 2011, Tanzania has been making IPT available in HIV clinics and encouraging its use through a nationwide clinician education program for the prevention of active TB development. Groups of individual clinics are selected in waves, and their clinicians receive educational training encouraging IPT prescription for *all* HIV patients not yet diagnosed with active TB. Clinics were enrolled in the program through 2017, incrementally increasing the number of clinics using IPT and the number of patients given the treatment nationwide. By the end of that period, more than a third of the 318 HIV clinics had been enrolled in the program. Prior to these trainings, although isoniazid was formally a standard part of care in these clinics for patients with active TB, we did not find any use of IPT. Following the trainings, while IPT was universally recommended, we find that it was still used only 25% of the time, in the clinics that adopted it at all. Table 1 shows the times at which IPT use effectively began, the TB development rates before and after this, and the subsequent levels of IPT uptake, in each of the 21 clinics included in the analysis below.

We have little information about the process by which certain clinics are chosen rather than others. More importantly, we cannot hope to obtain a full understanding of the process by which certain patients receive IPT while others do not. Despite having a medical doctor intimately

familiar with this program as an author of this paper, we see no hope for a defensible claim that conditioning on any set of observed covariates would render the treatment unconfounded.

When estimating the effect of IPT, in principle each clinic provides an opportunity for a clinic-specific estimate. Alternatively, we can aim to construct a single, nationwide effect estimate by pooling together patients across clinics. We discuss both estimates below, though we rely mainly on the pooled estimator as a consequence of limited sample sizes. In either case, the unit of interest is a patient, the outcome is whether or not the patient was eventually diagnosed with TB, and the treatment is the prescription of IPT.[5]

Finally, recall that SCQE will estimate an average effect of IPT on TB incidence, among those who opted to receive it. This is to be distinguished from efforts to estimate the *efficacy* of IPT in preventing TB (as in a randomized trial), or alternatively about the effectiveness of the program as a whole on all HIV-positive patients. The estimand from SCQE is thus most relevant for an analysis seeking to measure what the actual effect of IPT has been specifically on those prescribed it.

3.2 Inclusion criteria and coding rules

Our initial dataset consists of electronic medical records from all Tanzanian HIV clinics, from 2012 to September 2017, covering over 5.9 million patient visits. A number of choices must be made to determine the time at which treatment became available, to construct cohorts of patients at each clinic who are observed before or after IPT is available, and finally to determine the treatment status and outcome for each individual.

We must first know when each clinic began prescribing IPT. We were not given, and there does not appear to exist, any record of when staff from specific clinics attended trainings on the use of IPT, or when they returned to their clinics to begin using it. Instead, this had to be inferred from the data by observing each clinic-specific time-series of individual IPT prescriptions. In each clinic, IPT prescriptions did begin fairly abruptly, but with occasional cases of IPT being used much earlier, perhaps due to coding errors. We chose the 2^{nd} percentile of IPT prescription dates as our indicator for when IPT began, which effectively aligned this date with the clear spike in initial use at each clinic, and removed all IPT uses before that date.

[5]Throughout the paper, the TB outcome refers to development of *active* TB rather than acquisition of latent TB. Prescriptions of IPT after a diagnosis of, and treatment for, active TB were excluded.

If any of these "ignored" IPT uses before that date were genuine, some effect of the treatment could be improperly experienced in the $T = 0$ period. Likewise, there can be coding errors whereby a patient is coded as treated in $T = 1$ when they should not have been. Both of these problems would lead to an understatement of the ATT.[6]

In constructing the cohorts, it is important to avoid procedures that would generate compositional differences between these groups, which could generate differences in their average non-treatment outcomes. With this in mind, we limit our data to (a) the first year of clinic visits for patients, who (b) are found in the data for at least that year, (c) that whole year of which was contained within either the pre-IPT or post-IPT period. Rules (a) and (b) ensure that we are looking at "new patients" in both cohorts, for whom a year's worth of data are available. This avoids picking systematically "older" patients in the latter cohort and prevents differential left- or right-censorship in the two cohorts. Rule (c) ensures patient-observations are limited to time periods entirely before or entirely after the introduction of IPT to avoid "crossovers."

To determine each patient's outcome, we use a follow-up period of M_Y months. We must also define an eligibility window during which we will consider a patient to have received the treatment, M_D. Here we set $M_D = M_y = 12$ months, which informed the use of the 1-year periods mentioned above in cohort construction. That is, we follow a patient for a year from their first visit to determine whether they receive IPT in that time, and whether they develop TB in that time.[7]

Finally, to be eligible for analysis, a clinic must have at least 100 patients in each cohort, with at least 10% of patients in the post-treatment cohort receiving IPT. We thus required that at least 10 patients were treated in the post-IPT period in each included clinic.[8]

[6]The effect estimates presented here showed no appreciable change when using the 1^{st} and 5^{th} percentile of dates to determine IPT eligibility, and in practice these dates are very close to each other as the amount of IPT-use increases very rapidly once it truly begins in a clinic.

[7]In some cases the analyst may wish to set M_D to be shorter than M_Y, e.g. they may code an individual's treatment status based on their first month ($M_D = 1$ month), but allow an additional 11 months ($M_Y = 12$ months) of observation on the outcome. The tradeoff is that this would give a longer follow-up window relative to treatment, but codes somebody who received the treatment in month 2 or later as if they are untreated. Our choice of $M_D = M_Y = 12$ months has the downside of allowing some individuals a very short time to see a benefit after taking treatment, but the upside of not coding those who received treatment before the end of the year as untreated. Other choices may be reasonable as well. In Figure A.1 in the Appendix we show the results we would obtain had we instead used $M_D = 6$ months and $M_Y = 12$ months. These results do not materially differ, though this does reduce the number of clinics with sufficient data from 21 down to 15.

[8]With these criteria, the minimum possible F-statistic one would get for a first-stage regression (of treatment status on the post-treatment indicator) would be 11, favorably comparing to the traditional guideline of 10 (Stock and Yogo, 2002). In practice our clinic-level F-statistics ranged from 26 to 1356.

Table 1: Clinic Implementation of IPT

Clinic Number	Total Patients	Pre-Implementation: TB Rate	Post-Implementation: TB Rate	Post-Implementation: IPT Rate	IPT Implementation Date
1	3,031	0.15	0.18	0.16	2014-06-06
2	2,457	0.19	0.21	0.25	2014-06-10
3	2,448	0.14	0.12	0.26	2014-05-27
4	2,053	0.12	0.09	0.27	2014-06-26
5	1,610	0.04	0.04	0.23	2014-09-15
6	1,602	0.12	0.11	0.20	2015-09-03
7	1,569	0.18	0.12	0.26	2015-05-13
8	1,406	0.01	0.05	0.39	2015-01-19
9	1,382	0.13	0.14	0.23	2014-06-23
10	1,186	0.27	0.18	0.23	2015-09-22
11	1,035	0.00	0.01	0.14	2015-09-14
12	962	0.06	0.08	0.16	2015-03-16
13	946	0.17	0.10	0.22	2015-03-23
14	895	0.21	0.15	0.33	2015-11-17
15	818	0.03	0.04	0.15	2015-12-18
16	688	0.12	0.16	0.51	2015-03-23
17	638	0.06	0.07	0.73	2016-01-04
18	591	0.30	0.13	0.24	2015-03-13
19	490	0.10	0.07	0.32	2015-03-10
20	485	0.13	0.08	0.19	2015-09-14
21	423	0.02	0.02	0.11	2015-01-19
Mean	1,272	0.13	0.12	0.25	
Total	26,715				

Note: Implementation details for the 21 clinics that qualify for an ATT estimate, as defined in Section 3.2. Total patients is the number of pre- and post-implementation patients used, following the same criteria.

3.3 Specifying plausible ranges for δ

Where should our beliefs about δ come from? Ideally, domain knowledge would provide a strong claim as to the appropriate value or range of values. One of the authors (Dr. Maokola) is an expert on this topic and we documented his beliefs about δ prior to examining the data. He noted that there were no known changes in TB incidence rates in recent years or any medical or epidemiological reasons to expect a change, but that ongoing efforts to increase the amount of testing and treatment for TB in the population we studied could have increased reporting rates slightly. His best guess for the non-treatment outcome trend was an absolute increase of 0.5pp to 1pp per year. He indicated that he was not confident in the coverage of this range, though, and encouraged a data-driven approach.

Note that rather than impose a distributional belief over δ, we consider only its extremes by repeating the analysis at the high- and low-ends of the allowed interval. If one thinks of δ as a random variable with a distribution, then one could instead determine an appropriate posterior distribution for the ATT. This may be a reasonable approach that appeals in some contexts, depending on one's purpose. For present purposes, we want a result that communicates the extreme boundaries of what ATT estimates users must consider plausible given their assumptions on δ, not a distribution. We thus rely only upon assumed boundary points that define the range of δ considered plausible.

