UNIVERSITY OF CALIFORNIA

Los Angeles

Causal Inference Outside of Randomized Trials with the Stability-Controlled Quasi-Experiment: Extensions and Considerations

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Statistics

by

David Amichai Wulf

2021

ABSTRACT OF THE DISSERTATION

Causal Inference Outside of Randomized Trials
with the Stability-Controlled Quasi-Experiment:
Extensions and Considerations

by

David Amichai Wulf
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2021
Professor Chad J Hazlett, Chair

In many non-randomized, observational settings, treatment assignment mechanisms are opaque and researchers may not trust an assumption of conditional ignorability or the covariate-adjustment-based identification it allows. The Stability-Controlled Quasi-Experiment (SCQE, 2019) avoids this adjustment approach and makes no direct comparison between treated and untreated units. In settings with a change in the usage rate of some treatment of interest between successive cohorts, SCQE transforms an assumption about cohort-wide counterfactual outcome trends into estimates of the treatment's effect on those units that were impacted by the treatment change.

For example, imagine two consecutive cohorts of patients, the latter featuring non-randomized use of a new treatment. SCQE can identify and estimate an Average Treatment Effect among the Treated (ATT). The assumption required is a proposed value of δ, the unobservable difference between the cohorts' average outcomes that we would have seen *had the treatment not been introduced*.

SCQE is a partial identification strategy, presenting these effects across values of δ. In doing so, it ties claims about beneficial, harmful, or null effects to the corresponding δ values we would need to defend in order to support them, helping to avoid overconfidence in “suggestive” point estimates. In many applications, δ may also be more intuitive and easy to reason with than sensitivity analyses for conditional-ignorability-based results.

In this thesis, I (1) present new inferential tools for SCQE, including summary-statistic-based inference for use when unit-level data is unavailable or limited by data sharing restrictions; (2) extend the method to accommodate expert-informed prior distributions on δ, applications with only one cohort, and existing treatments experiencing changes in usage rates; (3) apply the method to evaluations of a tuberculosis prophylactic in Tanzania, early COVID-19 treatments, and in-hospital rapid response alerts; and (4) discuss important considerations, conceptual guidance, and best practices for practitioners.

The dissertation of David Amichai Wulf is approved.

Erin K Hartman

Mark Stephen Handcock

Onyebuchi Aniweta Arah

Chad J Hazlett, Committee Chair

University of California, Los Angeles

2021

To my mother, who told a crying 6th grader that he couldn't give up on writing

To my father, who gave me everything

To my brother, who has shown me what it means to be driven

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Chapter 2 of this thesis is a reprint of a published, co-authored article:

> Chad Hazlett, Werner Maokola, and David Ami Wulf (2020). Inference Without Randomization or Ignorability: A Stability Controlled Quasi-Experiment on the Prevention of Tuberculosis, *Statistics in Medicine*, Wiley Publishing, Volume 39, Issue 28, pages 4169-4186, DOI 10.1002/sim.8717

The article is in its published form, save for two minor changes in notation to ensure continuity with the rest of this thesis. The contributions of each author are as follows: C.H. was PI; led study design, manuscript writing, and revision; and had a near-equal role in analysis, result interpretation, and visualization. W.M. gathered the data and provided a domain framework and values for δ. D.A.W. led analysis, result interpretation, and visualization; and had a near-equal role in study design, manuscript writing, and revision.

With immense appreciation, I acknowledge the central role of Dr. Chad Hazlett in the completion of this thesis. His guidance, encouragement, and welcoming openness was the reason I pursued a PhD. His mentorship, generosity, and interest in my well-being made this process as stress-free as I could have hoped. I was lucky to have had him as my advisor, and the fields of Statistics and Political Science are lucky to have him among their ranks.

Thank you to my other committee members, Drs. Onyebuchi Arah, Mark Handcock, and Erin Hartman, who provided insightful comments, edits, and suggestions and accommodated particularly tight deadlines.

Thank you to my mother, who magically edited this thesis even though she didn't understand it.

Thank you to the Causal Inference Reading Group at UCLA for multiple rounds of edits, suggestions, and critiques along the way.

Thank you to my classmates from early in the program, and in particular to Sam, Sydney, and Steph, without whom I never would have survived my first quarter.

Thank you to Abby for helping me turn what could have been a COVID experience spent languishing into a productive, safe, and happy 2020.

Thank you to Patricia, who introduced me to statistics 8 years ago, and to her, Vinnie, and Gabriel, who encouraged me to go to graduate school.

VITA

2017–2019	**M.S.**, Statistics, Most Outstanding Masters Student Award, *University of California, Los Angeles*
2011–2015	**B.S.**, Mathematics, Minor in Computer Science, Cum Laude, *Tufts University*

EMPLOYMENT

2020	**Applied Scientist Intern**, Inference and Marketplace, *Etsy*
2020	**Teaching Assistant**, Department of Statistics, *UCLA*
2019–Present	**Student Consultant**, Statistical Consulting Center, *UCLA*
2019	**Data Scientist Intern**, *BuzzBuzzHome*
2015–2017	**Consulting Data Analyst**, Division of Research, *Kaiser Permanente*

PUBLICATIONS

Pre-existing Conditions in Hispanics/Latinxs that are COVID-19 Risk Factors, *iScience*, with Timothy Chang, et al. (2021)

Inference Without Randomization or Ignorability: A Stability Controlled Quasi-Experiment on the Prevention of Tuberculosis, *Statistics in Medicine*, with Chad Hazlett, et al. (2020)

Impact of Emergency Physician-Provided Patient Education about Alternative Care Venues, *American Journal of Managed Care*, with Pankaj Patel, et al. (2018)

The Impact of Acute Organ Dysfunction on Long-Term Survival in Sepsis, *Critical Care Medicine*, with Alejandro Schuler, et al. (2018)

Development and Validation of an Electronic Medical Record-Based Alert Score for Detection of Inpatient Deterioration Outside the ICU, *Journal of Biomedical Informatics*, with Patricia Kipnis, et al. (2016)

Credible Learning of Hydroxychloroquine and Dexamethasone Effects on COVID-19 Mortality Outside of Randomized Trials, *Under Review*, with Chad Hazlett, et al.

What Can We Learn from Observational Studies of New Treatments? Credible Conclusions for the Effects of Remdesivir for COVID-19, *In Preparation*, with Chad Hazlett, et al.

INVITED PRESENTATIONS

Causal Inference Without Randomization or Ignorability: A Stability-Controlled Quasi-Experiment on the Prevention of Tuberculosis, Poster Presentation, *International Conference of Health Policy Statistics*, San Diego, California (2020)

The Stability-Controlled Trial and Quasi-Experiment: Learning Effects of New Treatments Without Randomization (or Ignorability), *Department of Statistical Sciences Seminar*, University of Cape Town, South Africa (2019)

Using Observational Data Analysis to Understand the Impact of TB Treatments, with Chad Hazlett, *CEGA EASST Summit*, Nairobi, Kenya (2019)

CHAPTER 1

Introduction

Researchers often wish to use observational data, rather than or in addition to randomized trials, to estimate the causal effect of some treatment of interest on some outcome of interest. There are many observational causal inference strategies, each one deriving claims from their own set of required assumptions and each one implementable within particular study designs. Often, these methods adjust for a variety of covariates in order to account for baseline differences between treated and untreated units. They then rely on an assumption of "no unobserved confounding," i.e., that this covariate adjustment was successful and allows for identification of the true treatment effect. One recently proposed method, however, the Stability-Controlled Quasi-Experiment (SCQE, Hazlett, 2019), does not rely on covariate adjustment and instead invokes a different assumption to identify causal effects. When a research question is amenable to SCQE's required design, as many medical applications are, SCQE offers a useful opportunity for sensible inference and clear reasoning about identification uncertainty.

SCQE requires two cohorts of units with different usage levels of some treatment of interest. The method asks researchers to reason about and quantify unobservable baseline differences in the expected outcome between the cohorts. These differences, in other words, are how the two cohorts would compare on their units' average value of that outcome, *had there not been a difference in the treatment usage rate* between the cohorts. For any such difference, the SCQE can identify the effect of the treatment on those units affected by the treatment usage change. The assumption that SCQE requires, then, are the values of this counterfactual difference, the "baseline trend," that the researcher deems plausible. In return, SCQE provides the corresponding range of effect estimates that we should accept as plausible.

For a simple example, imagine a new medical treatment is introduced, and we observe some outcome of interest in the patients admitted before and after the introduction. Our baseline trend assumption would describe how, if no treatment had been introduced, the average outcome in the later cohort might reasonably be expected to differ from the average outcome in the earlier cohort. Whatever degree of confidence that domain knowledge or outside data permits us in reasoning about this trend, SCQE gives us a set of effects that the treatment may have had on those given the treatment, about which we should be equivalently confident.

In this dissertation, we conduct a thorough exposition of SCQE. In Chapter 2, we reprint our first published application of the method, in which we evaluated the effectiveness of a tuberculosis prevention therapy introduced in Tanzania. The paper, published in Statistics in Medicine in 2020 under the title "Inference without randomization or ignorability: A stability-controlled quasi-experiment on the prevention of tuberculosis," was co-authored by Chad Hazlett, Werner Maokola, and myself, and is left in its original form save for minor notational adjustments to avoid conflicts with the rest of this work. Here, it provides a more detailed introduction to SCQE and demonstrates our presentation of the method to an experienced, academic audience.

In Chapter 3, we introduce new technical advances, allowing for robust inference and implementation without unit-level data, accompanied by an application estimating the impact of in-hospital early warning systems. In Chapter 4, we present several methodological elaborations on SCQE, broadening its applicability and adapting it to handle alternative assumptions. We also demonstrate these extensions through two applications, one estimating the effectiveness of hydroxychloroquine for COVID-19 treatment and one conducting a re-analysis of data from Chapter 2. Finally, in Chapter 5, we share procedural guidance, important considerations, and best practices for practitioners interested in SCQE.

CHAPTER 2

An SCQE application on the prevention of tuberculosis

2.1 Introduction

In many cases we wish to learn the effect of a treatment that was made available without randomization. Examples include cases where randomization of life-saving treatments may not be ethical, where the "treatment" in question is a behavior adopted by patients (e.g. smoking), or where treatments are approved therapies but we wish to investigate their "real world effectiveness" in a different population.

Existing research designs to assess the effectiveness of treatments outside of a conventional randomized control trial (RCT) fall mainly into two categories. The first is a variety of designs that allow for partial self-selection, including "comprehensive cohort studies" and "patient preference trials" (Olschewski and Scheurlen, 1985; 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, 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 (Knox et al., 2019). These designs seek to solve the representational shortcomings of RCTs: The population that is randomized into treatment likely differs from those who would choose to receive it outside of a trial. They do not, however, avoid the need for randomization, and thus do not aid in understanding the effects of treatments already given outside of randomization, or for which randomization may be infeasible or unethical.

When randomization is not an option at all, however, observational studies are employed. Most commonly these call upon covariate-adjustment approaches. While these adjustment or conditioning techniques (e.g. regression, weighting, matching, and sub-classification) can differ in their requirements and in the estimands they commonly target, at minimum they each ask the investigator to assume that there are "no unobserved confounders." The central concern with all of these approaches is their susceptibility to arbitrarily large biases in either direction due to unobserved confounding (see Uddin et al., 2016 for a discussion of approaches under confounding).

This article illustrates how the stability-controlled quasi-experiment (SCQE; Hazlett, 2019) offers an alternative or complement to covariate-adjustment strategies, particularly when a treatment sees large increases or decreases in its use over time, and when the "no unobserved confounding" assumption is difficult to defend. This approach rests on an assumption of what the average outcome would have been for the entire group being studied (some of whom take the treatment), had nobody in that group been treated. This can be arrived at through several comparisons depending on the setting. In this case, we observe a prior cohort of patients, for whom no treatment is available, followed by a second cohort in which treatment is available without randomization. The assumption is then made on how much the *average non-treatment* outcome could have changed between cohorts (i.e. the "baseline trend") had there been no change in treatment. Such an assumption alone is sufficient to identify the average treatment effect on the treated (ATT), *regardless of unobserved confounding.* Though the ATT is not always an investigator's target quantity, it is a particularly valuable one when treatment is selectively given, as it tells us the average treatment effect over those who chose (or were chosen) to receive the treatment. This is especially relevant, for example, when retrospectively assessing the "real world effectiveness" of a treatment on those who actually received it.

In some cases the combination of effect size, sample size, and the ability to support a narrow assumption on the baseline trend result in 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 in which the treatment was introduced, a sharp estimate may be possible and credible. In cases where results are not as decisive, the approach nevertheless reveals what cannot reliably be concluded, and 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 be convincingly ruled out, we learn that we cannot justify a conclusion as to the sign of the estimate without further assumptions or argument. By contrast, comparisons based on covariate adjustment typically report results as if the assumption of zero confounding holds precisely, risking overconfident conclusions despite unknown bias.

This suggests several use cases for the SCQE approach. First, in cases where randomization is not possible, SCQE may offer a valuable way forward that reveals what can be claimed as a consequence of any assumption made on the baseline trend. Second, even when a randomized trial exists (or will later exist) for a given treatment, we may wish to study the real world effect that treatment has had on a population that actually used it. Finally, the approach may be useful not just retrospectively but in the design of a new kind of trial for cases in which randomization may be undesirable, such as with treatments for emerging or highly fatal diseases. In these settings, a trial could be designed whereby SCQE is implemented intentionally by ensuring that, when a new treatment is made available (by choice), there are not other changes in cohort composition or treatments. We call this design-based version of the approach the Stability Controlled Trial (SCT).

This article is the first demonstration of the SCQE approach, providing a practical test of this method, and exploring its differences and equivalences to a number of other approaches. 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 SCQE 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. We find a stark contrast between what policymakers are likely to conclude based on simple comparisons or covariate adjustment approaches, and what can be said after applying SCQE. Specifically, a naive (cross-sectional) comparison contrasts a TB incidence rate of 16% among the untreated with 0.5% among the treated in the same time period, arriving at an effect estimate of -15.5 percentage points (pp). Approaches that adjust for observables make little change to this estimate. For example, simple linear regression with all available covariates produces an estimate of -15.0 pp with a narrow confidence interval (-16.4, -13.7) and a t-statistic of -22.6. Other covariate-adjustment methods that target the ATT and that avoid restrictive specifications (matching, regression imputation using linear models, or flexible models such as extreme gradient boosting) all return similar results between -15.3 and -14.0 pp. By contrast, SCQE conveys results over a range of assumptions, showing that a beneficial effect is possible but uncertain. Under the assumption that the baseline trend is "flat," for example, the ATT estimate is -3.2 pp, but with a wide confidence interval including zero (-16.3, 8.4). Furthermore, under a reasonable range of assumptions on the baseline trend, either a beneficial or harmful effect can be found.

In what follows, Section 2.2 describes the proposed method and the inferential extensions developed here. Section 2.3 describes the application in greater detail and gives results. Section 2.4 discusses, compares the approach to other identification strategies, and concludes.

2.2 Stability-controlled quasi-experiments

2.2.1 Setup

We use the potential outcomes framework (Neyman, 1923; Rubin, 1974) in which each individual indexed by i from 1 to N has a (potential) outcome under treatment ($Y_i(1)$) and under non-treatment ($Y_i(0)$), regardless of their actual realized treatment status ($D_i \in \{0, 1\}$). The observed outcome, Y_i, is the treatment potential outcome for units taking treatment and the non-treatment potential outcome for untreated units, i.e. $Y_i = Y_i(0)(1 - D_i) + Y_i(1)D_i$. We consider two time periods: $Z_i = 0$ for those individuals observed before the treatment becomes available, and $Z_i = 1$ 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) $\{Y_i(1), Y_i(0), D_i, Z_i\}_{i=1}^{N}$ is assumed to be drawn independently from common joint density $p(Y(1), Y(0), D, Z)$.

The assumption that potential outcomes can be sufficiently indexed with a single treatment indicator (that of unit i), implies the assumption that only an individual's own treatment status matters and not that of others, often referred to as the Stable Unit Treatment Value Assumption (SUTVA; Rubin, 1990). For notational ease we often suppress the index i when referring to the common distribution, e.g. $\mathbb{E}[Y(0)|Z = 1]$. Finally, we denote the proportion of individuals taking the treatment at time $Z=0$ as $\pi_0 \equiv Pr(D=1|Z=0)$ and the proportion taking it at time $Z=1$ as $\pi_1 \equiv Pr(D=1|Z=1)$. 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 ($\pi_0 = 0$). Hazlett (2019) also generalizes this method to cases where $\pi_0 > 0$.

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,

$$\delta \equiv \mathbb{E}[Y(0)|Z=1] - \mathbb{E}[Y(0)|Z=0]. \tag{2.1}$$

While various sources of information may be useful to inform beliefs about δ, some 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. Delaying estimation and uncertainty concerns, adding δ to the mean observed (non-treatment) outcome in time period zero ($\mathbb{E}[Y(0)|Z = 0]$) gives the non-treatment outcome among the whole group found in time period one ($\mathbb{E}[Y(0)|Z = 1]$). This average is a weighted combination of two other averages: the average non-treatment outcome among the untreated in this time period ($\mathbb{E}[Y(0)|D{=}0, Z{=}1]$), which we observe, and the average non-treatment outcome among the treated ($\mathbb{E}[Y(0)|D{=}1, Z{=}1]$), for which we can now solve by applying the law of iterated expectations. That is,

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

which we can re-arrange to identify the unobservable non-treatment outcome among the treated ($\mathbb{E}[Y(0)|D{=}1, Z{=}1]$) in terms of observables,

$$\begin{aligned}\mathbb{E}[Y(0)|D{=}1, Z{=}1] &= \frac{\mathbb{E}[Y(0)|Z{=}0] - \mathbb{E}[Y(0)|D{=}0, Z{=}1](1 - \pi_1) + \delta}{\pi_1} \\ &= \frac{\mathbb{E}[Y|Z{=}0] - \mathbb{E}[Y|D{=}0, Z{=}1](1 - \pi_1) + \delta}{\pi_1}. \qquad (2.2)\end{aligned}$$

where we can replace $Y(0)$ in the conditional expectations on the right hand side by the observed outcome, Y, because these expressions involve only untreated units for whom the observed outcome is the non-treatment potential outcome. Note that in some cases intermediate quantities such as $\mathbb{E}[Y(0)|Z{=}1]$ could take on seemingly “impossible” values (say, a negative $\mathbb{E}[Y(0)|Z = 1]$ for a binary outcome) in finite samples. This is not unique to SCQE—one could see similar values for intermediate quantities computed under difference-in-differences assumptions, for example—and does not jeopardize the unbiasedness of the final ATT estimate.

Identification of the average non-treatment outcome among the treated in equation (2.2) can be of direct interest 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, $\mathbb{E}[Y(1)-Y(0)|D\!=\!1,Z\!=\!1]$,

$$\begin{aligned}\widehat{\text{ATT}} &= \mathbb{E}[Y(1)|D\!=\!1,Z\!=\!1]-\mathbb{E}[Y(0)|D\!=\!1,Z\!=\!1]\\ &= \mathbb{E}[Y|D\!=\!1,Z\!=\!1]-\left(\frac{\mathbb{E}[Y|Z\!=\!0]-\mathbb{E}[Y|D\!=\!0,Z\!=\!1](1-\pi_1)+\delta}{\pi_1}\right). \qquad (2.3)\end{aligned}$$

2.2.1.1 Optional uses of covariates in SCQE

This approach requires no covariates nor specification assumptions beyond those implied by the construction of δ as an additive quantity. However, covariates could be put to use in several ways. In principle covariates could simply be included in the equivalent, modified instrumental variable approach described below (Section 2.2.2). However, the more transparent use of covariates that we favor is to employ them as aids in forming beliefs about δ. The simplest such use is to see if there are large enough differences in the covariates' means (or distributions more generally) between the cohorts to warrant revision of our beliefs about compositional changes over time, that could in turn suggest a different range of δ to be considered.[1] Here for example, the mean ages of the first and second cohorts are 29.3 and 31.6 respectively; both are 34-35% female, and the mean WHO HIV disease stage is 2.3 for both groups. These differences do not suggest any change to the range of δ values considered plausible below. We note that a further way of using covariates, when their distribution does differ non-trivially between cohorts, would be to model changes in the non-treatment outcome so as to inform δ. For example, one could train a model that uses pre-treatment covariates X_i to predict $Y_i(0)$ only 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 those individuals. Taking the

[1] We thank an anonymous reviewer for suggesting this.

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. $Y(0) \perp\!\!\!\perp Z_i \mid X_i$. Note that, Z_i indicating time and not treatment status, this assumption is milder than the $Y(\mathrm{d}) \perp\!\!\!\perp D_i \mid X_i$ 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. Nevertheless, the non-necessity of covariates in SCQE and absence of related specification assumptions is also a strength, and experts may be better able to reason about an "all-inclusive" baseline trend than one that has removed the effects of observables.

2.2.2 Relationship to a modified instrumental variable approach

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,

$$\begin{aligned}
\widehat{\text{ATT}} &= \mathbb{E}[Y|D=1, Z=1] - \left(\frac{\mathbb{E}[Y|Z=0] - \mathbb{E}[Y|D=0, Z=1](1-\pi_1) + \delta}{\pi_1}\right) \\
&= \frac{1}{\pi_1}\left(\pi_1\mathbb{E}[Y|D=1, Z=1] + (1-\pi_1)\mathbb{E}[Y|D=0, Z=1] - \mathbb{E}[Y|Z=0] - \delta\right) \\
&= \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0] - \delta}{\pi_1}. \qquad (2.4)
\end{aligned}$$

This formulation suggests that time (i.e. cohort) can be thought of as an instrumental variable, or more intuitively an "encouragement" to receive the treatment, because the cohort at time $Z=1$ has a higher probability of taking treatment than the cohort at time $Z=0$. We refer readers to Baiocchi et al. (2014) for a tutorial on instrumental variables (IV) in medicine.

A conventional instrumental variable must satisfy the exclusion restriction: the instrument (time) can not cause the outcome to change except through treatment. Further, the relationship between the instrument (time) and the outcome can not be confounded by unobserved common causes. Violating either is problematic because it would have the consequence that $\mathbb{E}[Y(0)|Z=1] \neq \mathbb{E}[Y(0)|Z=0]$. It is this relationship that is relaxed by incorporating δ, instead allowing $\mathbb{E}[Y(0)|Z=1] = \mathbb{E}[Y(0)|Z=0] + \delta$. 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 a direct effect of time on the outcome or confounding of the outcome with time. A further assumption of IV is that the instrument is "relevant," i.e. influences treatment uptake. SCQE is applicable only when there is a large over-time shift in use of a treatment, ensuring relevance. We discuss the "strength" of the instrument in the applied example below. Finally, IV invokes a "monotonicity" assumption that holds by construction in this example: units have zero probability of taking the treatment at time $Z=0$, so their probability of taking it at time $Z=1$ can only be higher.

What follows is an algorithm for implementing SCQE by making adjustments to the standard IV machinery. Intuitively, because the non-treatment potential outcomes are "too high" by δ in the second period for the IV assumptions to hold, we must subtract δ off all observed outcomes in the second period. More formally, we define a new pseudo-outcome, $\tilde{Y} = Y_i - \delta Z_i$. The non-treatment potential outcome for $\tilde{Y}_i$ would thus be $\tilde{Y}_i(0) = Y_i(0) - \delta Z_i$. Because this adjusts for the shift between the mean of $Y(0)$ in the first and second cohorts, $\tilde{Y}_i(0)$ and Z_i are mean independent. Algebraically, $\mathbb{E}[\tilde{Y}_i(0)|Z_i=1] = \mathbb{E}[Y_i(0)|Z_i=1] - \delta = \mathbb{E}[Y_i(0)|Z_i=1] - (\mathbb{E}[Y_i(0)|Z_i=1] - \mathbb{E}[Y_i(0)|Z_i=0]) = \mathbb{E}[Y_i(0)|Z_i=0] = \mathbb{E}[\tilde{Y}_i(0)|Z_i=0]$. The same applies to the treatment potential outcomes. Put differently, any difference seen in the expected $\tilde{Y}_i$ between cohorts cannot be due to differences in their average non-treatment values (δ), but rather is generated by the treatment effect. The "reduced form" effect of time (difference in δ-adjusted average outcomes between $Z=1$ and $Z=0$) is the numerator in equation (2.4). Having restored the exclusion restriction, all of this difference must be caused only by the subset of units that were treated, and 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.

More generally, IV 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. This can be seen most simply by recalling that the ATT provided by the SCQE proves numerically equal to the "δ-adjusted" IV estimate. Equivalently, in the jargon of IV, the assumption that the treatment is newly available ($\pi_0 = 0$) implies "one-sided non-compliance," which makes the complier average treatment effect equal to the ATT (see e.g. Angrist and Pischke, 2008).