In addition to domain knowledge, data may inform beliefs about δ. In estimating such a "prior baseline trend," we can either use data from clinics that did not employ IPT but in the same time range as those that did, or we can employ data from the clinics that did use IPT, but looking only to the time points prior to IPT. The former is more informative if we believe that secular or "calendar time" trends in TB are being experienced similarly by all clinics. The latter is more informative if we believe that while IPT-adopting and non-adopting clinics may differ in their trends, one finds stable trends over time within the clinics that adopt IPT. Absent any strong assumption on which of these is preferable, we choose to combine all available data in order to capture as much information as possible.

This lack of certainty about how a baseline trend should be estimated to best approximate δ emphasizes that δ is not identifiable from the data without additional assumptions. Rather, the data can aid in determining a range of choices for δ that is consistent with prior observation. Notably, Dr. Maokola's belief that there were no major changes (besides the treatment) over these years in TB prevention or management, other than a potential increase in reporting, adds reason to believe that the trends found in TB rates elsewhere in the data might be informative. However, even without trusting that a data-driven choice of δ is correct, simply knowing it is reasonable tells us that the corresponding ATT estimate cannot easily be ruled out. For instance, if the data showed a 1pp increase in TB per year, the assumption that the true δ is near 1pp per year is certainly not unreasonable, and neither is the resultant ATT estimate. In this way, data can be a powerful tool in determining which ATT estimates cannot be ruled out without further argument.

Beyond choosing which clinics and timepoints to include in our estimates of the prior baseline trends, we must also make parametric modeling decisions. We can conceive of the trend over time

as a linear one, or as an exponential or other non-linear rate (though the latter is then translated back into an absolute shift to apply the appropriate δ). An exponential trend is particularly reasonable given that the TB incidence rate is near zero in some clinics. For the linear estimate, we regressed a binary indicator of whether a patient developed TB or not on the date of their first visit. This was done for all non-implementing clinics, and all implementing clinics prior to the date of implementation. Clinic level intercepts were included as fixed effects. The resulting estimate was a yearly shift of -0.0029 (i.e. almost a 0.3pp drop), with a 95% CI of [-0.00068, -0.0052]. By multiplying the linear estimate by the time between the pre-implementation and post-implementation periods, we obtained a value of δ to be used in Equation (3). For the exponential decay estimate, we ran a binomial regression with a log link using the same terms as the linear regression, which produced a daily decay rate of 0.99980, or a yearly decay rate of 0.93 (95% CI [0.89, 0.97]).[9] Exponentiating the decay rate by the pre-to-post time, we got the correct effective decay rate for $\mathbb{E}[Y_0]$, which was be combined with $\mathbb{E}[Y_0|T = 0]$ to obtain the corresponding value of δ for use in Equation (3). For example, given the average of the TB development rates in the pre-implementation period of 0.13, the decay maps to a one-year absolute change in $\mathbb{E}[Y_0]$ of -0.0088 (95% CI [-0.0038, -0.0136]).[10]

3.4 Clinic level estimates

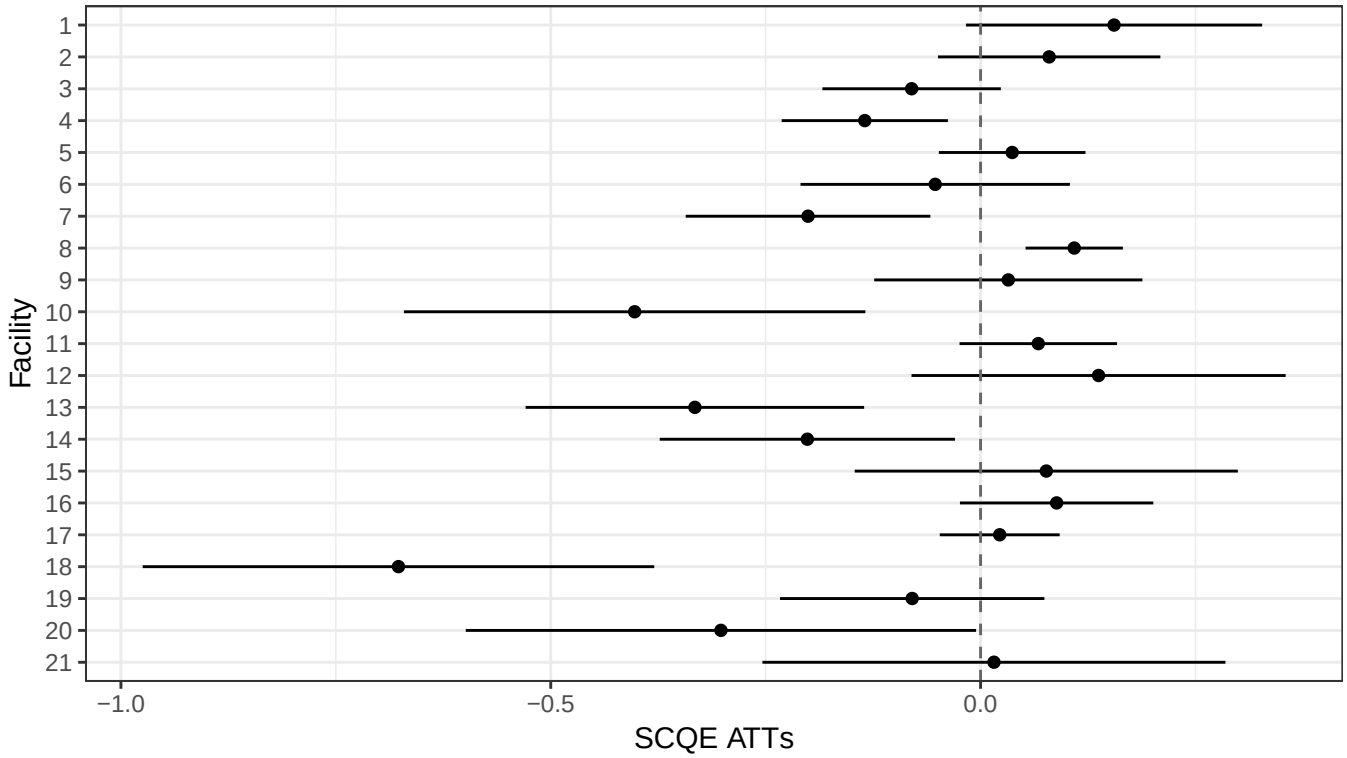
We begin with clinic-level estimates that, while relying on small samples, are useful to show variability across clinics. We first generate an ATT at each clinic from Equation (3) with the appropriately-scaled values of δ and obtain standard errors using the IV approach. Note that the standard errors constructed at a given choice of δ account only for statistical uncertainty and not for lack of information about δ itself. For illustration, we show the clinic-level estimates at $\delta = 0$ in Figure 1. At this value, seven of the 21 clinics show a negative (beneficial) estimate with 95% confidence intervals excluding zero (i.e. two-sided $p < 0.05$); one clinic shows a statistically significant positive (harmful) effect; and the remaining 13 having confidence intervals that include zero.

If δ is positive, as proposed by Dr. Maokola due to increased reporting, these results would

[9] We call these multipliers "decay" rates because the data produced estimates of less than 1, but they could have represented "growth" rates had they been above 1.

[10] The results of these regressions vary depending on which subsets of informative data we use. See Table A.2 in the Appendix for details.

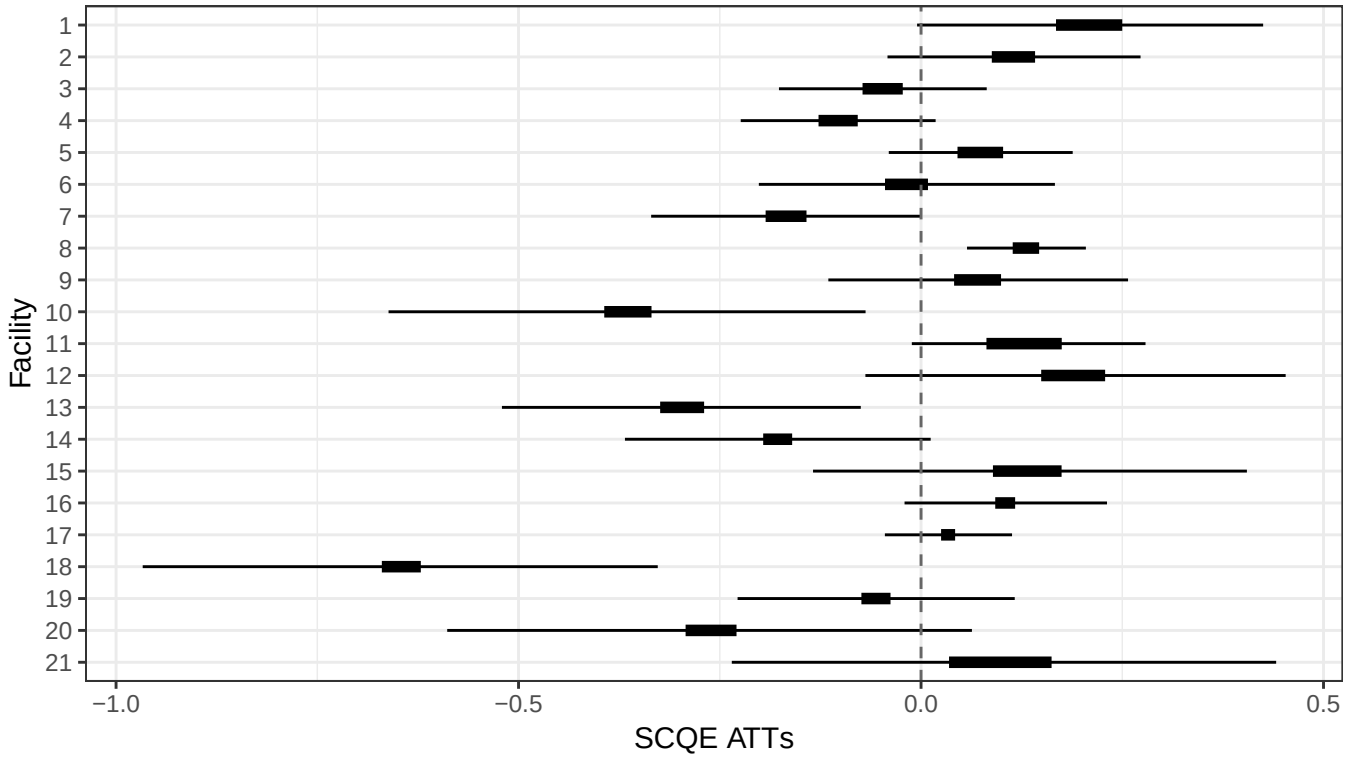
Figure 1: Clinic-Specific ATTs using $\delta = 0$



Note: ATT estimates for each clinic given δ is assumed to be 0. The whiskers represent the 95% CI from the IV estimator for the standard errors. The results appear to be significantly and substantively beneficial in seven clinics (those to the left of zero line, with the 95% confidence interval excluding zero); in one it is significant in the opposite direction; and in the remaining 13 the confidence interval includes zero. Clinics are ordered by total number of patients.

generally move to the left, pushing more clinics to show beneficial results. If the true δ is negative (declining TB incidence), then the results move right, and we might find fewer clinics with clearly beneficial effects. More generally, we can consider a range of values for δ. It is then convenient to construct estimates with confidence intervals reflecting both "identification uncertainty" (due to the range of δ values) and the usual statistical uncertainty. To do this, we construct the lower and upper CIs around the lowest and highest point estimates produced by a given range of δ. This results in a "range-and-whisker" rather than a "dot-and-whisker" visualization of clinic-level ATT estimates. Figure 2 shows such ATT estimates using a range of δ based on the linear trend estimate's 95% CI. The thick band connects the highest and lowest point estimates obtained over this range of δ. The whiskers then show the lower or upper portions of the confidence intervals extending from these. In four clinics, the effect estimate is in the beneficial direction with the augmented 95% CI excluding zero, in one it is significant in

Figure 2: Clinic-Specific ATTs using δ suggested by linear trends



Note: ATT estimates for each clinic, using the range of δ implied by learning the linear trend over untreated periods, and constructing estimates using the upper and lower 95% confidence interval of that δ, together with the 95% confidence interval around the ATTs from each of those. The results appear to be significantly and substantively beneficial in four clinics (those to the left of zero line, with the augmented 95% confidence interval excluding zero); in one it is significant in the opposite direction; and in the remaining 16 the augmented confidence interval includes zero. Clinics are ordered by total number of patients. See Figure A.2 and Figure A.3 for similar plots but using choices of δ generated from different sources.

the opposite direction, and in the remaining 16 the augmented confidence interval includes zero.

In the Appendix, we show similar plots with both the range of δ suggested by Dr. Maokola of a 0.5pp to 1pp increase per year (Figure A.2), and the range obtained by using an exponential decay rate to learn from trends in the non-IPT data (Figure A.3). The first are more optimistic, with 8 clinics showing augmented CIs that exclude zero in the beneficial direction and one showing significant estimates in the other direction. The latter is the most pessimistic: three clinics appear to have positive (harmful) effects of IPT with augmented CIs excluding zero and only one shows evidence of a significant beneficial effect.

While δ is not identifiable from data, we have no arguments with which to reject the values proposed by Dr. Maokola or those determined by examining baseline trends elsewhere in the data. This leaves us unable to rule out any of the ATT estimates just discussed as plausible.

We have thus learned principally what we do *not* know about the effect of IPT with confidence; that the clinic-level ATT estimates are fragile and not defensibly positive or negative in most clinics.

We note several finite sample considerations that can arise using this approach, particularly at lower levels of aggregation such as the clinic-level estimates. First, while we do not face the equivalent of a "weak instrument" problem (too small a π_1), that could be an issue in other applications or if we included other clinics where uptake of IPT was lower. Second, we also came to find that small samples can give rise to estimated intermediate quantities that can take on "impossible," negative values. This can occur in cases such as: (i) when the pre-treatment incidence rate was already low and a negative δ was then applied (resulting in a negative estimate of $\mathbb{E}[Y(0)|T=1]$ for that clinic), or (ii) when the estimate of $\mathbb{E}[Y(0)|T=1,D=0]$ is higher than the imputed $\mathbb{E}[Y(0)|T=1]$, forcing the estimate of $\mathbb{E}[Y(0)|T=1,D=1]$ to be below it, and possibly below zero. While disconcerting at first, these are not really "impossible" or problematic: they arise due to finite sample error.[11]

3.5 Pooled results

Our primary estimate of interest pools across clinics so as to attain a single estimate that is as precise as possible. At any level of aggregation there exists a choice of δ applicable to the population in question, and for every choice of δ there is a consequent ATT estimate for that population.

From each clinic we record estimates of four expectations, $\mathbb{E}[Y|T=0]$, $\mathbb{E}[Y|D=1,T=1]$, $\mathbb{E}[Y|D=0,T=1]$, and π_1, as well as the time-gap between $T=0$ and $T=1$. The time-gap is needed at the clinic level because, for a given choice of baseline annual trend, the actual value of δ depends upon the gap in time between $T=0$ and $T=1$, which varies slightly by clinic. By determining the correct δ for each clinic, we can then construct the clinic level estimate of $\mathbb{E}[Y(0)|T=1]$. We then pool data across clinics, constructing a weighted average of the

[11] Other methods can also produce such "impossible" implied values when applied to smaller sampling units. For example, in a difference-in-difference approach, one selects a particular unit's outcome in the pre-treatment period, and shifts it by the common "over-time" shift (learned from the difference between the average of the control group outcome before and after treatment). That result, too, could easily take on an "impossible" value (e.g. an employment level below zero), despite representing an imputed non-treatment value for that unit in the later time period. And yet, this is a consequence of idiosyncratic sampling error, and averages out to yield a valid treatment effect estimate across units.

expectations required to compute the pooled ATT,

$$ATT = \mathbb{E}[Y|D=1,T=1] - \left(\frac{\mathbb{E}[Y(0)|T=1] - \mathbb{E}[Y|D=0,T=1](1-\pi_1)}{\pi_1}\right). \quad (7)$$

Note that Equation (7) is simply Equation (3) but in which the δ has been added to $\mathbb{E}[Y|T=0]$ to form $\mathbb{E}[Y(0)|T=1]$ within each clinic first.