The re-expression of SCQE as modified IV also suggests a natural approach to variance estimation. In IV estimation, standard errors are derived by reference to the fitted "IV regression," which when written using our pseudo-outcome $\tilde{Y}_i$ becomes

$$\tilde{Y}_i = \hat{\beta}_0 + \hat{\beta}_{IV} D_i + \hat{\mu}_i. \tag{2.5}$$

In this expression, $\hat{\beta}_{IV}$ is the (δ-adjusted) IV estimate in which time (Z) is 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

$$\widehat{SE}(\hat{\beta}_{IV}) = \frac{\hat{\sigma}_\mu}{\sqrt{N}\, \hat{\rho}_{D,Z}\, \hat{\sigma}_D}, \tag{2.6}$$

where $\hat{\sigma}_\mu = \sum_i \hat{\mu}_i^2/(n-3)$, $\hat{\rho}_{D,Z}$ is the sample correlation of the treatment and time indicators, and $\hat{\sigma}_D$ is the sample standard deviation of the treatment indicator (Wooldridge, 2009).

Further, the more general estimator for the standard error, without assuming $\Sigma = \sigma^2 I$, is given by the “sandwich estimator” form,

$$\widehat{Var}(\hat{\beta}) = (\mathbf{Z}^\top \mathbf{D})^{-1} \mathbf{Z}^\top \hat{\mathbf{\Sigma}} \mathbf{Z} (\mathbf{Z}^\top \mathbf{D})^{-1} \tag{2.7}$$

where $\mathbf{Z} = [1 \ \ Z]$ 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, we employ the heteroskedasticity-robust form $\hat{\mathbf{\Sigma}} = diag(\hat{\mu}^2)$ below when we produce estimates within clinic (Cameron and Trivedi, 2005).

2.2.3 Pooled estimates and block bootstrap

As will often be the case for treatments attempted at multiple sites or other groupings, here we are interested in both the clinic-level estimates and in pooled estimates representative of all those treated in the study sample. To construct the pooled ATT estimate, we first obtain the moment estimates called for in equation 2.3 from each clinic m, from 1 to M, adjusting by δ at the clinic-level. From these clinic-level mean estimates, we can reconstruct the moment estimates for the entire pooled sample by weighted averaging. For example, the pooled estimate for the mean Y among all those in $Z\!=\!0$ is given by $\sum_{m=1}^{M} \overline{Y}_m \ Pr(\text{clinic}\!=\!m|Z\!=\!0)$. Each of the required pooled sample means is constructed in this way, weighting by the proportion of observations falling into that clinic among those in the relevant conditioning set. The ATT is then estimated for the entire pooled sample.

For the standard error of this pooled estimate, we wish to avoid assumptions of independence within clinics. In principle, one option would be to employ the cluster-robust standard error under the IV approach, using the observation-level data. However, such estimates rely on having a sufficient number of clusters (clinics), and so we instead employ a block bootstrap approach. Specifically, we bootstrap the pooled ATT estimate, resampling M clinics with replacement from the original M, generating a new ATT estimate on each iterate. After 100,000 iterations we construct the 95% confidence intervals around the full estimate from the 2.5^{th} and 97.5^{th} (empirical) percentiles of the bootstrap estimates.

2.3 Application: Evaluating IPT for TB prevention

2.3.1 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 et al., 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 recommends 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). Evaluations of the actual *effectiveness* of national IPT-promoting policies and programs have thus far relied on covariate adjustment to control for substantial differences between those patients prescribed and not prescribed IPT (Geremew et al., 2019; Sabasaba et al., 2019).

Beginning in 2011, Tanzania has made IPT available in HIV clinics and encourages its use through a nationwide clinician education program. Groups of individual clinics were selected in waves, and their clinicians received educational training encouraging IPT prescription for *all* HIV patients not yet diagnosed with active TB. Clinics were enrolled in these trainings 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 were enrolled. 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 prescribed only 25% of the time, in the clinics that adopted it at all. Table 2.1 shows, for each of the 21 clinics included in the analyses below, the time at which

IPT use effectively began, the TB development rates before and after introduction, and the level of IPT uptake.

No information is available about the process by which certain clinics were chosen rather than others. More importantly, there is virtually no information about the process by which certain patients were prescribed IPT while others were 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. This is exacerbated by the fewness of pre-treatment covariates—the clinic identifier, age, gender, and WHO HIV disease stage. With so little hope of understanding the treatment assignment process well enough to adjust for important confounders, covariate adjustment approaches are untrustworthy in this case; we employ them below for sake of comparison to standard practice and to contrast the types of conclusions one would draw under each method.

In principle each clinic provides an opportunity for a clinic-specific estimate. Alternatively, we can 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. 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.

Finally, recall that SCQE will estimate an average effect of IPT on TB incidence, among the treated. 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.

Table 2.1: Clinic Implementation of IPT

Clinic	Total	Pre-Implementation:	Post-Implementation:		IPT Implementation
Number	Patients	TB Rate	TB Rate	IPT Rate	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	0.24	
Total	26,715				

Note: Implementation details for the 21 clinics that qualify for an ATT estimate, as defined in Section 2.3.2. Total patients is the number of qualifying pre- and post-implementation patients, following the same criteria.

2.3.2 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. This was not recorded and thus had to be inferred. In each clinic, IPT prescriptions began suddenly, but with occasional prescriptions appearing 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 as erroneous. If any of these "ignored" IPT uses before that date were genuine, some effect of the treatment could be improperly experienced in the $Z = 0$ period. Likewise, there can be coding errors whereby a patient is coded as treated in $Z = 1$ when they should not have been. Both of these problems would lead to an understatement of the ATT. The effect estimates presented here showed no appreciable change when using the 1^{st} and 5^{th} percentile of prescription dates to determine IPT eligibility instead, 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.

In constructing the cohorts, we avoid coding procedures that could artificially generate compositional differences between the groups, which could influence the difference in non-treatment outcomes for these groups in ways the investigator is unlikely to consider when choosing an appropriate range for δ. We limit our data to (a) the first year of clinic visits for patients (b) who are found in the data for at least one year, (c) that whole year of which was contained within either their clinic's pre-IPT or post-IPT period, regardless of whether the patient received treatment. 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 crossover individuals—those coded to time $Z = 0$ but who later receive treatment during $Z = 1$ before having their outcome measured.

To determine each patient's outcome, we use a follow-up period of M_Y months. We also define an eligibility period 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 one-year periods 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. The choice of one year was based on the up-to 18 months of protection from TB that implementers expected IPT to provide (National AIDS Control Program, 2009). We also ran the same analyses using $M_D = M_Y = 18$ months, which reduced the number of clinics from 21 to 17 (and reduced the number patients within them) with available follow-up time in our data, but produced substantively similar conclusions (Figure A.1). Note that one could in principle set M_D to be shorter than M_Y, e.g. coding an individual's treatment status based on their first month ($M_D = 1$ month), but allowing 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 short time to see a benefit after taking treatment, but the upside of not coding anybody who received treatment before the end of the year as untreated. Figure A.2 in the Appendix shows results had we instead used $M_D = 6$ months and $M_Y = 12$ months. The results do not materially differ, but this reduces the number of clinics with sufficient data from 21 down to 15.

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. One virtue of these rules is that they avoid a scenario in which π_1 is too small, which can introduce bias referred to as the "weak instrument" problem in the IV literature. 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 to protect against weak instruments (Stock and Yogo, 2002). In practice our clinic-level F-statistics ranged from 26 to 1356.

2.3.3 Specifying plausible ranges for δ

While δ is unidentifiable, beliefs about its value can be informed by expert/domain knowledge or by data. Beginning with expert knowledge, one of the authors (Dr. Maokola) is a leading expert on IPT and TB in Tanzania. Prior to examining the data, we documented his beliefs about δ, which were that (a) there were no known changes in TB incidence rates in recent years or any medical or epidemiological reasons to expect a change, but (b) the *reported* TB incidence may have increased by 0.5pp to 1pp per year, due to improved surveillance. However, Dr. Maokola indicated low confidence in the coverage of this range. Note that some users may be interested in constructing an entire prior distribution over δ, but we prefer here to simply ask what values of δ are defensible, after which one can ask what ATT estimate would be implied by each such value.

Turning to data, we can also inform our beliefs by looking to trends from clinics that did not employ IPT over this time range, or to trends at a prior time in the clinics that did employ 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 outcome trends under non-treatment would have remained stable 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.

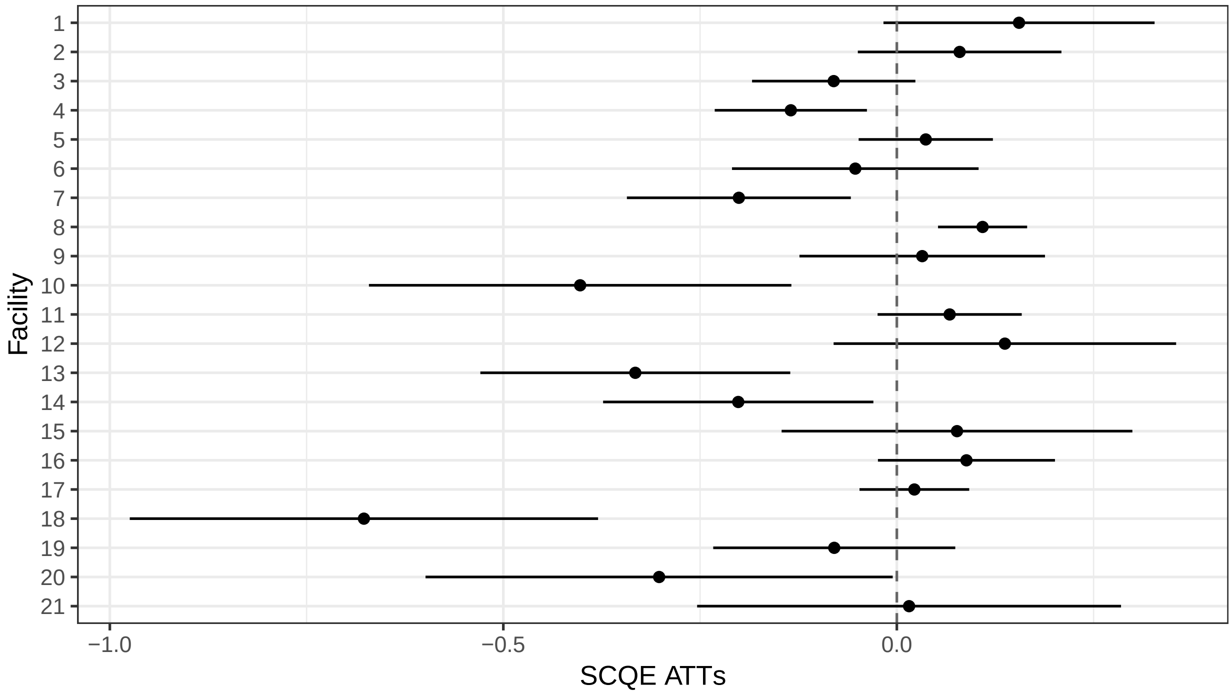
We can conceive of the trend over time as a linear one, or as an exponential or other non-linear rate (though the latter is then translated back into an absolute shift to apply the appropriate δ). An exponential trend is particularly reasonable given that the TB incidence rate is near zero in some clinics. 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 0.3 pp drop), with a 95% CI of [-0.0052, -0.0007]. By multiplying the linear estimate by the time between the pre-implementation and post-implementation periods in each clinic, we obtained that clinic's effective value of δ to be used in equation (2.3). In our data, this length of time between patients in the two cohorts averages to about three years. 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]). 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. 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 $\mathbb{E}[Y_0|Z=0]$ to obtain the corresponding value of δ for use in equation (2.3). 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 [-0.0136, -0.0038]). The results of these regressions vary depending on which subsets of informative data we use. See Table A.1 in the Appendix for details.

2.3.4 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 (2.3) with the appropriately-scaled values of δ. Standard errors are computed using the heteroskedasticity-consistent form of the adjusted IV approach as in equation (2.7). Note that the standard errors constructed at a given choice of δ account only for statistical uncertainty at that fixed δ. For illustration, we show the clinic-level estimates at $\delta = 0$ in Figure 2.1. 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 have confidence intervals that include zero.

If δ is positive, as proposed by Dr. Maokola due to increased reporting, these results would move to the left (more clinics would show beneficial results). If the true δ is negative (declining TB incidence), then the results move right, and we might find fewer clinics with

Figure 2.1: Clinic-Specific ATTs using $\delta = 0$

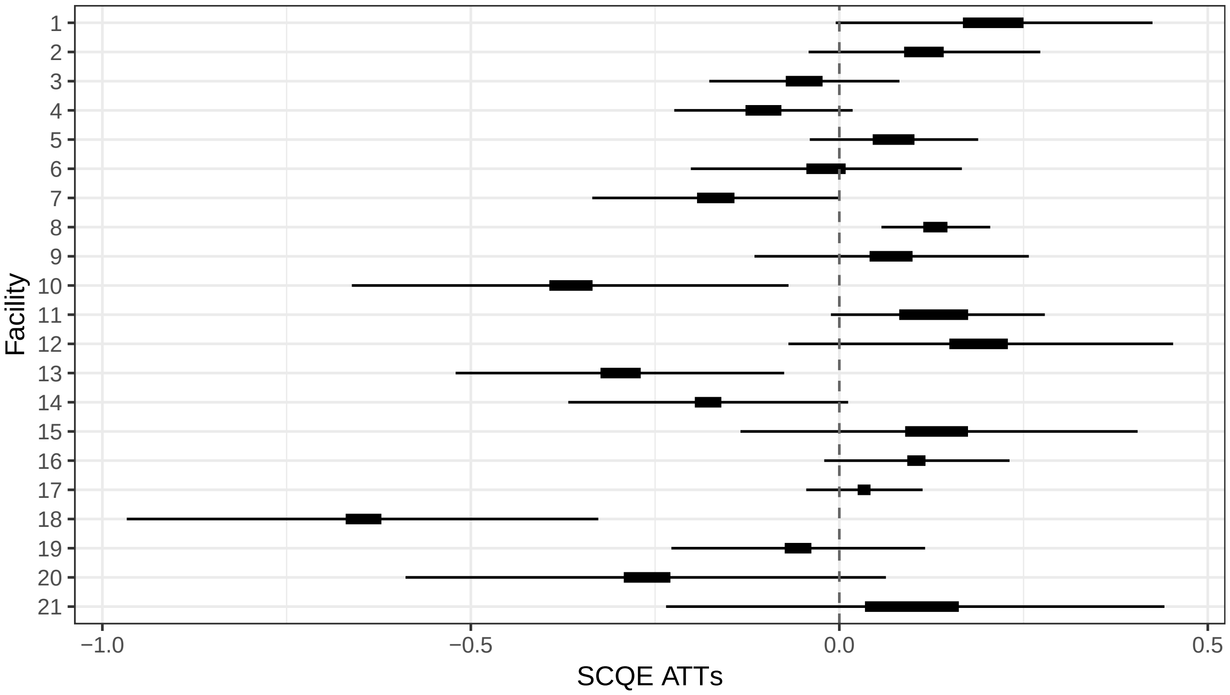


Note: ATT estimates for each clinic given δ is assumed to be 0. The whiskers represent the 95% CI using the IV estimator for the standard errors. The results appear to be significantly and substantively beneficial in seven clinics (those to the left of zero line, with the 95% confidence interval excluding 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.

clearly beneficial effects. More generally we advise considering a range of values for δ. A convenient graphical device for doing so is a "range-and-whisker" plot, which incorporates both identification uncertainty (due to the range of δ values) and the usual statistical uncertainty. Figure 2.2 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 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.3), and the range obtained by using an ex-

Figure 2.2: 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.3 and Figure A.4 for similar plots but using choices of δ generated from different sources.

ponential decay rate to learn from trends in the non-IPT data (Figure A.4). 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 harmful effects of IPT with augmented CIs excluding zero and only one shows evidence of a significant beneficial effect.

We have no arguments with which to reject the values of δ proposed by Dr. Maokola, nor 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. The clinic-level ATT estimates are fragile and not defensibly positive or negative in most clinics. This stands in stark contrast to estimates from naive comparison or covariate adjustment, which provide a large and confident estimate but under unknown amounts of confounding.

2.3.5 Pooled results

Our primary estimate of interest pools across clinics. From each clinic we record estimates of four expectations, $\mathbb{E}[Y|Z=0]$, $\mathbb{E}[Y|D=1,Z=1]$, $\mathbb{E}[Y|D=0,Z=1]$, and π_1, as well as the time-gap between $Z=0$ and $Z=1$. 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 $Z=0$ and $Z=1$, which varies slightly by clinic. By determining the correct δ for each clinic, we can then construct the clinic level estimate of $\mathbb{E}[Y(0)|Z=1]$. We then pool data across clinics, weighting as described above, to construct the sample moments required to compute the pooled ATT,

$$\widehat{\text{ATT}} = \mathbb{E}[Y|D=1,Z=1] - \left(\frac{\mathbb{E}[Y(0)|Z=1] - \mathbb{E}[Y|D=0,Z=1](1-\pi_1)}{\pi_1} \right). \tag{2.8}$$

Note that equation (2.8) is simply equation (2.3) but in which the δ has been added to $\mathbb{E}[Y|Z=0]$ to form $\mathbb{E}[Y(0)|Z=1]$ within each clinic first.

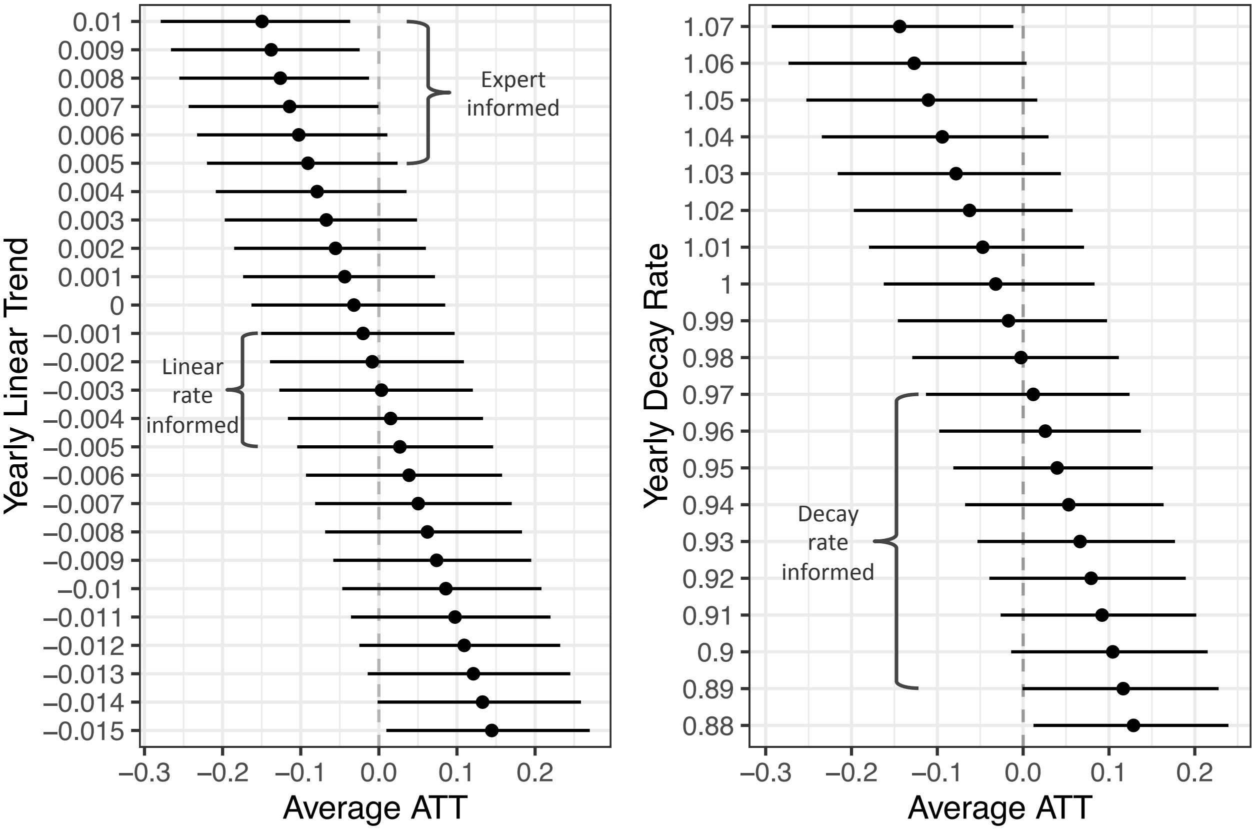
We find that prior to IPT, the pooled average TB incidence rate was 13%. After IPT became available, 24% of patients were prescribed IPT. The observed average TB incidence rate for those who did not receive IPT after it was introduced was 16%, slightly higher than the 13% among the overall pre-IPT cohort. But the incidence was radically lower for those who received IPT, at 0.5%. Was the large difference in TB rates between those who received IPT (0.5%) and those who did not (16%) in the later period due to an effect of IPT, or a selection process? Suppose momentarily we employ $\delta = 0$. Applying equation (2.8) we get an ATT estimate of -3 pp, with a 95% confidence interval widely including zero.

To reinforce intuition, we can produce the same estimate logically as follows. Supposing $\delta = 0$, the expected non-treatment outcome over *everybody* at time $Z=1$ is 13%. To maintain this while observing the outcome of 16% among 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 3.5%. Comparing this to the observed outcome of the treated (0.5%) gives our ATT (-3 pp).

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 ($0.5\% - 16\% = -15.5$ pp) into a point estimate for the ATT (-3 pp) and "selection bias" that tells us how the treated and untreated differ on their non-treatment outcomes ($3.5\% - 16\% = -12.5$ pp). That is, the group receiving treatment was 12.5 pp less likely to have developed TB even in the absence of the treatment. Relaxing the $\delta = 0$ assumption, we would find that as long as the baseline trend was not 1 pp per year or larger, IPT must have been directed towards those who were less likely to develop TB (as, over the roughly three year gap between the cohorts, this trend would bridge the gap between the two cohorts' observed untreated outcomes).

We have thus dealt with selection concerns not by assuming the observability of all confounders, but through an assumption on δ. Figure 2.3 is more comprehensive and our preferred means of reporting results, visualizing the estimates obtained under varying choices of δ. The left panel of Figure 2.3 shows how assumptions on a linear trend in the non-treatment outcome generate varying estimates. 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 (0.5 pp) to 0.01 per year, the resultant ATT estimates range in (beneficial) non-significant to significant effects on the TB incidence rate of -9 to -15 pp. By contrast, the data-informed choice of δ based on linear trends in the non-IPT data suggests a range of -0.005 to -0.001 per year. These correspond to small and non-significant estimated ATTs. Finally, the right panel of Figure 2.3 indexes estimates by the annualized decay rate used to formulate δ. The data-driven assumption that decay rates vary from 0.89 to 0.97 produces ATT estimates

Figure 2.3: 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. The equivalent effective values of δ are shown in the corresponding tables A.2 and A.3, respectively. Confidence intervals were generated using the block bootstrap method described in Section 2.2.2. For each assumption on the baseline trend (vertical axis) there is a consequent ATT estimate with its 95% confidence interval. Under the "expert informed" assumption that the non-treatment average TB incidence would rise by 0.5 to 1 pp (0.005 to 0.010) per year, we see a combination of non-significant and significant negative (beneficial) ATT estimates (*left*) ranging from -9 to -15 pp. A data-assisted choice of δ under an linear model suggests annualized trends of -0.005 to -0.001, still on the *left*, which correspond to non-significant ATTs. A data-assisted choice of δ under an exponential decay model suggested annualized decay rates of 0.89 to 0.97. On the *right*, we see these decay rates correspond to non-significant positive (harmful) estimates.

ranging from a 12 pp harmful (though non-significant) effect of IPT down to an estimate of approximately zero.

2.4 Relationships to other approaches

In what follows, we compare SCQE to other commonly used identification approaches for observational evaluations.

2.4.1 Covariate adjustment under "no unobserved confounding"

A common identification strategy for making causal claims from observational data is to assume that covariates account for all confounders, and thus there is "no unobserved confounding". In the potential outcomes tradition, this is to assume the independence of the potential outcomes with the treatment conditionally on (pre-treatment, non-colliding) covariates $\mathbf{X}$, i.e. $\{Y(1), Y(0)\} \perp\!\!\!\perp D \mid \mathbf{X}$ (Rubin, 1990). Equivalently, with structural causal models or their graphical representations, $\mathbf{X}$ must satisfy the "backdoor criterion" of Pearl (2009), which requires that conditioning on $\mathbf{X}$ is able to rule out non-causal paths from the treatment to the outcome while avoiding creating new such paths. To employ this identification strategy requires an estimation procedure that attempts to implement the conditioning on $\mathbf{X}$. Examples include regression approaches that estimate treatment effects as coefficients, regression-based imputation approaches, matching, weighting for covariate balance, weighting/matching with propensity scores, and stratification/sub-classification estimators. We refer to these broadly as covariate adjustment procedures.