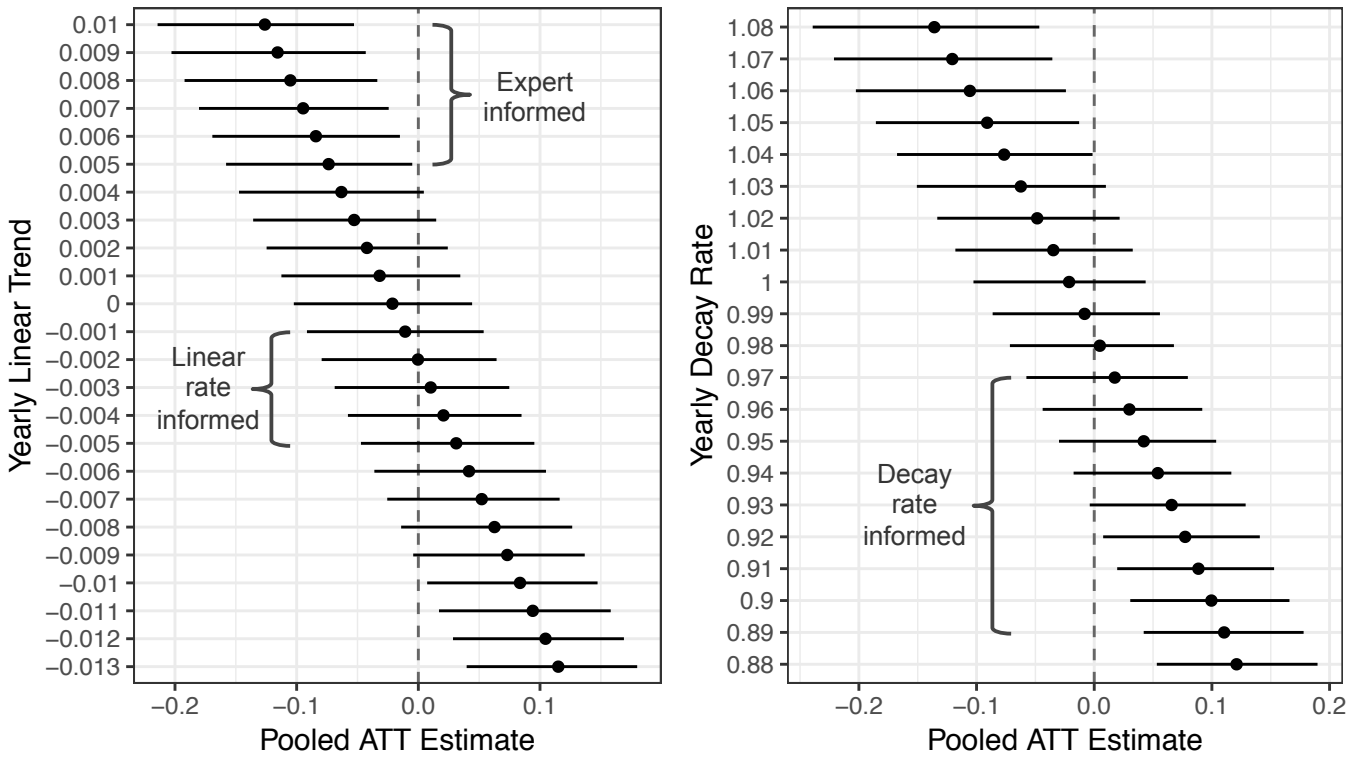
We find that prior to IPT, the pooled average TB incidence rate was 13%. After IPT became available, 25% of patients were prescribed IPT. The observed average TB incidence rate for those who did not receive IPT after it was introduced was 16%, slightly higher than the 13% among the overall pre-IPT cohort. But the incidence was radically lower for those who received IPT, at just below 1%.

The question to be answered is whether the large difference in TB rates between those who received IPT (1%) and those who did not (16%) in the later period was due to an effect of IPT, or a selection process. Suppose momentarily we employ $\delta = 0$, which is to propose that nothing important has changed over time in the non-treatment expected TB incidence rate. Applying Equation (7) we get an estimated ATT of -2pp, with a 95% confidence interval from the block bootstrap method of [-10, 5].

We can arrive at the same estimate logically as follows. Supposing $\delta = 0$, the expected non-treatment outcome over *everybody* at time $T = 1$ is 13%. To maintain this while observing the outcome of 16% among just the non-treated requires the non-treatment average of the treated to fall well below 13%. The law of iterated expectations tells us exactly what it must be, at 3%. Once we know the average non-treatment outcome for the treated group (3%), simply comparing it to the observed outcome of the treated (1%) gives our ATT (-2pp).

This simple analysis also says a great deal about the selection process and bias. Continuing with the $\delta = 0$ assumption for a moment, we can decompose the naive comparison (1%-16% = -15pp) into a point estimate for the ATT (-2pp) and "selection bias" that tells us how the treated and untreated differ on their non-treatment outcomes (3% -16%= -13pp). That is, the group receiving treatment was 13pp less likely to have developed TB in the absence of the treatment. Relaxing the $\delta = 0$ assumption, we would find that IPT was selectively given to those who were less likely to develop TB, as long as the baseline trend did not increase by approximately 1pp per year or more.

Figure 3: Pooled ATTs, by δ



Note: Pooled estimates for the ATT under varying assumptions on δ, re-expressed here in terms of yearly trends (*left*) or decay rates (*right*) for ease of interpretation. Confidence intervals were generated using the block bootstrap method described in Section 2.2. For each assumption on the baseline trend, there is a consequent ATT estimate. Under the expert-based assumption that the non-treatment average TB incidence would rise by 0.005 to 0.01 per year, we see negative (beneficial) ATT estimates (*left*), distinguishable from zero and ranging from -7pp to -13pp. A data-assisted choice of δ under an linear model suggests annualized trends of -0.001 to -0.005, still on the *left*, which correspond to non-significant pooled ATTs. A data-assisted choice of δ under an exponential decay model suggested annualized decay rates of 0.97 to 0.89. On the *right*, we see these decay rates correspond to a combination of non-significant and significant positive (harmful) estimates.

We have thus dealt with selection concerns not through assuming the observability of all confounders and adjusting for them, but through an assumption on δ. Figure 3 is more comprehensive and our preferred means of reporting results, visualizing the estimates obtained under varying choices of δ. The left panel of Figure 3 shows how assumptions on a linear trend in the non-treatment outcome generate varying estimates.[12] Those values produce the ATT estimates plotted, with 95% confidence intervals produced by the block bootstrap method. Under the "domain knowledge" assumption that reported TB rates would have risen by 0.005 to 0.01 per year, the resultant ATT estimates range from a (beneficial) effect on the TB incidence rate of -7pp to -13pp. By contrast, the data-informed choice of δ based on linear trends in the non-IPT

[12]The time gap between the average pre- and post-IPT patients was close to three years for all clinics.

data suggests a range of -0.001 to -0.005 per year. These correspond to small and non-significant estimated ATTs. Finally, the right panel of Figure 3 indexes estimates by the annualized decay rate used to formulate δ. The data-driven assumption that decay rates vary from 0.97 to 0.89 produces ATT estimates ranging from a 11pp harmful (and significant) effect of IPT down to an estimate of approximately zero.

4 Discussion

We regard the main strengths of SCQE relative to other methods to be: the ease of understanding an assumption on the baseline trend, the possibility in some scenarios that such assumptions can be made more defensibly than assumptions involving treatment assignment (such as unconfoundedness), and the protection against over-confidence that arises from constructing estimates over a range of assumed baseline trends rather than providing a single estimate and warning users the assumptions may not be valid. These benefits are limited, of course, to situations to which SCQE applies—those with a rapid change in treatment uptake, and the required measurements on the average outcomes before and after that uptake. Here we describe how SCQE compares in greater detail to a number of approaches that are either broadly used for non-experimental data (i.e. covariate adjustment), or that bear some resemblance or equivalence to SCQE.

4.1 Covariate adjustment

The most common strategy when dealing with non-randomized treatments is to employ covariate-adjustment procedures able to produce unbiased causal estimates only in the absence of unobserved confounding. These methods include regression-based approaches (linear or otherwise), covariate or propensity score matching or weighting, and stratification/sub-classification. In each, it is hoped that observed (pre-treatment, non-colliding) covariates $\mathbf{X}$ included in the analysis account for all confounding, which results in ignorability of the treatment conditional on covariates, $(\{Y(1), Y(0)\} \perp\!\!\!\perp D) \mid \mathbf{X}$. Equivalently, $\mathbf{X}$ must satisfy the "backdoor criterion" of Pearl (2009). These methods differ only in how they achieve the conditioning required to exploit this assumption during estimation. In this study, as is very often the case, we are unable to defend the assumption that all confounding variables have been observed. The usual difficulty

of such a claim is heightened by the inability of even experts on this topic to know why IPT was prescribed or not prescribed to each patient, and the limited number of reliably-measured covariates available to us: patient age, sex, date of first visit, and HIV severity at first visit.

Despite these concerns, as covariate adjustment remains standard, we ask what such an approach would communicate to the reader in our case. Recall that the naive cross-sectional comparison in the post-treatment period shows a 15 pp lower TB incidence among those taking IPT compared to those not taking it. A simple (OLS) regression in the post-treatment period of our data, adding only clinic fixed effects to this formulation, suggests an estimate in line with that comparison (-15.5 [-14.3, -17.0] pp). Adding all available covariates to this, the result remains similar, at -15.1 [-13.7, -16.4] pp.[13] Given that the underlying assumption of no-unobserved-confounding is highly suspect here, presenting such results for public consumption is problematic. To label the result as merely "suggestive" may not be sufficient to prevent abuse. Indeed, a covariate-adjusted result may be even more vulnerable to abuse than a naive one, if the statistical adjustment technique is believed by readers to add sophistication and crediblity to the result. Further, the substantively large and highly statistically significant result, with a t-statistic over 22, may contribute to over-confidence in this case.

How does this compare to SCQE? The most critical difference in our view lies in the credibility of the underlying assumptions and the risk of over-confidence in unfounded results. While we may be uncertain as to the true value of δ, results can be presented over a sufficient range to either change the result or cover all values argued to be reasonable. By contrast, as commonly practiced at least, regression and other covariate-adjustment procedures report an answer consistent only with the assumption of precisely zero confounding. In our case, the answer provided by regression (-15 pp) proves to be more extreme then the point estimates produced by SCQE over the entire range of δ considered in Figure 3. Further, SCQE informs us that in order to achieve a true ATT point estimate as extreme as the regression estimate, the baseline rate of TB would have to rise by more than 1.2 percentage points per year linearly (or have a growth rate of more than 1.08). This is a relatively extreme trend to defend, more positive than the expert opinion and with the opposite sign of the empirical trends seen in the untreated outcomes.