In this application, as in many others, investigators would struggle to defend the assumption that all confounding variables have been observed. Even experts on this case cannot claim to know all confounding factors. Further, the set of reliably-measured covariates is small (patient age, sex, date of first visit, and HIV severity at first visit). Yet, the indefensible nature of the "no unobserved confounders" assumption has not prevented investigators from turning to covariate adjustment approaches here as in many other settings. Using a similar cohort in Tanzania to study the effects of IPT on TB incidence, Sabasaba et al. (2019) obtain an incidence rate ratio of 0.52 after covariate adjustment. Geremew et al. (2019), Assebe et al. (2015), and Temesgen et al. (2019) similarly produce estimates ranging from 0.04 to 0.50 in analogous studies in Ethiopia.

Given that covariate-adjustments approaches remain standard in such settings despite the serious threat of unobserved confounding, we provide a series of covariate-adjustment estimates here to show how the results compare to those of SCQE and, more importantly, to illustrate the risks they pose in terms of generating overconfident inferences around an arbitrarily biased estimate.

Recall first that the naive cross-sectional comparison in the post-treatment period shows a 15.5 pp lower TB incidence among those taking IPT compared to those not taking it. A simple linear regression in the post-treatment period of our data, adding only clinic fixed effects to this formulation, produces a similar estimate (-15.6 [-17.0, -14.3] pp, t = -22.5). Adding all available covariates, the result remains similar, at -15.0 [-16.4, -13.7] pp, t = -22.6.

Further, while all of these estimators are employed under the same identification strategy—no unobserved confounding—they can vary in their specification assumptions, model dependency, requirements for overlap/common support, and their default choice of estimand (e.g. ATT, ATE, or an assumption of constant effects). To focus on differences in identification strategies rather than these other features, we also employ a range of covariate-adjustment approaches that (i) weaken or vary the specification assumptions, and (ii) estimate the ATT (rather than the ATE, for example), as this is the estimand targeted by SCQE.[2] First, to extract ATT estimates from OLS we use a (linear) regression imputation estimator. This is done by separately fitting two models, $\hat{\mu}_1(X) \approx \mathbb{E}[Y(1)|X, D=1]$ and $\hat{\mu}_0(X) \approx \mathbb{E}[Y(0)|X, D=0]$; predicting outcomes from both models for all treated units; and computing the average difference to obtain an ATT. Next, we can construct a similar imputation regression estimator for the ATT but fitting μ_0 and μ_1 using a more flexible machine learning method rather than linear regression. We chose extreme gradient boosting (XG-

[2] We thank an anonymous reviewer for suggesting this.

Boost; Chen and Guestrin, 2016), an ensemble method that employs a series of regression trees, each trained on the error of the preceding one. Third, another way to relax specification assumptions is to simply use a matching estimator, also targeting the ATT, pairing each treated individual to a non-treated one at the same clinic and with similar characteristics.

All of these examples of covariate adjustment techniques produce point estimates between -15.3 and -14.0 pp. Details are provided in Appendix A.1.1. By comparison, SCQE does not emphasize a single estimate and confidence interval, but instead shows a range of estimates that result from plausible assumptions on δ as in Figure 2.3. The results of SCQE at many plausible values of δ are far more modest than those of the covariate adjustment methods, and often not distinguishable from zero. SCQE also reveals the implicit assumption on the baseline trend being made under covariate adjustment: to argue that the covariate adjustment result is correct requires arguing that in the absence of IPT the incidence of TB would have risen by roughly 1 pp (or 7%) per year. This corresponds to the most extreme choice offered by our expert, and falls far from what is suggested by the empirical trends from untreated outcomes. While possibly correct, we do not have information or beliefs that would lead to defending such a claim with certainty.

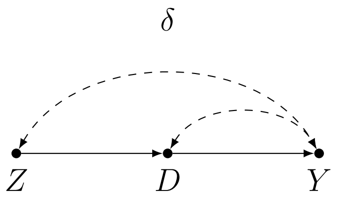
We note that regression, matching, and other adjustment approaches can be subjected to sensitivity analyses (see Zhang et al., 2018 for an overview), which we endorse. One virtue of SCQE, where it applies, is that the assumption it invokes and allows sensitivity analysis on (the baseline trend assumption, δ) is easy to describe and contemplate. Another is that it avoids privileging one estimate based on a frail assumption and instead requires the user to defend an assumption to defend an estimate. Nevertheless, whether more is to be gained by considering assumptions on δ or on how strongly imagined confounding is related to treatment and to the outcome (as many sensitivity analyses do, e.g. Arah, 2017; VanderWeele and Ding, 2017; Cinelli and Hazlett, 2020), will depend on the application. SCQE and other approaches invoking sensitivity analyses may be complementary in some contexts as they provide separate views into the assumptions under which a particular conclusion can be reached.

2.4.2 Instrumental variables

As shown above, SCQE is equivalent to 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, Figure 2.4) shows time influencing treatment uptake, and (exclusively) through it the outcome, while the treatment and outcome may be connected by unobserved confounding. To account for $\delta \neq 0$ we annotate the IV graph with an additional bi-directed arc between Z and Y, breaking the exclusion and exogeneity restrictions. We deviate from conventional graph notation, first, by allowing this bidirected arc to stand in for either common-cause confounding of Z and Y, or an effect of Z on Y not through D. Second, we label this edge with δ, whose precise meaning is given by the (parametric) condition $\mathbb{E}[Y(0)|Z=1] - \mathbb{E}[Y(0)|Z=0] = \delta$, or in the *do* notation of Pearl (2009), $\mathbb{E}[Y|do(D=0), Z=1] - \mathbb{E}[Y|do(D=0), Z=0] = \delta$. We also note that, as with conventional IV, there is an additional assumption of no-defiers (montonicity) that is not represented on the DAG because it is a parametric assumption. This assumption holds mechanically in our application because no person at time $Z=0$ can get the treatment.

Though the idea of using time as an instrument may be awkward because time is not a literal cause, it has been exploited elsewhere in the health sciences (Johnston et al.,

Figure 2.4: Graphical representation of the SCQE approach



Note: DAG for SCQE. $Z \in \{0,1\}$ is the time period, $D \in \{0,1\}$ is treatment status, and Y is the outcome. The addition of δ and the edge it marks conveys that the usual exclusion and exogeneity assumption required in the IV setting does not hold.

2008; Cain et al., 2009; Shetty et al., 2009; Mack et al., 2015; Gokhale et al., 2018, see also Brookhart et al., 2010; Streeter et al., 2017 for discussion). Nevertheless, the description of the strategy here—i.e. using the observed outcome for the untreated and an assumption on the baseline trend to back-out the non-treatment outcome among the treated—seems not to have been offered in previous treatments of time as an instrument. Further, whereas IV results are conventionally reported as if the exclusion and exogeneity assumptions hold exactly and with certainty, SCQE both allows deviation from this ($\delta \neq 0$) and encourages consideration over the plausible range of δ values (see Ji et al., 2017 for an alternative proposal for modifying IV assumptions).

2.4.3 Difference-in-differences

The use of over-time comparisons may also call to mind the difference-in-differences (DID) approach. DID begins by assuming "parallel trends," i.e. that the change in non-treatment outcomes for a group that never received treatment, $\mathbb{E}[Y(0)|Z=1,D=0]-\mathbb{E}[Y(0)|Z=0,D=0]$, equals the (unobservable) change in non-treatment outcomes for the group who received treatment in the second period, $\mathbb{E}[Y(0)|Z=1,D=1]-\mathbb{E}[Y(0)|Z=0,D=1]$ (Meyer, 1995). The DID estimand is the over-time change in outcomes for the treated group minus the over-time change in outcomes for the untreated group, i.e. $(\mathbb{E}[Y(1)|Z=1,D=1]-\mathbb{E}[Y(0)|Z=0,D=1])-(\mathbb{E}[Y(1)|Z=1,D=0]-\mathbb{E}[Y(0)|Z=0,D=0])$. Under the parallel trends assumption, this identifies the ATT.

The first distinction to make between DID and SCQE regards the data structure required for DID to be possible. In the DID setting the investigator must be able to label each observation as one that "would get treated" had it appeared in the second cohort. This is possible either with panel data, or with cross-sectional data when individuals belong to larger units (e.g. clinics, here) and *everybody* in that larger unit is treated in the second period. By contrast in other settings, including this application, we do not have this structure: for an individual observed at time $Z=0$, we cannot group them as "would be treated" or "would not be treated". Thus, SCQE is possible but DID is not in this case. More generally, SCQE is particularly appropriate in cases such as the introduction of a new treatment or policy with individualized self-selection, where the DID data structure does not typically exist.

Second, in settings where DID is possible, it can be understood as a special case of SCQE. The parallel trends assumption is often difficult to defend, because treated units may have selected treatment for reasons that relate to their trends in $Y(0)$. In an attempt to relax this assumption, methods such as the differentially-adjusted DID allow users to specify differing trends for the treated and control groups, based on pre-treatment histories (Bell et al., 1999). SCQE instead asks how the earlier and later *cohorts* differ on their mean $Y(0)$. Any such assumption can be algebraically transformed into a difference-in-trends assumption and vice-versa. The virtue of focusing on the change in mean $Y(0)$, however, is that it avoids asking for an assumption regarding how the treated and control differ in their trends, which is complicated by the selection process. Further, by varying δ, the SCQE approach has a built-in sensitivity analysis to avoid overconfidence in any single assumption.

2.4.4 Interrupted time-series and (fuzzy) regression discontinuity

Two other approaches that may seem to be related include the interrupted time-series (ITS, see Hudson et al., 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). Both seek to identify causal effects by virtue of a rapid change in the probability of treatment and require assumptions on the continuity, smoothness, or function space for $\mathbb{E}[Y(\mathrm{d})|Z]$, the expected potential outcome conditional on time. For example, a simple ITS estimator for an intervention introduced at time $Z = z_D$ may take the form of the regression model

$$Y = \beta_0 + \beta_1 Z + \beta_2 \mathbb{1}_{z_D} + \beta_3 (Z - z_D) \mathbb{1}_{z_D}, \tag{2.9}$$

where $\mathbb{1}_{z_D}$ is an indicator for $Z \geq z_D$, and β_2 and β_3 represent changes in the level and trend of the outcome, respectively, due to the intervention (Wagner et al., 2002). In RDD in time, the treatment effect locally at the time of intervention, $\mathbb{E}[Y(1) - Y(0)|Z = z_D]$ is identified by an assumption such as continuity of the potential outcomes in the interval around $Z = z_d$ (Lee and Lemieux, 2010), and is estimated by an estimator of the form

$$\lim_{\epsilon \downarrow 0} \mathbb{E}[Y(1)|\, Z = z_D + \epsilon] \;-\; \lim_{\epsilon \uparrow 0} \mathbb{E}[Y(0)|\, Z = z_D + \epsilon], \tag{2.10}$$

where each conditional expectation is typically fitted by a (separate) flexible, local model.

One major difference between these approaches and SCQE is again in the contexts in which they apply. These approaches are best suited to data in which we can code 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 precise moment in time, this would 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. Such scenarios instead require "binning" observations into wide cohorts, 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" stems from precisely this type of concern, viewing an indicator for being in the post-treatment period as an encouragement (instrument) for treatment (Trochim, 1984). In this sense, SCQE is a version of fuzzy RDD, but (a) comparing wider time bins and (b) allowing δ to account for possible shifts in $Y(0)$ over time between these bins rather than depending on a model, except possibly to inform the choice of δ.

2.5 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, just 0.5% of patients on IPT developed active TB compared to 16% of patients who were not. Covariate adjustment by regression similarly produces an estimate of -15.0 pp ($t = -22.6$) which may appear convincing both because it accounts for covariates and is highly statistically significant. Other adjustment approaches (e.g. matching, regression imputation with flexible models) produce similar estimates. Our concern is that despite any warnings authors invoke that such a result is "only suggestive," it is reasonable to expect that even sophisticated consumers of such analyses will see such a result as their best means of using data to inform policy, in the absence of other information. But to construct 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 baseline 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 a flat baseline trend ($\delta = 0$) immediately suggests the evidence is not strong, producing an ATT estimate that is not distinguishable from zero (-3 pp, [-18,6]). Further, 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 0.7 pp 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 1.5 pp 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 $\delta \leq 1$ pp 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 useful information for policymakers who may wish to improve this program's reach. Further, it is good news in the sense that the benefit of IPT can potentially be amplified through programmatic changes that reach higher risk individuals who may thus have more to gain from IPT than those who have taken it thus far.

In conclusion, the SCQE relies not on the absence of unobserved confounding, but instead on a user-chosen assumption on the baseline trend that is both easy to understand and allowed to vary. This makes it valuable as a complement or substitute for covariate adjustment approaches when investigating the real world effects of newly introduced treatments or in other observational settings where unconfoundedness can be difficult to defend. The SCQE offers an estimate of the ATT only when a treatment increases or decreases in popularity substantially between two periods, and when we can measure the outcome rates before and after the increase. It may produce sharp and definite conclusions, particularly where effects are strong and/or narrow assumptions on the range of δ can be supported due to the nature of the application. In other cases, as here, the assumptions we can make on δ may support a 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 and false confidence. It instead describes what assumptions about the baseline trend would have to be ruled out (or in) to argue that there was a beneficial or harmful effect.

CHAPTER 3

Inference

Since the publication of Chapter 2 (as Hazlett et al., 2020), we have developed several improvements to, extensions for, and generalizations of SCQE. This chapter covers the technical inferential improvements underlying some of these advances, and introduces the first implementation and application they allow. We briefly reiterate the parts of Chapter 2 that are relevant to these improvements and further detail the SCQE-IV relationship in Section 3.1. In Section 3.2, we introduce our first improvement: a newly derived inferential tool in the form of confidence sets that are robust to "weak instruments." In Section 3.3, we demonstrate that when the cohort, treatment, and outcome are all dichotomous, SCQE's estimate and these new confidence sets can be calculated from 8 or fewer summary statistics (counts or proportions) from the data, rather than requiring unit-level data. This finding is particularly helpful in the face of data sharing restrictions, data access problems, and privacy concerns. We also introduce a web application implementing this count-based inference. Finally, in Section 3.4, we implement both of these advances in an application, re-analyzing a study that estimates the impact of in-hospital rapid response systems on clinical deterioration.

3.1 Reintroducing SCQE and its relationship to IV

Inferring causal effects necessarily requires one or more assumptions, and the most reliable tool to satisfy a sufficient set of assumptions is treatment randomization. In observational settings, however, the choice of assumptions and defense of their verity become the central steps in the inferential process. The most commonly employed methods use the conditional ignorability assumption, but we know that any residual confounding that is improperly assumed away can nullify or even reverse the sign of the posited causal effect, which makes clear the importance of using these assumptions carefully, accompanied by justification. Numerous strategies for sensitivity analyses of these effect estimates have been proposed (e.g. VanderWeele and Ding, 2017; Cinelli and Hazlett, 2020; Broderick et al., 2020) and are widely used.

In Chapter 2, we described the Stability-Controlled Quasi-Experiment (SCQE) introduced by Hazlett (2019), which does not rely on the conditional ignorability assumption. SCQE can be applied in its simplest form when a treatment of interest is introduced between two cohorts of units, and relies instead on an assumed set of cohort-to-cohort baseline trends, δ, that the researchers deem plausible. For any assumption about how an outcome of interest's expected value would have shifted over time had the treatment usage change not occurred, SCQE can provide the consequent ATT estimate for the units that were treated. The range of the effect estimates produced by a plausible range of δ values tells us what causal claims (beneficial, harmful, or unknown) are defensible. SCQE is thus a "partial identification" strategy: our conclusions are dependent on the value of δ, and if we want to defend a given conclusion we must argue for the corresponding range of δ (see Section 5.7 for more).

We start with similar notation to that of Section 2.2, but make a few additions. First, we explicitly define the $Z = 0$ cohort as the one with less treatment use and call it the "low-use cohort," while we call $Z = 1$ the "high-use cohort." We remain focused on the case with no treatment in the low-use cohort ($\pi_0 = 0$), making it a "no-use cohort" instead, but the adjustments necessary for the $\pi_0 > 0$ case are addressed in Section 4.1. Second, we introduce notation for specific treatment regimes: just as we can imagine a unit's potential outcome under treatment ($Y_i(1)$) or non-treatment ($Y_i(0)$), we can also refer to potential outcomes under the specific treatment practices used in a low-use cohort ($Y_i(\mathrm{d}_{Z=0})$) or high-use cohort ($Y_i(\mathrm{d}_{Z=1})$), or under an unspecified treatment regime from among or beyond these four (Y_i(d)). Third, although the central assumption is again defined as

$$\delta \equiv \mathbb{E}[Y(0)|Z\!=\!1] - \mathbb{E}[Y(0)|Z\!=\!0], \tag{3.1}$$

the assumption of additive treatment effects mentioned in Section 2.2.1.1 means that δ can alternatively be defined using potential outcomes under any fixed assignment regime $Y(\mathrm{d})$, not just $Y(0)$.

We formally define our ATT estimand, the difference between the treated units' average outcome under treatment and their average outcome under non-treatment, with

$$\mathbb{E}[Y(1)|D\!=\!1, Z\!=\!1] - \mathbb{E}[Y(0)|D\!=\!1, Z\!=\!1] \tag{3.2}$$

Like all causal questions this presents a missing data problem and requires the use of some assumption, as only the first of these two terms is observable. We derived one solution to this problem, the SCQE ATT estimator, in Section 2.2:

$$\widehat{\mathrm{ATT}} = \mathbb{E}[Y(1)|D\!=\!1, Z\!=\!1] - \left(\frac{\mathbb{E}[Y(0)|Z\!=\!0] - \mathbb{E}[Y(0)|D\!=\!0, Z\!=\!1](1\!-\!\pi_1) + \delta}{\pi_1}\right) \tag{3.3}$$

This formulation is our preferred means of expressing the SCQE estimator because it matches the form of the estimand (equation 3.2), thereby not only providing the estimated treated effect (its magnitude) but placing that effect as well (its location). In other words, the two sides of the minus sign show the expected outcome rates that this treatment effect brought the treated units *to* and *from*, not just the difference between those values (this is discussed further in Section 4.1). However, as noted in Chapter 2, there are other benefits of rewriting our estimator as

$$\widehat{\mathrm{ATT}} = \frac{\mathbb{E}[Y|Z\!=\!1] - \mathbb{E}[Y|Z\!=\!0] - \delta}{\pi_1} \tag{3.4}$$

or, when $\delta = 0$, as

$$\widehat{\text{ATT}} = \frac{\mathbb{E}[Y|Z\!=\!1] - \mathbb{E}[Y|Z\!=\!0]}{\pi_1} \tag{3.5}$$

Equation 3.5 is the Wald estimator, used as the IV estimator for the ATT of D on Y given an instrument Z. Though this form is less amenable to effect placement than equation 3.3, we begin to see its usefulness for inference when we investigate the standard assumptions needed for Z to be a valid instrument:

1. The relevance assumption: Z must have a non-zero effect on D.
2. The exogeneity assumption: $Y(\text{d}) \perp\!\!\!\perp Z$, which is often divided into two separate assumptions:
   (a) The exclusion restriction: Z cannot affect Y except by way of affecting D.
   (b) The exchangeability assumption: Z and Y have no common causes.
3. The monotonicity, or no defiers, assumption: A unit that would have received the treatment had it not received the instrument will necessarily receive the treatment if it receives the instrument.[1]

Notably, we can restate the IV exogeneity assumption as a claim that there is no cohort-to-cohort difference between outcomes under some set treatment practice $Y(\text{d})$. The corresponding SCQE assumption of $\delta = 0$, then, guarantees a central requirement for IV validity, and allows us to borrow the various inferential tools available in the IV literature. If $\delta \neq 0$ however, the exogeneity assumption necessarily does not hold, and we lose such a parallel. Section 2.2.2 allows us to retain the useful equivalence by using δ to adjust our outcome of interest, not indirectly through its inclusion in equation 3.4, but by directly altering the outcome used in equation 3.5; instead of using Y itself, we use the δ-adjusted

[1] There are other assumptions that can take the place of the monotonicity assumption to produce valid IV, but here we only consider this version of the method. This assumption holds by definition given our simplifying assumption of $\pi_0 = 0$, but must be defended in the alternative case discussed in Section 4.1.

$\tilde{Y} = Y - \delta Z$. Replacing Y with $\tilde{Y} + \delta Z$ in equation 3.1, we find that

$$\begin{aligned}
\delta &= \mathbb{E}[\tilde{Y}(0) + \delta Z | Z{=}1] - \mathbb{E}[\tilde{Y}(0) + \delta Z | Z{=}0] \\
\delta &= \mathbb{E}[\tilde{Y}(0) | Z{=}1] + \delta - \mathbb{E}[\tilde{Y}(0) | Z{=}0] \\
\mathbb{E}[\tilde{Y}(0) | Z{=}0] &= \mathbb{E}[\tilde{Y}(0) | Z{=}1]
\end{aligned}$$

This equivalence shows that, for our assumed δ value, the IV exogeneity assumption will hold if we use $\tilde{Y}$ instead of Y as our outcome, which allows us to port IV's tools for inference into SCQE.

In utilizing this parallel, the partial identification strategy in Chapter 2 – expressing SCQE estimates across a plausible range of assumptions – applied our transformation of Y to $\tilde{Y}$ repeatedly, at each of a series of δ values that span such a range. As published, however, that paper suggested and used tools for inference (the "sandwich" estimator and a bootstrapping procedure) that are vulnerable to violations of a different IV requirement, the relevance assumption. When faced with a "weak instrument" (when the effect of the instrument on the treatment is insufficiently powerful), these classical tools may produce confidence intervals with insufficient coverage. As an alternative we introduce Anderson-Rubin confidence sets, which are robust to weak instruments, in Section 3.2 and we demonstrate how they may be generated in the SCQE setting (Anderson et al., 1949).

Before doing so, we note that many sensitivity analyses have been developed in the IV literature (e.g. Small and Rosenbaum, 2008; Conley et al., 2012; van Kippersluis and Rietveld, 2018; Cinelli and Hazlett, 2021). While in many cases IV assumptions may be more believable than the conditional ignorability of the treatment, a violation of the IV assumptions can similarly cause an estimated treatment effect to diverge arbitrarily far from the true effect. IV sensitivity analyses are built to address these concerns in a variety of general IV settings. In specific applications where SCQE can be sensibly implemented, however, we believe it offers a sensitivity analysis that uniquely utilizes a human ability to reason about over-time trends (see Section 5.1 for more).

3.2 Robust inference in SCQE

3.2.1 Adapting Anderson-Rubin confidence sets to SCQE

3.2.1.1 Concepts behind these confidence sets

At a high level, the construction of Anderson-Rubin (AR) confidence sets in the standard IV setting works as follows: Given a valid instrument, Z, we know that D's effect on Y is the only means by which Z can affect Y. If we knew that τ was the true effect of D on Y, adjusting Y by removing that effect would render Z unrelated to the adjusted Y. Thus, by removing some *presumed* effect, τ_0, from Y and testing whether we can reject a null hypothesis that no relationship between Z and the adjusted Y remains, we can decide whether to reject τ_0's potential as the true treatment effect τ. The AR confidence set is then the collection of τ_0 values for which we cannot reject such a test. The appearance of a confidence set depends on the data itself; while the sets will often resemble a standard confidence interval, they may at times be disjoint or contain the entire real line.

A potential source of confusion to note before introducing notation is that we are now making a second adjustment to Y. The δ adjustment, central to the SCQE design, adjusted Y in terms of its relationship to Z, and the τ_0 adjustment, central to the AR process, adjusts Y in terms of its relationship to D. The former is an acknowledgement that the exogeneity assumption is not valid with respect to an unadjusted Y, while the latter is simply a tool used to estimate a confidence set.

Just as standard errors were calculated conditionally on a value of δ in Chapter 2, we now proceed through the derivation (and, in our applications, the calculation) of AR confidence sets conditionally on a fixed δ, and thus use $\tilde{Y}$ as our outcome rather than Y itself.

3.2.1.2 Finding our test statistic

There are three estimated regressions to note. The first two are the standard regressions of interest in IV, measuring the relationship between instrument and treatment (the "first stage regression") and between instrument and outcome (the "reduced form regression"):

$$D = \hat{\theta}_0 + \hat{\theta} Z + e_{\text{FS}}$$
$$\tilde{Y} = \hat{\tilde{\lambda}}_0 + \hat{\tilde{\lambda}} Z + \tilde{e}_{\text{RF}}$$

The third estimated regression is specific to the AR approach. For some presumed treatment effect, τ_0, we create a new outcome from which we have removed the effect of the treatment, $\tilde{Y}_{\tau_0} = \tilde{Y} - \tau_0 D$, and regress it on the instrument:

$$\tilde{Y}_{\tau_0} = \hat{\phi}_0 + \hat{\phi}_{\tau_0} Z + e_{\tau_0},$$

Our hypothesis test is focused on the estimated coefficient for the instrument in this regression. The corresponding t-statistic for such a test of $\text{H}_0 : \hat{\phi}_{\tau_0} = 0$ is:[2]

$$t_{\hat{\phi}_{\tau_0}} = \frac{\hat{\phi}_{\tau_0}}{\widehat{\text{sd}}(\hat{\phi}_{\tau_0})} \tag{3.6}$$

We reject the null at the α level if $|t_{\hat{\phi}_{\tau_0}}| > t^*$, i.e. if $t^2_{\hat{\phi}_{\tau_0}} > t^{*2}$, where t^* is the cutoff value $t_{\alpha,n-1}$. Thus, the confidence set is $CS_\alpha(\tau_{AR}) = \{\tau_0; t^2_{\hat{\phi}_{\tau_0}} \leq t^{*2}\}$.

Rewriting the test statistic (3.6) in terms of the standard IV coefficients is a first step

[2] Note that many sources construct a test statistic based on the F distribution rather than a t distribution, but we follow Cinelli and Hazlett (2021) here and note that the rejection criteria generate the same confidence sets in both strategies.

towards actually using AR confidence sets. We do so by expanding both the point estimate:

$$\begin{aligned}\hat{\phi}_{\tau_0} &= \frac{\text{Cov}(\tilde{Y}_{\tau_0}, Z)}{\text{Var}(Z)} \\ &= \frac{\text{Cov}(\tilde{Y} - \tau_0 D, Z)}{\text{Var}(Z)} \\ &= \frac{\text{Cov}(\tilde{Y}, Z)}{\text{Var}(Z)} - \tau_0 \frac{\text{Cov}(D, Z)}{\text{Var}(Z)} \\ &= \hat{\hat{\lambda}} - \tau_0 \hat{\theta}, \end{aligned} \tag{3.7}$$

as well as it's variance:

$$\begin{aligned}\widehat{\text{Var}}(\hat{\phi}_{\tau_0}) &= \widehat{\text{Var}}(\hat{\hat{\lambda}} - \tau_0 \hat{\theta}) \\ &= \widehat{\text{Var}}(\hat{\theta})\tau_0^2 - 2\widehat{\text{Cov}}(\hat{\hat{\lambda}}, \hat{\theta})\tau_0 + \widehat{\text{Var}}(\hat{\hat{\lambda}}),\end{aligned}$$

which gives us the expanded test statistic

$$t_{\hat{\phi}_{\tau_0}} = \frac{\hat{\hat{\lambda}} - \tau_0 \hat{\theta}}{\sqrt{\widehat{\text{Var}}(\hat{\theta})\tau_0^2 - 2\widehat{\text{Cov}}(\hat{\hat{\lambda}}, \hat{\theta})\tau_0 + \widehat{\text{Var}}(\hat{\hat{\lambda}})}}$$

3.2.1.3 Solving for our confidence set

We can rewrite the inequality that defines the confidence set, $t^2_{\hat{\phi}_{\tau_0}} \leq t^{*2}$, as the quadratic inequality

$$(\hat{\theta}^2 - \widehat{\text{Var}}(\hat{\theta})t^{*2})\tau_0^2 + 2(\widehat{\text{Cov}}(\hat{\hat{\lambda}}, \hat{\theta})t^{*2} - \hat{\hat{\lambda}}\hat{\theta})\tau_0 + (\hat{\hat{\lambda}}^2 - \widehat{\text{Var}}(\hat{\hat{\lambda}})t^{*2}) \leq 0,$$

where the three terms defining this quadratic's roots are:

$$\begin{aligned}a &= \hat{\theta}^2 - \widehat{\text{Var}}(\hat{\theta})t^{*2} \\ b &= 2(\widehat{\text{Cov}}(\hat{\hat{\lambda}}, \hat{\theta})t^{*2} - \hat{\hat{\lambda}}\hat{\theta}) \\ c &= \hat{\hat{\lambda}}^2 - \widehat{\text{Var}}(\hat{\hat{\lambda}})t^{*2}\end{aligned}$$

These roots correspond to the critical points in the Anderson-Rubin confidence set. The construction of the confidence set from these points depends on the signs of both the a term and the $b^2 - 4ac$ term. The a term represents the strength of the instrument. A positive a indicates a strong instrument, and results in a confidence set that takes the form of a standard confidence interval. If a is negative, the instrument is weak and the confidence set will either be a union of two intervals (extending from $-\infty$ to one root, and from the other root to ∞) or will span the entire real line, depending on the sign of $b^2 - 4ac$. Further details are discussed elsewhere (e.g. Andrews et al., 2019).

As a sanity check for the confidence set, we show that the estimated treatment effect is sensible and is included in the set. Because the set is defined by those τ_0 values for which we do not reject our null hypothesis, the value of τ_0 that produces exactly $\hat{\phi}_{\tau_0} = 0$ will necessarily be included. Solving for that particular τ_0 using equation 3.7 produces:

$$\widehat{\tau_{AR}} = \frac{\hat{\tilde{\lambda}}}{\hat{\theta}}.$$

This particular value – the reduced form effect divided by the first stage effect – is in fact equivalent to the treatment effect estimate we get from SCQE (or from an IV using our δ-adjusted outcome $\tilde{Y}$).

3.2.2 Estimating a confidence set from observable data:

We now have an estimator for the point estimate as well as the roots that define the bounds and shape of our confidence set – all of which are conditional on a single value of δ – expressed in terms of $\hat{\theta}$, $\hat{\tilde{\lambda}}$, and estimators of their variances and covariance. However, to allow for implementation, we need to rewrite these five estimators in terms of our observable data, Z, D, and Y.[3]

[3] Note that although we have been working with $\tilde{Y}$ in deriving these terms, we now want our estimators expressed in terms of the observable Y for ease of implementation.

We start with the estimators for θ and $\tilde{\lambda}$. Because $\hat{\theta}$ is the coefficient of a binary variable in an OLS regression, it must be the case that:

$$\hat{\theta} = \overline{D_{Z=1}} - \overline{D_{Z=0}} \tag{3.8}$$

and similarly:

$$\begin{aligned}
\hat{\tilde{\lambda}} &= \overline{\tilde{Y}_{Z=1}} - \overline{\tilde{Y}_{Z=0}} \\
&= \overline{(Y - \delta Z)_{Z=1}} - \overline{(Y - \delta Z)_{Z=0}} \\
&= \overline{Y_{Z=1}} - \delta \cdot 1 - (\overline{Y_{Z=0}} - \delta \cdot 0) \\
&= \overline{Y_{Z=1}} - \overline{Y_{Z=0}} - \delta
\end{aligned} \tag{3.9}$$

Again, this shows the AR point estimate is simply equation 3.4 in the $\overline{D_{Z=0}} = \pi_0 = 0$ case:

$$\widehat{\tau_{AR}} = \frac{\hat{\tilde{\lambda}}}{\hat{\theta}} = \frac{\overline{Y_{Z=1}} - \overline{Y_{Z=0}} - \delta}{\overline{D_{Z=1}}}$$

Continuing to the variance terms, we start with the variance of θ:

$$\begin{aligned}
\widehat{\mathrm{Var}}(\hat{\theta}) &= \frac{\widehat{\sigma_{\text{FS}}}}{S_{ZZ}} \\
&= \frac{\widehat{e_{\text{FS}}}^{\top} \widehat{e_{\text{FS}}}}{(N-1) S_{ZZ}} \\
&= \frac{(D - \hat{\theta} Z)^{\top} (D - \hat{\theta} Z)}{(N-1)^2 \mathrm{Var}(Z)} \\
&= \frac{(N-1) \mathrm{Var}(D - \hat{\theta} Z)}{(N-1)^2 \mathrm{Var}(Z)} \\
&= \frac{\mathrm{Var}(D) + \hat{\theta}^2 \mathrm{Var}(Z) - 2\hat{\theta} \mathrm{Cov}(D, Z)}{(N-1) \mathrm{Var}(Z)} \\
&= \frac{\mathrm{Var}(D) + \hat{\theta}^2 \mathrm{Var}(Z) - 2\hat{\theta}^2 \mathrm{Var}(Z)}{(N-1) \mathrm{Var}(Z)} \\
&= \frac{\mathrm{Var}(D) - \hat{\theta}^2 \mathrm{Var}(Z)}{(N-1) \mathrm{Var}(Z)}
\end{aligned}$$

To work with $\widehat{\text{Var}}(\hat{\hat{\lambda}})$, we first note from equation 3.9 that $\hat{\hat{\lambda}}$ is δ less than $\overline{Y_{Z=1}} - \overline{Y_{Z=0}}$. This difference is simply $\hat{\lambda}$, the estimate we would obtain by regressing Y on Z rather than $\tilde{Y}$ on Z in the reduced form:

$$Y = \hat{\lambda}_0 + \hat{\lambda} Z + e_{\text{RF}}$$

That is, $\hat{\lambda} = \hat{\hat{\lambda}} + \delta$. Because δ is a constant, $\widehat{\text{Var}}(\hat{\hat{\lambda}}) = \widehat{\text{Var}}(\hat{\lambda} - \delta) = \widehat{\text{Var}}(\hat{\lambda})$, which allows us to run through the same steps here as we did for $\widehat{\text{Var}}(\hat{\theta})$, substituting $\hat{\lambda}$ (rather than $\hat{\hat{\lambda}}$) for $\hat{\theta}$ and Y (rather than $\tilde{Y}$) for D:

$$\begin{aligned}
\widehat{\text{Var}}(\hat{\hat{\lambda}}) &= \widehat{\text{Var}}(\hat{\lambda}) \\
&= \frac{\text{Var}(Y) - \hat{\lambda}^2 \text{Var}(Z)}{(N-1)\text{Var}(Z)}
\end{aligned}$$

Next, for the covariance term, we can again first remove the δ term, since $\widehat{\text{Cov}}(\hat{\hat{\lambda}}, \hat{\theta}) = \widehat{\text{Cov}}(\hat{\lambda} - \delta, \hat{\theta}) = \widehat{\text{Cov}}(\hat{\lambda}, \hat{\theta})$, and working from there,

$$\begin{aligned}
\widehat{\text{Cov}}(\hat{\theta}, \hat{\hat{\lambda}}) &= \widehat{\text{Cov}}(\hat{\theta}, \hat{\lambda}) \\
&= \frac{\text{Cov}(D^{\perp Z}, Y^{\perp Z})}{(N-1)\text{Var}(Z)} \\
&= \frac{\text{Cov}(D - Z\hat{\theta}, Y - Z\hat{\lambda})}{(N-1)\text{Var}(Z)} \\
&= \frac{\text{Cov}(D, Y) - \text{Cov}(D, Z\hat{\lambda}) - \text{Cov}(Y, Z\hat{\theta}) + \text{Cov}(Z\hat{\theta}, Z\hat{\lambda})}{(N-1)\text{Var}(Z)} \\
&= \frac{\text{Cov}(D, Y) - \hat{\lambda}\text{Cov}(D, Z) - \hat{\theta}\text{Cov}(Y, Z) + \hat{\theta}\hat{\lambda}\text{Var}(Z)}{(N-1)\text{Var}(Z)} \\
&= \frac{\text{Cov}(D, Y) - \hat{\lambda}\hat{\theta}\text{Var}(Z) - \hat{\theta}\hat{\lambda}\text{Var}(Z) + \hat{\theta}\hat{\lambda}\text{Var}(Z)}{(N-1)\text{Var}(Z)} \\
&= \frac{\text{Cov}(D, Y) - \hat{\theta}\hat{\lambda}\text{Var}(Z)}{(N-1)\text{Var}(Z)} \qquad (3.10)
\end{aligned}$$

Finally, we can rewrite the roots of the quadratic in terms of observable data (with $\hat{\lambda}$ and $\hat{\theta}$ not expanded for readability):

$$a = \hat{\theta}^2 - \widehat{\text{Var}}(\hat{\theta})t^{*2} = \hat{\theta}^2 - \left(\frac{\text{Var}(D) - \hat{\theta}^2\text{Var}(Z)}{(N{-}1)\text{Var}(Z)}\right) t^{*2}$$

$$b = 2(\widehat{\text{Cov}}(\hat{\hat{\lambda}}, \hat{\theta})t^{*2} - \hat{\hat{\lambda}}\hat{\theta}) = 2\left(\frac{\text{Cov}(D, Y) - \hat{\theta}\hat{\lambda}\text{Var}(Z)}{(N{-}1)\text{Var}(Z)} t^{*2} - \hat{\theta}(\hat{\lambda} - \delta)\right)$$

$$c = \hat{\hat{\lambda}}^2 - \widehat{\text{Var}}(\hat{\hat{\lambda}})t^{*2} = (\hat{\lambda} - \delta)^2 - \left(\frac{\text{Var}(Y) - \hat{\lambda}^2\text{Var}(Z)}{(N{-}1)\text{Var}(Z)}\right) t^{*2}$$

3.3 Inference using only summary statistics

We have thus far considered the case of a dichotomous instrument and a dichotomous treatment, but have not yet placed any restrictions on the outcome.[4] However, in the special though not uncommon scenario of a dichotomous outcome, the AR confidence set formulas can be expressed in terms of a limited number of summary statistics. This "count-based SCQE" allows us to run the same method and produce the same results whether or not we have access to unit-level data.

The motivations for our development of a summary-statistic-only SCQE are two fold: First, eliminating the need for unit-level data may help practitioners avoid several data-sharing and data-privacy concerns. For applications in healthcare, HIPAA compliance requirements, which often prevent medical records from being shared without in-depth de-identification efforts, are far less likely to stymie the sharing of these summary statistics. Even if legal or privacy concerns do not prevent data sharing, those with proprietary data

[4] We have considered only a dichotomous treatment for convenience; none of the formulas derived thus far place any such restrictions on the treatment. However, in order to use the count-based inference presented in this section the treatment does need to be dichotomous. When a continuous treatment is feasible, a continuous treatment changes the interpretation of the estimated treatment effect as it would in standard IV settings, and adjustments need to be made to formulas like the Y_i switching function.

may be more willing to share high-level counts and proportions than to grant full data access. Secondly, this implementation allows for post-publication re-analyses by researchers, not only the original authors. Many potential applications for SCQE, already studied using different methodologies, have been published in papers that contain the necessary summary statistics to run this count-based version. The required counts or proportions are regularly presented in the most commonly seen tables or directly within the text, and even when some of the data points are not presented, authors who may not otherwise be willing to provide the extensive data necessary for many re-analyses may be more receptive to filling in only minor gaps. The applications in Sections 3.4 and 4.2.2 demonstrate such post-publication re-analysis in settings where data sharing and privacy concerns would otherwise be prohibitive.

It takes only 8 counts to calculate the series of averages that will be utilized in our rewritten count-based formulas: $N_{Z=0,D=0}$, $N_{Y=1,Z=0,D=0}$, $N_{Z=0,D=1}$, $N_{Y=1,Z=0,D=1}$, $N_{Z=1,D=0}$, $N_{Y=1,Z=1,D=0}$, $N_{Z=1,D=1}$, and $N_{Y=1,Z=1,D=1}$.[5] From these counts, we can construct several quantities of interest: $\overline{D}$, $\overline{Y}$, and $\overline{Z}$; $\overline{D_{Z=z}}$ and $\overline{Y_{Z=z}}$ for $z \in \{0,1\}$; $\overline{Z_{D=d}}$ for $d \in \{0,1\}$; and $\overline{Y_{Z=z,D=d}}$ for $z \in \{0,1\}$ and $d \in \{0,1\}$. Armed with these quantities, we can once again rewrite our five estimators.

The estimators for θ and λ are unchanged from equations 3.8 and 3.9.

Our count-based reformulation of the variance terms are also relatively simple:

$$\begin{aligned}\widehat{\text{Var}}(\hat{\theta}) &= \frac{\text{Var}(D) - \hat{\theta}^2\text{Var}(Z)}{(N{-}1)\text{Var}(Z)} \\ &= \frac{\overline{D}(1-\overline{D}) - \hat{\theta}^2\overline{Z}(1-\overline{Z})}{(N{-}1)\overline{Z}(1-\overline{Z})} \\ &= \frac{\overline{D}(1-\overline{D})}{(N{-}1)\overline{Z}(1-\overline{Z})} - \frac{\hat{\theta}^2}{N{-}1}\end{aligned}$$

[5] The averages can be calculated from any number of combinations of 8 units of information (both counts and proportions), and we focus on these counts due to convenience and their avoidance of the imprecision of proportions.

$$\widehat{\mathrm{Var}}(\hat{\hat{\lambda}}) = \frac{\mathrm{Var}(Y) - \hat{\lambda}^2\mathrm{Var}(Z)}{(N{-}1)\mathrm{Var}(Z)}$$
$$= \frac{\bar{Y}(1-\bar{Y})}{(N{-}1)\bar{Z}(1-\bar{Z})} - \frac{\hat{\lambda}^2}{N{-}1}$$

For the covariance (equation 3.10) we first look at the $\mathrm{Cov}(D, Y)$ term, which is equal to $\mathrm{Var}(D)$ multiplied by $\hat{\beta}_{D,Y\sim D}$, the effect of D estimated when regressing Y on D. Thus, our count-based covariance estimator is:

$$\widehat{\mathrm{Cov}}(\hat{\theta}, \hat{\lambda}) = \frac{\mathrm{Cov}(D, Y) - \hat{\theta}\hat{\lambda}\mathrm{Var}(Z)}{(N{-}1)\mathrm{Var}(Z)}$$
$$= \frac{\mathrm{Var}(D) \cdot \hat{\beta}_{D,Y\sim D} - \hat{\theta}\hat{\lambda}\mathrm{Var}(Z)}{(N{-}1)\mathrm{Var}(Z)}$$
$$= \frac{\bar{D}(1{-}\bar{D}) \cdot \hat{\beta}_{D,Y\sim D}}{(N{-}1)\bar{Z}(1{-}\bar{Z})} - \frac{\hat{\theta}\hat{\lambda}}{N{-}1}$$

where the count-based $\beta_{D,Y\sim D}$ estimator is:

$$\begin{aligned}
\hat{\beta}_{D,Y\sim D} &= \mathbb{E}[Y|D{=}1] - \mathbb{E}[Y|D{=}0] \\
&= \mathbb{E}_{Z|D=1}\big[\mathbb{E}[Y|D{=}1, Z]\big] - \mathbb{E}_{Z|D=0}\big[\mathbb{E}[Y|D{=}0, Z]\big] \\
&= \mathbf{P}(Z{=}1|D{=}1)\mathbb{E}[Y|D{=}1, Z{=}1] + \mathbf{P}(Z{=}0|D{=}1)\mathbb{E}[Y|D{=}1, Z{=}0] \\
&\quad - \Big(\mathbf{P}(Z{=}1|D{=}0)\mathbb{E}[Y|D{=}0, Z{=}1] + \mathbf{P}(Z{=}0|D{=}0)\mathbb{E}[Y|D{=}0, Z{=}0]\Big) \\
&= \overline{Z_{D=1}} \cdot \overline{Y_{Z=1,D=1}} + (1 - \overline{Z_{D=1}}) \cdot \overline{Y_{Z=0,D=1}} \\
&\quad - \overline{Z_{D=0}} \cdot \overline{Y_{Z=1,D=0}} - (1 - \overline{Z_{D=0}}) \cdot \overline{Y_{Z=0,D=0}}
\end{aligned}$$

Finally, we can rewrite each of the quantities that define our confidence set, now in terms of simple summary statistics:

$$
\begin{aligned}
a &= \hat{\theta}^2 - \widehat{\mathrm{Var}}(\hat{\theta})t^{*2} = \hat{\theta}^2 - \frac{t^{*2}}{N{-}1}\left(\frac{\bar{D}(1{-}\bar{D})}{\bar{Z}(1-\bar{Z})} - \hat{\theta}^2\right)\\
b &= 2(\widehat{\mathrm{Cov}}(\hat{\hat{\lambda}}, \hat{\theta})t^{*2} - \hat{\hat{\lambda}}\hat{\theta}) = \frac{2\,t^{*2}}{N{-}1}\left(\frac{\bar{D}(1{-}\bar{D}) \cdot \hat{\beta}_{D,Y\sim D}}{\bar{Z}(1{-}\bar{Z})} - \hat{\theta}\hat{\lambda}\right) - 2\hat{\theta}(\hat{\lambda} - \delta)\\
c &= \hat{\hat{\lambda}}^2 - \widehat{\mathrm{Var}}(\hat{\hat{\lambda}})t^{*2} = (\hat{\lambda} - \delta)^2 - \frac{t^{*2}}{N{-}1}\left(\frac{\bar{Y}(1{-}\bar{Y})}{\bar{Z}(1-\bar{Z})} - \hat{\lambda}^2\right)
\end{aligned}
$$

We have built a web application that implements the above count-based formulas, and have made it available for use at `http://amiwulf.shinyapps.io/SCQE_demo`. The web tool takes in the 8 required counts along with a range of assumption values and generates our preferred visualization of SCQE's results, as well as some suggested causal statements that are demonstrated by the method to be valid.

SCQE's mechanics are attractively simple, relative to many commonly used adjustment procedures, without *over*simplifying the inferential task at hand. The count-based SCQE, and the web app implementing it, reflect that simplicity. We hope it signals to technical researchers that the method is straightforward and encourages those who are less technical to consider SCQE as an approachable analysis strategy.

3.4 Rapid Response Systems

In this section, we apply the count-based SCQE in a re-analysis of an article investigating an important, open question in the literature about hospital quality and safety. The use of SCQE here should be understood primarily as a demonstration of the method in practice, as well as the usefulness of the count-based implementation; we do not claim SCQE is a silver bullet for the entire open question, only that it offers an important new way forward in ongoing investigations.

Rapid Response Systems (RRS) are common hospital programs implemented to identify

– and direct resources to – patients at risk of impending clinical deterioration, often defined as changes in clinical status that necessitate a transfer into an intensive care unit (ICU). The systems vary by the real time alert triggers they use, such as clinician gestalt, certain vital sign measurements, or complex, predictive, risk-scoring models (Lyons et al., 2018), as well as by the particular response-team interventions provided to "alerted" patients. Though RRSs have become increasingly common since their introduction decades ago, the literature on their effectiveness has not produced clear evidence of their value. The most comprehensive randomized trial run to date produced null results (MERIT Study Investigators, 2005), yet hospitals' acceptance of RRSs as self-evidently beneficial means that further randomized trials are seen as unethical to the control group rather than an important test of whether system resources are being wasted without helping patients. Bolstering the assumption that RRSs can lower mortality or cardiac arrest rates are a series of observational studies that, while providing several null results, evaluate the systems' impacts more favorably on average than the few existing randomized trials (Maharaj et al., 2015; Alam et al., 2014; Chan et al., 2010).

Any observational study based on defending an assumption of conditional ignorability faces a particularly daunting task. The more standardized and uniformly applied a given RRS trigger is, the less overlap we expect to see between alerted and non-alerted patients, and the less we should trust adjustment strategies to produce theoretically grounded results. Indeed, in the case of deterministic, algorithmically triggered RRSs, this overlap is by definition nonexistent, and thus inference necessarily requires extrapolation. We note, however, that deterministic alerts may provide opportunities for inference as well, as they did in a recent (multi-hospital, stepped-wedge design) RRS rollout that ran the alert score on "silent" at each hospital until their local response team was trained and began receiving the alerts (Escobar et al., 2020).

As a result of such a clear threat to conditional ignorability, many published studies are forced to make weaker claims, not asking about the relationship between an RRS alert and the alerted patient's outcome, but whether implementation of the RRS as a whole was associated with a decrease in mortality hospital wide. They take on pre-post designs to accommodate this question, and covariate adjustment is used in an attempt to identify the "effect" of being in the post-implementation period, after controlling for other possible changes over time. Skeptical readers will understandably question whether estimates from these studies provide useful or generalizable information about the impact of an alert system, and may instead interpret results as a general statement about the effectiveness of localized culture change, as one such study suggests explicitly (Buist et al., 2002).

What these studies do offer is an opportunity for SCQE re-analysis, as they often include summary statistics about both populations and may share over-time trends in their outcome of interest that help to inform our bounds on δ, the unobserved over-time shifts in hospital-wide outcomes. In these instances, SCQE can generate claims that do not rely on a questionable assertion of conditional ignorability and yet still target the treatment and population of interest in the desired fashion. When applied to some such studies, SCQE may produce affirmative statements about a system's benefit or harm to alerted patients, while in others SCQE serves to quantify uncertainty and avoid deceptively overconfident claims. Neither result necessarily constitutes a better setting for the method, as both allow for transparent evaluation of our assumptions. Below, we run our newly-derived count-based SCQE on an observational RRS study as an demonstration.