[13] We report OLS results for simplicity, but recognize that some disciplines prefer not to use OLS with binary outcomes. Had logistic regression been used instead, the coefficient estimate would be -3.90 [-4.48, -3.33] (odds ratio of 0.02) with a t-statistic of -13.2. As with OLS, this risks producing high confidence in a substantively significant result that fails to communicate concerns about the underlying assumption of no unobserved confounding.

We note that regression and related approaches can be subjected to sensitivity analyses (e.g. Arah, 2017; VanderWeele and Ding, 2017; Oster, 2017), which we endorse, particularly if no other identification strategy such as SCQE is available. However where SCQE is possible it brings in additional information by considering time and δ. Further, the sensitivity analysis automatically built in (by examining values over a range of δ) has two useful features. The first is that users can easily imagine what is meant by the "baseline trend", whereas the quantities required for sensitivity analyses in other contexts can be more difficult to understand and debate (e.g. the partial variance in the outcome or treatment explained by unobserved confounding, as in Cinelli and Hazlett, 2018). Second, regression-based sensitivity analyses draw attention to a single point estimate presuming no confounding, and ask what strength of confounding would be required to alter this conclusion. By contrast, SCQE shows what effects would be estimated under a range of assumptions, requiring the user to actively argue for or against certain ranges of δ in order to argue for or against an effect in either direction.

4.2 Instrumental variables

As noted above, SCQE can be understood as an IV approach in which time is regarded as an instrument and $\delta \neq 0$ allows a prescribed deviation from the exclusion and exogeneity assumptions. Accordingly, the directed acyclic graph (DAG) for SCQE can be drawn as it is commonly drawn for IV: Time influences treatment uptake, and (exclusively) through it the outcome, while the treatment and outcome may be connected by unobserved confounding (see Hazlett, 2019). This holds when $\delta = 0$; otherwise one must modify the DAG to account for δ. The idea of using time as an instrument strikes some investigators as unusual, perhaps because time is awkward to think of as a literal cause (or here, encouragement), but it has been exploited elsewhere in the medical and epidemiological literature (Johnston et al., 2008; Cain et al., 2009; Shetty, Vogt and Bhattacharya, 2009; Mack et al., 2015; Gokhale et al., 2018, see also Brookhart, Rassen and Schneeweiss, 2010 for discussion). Nevertheless, we note several differences between SCQE and IV. The first is conceptual. The description of the strategy here—i.e. using an assumption on the baseline trend and the observed outcome for the untreated to back-out the non-treatment outcome among the treated—seems not to have been offered in previous treatments of time as an instrument. Second and most importantly, whereas an IV analysis reports its result as if the exclusion and exogeneity assumptions hold exactly and with

certainty, SCQE both allows deviation from this ($\delta \neq 0$) and encourages consideration of ranges of δ due to uncertainty. In this application, and many others in which these IV assumptions most likely do not hold, SCQE can provide an alternative option.[14] Finally, applications using the time-as-instrument approach have in practice attempted to buoy the required assumptions (i.e. exogeneity of the encouragements and the exclusion restriction) by adding covariates. The addition of covariates is possible in SCQE, which could be employed to improve our guess as to δ.[15] That said, we ultimately see the use of an "all-inclusive δ" as a strength of this approach: a choice of δ summarizes how the pre- and post-treatment cohorts differ on their expected $Y(0)$, without necessarily asking how much of this would remain conditionally on covariates. Such an assumed δ is all that is required for identification, it is transparent, and it is easy to understand and think about in real terms as a "baseline trend".

4.3 Difference-in-differences

The use of over-time comparisons may also call to mind the difference-in-differences (DID) approach. The first and most important distinction between the approaches is that SCQE can work in settings where the DID does not apply and estimates cannot even be computed. For DID we must either have panel data for individuals, or be able to place individuals into larger groupings that persist over time (such as states or clinics), with treatment then being assigned to all individuals within a subset of those groups in the second time period. That is, we use the groups individuals belong to as a means of labeling individuals as "would be treated" or "would-be-untreated". This allows us to draw the "two lines" required of DID: the trend in the would-be-treated group, and the trend in the would-be-untreated group. By contrast, SCQE conceives of entirely separate cohorts at separate time points, and works even if there is no way

[14] Varying sensitivity analyses and placebo testing approaches for IV approaches also exist, see Baiocchi, Cheng and Small (2014) for discussion.

[15] That SCQE does not in general require covariates is a useful feature. That said, covariates could be incorporated into SCQE in various ways and to various benefits. One natural approach would be to train a model that uses pre-treatment characteristics X_i to predict $Y_i(0)$ in the pre-treatment cohort, then apply the same model to all X_i in the post-treatment cohort to predict $Y_i(0)$ for them. Taking the difference in averages would then produce a value of δ that directly accounts for changes in covariates. Without further variation in δ, the result would be correct only under an assumption that time is ignorable for the non-treatment outcome given X_i, i.e. $Y(0) \perp\!\!\!\perp T_i \mid X_i$. Note that, T_i indicating time and not treatment status, this assumption is milder than the $Y(d) \perp\!\!\!\perp D_i \mid X_i$ assumption usually used in covariate adjustment because "selection into time period" may be more limited than selection into treatment-taking in many scenarios. However, in keeping with the spirit of our proposal, one could still vary δ above and below this value to capture possible "other (unobserved) causes of change in the non-treatment outcome." That is, just as one need not believe $\delta = 0$ in the typical use case, one need not believe that δ is driven only by changes in covariates in this extension.

to know if an individual observed at time $T = 0$ would have chosen or been assigned treatment had they appeared at time $T = 1$. SCQE is thus particularly appropriate in cases such as the introduction of a new treatment or policy with individualized self-selection, where the DID data structure may simply not exist.

Second, consider settings in which DID is possible, in either the cross-sectional or panel form. Here, the SCQE approach actually nests DID while accommodating uncertainty over the identification assumption. Specifically, DID requires the parallel trends assumption,

$$\mathbb{E}\Big[Y(0)|T=1,D=1\Big] - \mathbb{E}\Big[Y(0)|T=0,D=1\Big] \;=\; \mathbb{E}\Big[Y(0)|T=1,D=0\Big] - \mathbb{E}\Big[Y(0)|T=0,D=0\Big].$$

An important concern with asking users to make such an assumption is that in recognizing there is some selection into treatment, the analyst expects the average $Y(0)$ for the treated and control group to likely differ. This makes it tenuous in some cases to then argue with confidence that despite this admitted difference, the *trends* in $Y(0)$ are exactly equal in these two groups. Rather, we might reasonably suspect that whatever reason led one group to switch treatment status when it did could plausibly be related to the outcome trends. For this reason, SCQE does not ask for judgements about how treated and control groups compare, on either levels or trends. It instead relies upon an assumption on how the entirety of the cohorts at times $T = 0$ and at $T = 1$ differ, on mean $Y(0)$.

That said, for *any* assumption on the non-treatment trends in the treated and control groups, simply taking a weighted average of these trend lines produces a corresponding value of δ that can be used in SCQE. DID is thus a special case of SCQE, in which (i) we can label individuals observed at both time points as those that "would-be-treated" or "would-be-control," according to larger units they belong to and to which treatment is eventually assigned; (ii) we choose to "learn δ" from the over-time change in the units not eligible for treatment; and (iii) we assume the change in non-treatment outcomes is the same for both groups, i.e. parallel trends.

In our case, SCQE and DID cannot be directly compared because our example is one where *DID is not possible*: we cannot label individuals as would-be-treated and would-be-untreated in the first cohort. However we can instead consider an alternative form of DID. Since individuals are organized at the clinic level, suppose we redefine the treatment to be "going to a clinic that uses IPT," and attempt to estimate this effect by DID. Doing so requires the parallel

trends assumption. Once such an estimate is made, we could further posit a precise exclusion restriction: that clinics using IPT only improve patient outcomes on TB through observed IPT prescription. With this, we can rescale the overall effect of being in an IPT-clinic by the proportion treated in the second period (the compliance rate) to estimate the effect among compliers, in what is effectively an "IV-DID" approach. Such an approach would produce one of the estimates like those shown on Figure 2, where the assumed yearly linear trend (vertical axis) is set to equal the rate of change in the non-IPT clinics. Altogether, the benefits of SCQE over such an approach are that (i) δ can be chosen by means other than assuming parallel trends – and can be contemplated by considering a baseline trend in everybody's $Y(0)$ rather than attempting to compare treated and control groups; (ii) the choice of δ also avoids a separate exclusion restriction, by directly assuming how outcomes would compare across cohorts in the absence of treatment; and (iii) a range can be placed on δ so that reported results do not place full confidence in uncertain assumptions.