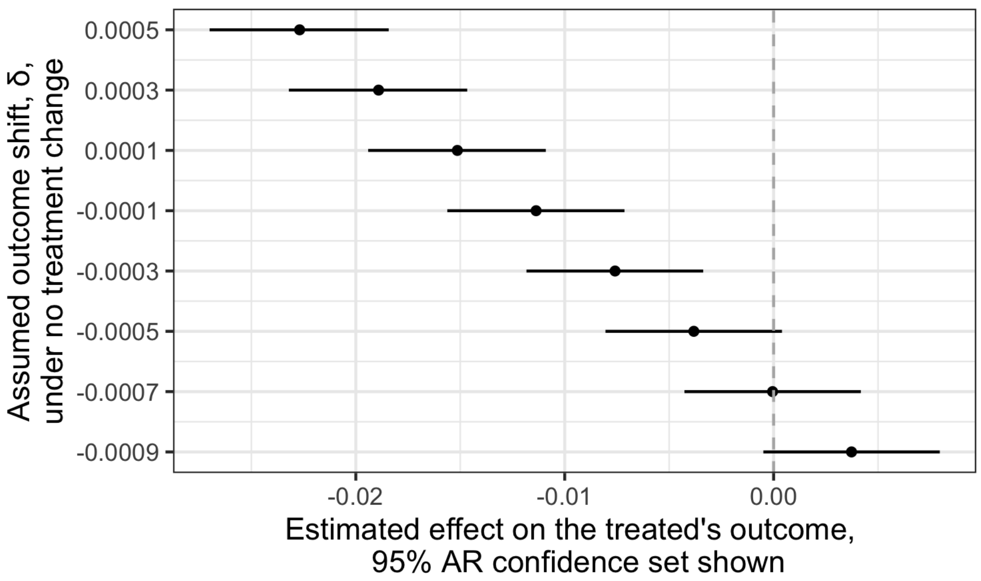
3.4.1 An illustrative RRS re-analysis

A 2012 article by Howell et al. describes the implementation of an RRS in a single hospital in 2005 and attempts to measure the association between system introduction and unexplained mortality (UM), which they define as death outside the ICU in patients without a "do not resuscitate" order. The alert response involved a formal, timely bedside meeting of several of the patient's caregivers. The authors use the 22 months prior to implementation (with a 0.09% UM rate among 66,496 admissions) as their pre-period, allow for a 6-month implementation gap, and use the next 31 months as a post-implementation period (which had a 0.02% UM rate among 90,045 admissions). The paper's primary analysis claimed that the post-implementation period was associated with an 80% reduction in UM, and a secondary interrupted time series (ITS) approach found a 65% reduction in UM by the end of the post-implementation period. They share monthly UM rates across the study period, which we can use to inform our choices of δ. In particular, the pre-period UM rates ranged from 0% to 0.25%, and a 6-month rolling average ranged from 0.05% to 0.14%. No trend was clearly visible in these unadjusted rates, though the trend they found through the adjusted ITS approach suggested that UM rates declined by just over 0.01% per year. The other required data for SCQE are listed in the text, giving a treatment rate in the high-use period of 5.3% and outcome rates for the treated and untreated patients of 0.09% UM and 0.016% UM, respectively.[6]

Taking into account our starting point of 0.09% UM, the natural bound of 0% UM, the slight downward trend in the no-use cohort, and the amount of variability we saw in the same cohort, we should feel comfortable setting a lower bound of -0.07 percentage points (pp), and an upper bound of +0.03pp. That is, we would declare a relative decrease in non-treatment outcomes of more than 77% over the course of 3 years to be implausible, and we would not expect the trend to reverse instead and rise by 33% relative to the pre-period over the same time. As we see in Figure 3.1, this range of δ values correspond to ATT estimates ranging from 0 to -1.89pp. The latter estimate would imply that RRS alerts were responsible for a 95% reduction in the treated patients' UM, bringing them from 1.98% down to the observed 0.09%, while the former clearly implies no impact of the RRS at all. In order to claim the RRS had a beneficial effect on UM at that level, we would need to defend a δ value that is positive, 0, or no more negative than -0.05pp. This is certainly a likely range, but we cannot defensibly rule out the lowest area of our plausible δ values, and thus we would not declare

[6] Given the large sample size in this application, a 5.3pp increase in treatment rate does not present a weak instrument problem (first-stage F = 3,721) and thus we would not expect the AR confidence sets here to differ from standard confidence intervals.

Figure 3.1: SCQE rapid response system re-analysis



Note: SCQE plot for the re-analysis of Howell et al. (2012). Count-based implementation used, with point estimates and Anderson-Rubin 95% confidence sets shown.

that the treatment was unambiguously beneficial. The results are far more conclusive in the other direction: even at the lowest mechanically possible δ value, -0.09pp, SCQE suggests a null effect estimate. This is not a weak claim about the difficulty of defending a range of harm-implying assumptions, but instead a recognition that no harm-implying assumption exists to be defended.

Precisely comparing SCQE's results to those from the original analysis is difficult given the differences in their estimands and treatment definitions, but we share two general points. First, though SCQE may be more emphatic than the original article in rejecting the possibility that the RRS was harmful, both methods agree substantively on this conclusion. Secondly, a particular feature of SCQE draws our attention to the effect mechanism, which highlights important uncertainty. We noted the wide span of ATTs that SCQE deems plausible, ranging from no effect to a 95% relative reduction in UM. The reason for this high sensitivity is the low treatment uptake rate, which we can see, by noting the ratio form of our SCQE estimator, will magnify uncertainty about δ in estimating the ATT. The methods used in the original article, however, are unaffected by the amount of treatment actually given. If 10% of patients rather than 5% of patients had been alerted, the original analysis would produce the same results, but the magnitude of SCQE's ATT estimates would be halved.

That Howell et al.'s effects are uncoupled with treatment rates is the rule rather than the exception among RRS evaluations. The authors even explicitly mention that they do not know the exact mechanism by which the RRS rollout may have lowered mortality. Still, evaluating this question even without domain knowledge can be instructive. If the RRS affects mortality rates overwhelmingly through its impact on alerted patients, SCQE addresses that effect explicitly and in doing so quantifies uncertainty that we see would be missed by standard approaches. If there are significant alternative means by which an RRS affects mortality, these standard approaches are better suited to measuring such effects, though they would need to honestly admit that a claim of effectiveness does not imply that the alert and response process itself is helping patients.

CHAPTER 4

Extensions

In Chapter 3, we introduced AR confidence sets for SCQE and derived a count-only implementation to expand the method's applicability. In this chapter, we develop three additional large extensions aided by that inferential grounding. The first, presented in Section 4.1, generalizes SCQE by allowing for a low-use cohort, rather than a no-use cohort, alongside the high-use cohort; as long as our two observed cohorts have meaningfully different treatment rates, SCQE can produce estimates and confidence sets. We first discuss the alternate estimand that this generalization implies and share our corresponding estimator. We then note a gap in the information SCQE produces in this expanded setting, and we propose a solution to recover the missing information by employing a secondary counterfactual trend assumption. In Section 4.2 we introduce our second extension, which enables SCQE's use in settings with only one cohort. By adjusting the assumption required from δ, the imagined shift between cohorts, to μ_{Y0}, the outcome rate in an imaginary non-treatment cohort, we are able to maintain the structure and inferential style of the original two-cohort SCQE. We then demonstrate the use of a count-based implementation of this one-cohort SCQE in Section 4.2.2, applying it in the re-analysis of four papers estimating the impact of hydroxychloroquine on COVID-19 outcomes. Our final extension, offered in Section 4.3, allows researchers to specify their beliefs about δ in the form of a distribution rather than by suggesting worst-case bounds. We discuss two inferential strategies when using these δ priors, along with their motivations, differences, and the kind of knowledge they produce. We then implement both strategies in Section 4.3.2 with an application to the tuberculosis prevention data from Chapter 2.

4.1 Low-use SCQE

In settings such as Chapters 2 (Hazlett et al., 2020) and 3, we made a convenient simplifying assumption: our low-use cohorts have been no-use cohorts, featuring treatment use not merely at a lower level than the high-use cohort, but with no treatment use at all. In Section 4.1.1, we discuss both a generalized estimator that accounts for non-zero treatment in the low-use cohort as well as its estimand. In Section 4.1.2, we explain a complication that arises: we can no longer place the location of our effect when using the low-use SCQE. We develop a solution based on the imposition of a second δ-like trend assumption and demonstrate how it recovers the effect's location.

4.1.1 Interpretation and Estimation

Imagine we have two cohorts, where $\mathbb{E}[D|Z=0] = \pi_0$, $\mathbb{E}[D|Z=1] = \pi_1$, and $\pi_0 < \pi_1$. As we did when our low-use cohort included no treatment use, we must by definition ascribe the difference in $\tilde{Y}(\mathrm{d})$ (the δ-adjusted outcome under some fixed treatment level) between cohorts to the increase in treatment. However, the units experiencing that increase are no longer synonymous with all treated units in the high-use cohort. Instead, only those units that would be untreated in the low-use cohort but treated in the high-use cohort carry the impact of the treatment increase. Our treatment effect is no longer an ATT, but rather is localized to this subset of units. More precisely, as we note through our comparison with Instrumental Variables, our effect is a Complier Average Treatment Effect (CATE), where "complier" is a standard descriptor for the treated-in-high-use, untreated-in-low-use units.

Just as potential outcomes notation refers to a unit's outcome under a given treatment, we use "potential treatment" notation here to refer to the treatment that a unit would receive under a given cohort membership. A complier, then, is defined through $D(0)=0$, $D(1)=1$. Formally, the estimand of interest here is:

$$\text{CATE} = \mathbb{E}[Y(1)|D(0)=0, D(1)=1, Z=1] - \mathbb{E}[Y(0)|D(0)=0, D(1)=1, Z=1], \tag{4.1}$$

where we refer to the effect in the high-use cohort, $Z=1$, rather than in the low-use cohort, without loss of generality. The parallel to IV reminds us of another requirement: we must assume that there exist no "defier" units, i.e. no units that would have received treatment in the low-use cohort but would not have received treatment in the high-use cohort. This assumption is untestable and must be argued for substantively. While the assumption was also required in the no-use version of SCQE, the lack of treated units in that low-use cohort guaranteed that it held. The remaining two types of units, "always-takers" and "never-takers," are unaffected by the increase in treatment and are the focus of neither an ATT nor a CATE. These four unit groupings based on potential treatments are often called principal strata (Angrist et al., 1996; Frangakis and Rubin, 2002).

If we can defend the no-defiers assumption, we are able to utilize once again our parallels to IV by removing all exogeneity violations. That is, the IV Wald estimator still provides a valid estimate (this time of the CATE rather than the ATT) when using our δ-adjusted outcomes, simply dividing the treatment-driven change in outcomes by the size of that treatment increase:

$$\begin{aligned}\widehat{\text{CATE}} &= \frac{\mathbb{E}[\tilde{Y}|Z=1] - \mathbb{E}[\tilde{Y}|Z=0]}{\pi_1 - \pi_0} \\ &= \frac{\mathbb{E}[Y - \delta Z|Z=1] - \mathbb{E}[Y - \delta Z|Z=0]}{\pi_1 - \pi_0} \\ &= \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0] - \delta}{\pi_1 - \pi_0} \end{aligned} \tag{4.2}$$

Fortunately, though the estimand here is conceptually different from the ATT in no-use/high-use SCQE, very little has changed for the method mechanically; the estimation and inference procedures developed in Chapter 3 are fully capable of handling a low-use cohort, and they treat a no-use cohort as merely a special case. If we have a dataset featuring a low-use cohort, we can adjust our Y to $\tilde{Y}$ as before and generate estimates and AR confidence sets. In those cases with a binary outcome, our count-based SCQE implementation also still works. Indeed, the web application mentioned above allows for non-zero treatment counts to be entered for the low-use cohort.

4.1.2 Losing our estimate, and how to find it

We note, however, that equation 4.2 corresponds to the Wald ATT estimator (equation 3.4), rather than to the SCQE ATT estimator (equation 3.3). Although these produce the same estimate, we noted in Section 3.1 that the latter allows us to *place* or *locate* the ATT. Either estimator will tell us that, for example, a treatment *lowered* the treated patients' mortality rate by 5 percentage points, but a drop from 7% mortality to 2% mortality may have policy implications that differ meaningfully the implications of a drop from 48% to 43%. The ATT estimand (equation 3.2) had one observed and one unobserved term, and we identified the unobserved term using a single δ assumption. But both terms of our CATE estimand (equation 4.1) are unobservable, since we don't know which units are compliers: among the non-treated units in the low-use cohort we can't tell apart compliers and never-takers, and among the treated units in the high-use cohort we can't tell apart compliers and always-takers. In order to identify both, we need to supplement δ with an additional assumed counterfactual outcome shift in some subpopulation, from which we can identify the compliers' outcomes under both treatment regimes and thus the implied CATE. Below, after introducing notation, we demonstrate this procedure visually and algebraically.

Our additional counterfactual assumption will be the shift for the always-takers:

$$\delta_{\text{AT}} := \mathbb{E}[Y(1) \mid D(0)=1, D(1)=1, Z=1] - \mathbb{E}[Y(1) \mid D(0)=1, D(1)=1, Z=0]$$

Given proposed shifts for the entire cohort and for the subset of always-takers, we can uniquely identify the shift among the remaining units, which we call the not-always-takers (a combination of never-takers and compliers):

$$\delta_{\text{nAT}} := \mathbb{E}[Y(0) \mid D(0)=0, Z=1] - \mathbb{E}[Y(0) \mid D(0)=0, Z=0]$$

In other words, because $\delta = \pi_0\,\delta_{\mathrm{AT}} + (1-\pi_0)\,\delta_{\mathrm{nAT}}$,[1] we can define $\delta_{\mathrm{nAT}} = (\delta - \pi_0\,\delta_{\mathrm{AT}})/(1-\pi_0)$. Therefore, though we will use δ_{AT} and δ_{nAT} as our two assumptions in the derivation below, in practice we may choose to propose values for only one of them as well as δ (based on which of the two we feel most able to reason with) and then solve for the second one.[2]

Here, we demonstrate how our assumed values of δ_{AT} and δ_{nAT} recover the location of the CATE; they will allow for two important decompositions, each of which expresses one unobservable half of our CATE estimand in terms of observable quantities.

First, we can decompose the expected outcomes of the non-always-takers as

$$\begin{aligned}\mathbb{E}[Y(0) \mid D(0)\!=\!0, Z\!=\!0] &= \mathbb{E}[Y(0) \mid D(0)\!=\!0, Z\!=\!1] - \delta_{\mathrm{nAT}} \\ &= \frac{(1-\pi_1)}{(1-\pi_0)}\mathbb{E}[Y(0) \mid D(0)\!=\!0, D(1)\!=\!0, Z\!=\!1] \\ &\quad + \frac{(\pi_1-\pi_0)}{(1-\pi_0)}\mathbb{E}[Y(0) \mid D(0)\!=\!0, D(1)\!=\!1, Z\!=\!1] - \delta_{\mathrm{nAT}},\end{aligned}$$

which we rearrange to express the unobservable non-treated outcome of the compliers in the high-use cohort in terms of observable quantities:

$$\begin{aligned}\mathbb{E}[Y(0)|D(0)\!=\!0, D(1)\!=\!1, Z\!=\!1] &= \frac{(1-\pi_0)\,(\mathbb{E}[Y(0)|D(0)\!=\!0, Z\!=\!0] + \delta_{\mathrm{nAT}})}{(\pi_1-\pi_0)} \\ &\quad - \frac{(1-\pi_1)\mathbb{E}[Y(0) \mid D(0)\!=\!0, D(1)\!=\!0, Z\!=\!1]}{(\pi_1-\pi_0)} \\ &= \frac{(1-\pi_0)\,(\mathbb{E}[Y|D\!=\!0, Z\!=\!0] + \delta_{\mathrm{nAT}}) - (1-\pi_1)\mathbb{E}[Y|D\!=\!0, Z\!=\!1]}{(\pi_1-\pi_0)}\end{aligned}$$

[1] Recall that the subscripts on π refer to the cohorts rather than to the treatment assignments, so, e.g., $\mathbb{E}[D\!=\!1 \mid Z\!=\!0]$ is referred to by π_0 rather than by π_1, and $\mathbb{E}[D\!=\!0 \mid Z\!=\!0]$ is referred to by $1-\pi_0$ rather than by π_0.

[2] Formally, we must choose to assert two shifts from among one of two population trios – either two from the whole cohort, the always-takers, and the combined compliers and never-takers, or two from the whole cohort, the never-takers, and the combined compliers and always-takers.

We can also decompose the expected outcome of the always-takers and compliers as:

$$\begin{aligned}
\mathbb{E}[Y(1) \mid D(1)=1, Z=1] &= \frac{(\pi_1-\pi_0)}{\pi_1}\mathbb{E}[Y(1) \mid D(0)=0, D(1)=1, Z=1] \\
&\quad + \frac{\pi_0}{\pi_1}\mathbb{E}[Y(1) \mid D(0)=1, D(1)=1, Z=1] \\
&= \frac{(\pi_1-\pi_0)}{\pi_1}\mathbb{E}[Y(1) \mid D(0)=0, D(1)=1, Z=1] \\
&\quad + \frac{\pi_0}{\pi_1}\left(\mathbb{E}[Y(1) \mid D(0)=1, D(1)=1, Z=0] + \delta_{\mathrm{AT}}\right),
\end{aligned}$$

which we rearrange to express the unobservable treated outcome of the compliers in the high-use cohort in terms of observable quantities:

$$\begin{aligned}
\mathbb{E}[Y(1) \mid D(0)=0, D(1)=1, Z=1] &= \frac{\pi_1\mathbb{E}[Y(1) \mid D(1)=1, Z=1]}{(\pi_1-\pi_0)} \\
&\quad - \frac{\pi_0\left(\mathbb{E}[Y(1) \mid D(0)=1, D(1)=1, Z=0] + \delta_{\mathrm{AT}}\right)}{(\pi_1-\pi_0)} \\
&= \frac{\pi_1\mathbb{E}[Y \mid D=1, Z=1] - \pi_0\left(\mathbb{E}[Y \mid D=1, Z=0] + \delta_{\mathrm{AT}}\right)}{(\pi_1-\pi_0)}
\end{aligned}$$

Together, these two unobservable terms define the Complier Average Treatment Effect:

$$\begin{aligned}
\widehat{\mathrm{CATE}} &= \mathbb{E}[Y(1) \mid D(0)=0, D(1)=1, Z=1] - \mathbb{E}[Y(0) \mid D(0)=0, D(1)=1, Z=1] \\
&= \frac{\pi_1\mathbb{E}[Y|D{=}1, Z{=}1] - \pi_0\left(\mathbb{E}[Y|D{=}1, Z{=}0] + \delta_{\mathrm{AT}}\right)}{(\pi_1-\pi_0)} \\
&\quad - \frac{(1-\pi_0)\left(\mathbb{E}[Y|D{=}0, Z{=}0] + \delta_{\mathrm{nAT}}\right) - (1-\pi_1)\mathbb{E}[Y|D{=}0, Z{=}1]}{(\pi_1-\pi_0)} \\
&= \frac{\pi_1\mathbb{E}[Y|D{=}1, Z{=}1] - \pi_0\mathbb{E}[Y|D{=}1, Z{=}0] - (1{-}\pi_0)\mathbb{E}[Y|D{=}0, Z{=}0]}{(\pi_1-\pi_0)} \\
&\quad + \frac{(1{-}\pi_1)\mathbb{E}[Y|D{=}0, Z{=}1]}{(\pi_1-\pi_0)} - \frac{\pi_0\ \delta_{\mathrm{AT}} + (1-\pi_0)\ \delta_{\mathrm{nAT}}}{(\pi_1-\pi_0)} \\
&= \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0] - \left(\pi_0\ \delta_{\mathrm{AT}} + (1-\pi_0)\ \delta_{\mathrm{nAT}}\right)}{(\pi_1-\pi_0)}
\end{aligned} \tag{4.3}$$

Plugging in δ for the terms enclosed by parentheses in the last line's numerator produces the simpler CATE formula from equation 4.2, and reminds us that adding the additional assumption needed for the estimator in equation 4.3 will not produce a different estimated effect. Instead, the benefit of the decomposed estimator is its ability to place the CATE by estimating both the untreated and treated outcomes for the compliers.[3]

For some applications, this added feature may be altogether unhelpful. For others, it may be helpful but the added assumption may not be intuitive to reason with. Indeed, proposing possible shifts for the outcomes of the always-takers, for example, may lead practitioners somewhat closer to the sort of treatment-dependent reasoning that SCQE seeks to avoid. Still, there are likely applications in which reasoning with and bounding the outcome shifts for two populations is straightforward, and in such cases this feature can be utilized for a clearer picture of the treatment's impact. In particular, this may be likely to occur in applications with a treatment level status quo that is both long standing and quite well understood, giving domain experts experience and guiding their expectations for plausible shifts among both the long-treated subset of patients (δ_{AT}) as well as the population as a whole (δ). Conversely, applications in which two differing treatment levels happen to be observed, without much understanding of these subpopulations, are unlikely to inspire confidence in assessing possible values of any two shifts. We can still produce these results with whatever confidence level we do have, thereby locating our effect, but asserting a precise placement in such settings is more likely to suggest an indefensibly narrow shift assumption than a robust estimate.

[3] We know that if δ is assumed to be 0, we get the unadjusted, IV Wald estimator for the CATE. Further, for a given value of δ there are an infinite number of valid δ_{AT} and δ_{nAT} value pairings. Together, these assertions prove that IV cannot uniquely place its effect without an additional assumption.

4.2 One-cohort SCQE

The setting for SCQE discussed thus far requires two cohorts with differing treatment rates, but in many cases we have access to only a single observed cohort with some level of the treatment of interest. In this section, we demonstrate that the same partial identification strategy used in the two-cohort SCQE can be borrowed to produce an ATT with only one cohort. Those with access to data from a second cohort should use it, as the one-cohort SCQE simply offers an expansion of the settings within which we can use this method rather than an improvement on its predecessor. We introduce tools for one-cohort inference, including a count-based implementation; discuss the implications of the difference between the two-cohort and one-cohort assumptions; and apply these tools to re-analyze four observational studies measuring the effect of hydroxycholorquine on COVID-19 mortality rates.

4.2.1 Interpretation, Estimation, and Inference

Conceptually, the two-cohort SCQE procedure begins by using the δ assumption to take a quantity we observe, the average outcome in the low-use cohort, and then shifting it by a specified amount. The term that results from this assumption, $\mu_{Y0} \equiv \mathbb{E}[Y|Z=0] + \delta$, is the imagined average non-treatment outcome in the high-use cohort. In the one-cohort version of SCQE, that term *is* the assumption required, rather than the result of a different assumption; we directly assume the average non-treatment outcome rather than grounding ourselves by using a low-use cohort's average outcome and assuming a shift away from it. From that step onward, the SCQE process is unchanged – we simply back out the counterfactual non-treatment outcome among the treated, and thus identify the ATT. As in the two-cohort SCQE, we express the ATT across a practitioner-supplied plausible range of assumption values, where the values are μ_{Y0} rather than δ.

In the one-cohort SCQE, we have a binary treatment, D, and an outcome, Y, for only a single observed cohort, as well as an assumed μ_{Y0}. In a standard IV approach, the estimated effect of the treatment is the ratio of the instrument's effect on the average outcome and the instrument's effect on the treatment rate. Here, the instrument is presence in our observed (encouraged) cohort rather than the counterfactual non-treatment cohort. The instrument's effect on the outcome is the difference between the average value of Y and the assumed μ_{Y0}, and its effect on the treatment is simply the proportion treated in the observed cohort, $\pi = \overline{D}$. Formally,

$$\widehat{\text{ATT}}_{1C} = \frac{\overline{Y} - \mu_{Y0}}{\pi} \tag{4.4}$$

This estimator is very similar to the Wald estimator in the two-cohort version (equation 2.4), with our single observed cohort taking the place of the high-use cohort, and μ_{Y0} taking the place of the low-use cohort's average outcome shifted by delta.

This latter substitution has important implications conceptually as well as for inference. We presume that when practitioners imagine an expectation of $Y(0)$ in supplying μ_{Y0}, they are imagining a point estimate. By contrast, when we fill in this expectation based on the estimated average $Y(0)$ of another group, as in the two-cohort version, the sampling uncertainty in that group is factored into the variance of our estimate. As a result, we leave it to the practitioner to employ a range of μ_{Y0} that is wide enough to account for differences between the real and imagined cohorts (of the type normally captured by δ) *in addition to* any sampling uncertainty from the statistics that led to a given choice of μ_{Y0} (if that choice was based on other observations rather than first principles).

Incorporating this difference in our inferential procedures, we might try to create a confidence set around ATT_{1C} using two approaches. The first is to implement a closed form estimator using the Fieller approach that underlies Anderson-Rubin sets (Fieller, 1954; Cinelli and Hazlett, 2021). While we believe this form exists, attempts to derive it have left us with estimators which do not behave as expected in certain edge cases.[4] The second

[4] We would be happy to share our attempts to this end with interested researchers.

approach is to generate a fake no-use cohort with an outcome rate of μ_{Y0} and then to use the 2-cohort SCQE to provide a borrowed confidence set.[5] The key consideration for this approach is the size of that imagined cohort. Our conceptualization of μ_{Y0} as a constant rather than a random variable can be accommodated here by making our generated no-use cohort as close to infinite sized as is feasible. In practice, we have found that as long as it is several orders of magnitude larger than the observed cohort, increasing the size by several more orders of magnitude makes no meaningful difference for the confidence set bounds.

To the web application mentioned in Section 3.3, we added the option to run this one-cohort version of SCQE. Unlike the two-cohort version, it takes only 4 counts to run SCQE in this scenario: $N_{D=0}$, $N_{Y=1,D=0}$, $N_{D=1}$, and $N_{Y=1,D=1}$. The plots produced are similar to those we display for the two-cohort version, but indexed by μ_{Y0} rather than by δ.

4.2.2 One-cohort application: Hydroxychloroquine for COVID-19

To demonstrate the one-cohort SCQE strategy, we evaluate what four cross-sectional observational studies on hydroxychloroquine's use as a COVID-19 treatment can tell us about its effectiveness. Early in the COVID-19 pandemic, hydroxychloroquine (HCQ) was touted by some as a potentially effective treatment for the SARS-COV-2 virus, to the consternation of most medical and public health experts (Cathey, 2020), and was granted an Emergency Use Authorization (EUA) by the Federal Drug Administration on March 28, 2020 (FDA Press Release, 2020). In order to provide data-driven evidence about HCQ's effects, several randomized trials were initiated, which later showed null effects on mortality and clinical status (RECOVERY Collaborative Group, 2020; Cavalcanti et al., 2020; Self et al., 2020). Before these trials closed, however, several retrospective observational studies were conducted, some published in top medical journals, in an attempt to learn as much as possible from ongoing use of HCQ under the EUA. Four of the most prominent of these articles, those

[5] For the count-based SCQE, this would mean specifying some imagined N patients and the proposed number of those with the outcome. If we have unit-level data, it would mean constructing the corresponding D, Z, and Y vectors.

by Geleris et al., 2020, Mahévas et al., 2020, Magagnoli et al., 2020, and Rosenberg et al., 2020, each showed null effects of HCQ on mortality using combinations of multivariable propensity scores (as weights or through regression splines) and proportional hazard models. In an editorial published alongside the Geleris et al., 2020 study in the New England Journal of Medicine, the Editor-in-Chief, the Executive Editor, and others wrote:

> In short, the authors used modern methods to rigorously analyze data that are available now, despite the well-understood limitations of observational studies . . . We have chosen to publish this report so that clinicians will have some information that is based on rigorous analyses of available observational data. However, this observational study is in no way a substitute for randomized, placebo-controlled trials . . . When we have little idea about appropriate therapy, we have an obligation to help by performing studies that will help us to learn together with our patients. (Rubin et al., 2020)

The academic medical establishment, though intimately familiar with the dangers of substituting observational studies for randomized trials, was signaling its belief that until trial results were released, capturing the "best" information available using covariate adjusted observational analyses was a net positive in the fight against COVID-19.

We agree with this assessment, but suggest that evaluation of the same data using SCQE would have provided an important alternate demonstration of the strength of evidence the data provides, or the lack thereof. Each of these studies feature a single cohort, and the published papers contain the data necessary for implementation using the count-based one-cohort SCQE. Doing so for each study links an assumed outcome rate for the whole cohort under no HCQ use (μ_{Y0}) to the ATT that assumption implies. In particular, we learn the non-treatment mortality rate required to claim that HCQ had a significantly beneficial, null, or harmful effect on those who took it. These results are presented in Table 4.1.

Mortality rates in the early months of the pandemic varied widely between hospitals, even after accounting for age, sex, and comorbidities (Block et al., 2021). Given that SCQE

Table 4.1: Published and re-analyzed one-cohort studies of hydroxychloroquine's effectiveness for COVID-19 treatment

Paper:	Key population features and inclusion criteria:	Original covariate-adjusted hazard ratio (HR), [95% CI]:	Baseline mortality assumption needed to conclude HCQ was:		
			beneficial	zero	harmful
Geleris, et al.	1,376 patients at an NYC hospital. Inclusion criteria: no discharge, intubation, or death within 24hr	IPW HR, composite outcome of intubation or mortality: 1.04, [0.82, 1.32]	>18.8%	16.90%	<14.9%
Mahevas, et al.	181 patients at four French tertiary hospitals. Inclusion criteria: oxygen needed, but no direct ICU admissions	IPW HR, 21-day mortality: 1.2, [0.4, 3.3]	>13.6%	9.40%	<5.1%
Magagnoli, et al.	368 male US Veterans Administration patients. Most HCQ patients also received AZC	HR, HCQ vs none: 1.14, [0.56, 2.32]; HR, HCQ + AZC vs none: 2.61, [1.10, 6.17]	>23.0%	19.00%	<15.0%
Rosenberg, et al.	1,438 patients at 25 New York state hospitals. Inclusion Criteria: No discharge within 24 hours	HR, HCQ vs none: 1.08, [0.63, 1.85]; HR, HCQ + AZC vs none: 1.35, [0.76, 2.40]	>22.4%	20.30%	<18.2%

Note: Four early-pandemic observational studies of hydroxychloroquine's (HCQ) effectiveness in reducing COVID-19 hospital mortality rates. The populations, outcome and treatment specification, and analysis strategies varied between studies. Each calculated covariate-adjusted hazard ratios (HR); in order to bolster identification claims, two used Inverse Propensity Weighting (IPW) and a third used propensity scores in regression splines. Each study used a single cohort, so we apply the one-cohort SCQE to each. This generates the value of μ_{Y0} we would need to defend in order to claim a study suggests beneficial, harmful, or exactly zero causal effect (at the α=0.05 level, using Anderson-Rubin confidence sets). AZC=azithromycin.

transforms counterfactual mortality rates into ATT estimates, this application is not one that lends itself to narrow bounds or confident conclusions. What SCQE does offer however is a clear expression of how our lack of knowledge should affect our claims. That lack of knowledge is not limited to mortality rates, and yet many methods that rely on domain knowledge in other areas (like the propensity scoring models used in three of the original analyses) express results in a way that does not acknowledge that complexity as clearly.

We should not expect the four cohorts to be comparable, especially given the differing patient types and inclusion criteria, and thus the true μ_{Y0} is likely to vary between them. Still, SCQE shows that the baseline mortality rates we would need to deem *implausible* in order to assert conclusively beneficial or harmful effects across all four studies were 23.0% or less and 5.2% or more, respectively, neither of which seems like an indefensible range. More importantly, a baseline mortality rate of, for example, 14% would imply a harmful effect in three of the study groups but a beneficial effect in the fourth.

For any one of the four studies then, our re-analysis with SCQE demonstrates the fruitlessness of defending or ruling out a conclusion in either direction; we see the range of possible mortality rates not just in the literature broadly, but in the clear examples of the other studies. Clinicians for these studies would be able to set clearer bounds on their population's baseline mortality than we can, but the observed variability would warrant caution. With the original analyses, some readers may misinterpret the adjusted estimates as evidence of no effect (as opposed to an inability to reject the null), but the SCQE results are clear in their suggestion that the data provides us with very little clarity. Given the urgency of the pandemic at the time, it is understandable that doctors were prepared to incorporate the results of the original analyses in their practice, but a critical part of evidence-based practice should be acknowledging how strong or weak that evidence is, and SCQE is useful to that end.

4.3 Accommodating distributional specifications for δ

The presence of informed and measured domain knowledge is the source of SCQE's insights as well as its sensible caution. For some methodologists interested in SCQE, particularly those partial to bayesian approaches, the use of simple plausibility bounds for δ may seem to unnecessarily discard much of that domain knowledge. Instead, they suggest, SCQE should capture and utilize a prior distribution on δ. In Section 5.5 we suggest resources for eliciting prior information from domain experts, and in Section 5.6 we discuss the pros and cons of using a distribution on δ and how those two should help practitioners choose between a distribution and plausibility bounds. Here, we develop SCQE to handle distributional specifications for δ, transforming that knowledge into probabilistic information about the resultant CATE. We present two options in Section 4.3.1, alongside differentiating factors to consider, and include two further options in Appendix A.2. The first option applies a broadly prior-driven perspective across a simulation-based version of our CATE estimator. The second allows for a prior on δ while maintaining as much of the original SCQE procedure as possible. In Section 4.3.2, we demonstrate the use of both options through an application to data from the TB prevention study in Chapter 2.

4.3.1 Utilizing δ priors

The first option, which we refer to as the "full-bayesian" approach, generates a large number of draws from equation 4.2, where each term is replaced by some distribution. The δ term is drawn from the prior elicited from domain knowledge. The other terms in the equation are means (whether average outcomes or treatment rates) calculated from the observed data, and we can construct posterior distributions from which to generate our draws by combining that data with priors for each. The natural choice for such priors is a Beta distribution, a conjugate prior for our binomial proportions. In applications with a continuous rather than a binary outcome, we may use instead an appropriately chosen continuous prior. For the binary outcome case, reasonable choices may be to use flat priors for $\mathbb{E}[Y|Z=0]$, π_0, and π_1, and a prior for $\mathbb{E}[Y|Z=1]$ that is weak and centered around the value of $\mathbb{E}[Y|Z=0]$ (to represent a null hypothesis of no treatment effect). In most cases with sizable data, the likelihoods will – and should – dwarf our priors in terms of their impact on our posteriors. Still, this strategy allows for flexibility in applications with smaller datasets or particularly strong priors. From the posteriors we produce, alongside our prior on δ, we draw a large number of samples and generate the resultant distribution on CATE. This distribution may then be displayed graphically or described through summary statistics, drawing attention to those effect estimates which the δ prior suggests are most likely. This strategy is not a standard bayesian design, since our "posterior" distribution of interest, the CATE, had neither a prior nor relevant observed data for a likelihood. The appellation of full-bayesian refers instead to the application of priors to every term in equation 4.2 when generating draws for our simulation, unlike the strategy below and the two alternatives in Appendix A.2. Bayesian practitioners may find this approach natural and clear and would likely stress the importance of widely applying priors. Other practitioners may be concerned, however, that the full-bayesian approach is no longer a partial identification strategy and risks losing much of SCQE's focus on avoiding overconfidence and unsupported conclusions. That is, in interpreting a (likely unimodal) distribution, or summary statistics of central tendency thereof, readers may not grasp that the results are not a point estimate surrounded by statistical uncertainty, but rather an expression of identification uncertainty.

The second option, which we refer to as the "weighted-SCQE" approach, re-runs the standard SCQE procedure while acknowledging the prior on δ. But how might our distributional input produce results expressed through SCQE's partial identification perspective? The weighted-SCQE approach generates a plot like Figure 2.3, but with each δ value's confidence set weighted – literally, through the thickness of the plotted line – by the probability of that δ value, according to the elicited prior. Formally, given some set of plotted δ values, the weight, w_i, of the plotted line for δ_i should be proportional to the fraction of the prior's cumulative distribution function, F, located nearer to δ_i than any other plotted value δ_j:

$$w_i \propto F(b) - F(a) \quad s.t. \;\; \forall \; x \in [a, b], \;\; |x - \delta_i| < |x - \delta_j| \tag{4.5}$$

This plot should display the range of δ values deemed plausible, imposed either through the use of a bounded prior on δ or a separate elicitation of such a range from a subject-matter expert. Proponents of this method may prefer its ability to incorporate a distribution through simple marginalization rather than through blanket application of priors. The weights demonstrate the relative likelihood of each effect, satisfying those partial to such focus; the edges of the plotted area make clear the conclusions that we can and cannot deem (im)plausible; and statistical uncertainty is captured through the size of each confidence set rather than through the variance of the sampling distributions for the non-δ terms.

4.3.2 δ priors in tuberculosis prevention evaluation

To demonstrate these two implementation options for using prior distributions of δ, we apply both to data from Chapter 2's study of IPT's effects on tuberculosis incidence. In that study, we drew information about δ from a subject-matter expert as well as from treatment-unaffected data. The means with which we incorporated the former, however, presents a useful opportunity for improvements through priors. Dr. Maokola expressed that recorded TB incidence may have increased by 0.5pp to 1pp per year due to improved surveillance, but also that his confidence in this range was weak. Instead of using this range as a δ bound, it would more sensibly be interpreted as the modal area of a wider distribution. Below, we choose a sensible distribution, plug it in to both options explained in Section 4.3, and compare their results with that of the original analysis. This section is meant to demonstrate the methodological advances rather than to reevaluate the study's results; we do not incorporate data-driven suggestions for δ or apply the methods to the entirety of the data.

In order to give the two options an opportunity to distinguish themselves from each other and the original results, we will use only data from the facility with the smallest sample size, facility 21 in table 2.1.[6] There are several ways to turn our expert's beliefs into a formal distribution, as discussed in Section 5.5; here we choose to utilize the Sheffield Elicitation Framework (Oakley, 2020), requesting the best fit distribution with the following restrictions:

- Must be bounded between -2.3pp and 17pp, the former representing a shift from the no-use cohort's TB rate of 0.023 down to 0, and the latter representing an increase equal to the largest observed cohort-to-cohort shift in TB rates across the facilities.

- Should have a cumulative distribution function with around a third of its density between 1.5pp and 3pp, representing the (weak) best guess range from our expert

[6] The no-use cohort had 219 patients, 5 of whom developed TB; the high-use cohort had 204 patients, 22 of whom were treated; of those treated patients, none developed TB; of the 182 untreated patients, 5 developed TB; just over 3 years elapsed between the two cohorts' mean patient arrival dates.

scaled to the three years between the cohorts.

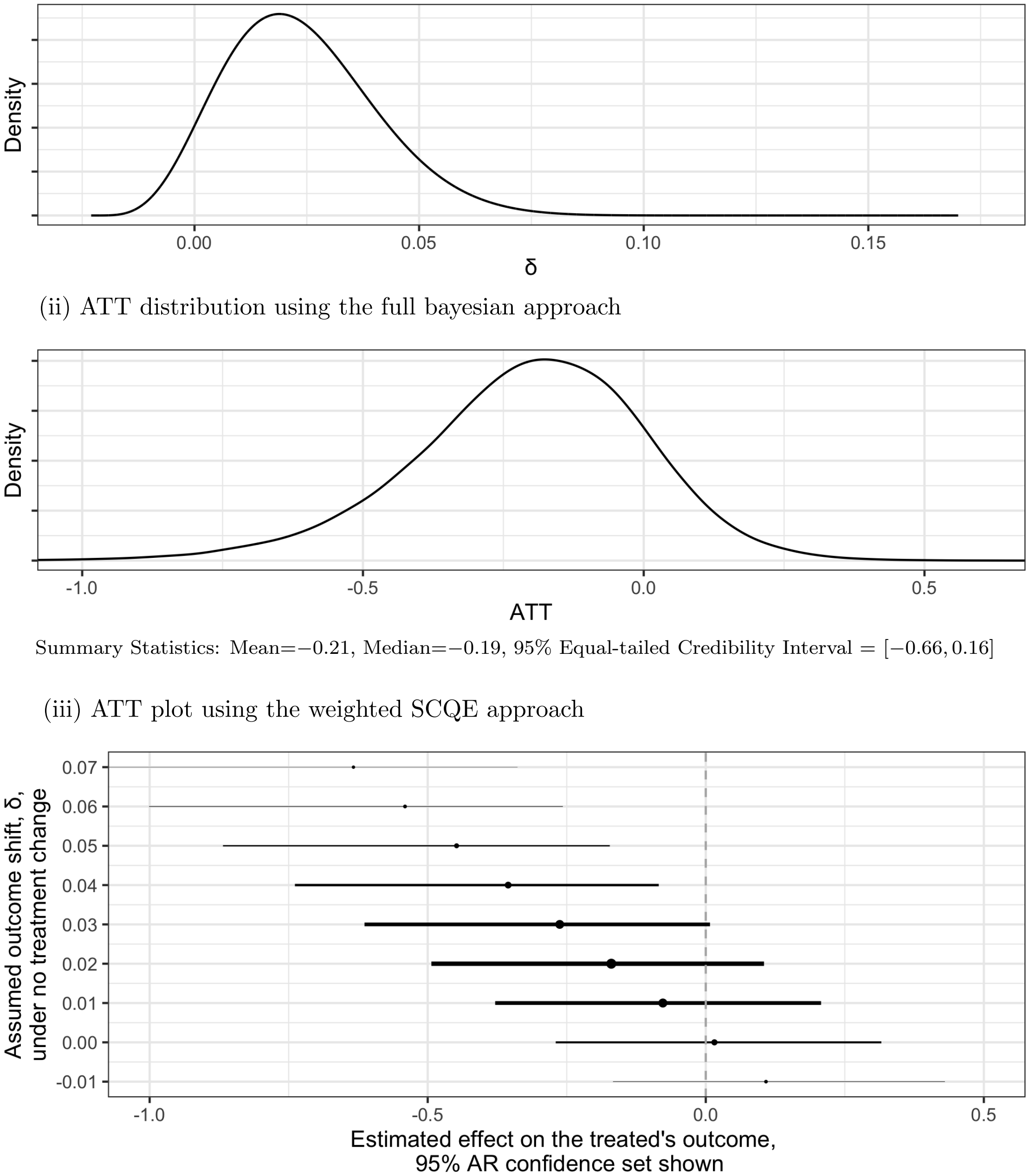
The suggested distribution was a $\delta \sim Beta(5.63, 17.7)$, scaled to the [-2.3,17.0] bounds. The distribution can be seen in Figure 4.1i, and seems to fit a reasonable shape.

For the full bayesian approach, as described above, we use a flat Beta prior ($Beta(1, 1)$) for $\mathbb{E}[Y|Z=0]$, π_0, and π_1, and a fairly weak Beta prior ($Beta(1, 42)$) for $\mathbb{E}[Y|Z=1]$ centered around the no-use cohort's outcome rate of 0.023. We calculate the simple closed-form posterior distributions given these Beta priors and the binomial proportion data we observe from the facility, and we generate 100,000 samples from each of them, as well as from our δ prior. Finally, plugging these samples into equation 4.2 gives us 100,000 samples of our ATT. We display this distribution in Figure 4.1ii, alongside some summary statistics.

For the weighted-SCQE approach, again as described above, we run a standard SCQE analysis with δ ranging from -2.3 to 17.0 and then we weight each fitted ATT confidence set by their likelihood according to the scaled $\delta \sim Beta(5.63, 17.7)$ prior. The resultant plot, shown in Figure 4.1iii, only displays values of δ between its 1st and 99th quantiles, to focus on those values with substantive likelihoods and weights. If we were instead to use those quantiles to form a plausibility range for δ rather than using the prior distribution itself, the standard SCQE plot produced would look just like Figure 4.1iii, but with uniformly-weighted lines.

The two outputs, Figures 4.1ii and 4.1iii, are displaying the same underlying knowledge, excepting differences between sampling coverage and AR set coverage that are not clearly visible in this case. Compared to the standard plausibility range plot (an unweighted version of Figure 4.1iii), both draw greater attention to ATTs around -0.5 to 0 and away from ATTs at the edges of the plausibility range. The simplicity of Figure 4.1ii and its summary statistics are enticing, but the cautious style of Figure 4.1iii presents identification uncertainty more clearly. It is up to the practitioner to decide which fits their preferences and inferential goals, though we suggest the weighted SCQE provides an attractive balance between capturing prior knowledge and avoiding overconfidence.

Figure 4.1: Two implementations of δ priors



(i) δ prior based on domain knowledge

(ii) ATT distribution using the full bayesian approach

Summary Statistics: Mean$=-0.21$, Median$=-0.19$, 95% Equal-tailed Credibility Interval $= [-0.66, 0.16]$

(iii) ATT plot using the weighted SCQE approach

Re-analysis of IPT's effectiveness in TB prevention from Chapter 2, only data from the smallest clinic, facility 21 in table 2.1. In 4.1i, we display a prior on δ that aligns with our domain expert's priors. We transform this δ prior and priors on the rest of the CATE estimand's quantities into a corresponding distribution on ATT using the full bayesian approach. That distribution is shown in 4.1ii alongside corresponding summary statistics and a credibility interval. We also use the prior to weight the standard SCQE results plot using the weighted SCQE approach, displaying it in 4.1iii.

CHAPTER 5

Concepts in SCQE

In the previous chapters, we have introduced SCQE and suggested that it may be a useful alternative to conditional-ignorability-based methods, especially for applications in which we have little understanding of the treatment assignment mechanism or little trust in adjusted comparisons between treated and untreated units. We extended the method in several ways and demonstrated how to apply it. Because of the novelty of SCQE, however, practitioners looking to use it have not had access to extensive guidance on best practices. In this chapter, we attempt to address that need.

In Section 5.1, we suggest features of potential applications that might make them a good fit for SCQE. Some features relate to how easy it is to reason with possible δ values, and thus how such reasoning should be conducted. We discuss those δ considerations in Section 5.2, offering an informal checklist of possible sources of cohort-to-cohort differences. In Section 5.3, we provide more details about the formal method for doing so – modeling $Y(0)$ as a function of covariates in order to estimate baseline risk – in addition to generalizing this strategy. We demonstrate in Section 5.4 the guidance, and warnings about bias, that DAGs can provide in any attempts to use data to inform guesses about δ.

Addressing our new accommodation of distributional δ specifications, we share general references and SCQE-specific guidance for elicitation of such priors in Section 5.5. We also explore the relative benefits and risks of using bounds or distributions for δ and how that choice might be informed by particular inferential goals in Section 5.6. Finally, in Section 5.7, we discuss SCQE's place within the broader landscape of identification and sensitivity analysis strategies.

5.1 What makes a great application for SCQE

Before establishing what makes an application a good fit for SCQE, it is helpful to reiterate first the two primary ways in which the method can be helpful. First, SCQE may generate narrow, substantively unambiguous effect estimates, giving practitioners a direct answer to questions of treatment impacts. Secondly, SCQE may demonstrate the weakness of our knowledge, mirroring the uncertainty we have about δ in the wide range of effect estimates it generates. While results of the first variety may be more satisfying or easier to communicate, we know that honest quantification of uncertainty is the cornerstone of statistical research and thus the second type of results can be just as helpful. In this section, however, we focus on settings in which SCQE is most likely to produce the first type of results: confident, targeted effect estimates with little uncertainty. Therefore, the "ideal" settings we discuss below should not be seen as the only settings in which SCQE can produce informative results. We present a summary of these suggestions in Table 5.1.

The design *requirements* for SCQE are the presence of two observed cohorts with different levels of some treatment of interest. In addition, the treatment of interest and the outcome should be identically defined in each cohort, and actively choosing definitions to ensure this holds is an important step in avoiding overlooked assumption complexities. Section 2.3.2 demonstrated a set of procedures motivated by this goal. In general, the more similar the cohorts – up to and including cohorts we can conceive of as being two draws from the same superpopulation – the more clearly that subject-matter experts will be able to consider and bound δ values, allowing for targeted effect estimates.

Most conducive designs will feature cohorts separated by time, and we have limited our discussion in this work to such cases because we believe it is particularly natural for experts to reason with outcome shifts in time. However, any definition of cohort separation (Z), such as location or another grouping, could theoretically be used. We present two examples that may fit non-time-separated SCQE well. First, we might compare student infection rates (Y) at two high schools (Z) in the same district during the COVID pandemic, only one of which offered optional in-person classes (D) because enough of their teachers agreed to do so. If school district administrators can bound the expected remote-learning-only outcome differences between the schools, taking into account the variety of socioeconomic and other differences between the student populations, SCQE can identify the effect the option of remote learning had on those students who took it. Second, say a sports-focused phone app offers in-app sports betting features (D) for residents of states in which gambling is legal (Z), and the developers wish to measure the effect that feature has on the average monthly time spent in-app (Y) by those using it. Given a range of possible state-to-state differences, informed by (among other things) the possibility that people who spend more time on a gambling app may also have chosen to live in states where gambling is legal, SCQE can estimate the effect of the gambling feature.

While these applications are methodologically feasible, the claims SCQE allows are likely to be insightful only when our understanding of the identification uncertainty is strong. We believe that is uniquely probable when cohorts are separated by time, and we continue our focus on time-separated SCQE moving forward. For other applications, a rich literature of sensitivity analyses for IV is available to practitioners, including more general methods that also focus on violations of the exogeneity assumption (Conley et al., 2012; van Kippersluis and Rietveld, 2018).

An ideal outcome should follow the treatment of interest as closely in time as possible for two reasons. First, we want to prevent the treatment increase or decrease in the later cohort from affecting the outcomes of units in the earlier cohort, though this cross contamination may not be possible in certain applications. Second, data-driven evaluations regarding δ can be threatened by the presence of post-treatment covariates that affect the outcome (see Section 5.4 for more); minimizing the time between treatment and outcome helps to avoid those threats.