4.4 Interrupted time-series and (fuzzy) regression discontinuity

Two other approaches that may seem to be related include the interrupted time-series (ITS, see Hudson, Fielding and Ramsay, 2019 for a recent review in medicine), and similarly the regression discontinuity (RDD) in time (see Hausman and Rapson, 2018 for a recent methodological review). In either approach, an assumption is imposed on the continuity, smoothness, or function space of the expected potential outcome conditional on time. A first difference between these approaches and SCQE is again in the context where they apply. These approaches require coding the "time" of each observation narrowly, e.g. to one day or perhaps one month. In cases where a treatment/non-treatment decision is made at a finite moment in time, this may be suitable. However in many cases, such as the one studied here, not only does the outcome require a suitably long follow-up window, but patients also have a wide window during which they may enter treatment or not. The wide treatment window in particular makes it problematic to code an observation (i.e. one unit with its treatment status and eventual outcome) to a precise moment in time. In such scenarios, the "binning" of observations into wide cohorts is instead required, bringing us back to an SCQE scenario.

A second major difference relates to whether everybody observed post-treatment actually gets treated, or just a subset as contemplated in SCQE. The ITS and the (sharp) RDD in time

typically apply where all units are considered treated in the post-treatment period, and none are considered treated before. This make sense when, for example, the treatment is a policy or a newsworthy event and we would like to know its effect on some attitude. It does not make sense when a treatment merely becomes available but we remain concerned about selection into it. By contrast, in the RDD tradition, the "fuzzy RDD" (Trochim, 1984) stems from precisely this type of concern, viewing an indicator for being in the post-treatment period as an encouragement (instrument) for treatment. In this sense, SCQE is most similar to a fuzzy RDD, but comparing wider time bins and allowing δ to account for possible shifts in $Y(0)$ over time between these bins rather than depending on a model of the expected outcome as a function of time — except optionally to inform the range chosen for δ.

4.5 Conclusions

Where investigators may otherwise rely upon naive or covariate-adjusted estimates, the SCQE approach allows users to extract valid causal information from observed data for the cost of an assumption on the baseline trend. In our application, with just 1% of patients on IPT developing active TB compared to 16% who were not, many policymakers would likely infer that the program definitively works. Covariate adjustment by regression similarly produces an estimate of -15 pp, which may appear more convincing both because it accounts for covariates and is highly statistically significant ($t > 22$). Despite any warnings statisticians may invoke that such a result is "only suggestive", it is reasonable to expect that even sophisticated consumers of such analyses will see such an result as their best means of using data to inform policy, in the absence of other information. To construct such a result under an assumption that is difficult to defend and call it "suggestive" actually says very little about what precisely can and cannot be concluded from the evidence. Our approach turns this problem around, pointing not to an invisible threat of confounding but rather requiring the reader to actively choose and defend an assumption on the trend (δ) if they wish to argue for a given result. In the process, it also shows how easily one could have drawn the opposite conclusion, encouraging skepticism.

In this case, first, a simple assumption of no baseline change ($\delta = 0$) immediately suggests an ATT estimate that is not distinguishable from zero (-2pp, [-10,5]). The program can be argued to be beneficial only if we can defend a claim that the (non-treatment) TB incidence rate was climbing by 0.5pp or more per year over this period. We cannot reject the possibility

that IPT is *harmful* unless we can rule out a downward trend in TB of 1pp per year or more. Our supposition is that policymakers are better off with this type of information than without it. An additional benefit of this approach is that it tells us something about who was selected into the treatment in terms of their non-treatment potential outcomes: As long as $\delta \leq$ 1pp per year, we can conclude that those who were prescribed IPT had lower chances of developing TB anyway. Knowing that this treatment is often assigned to those who are already "healthier" is likely useful information for assessing and redesigning the national effort to prevent TB.

In conclusion, the SCQE and SCT approaches have much to offer where randomized trials are infeasible or undesirable, or when we wish to examine the real world impact of a treatment already made available. They offer an estimate of the ATT only when a treatment increases in popularity rapidly and substantially, and when we can measure the outcome rates before and after the increase. The approaches may produce sharp and definite conclusions, particularly where effects are strong and/or narrow assumptions on δ can be supported due to the nature of the application. In other cases, as here, the assumptions we can make on δ may not be sharp enough to produce a narrow range of credible effect estimates. This may strike practitioners who expect a single point estimate as an insufficiently informative answer. In these cases, however, these methods aid in protecting against false conclusions, describe what assumptions on trends have to be ruled out (or in) to find a beneficial or harmful effect, and reveal information about those who are selected into treatment.

References

Akolo, C, I Adetifa, S Shepperd and J Volmink. 2010. "Treatment of latent tuberculosis infection in HIV infected persons." *Cochrane Database of Systematic Reviews* .

Arah, Onyebuchi A. 2017. "Bias analysis for uncontrolled confounding in the health sciences." *Annual review of public health* 38:23–38.

Baiocchi, Michael, Jing Cheng and Dylan S Small. 2014. "Instrumental variable methods for causal inference." *Statistics in medicine* 33(13):2297–2340.

Brewin, Chris R and Clare Bradley. 1989. "Patient preferences and randomised clinical trials." *BMJ: British Medical Journal* 299(6694):313.

Brookhart, M Alan, Jeremy A Rassen and Sebastian Schneeweiss. 2010. "Instrumental variable methods in comparative safety and effectiveness research." *Pharmacoepidemiology and drug safety* 19(6):537–554.

Cain, Lauren E, Stephen R Cole, Sander Greenland, Todd T Brown, Joan S Chmiel, Lawrence Kingsley and Roger Detels. 2009. "Effect of highly active antiretroviral therapy on incident AIDS using calendar period as an instrumental variable." *American journal of epidemiology* 169(9):1124–1132.

Cameron, A Colin and Pravin K Trivedi. 2005. *Microeconometrics: methods and applications*. Cambridge university press.

Cinelli, Carlos and Chad Hazlett. 2018. "Making Sense of Sensitivity: Extending Omitted Variable Bias." Working Paper.
URL: *https://doi.org/10.13140/RG.2.2.25518.56644*

Fox, Wallace, Gordon A Ellard and Denis A Mitchison. 1999. "Studies on the treatment of tuberculosis undertaken by the British Medical Research Council tuberculosis units, 1946–1986, with relevant subsequent publications." *The International Journal of Tuberculosis and Lung Disease* 3(10):S231–S279.

Freedman, David et al. 1984. "On bootstrapping two-stage least-squares estimates in stationary linear models." *The Annals of Statistics* 12(3):827–842.

Gokhale, Mugdha, John B Buse, Christina DeFilippo Mack, Michele Jonsson Funk, Jennifer Lund, Ross J Simpson and Til Stürmer. 2018. "Calendar time as an instrumental variable in assessing the risk of heart failure with antihyperglycemic drugs." *Pharmacoepidemiology and drug safety* 27(8):857–866.

Hausman, Catherine and David S Rapson. 2018. "Regression discontinuity in time: Considerations for empirical applications." *Annual Review of Resource Economics* 10:533–552.

Hazlett, Chad. 2019. "Estimating causal effects of new treatments despite self-selection: The case of experimental medical treatments." *Journal of Causal Inference* 7(1).

Hudson, Jemma, Shona Fielding and Craig R Ramsay. 2019. "Methodology and reporting characteristics of studies using interrupted time series design in healthcare." *BMC medical research methodology* 19(1):137.

Johnston, KM, P Gustafson, AR Levy and P Grootendorst. 2008. "Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to epidemiological research." *Statistics in medicine* 27(9):1539–1556.

Knox, Dean, Teppei Yamamoto, Matthew A Baum and Adam J Berinsky. 2019. "Design, identification, and sensitivity analysis for patient preference trials." *Journal of the American Statistical Association* pp. 1–27.

Mack, Christina DeFilippo, M Alan Brookhart, Robert J Glynn, Anne Marie Meyer, William R Carpenter, Robert S Sandler and Til Stürmer. 2015. "Comparative effectiveness of oxaliplatin vs. 5-flourouricil in older adults: an instrumental variable analysis." *Epidemiology (Cambridge, Mass.)* 26(5):690.

Neyman, Jerzy S. 1923. "On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Translated and edited by DM Dabrowska and TP speed, Statistical Science (1990), 5, 465-480." *Annals of Agricultural Sciences* 10:1–51.

Olschewski, M and H Scheurlen. 1985. "Comprehensive cohort study: an alternative to randomized consent design in a breast preservation trial." *Methods of information in medicine* 24(03):131–134.

Oster, Emily. 2017. "Unobservable selection and coefficient stability: Theory and evidence." *Journal of Business & Economic Statistics* pp. 1–18.

Pearl, Judea. 2009. *Causality*. Cambridge university press.

Shetty, Kanaka D, William B Vogt and Jayanta Bhattacharya. 2009. "Hormone replacement therapy and cardiovascular health in the United States." *Medical Care* pp. 600–605.

Smieja, Marek, Catherine Marchetti, Deborah Cook and Fiona M Smaill. 1999. "Isoniazid for preventing tuberculosis in non-HIV infected persons." *Cochrane Database of Systematic Reviews* .

Stock, James H and Motohiro Yogo. 2002. "Testing for weak instruments in linear IV regression.".

Trochim, William M. K. 1984. *Research Design for Program Evaluation: The Regression-Discontinuity Approach (Contemporary Evaluation Research)*. SAGE Publications, Inc.

VanderWeele, Tyler J and Peng Ding. 2017. "Sensitivity analysis in observational research: introducing the E-value." *Annals of internal medicine* 167(4):268–274.

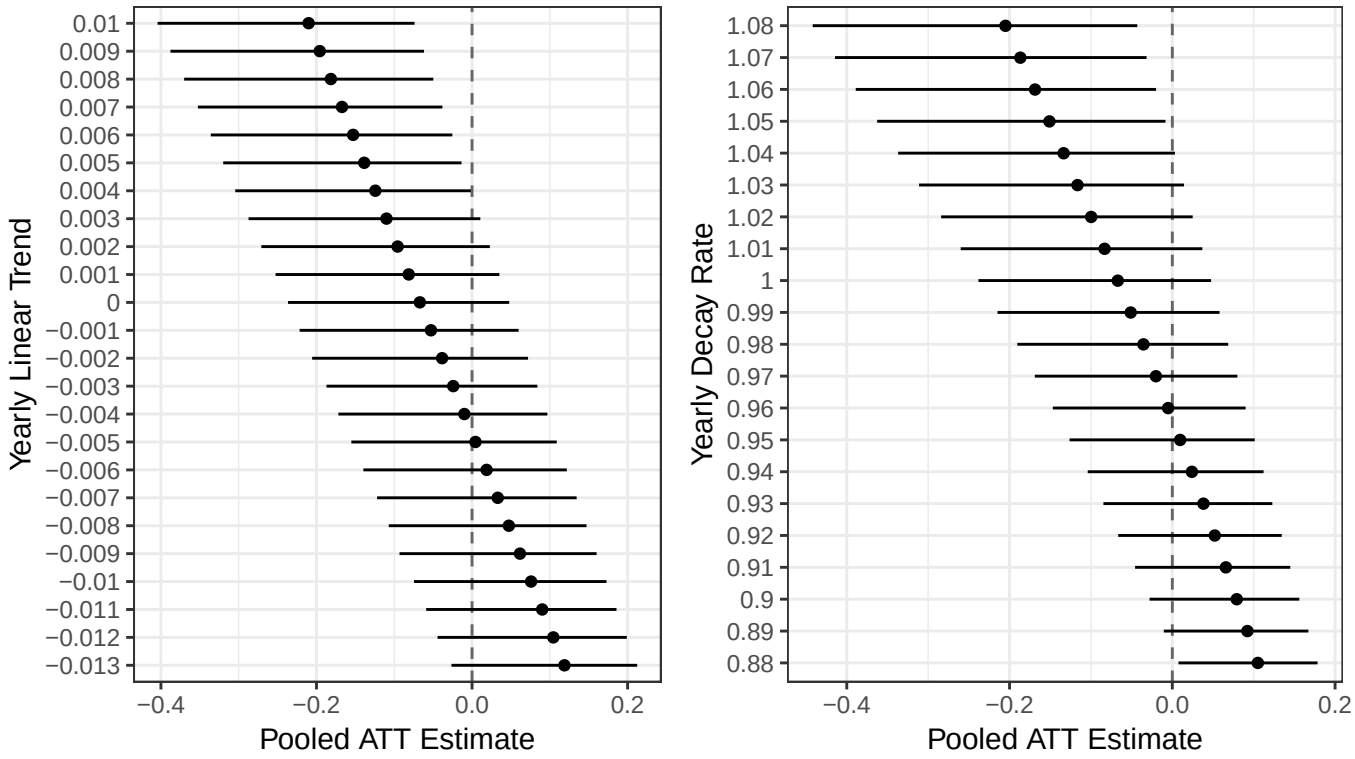
WHO. 2008. Isoniazid Preventive Therapy. In *Implementing the WHO Stop TB Strategy: A Handbook for National Tuberculosis Control Programmes*. chapter 7.

Wooldridge, J.M. 2009. *Introductory econometrics: A modern approach*. South-Western Pub.

Online Supplement for:
Inference without randomization or ignorability: A stability-controlled quasi-experiment on the prevention of tuberculosis

A Appendix

Figure A.1: Pooled ATTs, by δ, using $M_D = 6$ months



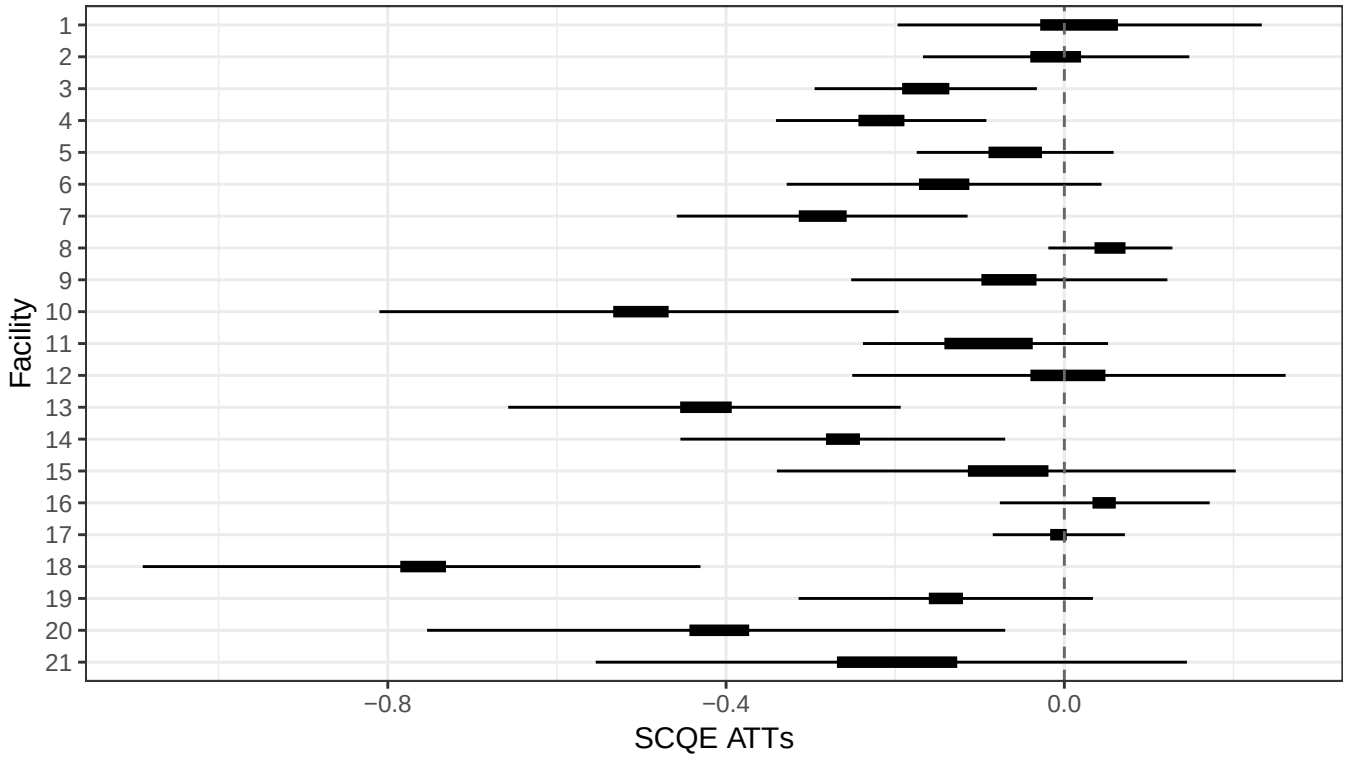
Note: Pooled ATT estimates as a function of δ, similar to Figure 3 in the main text. It shows those results if the period after a patient's first visit within which IPT administration is considered treatment is limited to 6 months instead of a full year. A notable consequence of this limit is that 6 of the 21 clinics included in those results have too few treated patients to reach the 10% treatment cutoff necessary for an ATT estimate here. The 15 remaining clinics generate these pooled ATT estimates and bootstrapped confidence intervals. Although these estimates and intervals suggest slightly more beneficial effects of IPT than those in Figure 3, the various ranges of δ suggest the same directionality and almost the same significance (or lack thereof).