Applications with larger changes in treatment usage are preferable to those with smaller such changes for three reasons. First, small changes may constitute a weak instrument

Table 5.1: Idealized settings for targeted, narrow SCQE results

Property	**Why it is ideal**
Cohorts separated in time	Outcome trends in time are particularly easy to reason about
Outcome follows treatment as closely as possible in time	(1) Avoids cross-cohort contamination of treatment practice (2) Evaluation of δ is complicated by post-treatment, pre-outcome covaraites
Large change in treatment usage rate	(1) Weak instruments accounted for by A-R sets, but disjoint or $\mathbb{R}$-spanning sets are hard to interpret (2) Small changes magnify errors in δ (3) Population of interest is larger
Pre-study trends available; Ideally flat (or at least low variance) or over a long period of time	Stable trends allow tighter arguments about δ, and a long pre-study period offers suggested empirical variance for δ
Y easy for domain experts to reason with	Familiarity allows for appropriately narrow bounds/distributions on δ
We know why treatment usage changed, and expected that change	If treatment changes were due to population shifts rather than practice changes, δ bounds are likely to be wider

problem, and though our use of Anderson-Rubin confidence sets guarantees correct coverage, a disjoint or real-line-spanning set will not provide precise information about any effect. Secondly, identification uncertainty in the form of δ misspecification will be increasingly magnified by smaller treatment changes.[1] Finally, our estimands refer to the units affected by the treatment change, and thus larger treatment changes may allow us to make claims about a broader portion of the population. For example, if researchers want use the estimated CATE as a proof of concept for a broader future treatment rollout, the larger the observed treated population the lower the expected bias in that generalization.

Certain designs are particularly conducive to utilizing data-driven knowledge. An ideal application will offer stable outcome and treatment levels over time before our first cohort, after which the treatment level will suddenly jump or drop. The stable outcome trend would ideally be flat, though a stable increase or decrease over time also lends credence to an

[1] We can prove this to ourselves using any of our estimators by separating the δ term into the true δ and a misspecification error. For example, in equation 4.1, removing this error term from the rest of the equation carries with it the $\pi_1 - \pi_0$ denominator, so the smaller the treatment difference the larger the effect the error has on our estimate.

assumption of counterfactual trend continuation. Even if the observed trend is not stable, a low outcome variance over time helps to limit the counterfactual changes we might expect. Pre-study trend data may also directly inform a proposed distribution of δ; in a theoretical example with a lengthy record of outcome data and a guarantee that the treatment usage change was the sole change during our cohorts, the series of placebo outcome shifts across the pre-study data provide an empirical distribution to use for δ.

If no data exist within which to observe a pre-study outcome trend, the outcome should be substantively understood beyond the scope of the observed data to facilitate reasoning about δ. In particular, the outcome under a given treatment regime, $Y(\mathrm{d})$, should ideally be a familiar, if not centrally important, measure for a subject-matter expert to reason with.

Ideally, practitioners should understand why the treatment rate changed between cohorts, as these reasons may affect our consideration of δ. That is, knowing that (and why) those assigning treatment made an effort to change treatment levels may suggest a smaller range of δ than if shifts in population composition lead to more "treatment-prone" units showing up in one cohort (further discussion in Section 5.2).

We reiterate that SCQE can still produce valid, informative results outside the bounds of these ideal settings. In such cases, the δ specification process should be accompanied by added caution, and special care should be taken to follow the best practices we introduce below.

5.2 Consideration and Decomposition of δ

We discuss here how practitioners should think about and reason with δ, and offer a particular framework within which to do so. We have defined δ as the difference in the expected values of the two cohort-wide average outcomes under a set treatment regime ($Y(\mathrm{d})$). Reasoning about this unidentifiable quantity is not and should not be simple; δ must theoretically account for shifts in any number of observed or unobserved factors with causal relationships to the outcome, as well as their interactions. That is, when we reason with possible values of δ, we must be honest about the complexities likely involved in the outcome-generating process and adjust our level of confidence accordingly.

In order to facilitate thorough evaluation about δ – whether for prior elicitation, plausibility bounding, or modeling strategies – we present a conceptual decomposition of δ to use as a general checklist of considerations. Practitioners are encouraged to conceptually decompose δ into three pieces: baseline covariate differences, other treatment differences, and context differences.[2] These pieces account for the various ways the outcome measure may be related to the cohort (other than through changes in our treatment of interest), and by including them in δ we remove their ability to bias the CATE estimate itself.

Baseline covariate differences refer to unit-level factors, whether observed or not, that differ between the cohorts. Some such differences are guaranteed to exist due to random selection, while others represent non-random shifts in the population composition. For an application measuring the effect of a medication on in-hospital mortality, these could include differences in covariates measuring demographics, comorbidities, and illness severity at admission. In another example, an online shopping application measuring the effect of an opt-in product-suggestion feature on the likelihood of purchase, we would include purchase history, the search term leading a user to the site, and whether they were on mobile or web.

Other treatment differences (where "other" refers to the exclusion of our treatment of interest) also refer to unit-level factors, but to those that are not fixed at the time of selection. In medical applications, these are often literal treatments, like medications or procedures, but may also refer to practice changes or treatment timing choices by caregivers.

[2] We draw this grouping in part from literature dealing with the decomposition of bias in treatment-effect estimation, specifically a recent external validity framework (Egami and Hartman, 2020). The authors' motivation, decomposing the aspects of a trial's external validity, differs meaningfully from our goal of decomposing baseline risk differences, but their decomposition offers a useful conceptual starting point. What they refer to as "Y-validity" and "T-validity" are requirements in the SCQE framework; in order for the method to be applicable, our definition of the treatment of interest and of the outcome must not differ between the cohorts. But violations of their "C-validity" and "X-validity" – cohort-to-cohort comparisons in contexts (settings) and populations (covariates), the latter of which we further subdivide – are exactly those differences about which we must reason in considering δ.

For an online shopping application, they may include discounts on individual items the user encounters, or an A/B test being run on the checkout process. A complication arises in our consideration of δ for certain post-selection examples: if these covariates are causally post-treatment – that is, if the usage of some other treatment differs between cohorts and is additionally affected by whether a unit gets the treatment of interest – we must separate the cohort-to-cohort changes from the changes driven by the treatment of interest and only consider the former in reasoning about δ. This challenge is explored further in Section 5.4.

Context differences represent cohort-level, rather than unit-level, differences in a given unit's expected outcome, i.e. differences dependent on which cohort they were in. If we could take two identical units – with the same observed and unobserved unit-level covariates and treatments and the same assignment for our treatment of interest – and place one in the low-use cohort and the other in the high-use cohort, context differences are those factors that would account for the inequality of their expected outcomes. For the hospital application, an increase or decrease in how busy or overworked caregivers were between cohorts would be a context difference. For the online shopping example, differing proximity to the holiday season would be a context difference.

We encourage practitioners to explicitly think through each of these three decomposed pieces of δ, as missing any one can lead to improper evaluation about a reasonable distribution or about plausible bounds. When one or two differences are particularly clear, using this decomposition as a checklist may help to look beyond them to a wider range of necessary δ considerations. However, even if we account for all such differences, the ways in which they could interact are also important, and the modeling strategy discussed below allows us to address that added complexity.

5.3 Modeling suggested δ values

Having discussed important ways of thinking about the components of what might affect δ, we now expand on a strategy to explicitly model observed covariates' impact on the outcome, in what we call the model-informed δ procedure. This strategy may be particularly useful in accounting for the complex interactive impacts of various inputs. As introduced in Section 2.2.1.1, which discusses the no-use/high-use form of SCQE, we start by modeling the outcome as a function of some set of covariates, X, in the no-use cohort, measuring their impact on $Y(0)$. We then assume that all units share the same X to $Y(0)$ relationship, regardless of their treatment status and thus their observed outcomes. Note that this does not assume the potential outcomes of the treated and non-treated units are the same, only that the relationship between X and their non-treatment potential outcomes are. Predicting $Y(0)$ for all units in both cohorts and comparing the average resultant $Y(0)$ estimates between the cohorts then provides an estimate of the cohort-to-cohort baseline risk difference, and thus of δ.

An extension to the low-use/high-use setting follows with minimal changes required. If we make the same assumption that modeling untreated units captures the X and $Y(0)$ relationship regardless of realized treatment status, we can estimate this model using only the subset of low-use cohort units that are untreated. Applying the resultant model to both the treated and untreated units in both the low-use and high-use cohorts once again provides estimates of each unit's baseline risk, of the average baseline risk in each cohorts, and thus of δ. While it may seem less natural to limit our modeling domain to a non-random subset of the low-use cohort's units, we did not require additional assumptions to do so. In other words, designs in which this strategy is biased are exactly those in which the no-use/high-use version of the same strategy is biased as well.[3]

As with any other modeling attempt, omitted variables and mis-specified models may produce biased estimates of $Y(0)$, and even with a well-chosen modeling strategy the quantity

[3] We further note that our modeling space need not be limited to the untreated units in the low-use (or no-use) cohort. Instead, we can estimate the relationship between X and $Y(0)$ using the untreated units in the high-use cohort as well, and then apply the resultant model to all units in both cohorts as before. This may be particularly helpful in settings with less data, as the larger modeled population can give more reliable estimates.

of interest in this process is unidentifiable.[4] Consequently, the results of this procedure should not be unquestioningly accepted as our best guess at δ. Instead, they should be considered as one suggestion, alongside those of domain knowledge experts, for plausible δ values. Just as the SCQE approach acknowledges that we doubt a model's ability to uncover the true treatment assignment mechanism, so too should we view our modeled counterfactual outcomes as a suggestion that is likely to contain bias.

5.4 Guidance, and warnings, from DAGs

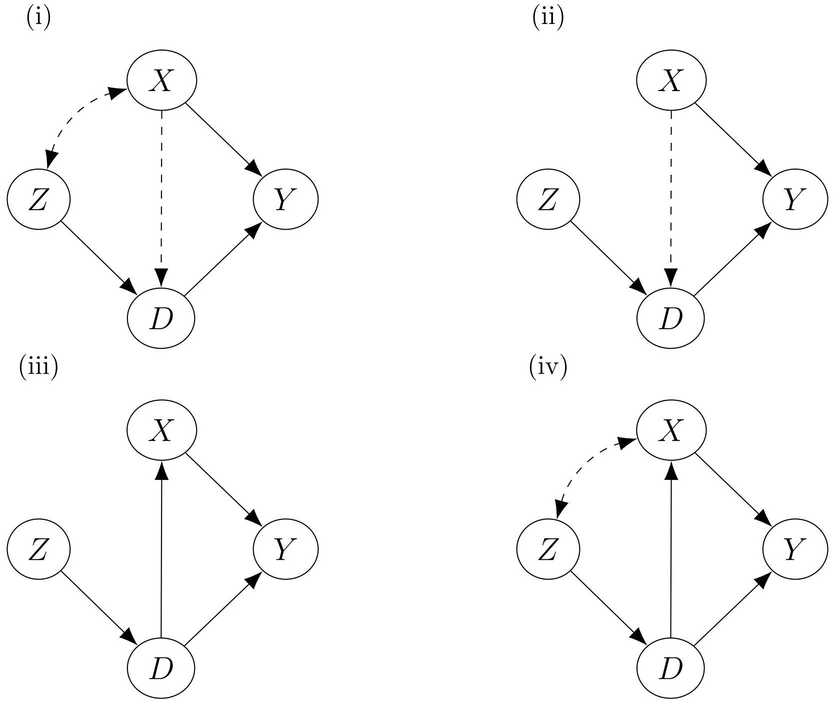
We have provided little guidance on which covariates should be included or excluded from general considerations of δ (Section 5.2) and our model-informed δ approach (Section 5.3). Here, we investigate four possible types of covariates, X, defined by the causal DAGs they form, in order to understand the complexities involved in answering this question. In real applications, there will be multiple covariates, likely of more than one type, but this simplification allows us to assess the implications of each separately. Where relevant, we note complications arising from the presence of multiple covariates. Guidance for covariate inclusion and exclusion below will reference the model-informed δ approach, but the guidance also applies to general considerations of δ unless stated otherwise. We limit ourselves to relevant contexts: those in which Z, our cohort designation, affects treatment use, D, which in turn affects the outcome, Y, and in which the covariate X affects Y as well.

Consider the first DAG (Figure 5.1i), in which the cohort and covariate are not independent.[5] Because this covariate differs between cohorts and affects the outcome, it (and, more generally, all X covariates with this causal graph) should be included in our model-informed

[4] This acknowledgement motivates our allowance for the use of any modeling procedure, from a linear regression to the most complex black box machine learning method. Indeed, this procedure provides a good opportunity for added complexity and non-parametric strategies, given our interest in complex covariate interactions.

[5] A clear causal direction may not exist between the two; this scenario simply implies we have no reason to believe the two have no relationship. Some DAG users would rewrite this bi-directed arrow as an unobserved variable U with arrows into X and Z.

Figure 5.1: Four possible DAGS for data-informed δ consideration



Possible Directed Acyclic Graphs (DAGs) for a covariate of interest, X. Solid arrows represent existing causal pathways; bi-directed arrows represent some unobserved confounding or association of unknown causal direction; and dotted arrows represent allowable, but not required, causal pathways. A covariate like the one in (i) should always be included in consideration about δ. Practically, it is advisable to include a covariate like the one in (ii) as well, as discussed in the text. DAGs (iii) and (iv) feature post-treatment covariates, which may cause problems whether included or excluded from model-based consideration of δ. As discussed, we suggest excluding such covariates and instead evaluating them individually afterward.

δ approach. The dotted arrow from X to D implies it is not substantively or procedurally relevant whether treatment assignment is influenced by this covariate.

In the second and third DAGs (Figures 5.1ii and 5.1iii), a human-knowledge-based assertion is being made about the independence of the covariate and cohort membership. (In the case of the third DAG, the independence is conditional on D.) This independence would imply that X should not be included in a model-informed δ approach. Indeed, if X were the only observed or unobserved covariate, these DAGs would qualify cohort membership as a valid instrument (or at least one that satisfies the exclusion restriction) without requiring any δ adjustment. Just as in a standard IV setting, it would then be up to the practitioner to make a convincing argument that the exogeneity assumption holds. This would likely include a necessary but not sufficient failure to reject the null in a test for correlation between the observed Z and X (tailored to the relevant independence argument — $Z \perp\!\!\!\perp X$ for the second DAG and $Z \perp\!\!\!\perp X \mid D$ for the third DAG). For real applications, covariates other than X are likely present, some of which are not independent of Z, and each of those should be addressed as suggested by the DAG they fit into.

The risks and rewards associated with our independence claim differ between the second and third DAGs. For the second DAG, if our assertion of independence was incorrect, we will improperly exclude X from the modeling procedure and will bias our model-informed δ (regardless of whether X affects D). We do not lose anything by incorrectly rejecting independence, however, so practitioners should err on the side of avoiding such a claim, leading them back to the first DAG.[6]

For the third DAG, however, there are risks in guessing wrong in either direction about the independence of Z and X, because X is a post-treatment covariate. If the independence assumption *is* true, X is simply a mediator of the effect of D on Y, and should not be included in our model-informed δ approach. If, in that case, we fail to make the independence claim and erroneously include X in our model, we would induce bias by falsely attributing a portion of D's effect to over-time trends. However, if the independence assumption *is not* true, we should look to the fourth dag instead, and below we explain how neither the inclusion nor the exclusion of X in our modeling procedure guarantees unbiasedness.

The fourth DAG (Figure 5.1iv) provides the biggest challenge to our model-informed δ approach, as X here is affected by cohort membership but is also post-treatment. When X is a function of D, we must distinguish values of X under treatment and non-treatment by introducing "potential covariate" notation: $X(0)$ and $X(1)$ are the covariate's values under non-treatment and treatment, respectively. By estimating our model among non-treated units only, we have been learning $\mathbb{E}[Y(0)|X(0)]$ in particular, in the hope of using this

[6] In both the first and the second DAGs, X can be included as an exogenous control variable to increase the precision of our estimate. SCQE can do this if we have unit level data, but not in the count-based version.

relationship to generate $Y(0)$ values for all units. When no arrow from D to X exists, $X(0)$ and $X(1)$ are equivalent and applying this learned model to treated units is appropriate because $\mathbb{E}[Y(0)|X(0)] = \mathbb{E}[Y(0)|X(1)]$. But because this covariate *is* affected by D, the $X(1)$ and $X(0)$ values differ, and in order to estimate the true baseline risk, we would need to remove the effect of D on X and use the resultant $X(0)$ values in our model-informed δ approach. However, we cannot do so, as there is no way to decompose the two effects on X: the relationship to Z and the impact of D. Excluding X from our model means ignoring the former, while including it ignores the latter. Either decision produces errors in our baseline risk assessment and thus biases our eventual treatment effect estimate.

These issues help to motivate our stated aversion in Section 5.1 to applications with impactful post-treatment covariates, but SCQE does not fail if this ideal setting is not met. Regardless of post-treatment covariates, the correct δ value will identify the correct CATE, so we need only to be more conservative when assessing what that value might be. The most straightforward approach to dealing with post-treatment covariates is to exclude them from the model-informed δ approach, separately reasoning with how they could affect baseline risk differences. In some settings we may have a clear sense (from outside the observed data) of how Z directly affected X. In others, this extra step may simply be widening our δ bounds. After all, SCQE relies on honest evaluation of what we can and cannot claim about δ, not on the correct specification of an individual model.

5.5 Prior and other knowledge elicitation

We step away from covariate consideration and adjustment now to address the elicitation of priors on δ. A robust literature, spanning a half century and several academic disciplines, addresses effective and sensible elicitation of prior knowledge from experts, and offers tools to do so (Winkler, 1967; Johnston et al., 2015; Stefan et al., 2020; Oakley, 2020). In addition to the suggestions therein, we offer three SCQE-specific additions here, which apply to elicitation of either prior distributions or simple bounds. First, ahead of this knowledge elicitation, we suggest gathering cohort summary statistics for a list of covariates provided to us beforehand by the expert, based on those factors they believe could influence the outcome. In order to ensure that list is complete, we suggest walking the expert through the δ decomposition steps from Section 5.2, asking for relevant baseline, treatment, and contextual covariates. Second, we suggest referring to the outcome under whichever treatment level, $Y(\mathrm{d})$, the domain knowledge expert feels most comfortable reasoning about. This can be asked of our expert explicitly, though it will often coincide with a treatment regime used over a long period of time; if the use of some medical treatment was constant at around 25% for many years, a clinician will likely be best at reasoning about possible outcome shifts at that treatment level rather than at $Y(0)$ or $Y(1)$.

Third, when posing questions to the expert, we suggest limiting the details provided in our questions to help avoid certain biases. To illustrate, imagine that our outcome, in-hospital mortality for stroke patients, was lower during a high-use period than an earlier no-use period, and we are inquiring of the hospital's chief attending neurologist about δ values. Say we tell them about the cohort demarcations, summary statistics for relevant covariates, the scale of treatment increase, and the observed falling outcome rates, and then ask how they would expect the non-treatment outcome rate to have shifted had there been no treatment introduced in the high-use cohort. The neurologist, well aware of the treatment practice change (that they likely orchestrated because of their support of the new treatment), will suggest that the counterfactual outcome rate would likely not have changed from one cohort to the next. Through no fault of their own, they have effectively provided their prior on the treatment effect rather than on the outcome shift, understanding the simple relationship between the two values and informing us that they believe in the treatment's effectiveness.

To avoid any such bias-inducing procedures, we should instead limit the information we provide. We can share all relevant covariate summary statistics (not including the treatment rate of interest) about our two cohorts, as well as the mortality rate in only the no-use cohort (without identifying it as such), and then ask about a range or distribution of mortality differences they might expect between that and the other cohort. The neurologist thereby provides an answer unencumbered by their priors, conscious or not, about the treatment effect. If some relevant information cannot be expressed without unblinding the neurologist to the cohorts (e.g. if they mentioned ahead of time that clinician busy-ness and the treatment of interest spiked at the same time), we can take a two-step approach. We might first provide the data without this factor and ask for their guesses at δ, only afterwards adding in the relevant information and explicitly asking how that particular factor might impact their already-stated belief about possible counterfactual outcome shifts (incorporating with suspicion any major changes that could instead be disguised biases).

As an alternative to these attempts to withhold data, we might consider inquiring of a less intricately involved expert (with knowledge of the setting, outcome, and trends, but not of this particular treatment change or its supposed impact) or of one with weaker priors on the treatment effect itself. The relative value of different experts will depend on the setting, and it is likely that we would do best to obtain priors from each and to combine those into a jointly informed prior (Genest et al., 1986; O'Hagan et al., 2006).

5.6 Considerations on Bounding vs. Distribution

Given the choice of providing bounds for plausible values of δ or a distribution as its prior, which should a practitioner choose to use? In this section, we explore the relative benefits and downsides of these strategies, and provide guidance conditional on practitioner goals.

The development of distribution-utilizing options was motivated by a concern that bounds on δ could draw undue attention to the edges of those bounds. Naturally, when we hear statements of the form, "The effect is between A and B" or "We feel confident dismissing any effects outside of the A to B range," the two effects we are most likely to think about are the bounds, A and B. However, these are precisely the two shift assumptions which the bound-setting practitioner thinks are *least* likely to produce the true effect. Moreover, among every possible descriptor of plausible δ values we might have asked them to estimate (e.g., asking for a single best guess), the estimation procedure with the highest variance is likely the request for such bounds. In other words, asking for bounds may give the least likely and most uncertain points too much weight.

A distribution on δ, on the other hand, is designed to place weight on those areas about which practitioners have the most confidence, utilizing all relevant domain knowledge to do so. Simple conclusions extracted from distribution-driven results reflect the highest, rather than the lowest, density areas among those effects we deem plausible. Similarly to the decision-making process for bounds, domain experts must not overstate their confidence, and responsible practitioners must be careful not to choose an under-dispersed distribution relative to that domain knowledge. It is important for practitioners to remember that relatively flat distributions, including ones with sharp bounds, are completely legitimate choices for δ if they reflect relevant knowledge and uncertainty. The particular risk of a distributional approach, still, is that a mean- or median-focused description of the resultant effect distribution engenders overconfidence in a single estimate. Even if the distribution itself is not overconfident, we must present results with an honest assessment of the variance that does exist. Unlike bound-driven SCQE results, a distribution presents a shiny object in the form of a central tendency parameter, and practitioners must be especially careful to minimize this risk when displaying results.

The choice between imposing a plausibility bound or imposing a distribution on δ is one that should depend on our inferential goals, as well as on the application at hand. If our primary goal is to avoid making unwarranted claims, whether due to philosophical preference or asymmetrical risk, we should use plausibility bounds we are ready to defend. If instead our goal is simply to map the best available guesses about δ to the effect estimates they imply, a distribution is clearly the appropriate choice. If our goals are somewhere between the two, we might choose to impose a bounded distribution on δ, or to use the weighted-SCQE approach introduced in Section 4.3.1 for a plot that only shows δ values for some bounded range. Both of these options draw attention to the effects we deem most likely, while signaling the importance of a given estimate being inside or outside our suggested realm of plausibility.

5.7 SCQE within the sensitivity analysis landscape

The field of causal inference has a particular focus on identification validity, and it understands the risks posed by unacknowledged uncertainty therein. Outside the bounds of randomized controlled trials, which provide a guarantee of point identification and well-defined measures of statistical uncertainty, a point assumption of conditional ignorability has no chance of holding exactly.[7] In observational settings, no matter how many covariates are included in an adjustment procedure, studies must be accompanied by sensitivity analysis or, quite rightly, they will not be accepted as valid. Which sensitivity analyses we choose to use, however, is a matter of researcher preference, application particulars, and inferential goals. These methods vary in the ways they capture and express identification uncertainty, as well as the particular "location" at which we must quantify that uncertainty. Below, we provide an informal categorization of sensitivity analyses, with the goal of understanding where SCQE fits and what role it has.

Some more informal sensitivity analyses take the procedural form of repeated attempts at point identification, in order to establish a rough field of possible effects. One such strategy is to run the same method multiple times with varying inputs. For example, a practitioner might use a simple regression adjustment approach, running it with several choices of X and several parameterizations of each choice of X (Vansteelandt et al., 2012). Another strategy is to run several alternate methods; for example, adding a propensity-weighting method and a matching method. One other approach is to vary the subset of the observations used for

[7] Technically this assumption could hold exactly in cases featuring human-imposed randomness within strata that could be fully uncovered through conditioning, but such cases are extremely rare and may be better categorized as complex randomization designs.

the model rather than varying the model or covariates (Broderick et al., 2020). In these methods, our quantification of identification uncertainty is based on our ability to cover a sufficiently broad set of models or model inputs.

Another set of sensitivity analyses, partial identification strategies, link the point estimate to some other parameter of interest. Each of these methods relies on our ability to reason with this alternative parameter; the more we can limit its value in a given application, the more precise our resultant effect estimate. A primary differentiating factor between partial identification strategies concerns their use or avoidance of point estimates. These strategies follow one of two general approaches.

The first type of partial identification strategies sits atop a point estimate generated through an identification assumption we acknowledge does not hold. Our goal then is to reason about violations of that assumption – to what degree it does not hold – as measured by the value of our alternative parameter. Several methods measuring sensitivity to unobserved confounding fall into this category; they take an estimate based on a conditional ignorability assumption and reason about the possible relationship an omitted variable may have with the treatment and the outcome (VanderWeele and Ding, 2017; Cinelli and Hazlett, 2020). The goal of these analyses are not to decide if the original point estimate is wrong (it certainly is), but to ask whether the general conclusion being made is defensible. An honest practitioner, finding that no realistic confounder could have nullified their beneficial effect estimate, does not claim this means their estimate is unbiased, only that the bias could not plausibly explain away the treatment's effect. And yet even for this analysis, which was conducted and reported sensibly, the original point estimate will often be the only thing internalized by readers, along with less associated uncertainty than warned about by the sensitivity analysis. Centering the original estimate in this procedure thus creates a risk for abuse and misinterpretation.