Table A.2: Suggestions for yearly trends by data subset

Trend type:	Clinics without IPT use:	Clinics with IPT use	
		All:	Only ATT qualifiers:
Linear	-0.0006 [-0.0031, 0.0020]	-0.0040 [-0.0070, -0.0011]	-0.0150 [-0.0241, -0.0060]
Exponential	0.9825 [0.8920, 1.0821]	0.9197 [0.8788, 0.9626]	0.8874 [0.8249, 0.9546]

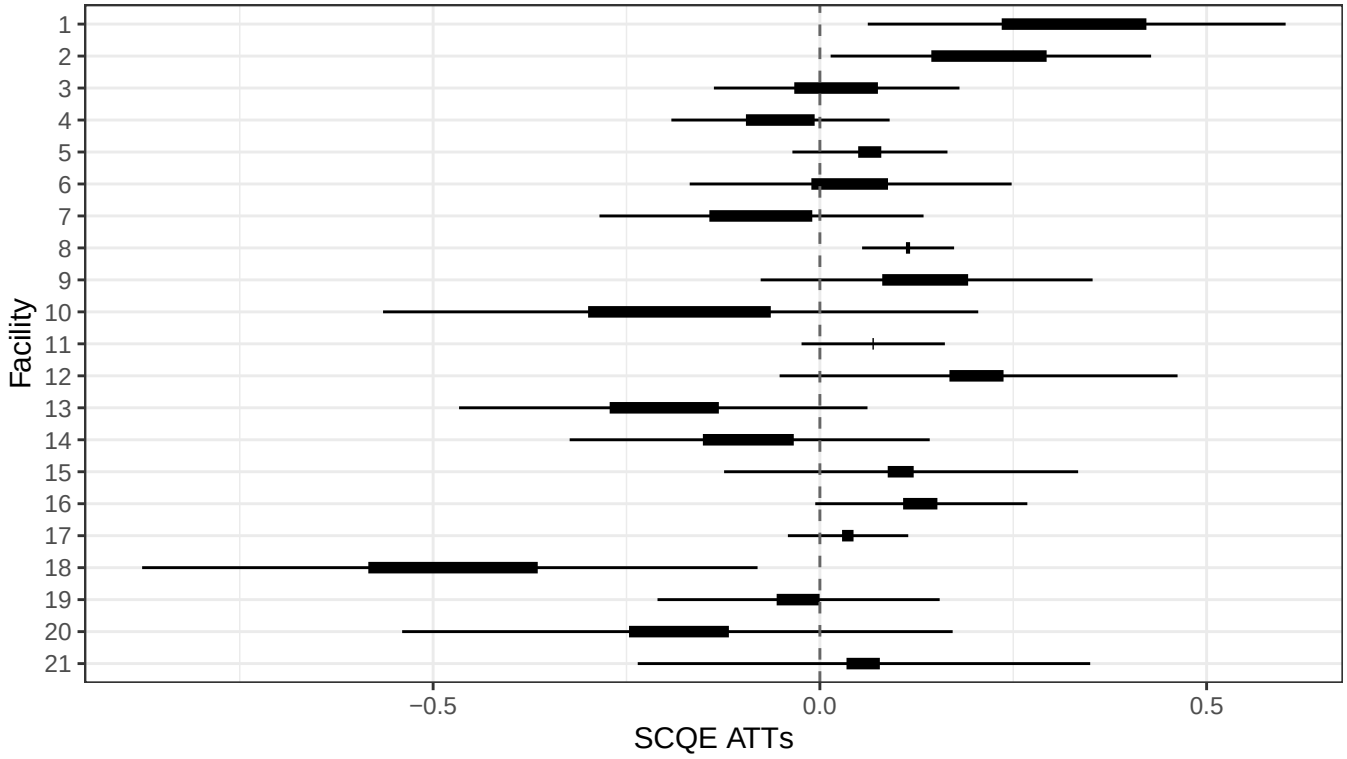
Note: Trends estimated from data to inform the range of δ can be learned from (i) clinics that never introduced IPT, or (ii) the pre-IPT periods in clinics that did. Further, (iii) the latter may be narrowed to only include clinics that qualify for inclusion in the analysis according to our criteria (Section 3.2). Although we have no reason to exclude any subset and instead combine of all of these data in our primary analysis, we present here the range of data-driven trends that result from each of these subsets. Trends are presented as point estimates in yearly linear and yearly decay form, with 95% confidence intervals. The first column shows results for group (i). This shows the trends over the same time periods when treatment was introduced in other clinics. It suggests flatter trends and, consequently, more beneficial imputed effects of IPT. The second column shows estimates for group (ii). This shows the trend at different times, but over the same clinics that adopt IPT. It shows a steeper TB rate decline, which would result in more harmful imputed effects of IPT. The third column shows results for group (iii), and aligns the clinics used to generate trends and effects. This suggests the steepest trends, implying the most harmful effect estimates.

Figure A.2: Clinic-Specific ATTs using δ suggested by domain knowledges



Note: ATT estimates for each clinic, using the "domain knowledge" range for δ, together with the 95% confidence interval around each end of that range. The results appear to be significantly and substantively beneficial in eight clinics; in none are they significant in the opposite direction; and in the remaining 13 the augmented confidence interval includes zero. These results are the most optimistic of the three sets of clinic-level estimates we generated. Clinics are ordered by total number of patients.

Figure A.3: Clinic-Specific ATTs for δ suggested by exponential trends



Note: ATT estimates for each clinic, using the δ implied by learning the exponential decay trend over untreated periods, and constructing estimates using the upper and lower 95% confidence interval of that δ, together with the 95% confidence interval around the ATTs from each of those. The results appear to be significantly and substantively beneficial in one clinic; in three they are significant in the opposite (harmful) direction; and in the remaining 17 the augmented confidence interval includes zero. These results are the most pessimistic of the three sets of clinic-level estimates we generated. Clinics are ordered by total number of patients.

Table A.3: Linear δ results

Yearly Linear Trend	Effective Used δ	Pooled ATT	Bootstrapped CI 2.5%	97.5%
0.010	0.029	-0.126	-0.214	-0.053
0.009	0.026	-0.116	-0.203	-0.043
0.008	0.023	-0.105	-0.192	-0.034
0.007	0.020	-0.095	-0.180	-0.024
0.006	0.017	-0.084	-0.169	-0.015
0.005	0.014	-0.074	-0.158	-0.005
0.004	0.011	-0.063	-0.147	0.005
0.003	0.009	-0.053	-0.136	0.015
0.002	0.006	-0.042	-0.125	0.024
0.001	0.003	-0.032	-0.113	0.034
0.000	0.000	-0.021	-0.102	0.044
-0.001	-0.003	-0.011	-0.092	0.054
-0.002	-0.006	-0.000	-0.079	0.064
-0.003	-0.009	0.010	-0.069	0.075
-0.004	-0.011	0.021	-0.058	0.085
-0.005	-0.014	0.031	-0.047	0.095
-0.006	-0.017	0.042	-0.036	0.105
-0.007	-0.020	0.052	-0.026	0.116
-0.008	-0.023	0.063	-0.014	0.126
-0.009	-0.026	0.073	-0.004	0.137
-0.010	-0.029	0.084	0.007	0.147
-0.011	-0.032	0.094	0.017	0.158
-0.012	-0.034	0.105	0.029	0.169
-0.013	-0.037	0.115	0.040	0.180

Note: Effective δ values used, and the resultant ATT estimates, for a range of proposed baseline (linear) trends. Those estimates, as well as the confidence intervals around them generated by bootstrapped resampling at the clinic level, produce the left half of Figure 3. The data-assisted choice of δ under this linear model ranged in yearly trends from -0.001 to -0.005, while the "domain knowledge" estimate was 0.005 to 0.01.

Table A.4: Exponential δ results

Yearly Decay Rate	Effective Used δ	Pooled ATT	Bootstrapped CI 2.5%	97.5%
1.08	0.031	-0.136	-0.239	-0.047
1.07	0.027	-0.121	-0.221	-0.036
1.06	0.023	-0.106	-0.203	-0.024
1.05	0.019	-0.091	-0.186	-0.013
1.04	0.015	-0.077	-0.168	-0.001
1.03	0.011	-0.062	-0.151	0.010
1.02	0.007	-0.048	-0.133	0.022
1.01	0.004	-0.035	-0.118	0.033
1.00	0.000	-0.021	-0.103	0.044
0.99	-0.004	-0.008	-0.086	0.056
0.98	-0.007	0.005	-0.072	0.068
0.97	-0.011	0.017	-0.058	0.080
0.96	-0.014	0.030	-0.044	0.092
0.95	-0.017	0.042	-0.030	0.104
0.94	-0.021	0.054	-0.018	0.117
0.93	-0.024	0.066	-0.004	0.129
0.92	-0.027	0.077	0.008	0.141
0.91	-0.030	0.089	0.020	0.153
0.90	-0.033	0.100	0.031	0.166
0.89	-0.036	0.110	0.042	0.178
0.88	-0.039	0.121	0.053	0.190

Note: Effective δ values used, and the resultant ATT estimates, for a range of proposed baseline (exponential) trends. Those estimates, as well as the confidence intervals around them generated by bootstrapped resampling at the clinic level, produce the right half of Figure 3. The data-assisted choice of δ under this exponential decay model ranged in yearly decay rates from 0.89 to 0.97.