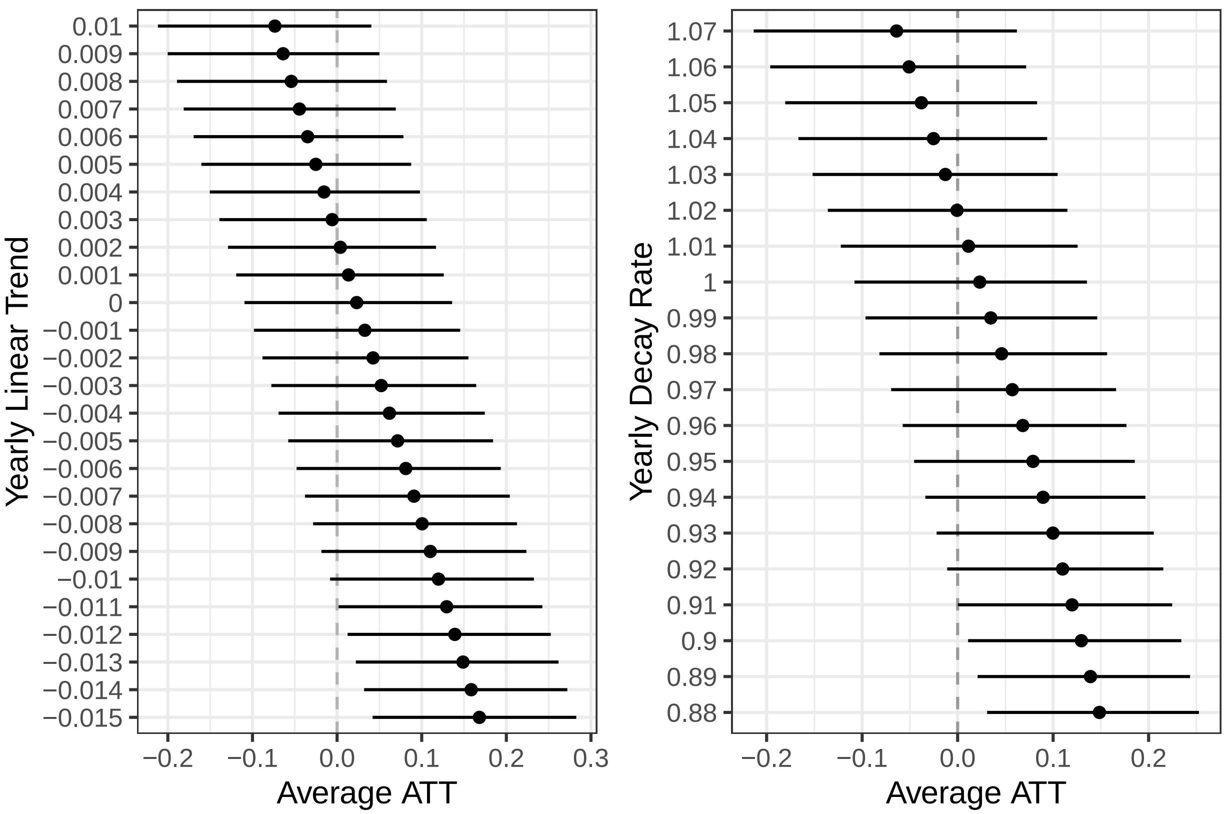
The second type of partial identification strategy makes it harder to misinterpret results in overconfident ways by eschewing the use of an initial estimate generated under an assumption of point identification. Instead, the identification assumption is an interval of possible assumptions. That is, instead of answering the question "How wrong could we be about this point estimate?" these strategies answer the question "What could the point estimate be?" SCQE falls into this interval-identified category, displaying estimates over our δ assumption and forcing practitioner and reader alike to internalize the range of implied estimates. Several bounding methods, whether providing worst case or assumption-based bounds (e.g. Manski, 1990; Lee, 2009; Manski, 2009), are driven by the same goals, if not by similar procedures.

How, then, do we choose from among these strategies for sensitivity analysis? Much of that decision will be driven by the particulars of an application, as well as the relative risks we ascribe to underconfidence and overconfidence. For practitioners with a strong desire to avoid identification overconfidence, we see interval-identified methods as a particularly effective tool. In some applications, SCQE's use of over-time reasoning to set an interval identification assumption will be natural. In others, where little informative knowledge or data exist, SCQE's focus on expressing that lack of knowledge clearly will be beneficial. And in other applications, undoubtedly, over-time trends will be difficult to reason with while other measures of identification uncertainty will be clear. In such cases, and even when SCQE is informative, we encourage the use of multiple sensitivity analyses. We hope SCQE provides a clear, credible addition to the methodological tool chests utilized by researchers.

APPENDIX A

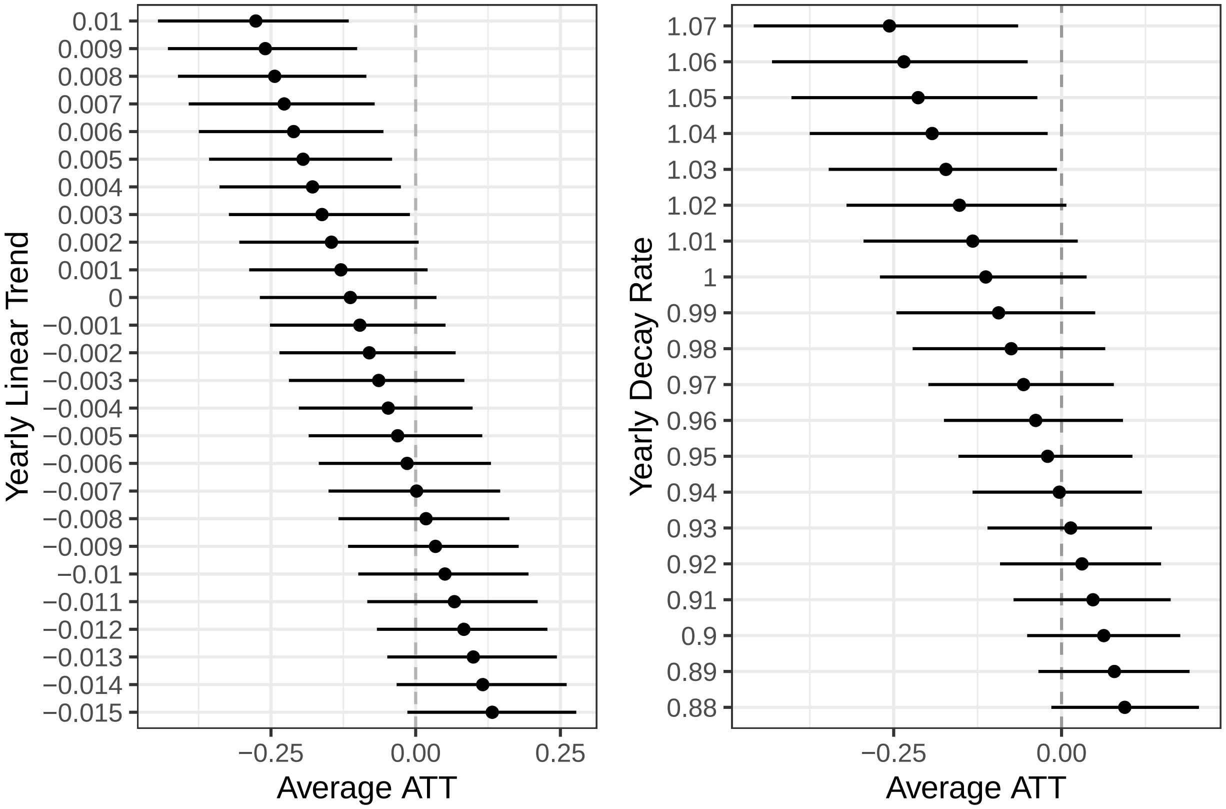
A.1 Tuberculosis application appendix

Figure A.1: Pooled ATTs, by δ, using $M_D = M_Y = 18$ months



Note: Pooled ATT estimates as a function of δ, similar to Figure 2.3 in the main text. It shows those results if the period after a patient's first visit within which both IPT administration is considered treatment and TB development is considered an outcome is extended to 18 months instead of one year. A notable consequence of this change is that 5 of the 21 clinics included in those results no longer qualify for an estimate, while 1 clinic which hadn't qualified now does. The 17 qualifying clinics generate these pooled ATT estimates and bootstrapped confidence intervals. Although these estimates and intervals suggest somewhat more harmful effects of IPT than those in Figure 2.3, most of the values of δ suggest the same directionality and significance (or lack thereof) of ATT estimates.

Figure A.2: Pooled ATTs, by δ, using $M_D = 6$ months



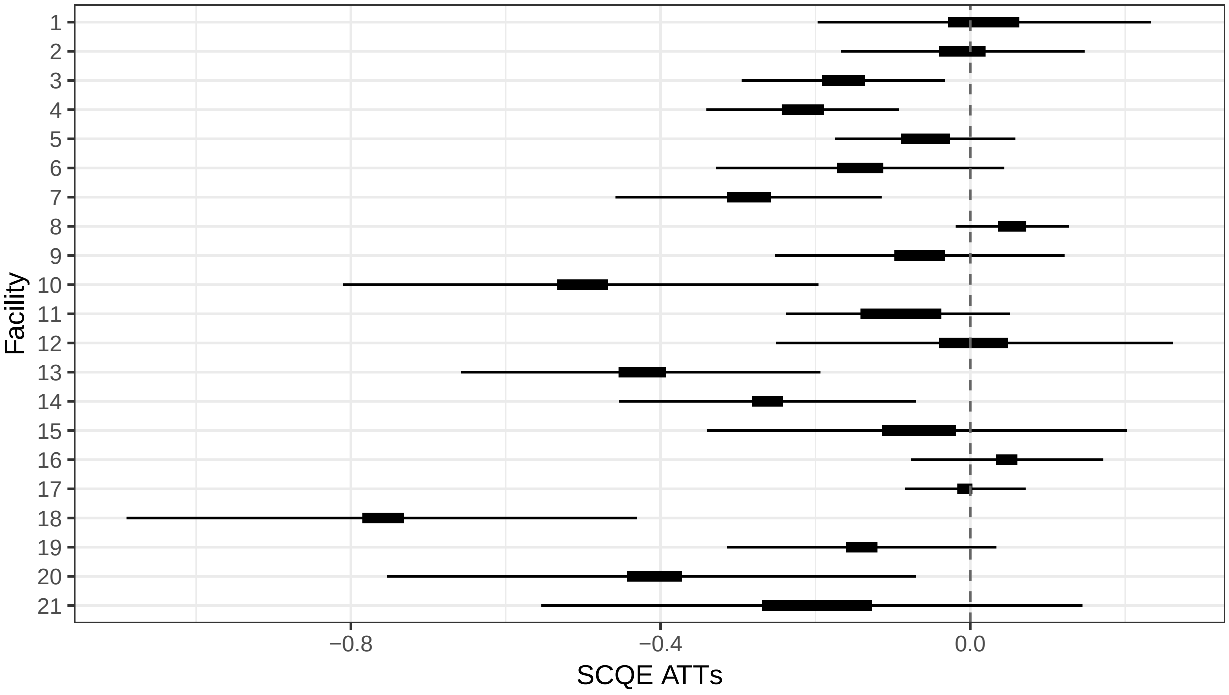
Note: Pooled ATT estimates as a function of δ, similar to Figure 2.3 in the main text. It shows those results if the period after a patient's first visit within which IPT administration is considered treatment is limited to 6 months instead of a full year. A notable consequence of this limit is that 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 somewhat more beneficial effects of IPT than those in Figure 2.3, most of the values of δ suggest the same directionality and significance (or lack thereof) of ATT estimates.

Table A.1: Suggestions for yearly trends by data subset

	Clinics without IPT use:	Clinics with IPT use	
Trend type:		All:	Only ATT qualifiers:
Linear	-0.0006 [-0.0031, 0.002]	-0.004 [-0.007, -0.0011]	-0.0150 [-0.0241, -0.006]
Exponential	0.9825 [0.8920, 1.0821]	0.9197 [0.8788, 0.9626]	0.8874 [0.8249, 0.9546]

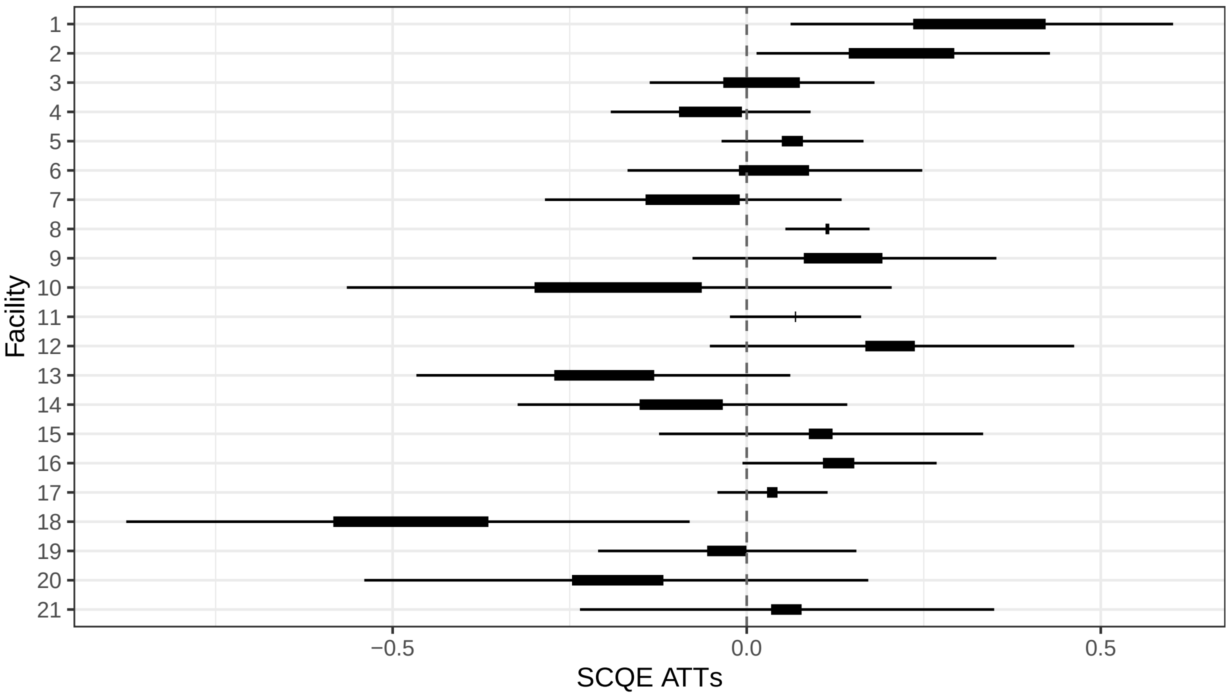
Note: Trends estimated from data to inform the range of δ can be learned from (i) clinics that never introduced IPT, or (ii) the pre-IPT periods in clinics that did. Further, (iii) the latter may be narrowed to only include clinics that qualify for inclusion in the analysis according to our criteria (Section 2.3.2). 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.3: 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.4: 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.2: Linear δ results

Yearly Linear Trend	Effective Used δ	Pooled ATT	Bootstrapped CI 2.5%	 97.5%
0.010	0.028	-0.150	-0.279	-0.037
0.009	0.026	-0.138	-0.266	-0.025
0.008	0.023	-0.126	-0.255	-0.013
0.007	0.020	-0.114	-0.243	-0.000
0.006	0.017	-0.103	-0.233	0.011
0.005	0.014	-0.091	-0.220	0.024
0.004	0.011	-0.079	-0.209	0.035
0.003	0.009	-0.067	-0.197	0.049
0.002	0.006	-0.055	-0.185	0.060
0.001	0.003	-0.044	-0.174	0.072
0.000	0.000	-0.032	-0.163	0.085
-0.001	-0.003	-0.020	-0.151	0.097
-0.002	-0.006	-0.008	-0.139	0.109
-0.003	-0.009	0.003	-0.127	0.120
-0.004	-0.011	0.015	-0.116	0.133
-0.005	-0.014	0.027	-0.105	0.146
-0.006	-0.017	0.039	-0.093	0.158
-0.007	-0.020	0.050	-0.082	0.170
-0.008	-0.023	0.062	-0.069	0.183
-0.009	-0.026	0.074	-0.058	0.195
-0.010	-0.028	0.086	-0.047	0.208
-0.011	-0.031	0.098	-0.036	0.220
-0.012	-0.034	0.109	-0.025	0.232
-0.013	-0.037	0.121	-0.014	0.245
-0.014	-0.040	0.133	-0.002	0.259
-0.015	-0.043	0.145	0.010	0.270

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 2.3. The data-assisted choice of δ under this linear model ranged in yearly trends from -0.005 to -0.001, while the “domain knowledge” estimate was 0.005 to 0.01.

Table A.3: Exponential δ results

Yearly	Effective	Pooled	Bootstrapped CI	
Decay Rate	Used δ	ATT	2.5%	97.5%
1.07	0.027	-0.144	-0.293	-0.011
1.06	0.023	-0.127	-0.273	0.004
1.05	0.019	-0.110	-0.253	0.016
1.04	0.015	-0.094	-0.235	0.030
1.03	0.011	-0.078	-0.216	0.044
1.02	0.007	-0.062	-0.197	0.058
1.01	0.004	-0.047	-0.180	0.071
1.00	0.000	-0.032	-0.162	0.083
0.99	-0.004	-0.017	-0.146	0.098
0.98	-0.007	-0.002	-0.129	0.112
0.97	-0.011	0.012	-0.113	0.124
0.96	-0.014	0.026	-0.098	0.137
0.95	-0.017	0.040	-0.081	0.151
0.94	-0.021	0.053	-0.068	0.164
0.93	-0.024	0.066	-0.053	0.177
0.92	-0.027	0.079	-0.039	0.189
0.91	-0.030	0.092	-0.026	0.202
0.90	-0.033	0.104	-0.014	0.215
0.89	-0.036	0.117	-0.001	0.228
0.88	-0.039	0.129	0.012	0.239

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 2.3. The data-assisted choice of δ under this exponential decay model ranged in yearly decay rates from 0.89 to 0.97.

A.1.1 Estimating ATTs using covariate adjustment methods

We first used the OLS imputation estimator, as described in Section 2.4.1. The models for both treated and untreated patients used the same covariates as the ATE estimate — age, sex, date of initial visit, WHO stage, and facility fixed effects — and gave an estimated ATT of -15.31pp.

The second imputation estimator used extreme gradient boosting as implemented in the xgboost package in R. The predictors used were similar, but because xgboost only accepts numeric variables, we chose to remove the facility label as a predictor for this model rather than one-hot encode it and overwhelm the model with 20 binary predictors. Separately for the treated and untreated patients, we ran repeated five-fold cross validation to tune the number of successive trees, as well as three tree booster parameters - eta, the learning rate, gamma, the minimum loss reduction to partition a node, and the maximum tree depth. For both patient groups, the algorithm chose to fit many weak, small trees, settling on 200 trees with a maximum depth of 2, an eta of 0.3, and a gamma of 0.5. Effectively, the resultant models converged within 30 trees rather than using all 200, and the imputed ATT was estimated as -13.97pp.

Finally, the third method we used to estimate an ATT was 1-to-1 matching of treated patients to untreated patients. Three matching estimates were generated:

Matching specification	ATT
Matching on age, sex, date of initial visit, WHO stage, and facility	-14.97pp, t = -17.6
Same parameters as above. Exact matching on sex, WHO stage, and facility. Age and date of initial visit matched to within 10 years and 1 year, respectively. 44 matches dropped due to these restrictions	-14.84pp, t = -17.9
Same parameters, exact matching, and ranges as above, but without including facility	-14.42pp, t = -19.0

A.2 Other δ distribution methods

In Section 4.3.1, we presented two ways of accommodating prior distributions on δ. Below, we suggest two additional approaches to doing so. Although both have merits, our choice of the two discussed in the main text was based on their coverage of most properties we imagine practitioners will find desirable; staunch bayesians will likely be satisfied by the full-bayesian approach, while those who want to maintain the reporting style of SCQE will find that in the weighted-SCQE approach. The two below present results in the form of distributions on the CATE (either plotted or described using its mean, median, 95% CI, etc.) as opposed to using the standard SCQE plot.

A.2.1 Binomial approach

The first additional option, which we call the binomial approach, can be used when the outcome is binary. It is similar to the full-bayesian approach in that we draw many samples from each term in our CATE estimand (equation 4.2) and observe the distribution of the CATEs calculated across draws. As in the full-bayesian approach, we draw the δ term from the expert-supplied prior, and the other terms are drawn in a way that represents their sampling uncertainty. It differs, however, in that it does not require that practitioners supply priors for each of the non-δ terms. Instead, each sample is drawn from the binomial distribution representing the observed proportion (either an outcome rate or a treatment rate). For example, if we had 100 units in the low-use cohort and 15 of them experienced the outcome, each draw of $\mathbb{E}[Y|Z=0]$ would be from the distribution $Bin(100, 0.15)/100$.

The results from this approach will often be very similar to those from the full-bayesian approach. Anecdotally, we have found the two will be most alike when the observed sample size is large (causing priors to be overwhelmed by data in the full-bayesian approach) or when the proportion is not very close to 0 or 1.[1] If a proportion is exactly 0 or exactly 1

[1] Boundary areas may increase the difficulty of obtaining precise estimates when using an uninformative prior like $Beta(1, 1)$, as the choice of which "uninformative" prior to use matters more when we find ourselves

we suggest using the full-bayesian approach instead, as this binomial sampling procedure will not introduce any uncertainty at all (Winkler et al., 2002). In summary, in binary outcome scenarios without proportions that are exactly 0 or 1, the binomial approach is a good alternative for practitioners repelled or confused by the full-bayesian approach's requirements for priors.

A.2.2 Compound distribution approach

The second additional option, which we call the compound distribution approach, utilizes a different derivation of the SCQE variance in combination with the weighting procedure of the weighted-SCQE (Section 4.3.1). This approach is motivated by a desire take each estimated CATE distribution for individual δ_i values and combine them into a single estimated CATE distribution marginalized across δ (weighted according to the prior on δ).[2] To do so requires a well-defined estimated distribution for CATE, which we have not yet provided. The AR confidence set procedure we introduced in Section 3.2 does not provide a distribution, only an (asymmetrical, possibly non-contiguous) set for a given confidence level which we cannot transform into a probability distribution function (pdf). However, if we derive an estimator for the variance of CATE and make a standard normality of errors assumption, the resultant CATE follows Gaussian distribution with that derived variance. We can then combine these distributions across δ to form a compound Gaussian probability distribution.

Formally, this compound CATE distribution's pdf is defined through the density of each δ-specific normally-distributed CATE, weighted by the prior probability of the δ generating that distribution. That is, if $p(\delta)$ is the pdf representing the δ prior, and $f_\delta(x)$ is the pdf of the Gaussian CATE estimator generated from a given δ_i, $\mathcal{N}\Big(\widehat{\text{CATE}}_i,\ \widehat{\text{Var}}(\widehat{\text{CATE}}_i)\Big)$, then

close to the zero-numerator problem (Winkler et al., 2002). The prior will carry far more weight and require more data to be overruled in these cases (Dixon et al., 2005), not because the prior itself is stronger, but because we will likely care more about small differences in our posterior when they lie near boundary areas.

[2] Each CATE distribution represents statistical uncertainty around the CATE estimate for a fixed δ value. That is, after removing all identification uncertainty by only considering a single δ value, this distribution represents only sampling uncertainty from our finite sample size.

the pdf of the compound CATE distribution is given by

$$f(x) = \int_\delta p(\delta)\, f_\delta(x)\, d\delta$$

In practice, we would only approximate this distribution by summing over many choices of δ rather than integrating over its full range:

$$f(x) \approx \sum_{\delta_i} w_i f_{\delta_i}(x)$$

Where w_i is the weight associated with δ_i, as defined in equation 4.5. We can then plot the distribution or calculate a median, CI, etc.

For our variance estimator, we use the formula presented in equation 2.6 that required a spherical error assumption. In Appendix A.3, we derive a count-based version of that estimator, as well as a variance estimator (and it's count-based version) for the one-cohort scenario. Because these variances, unlike an AR confidence set, are not robust to weak instruments, the compound distribution approach should not be used when the change in treatment rates is small.[3]

A.3 CATE variance derivation

A.3.1 Two-cohort variance

We begin with the two-cohort scenario. As detailed in Section 2.2.2, Wooldridge (2009) provides an estimator for the the standard error of the IV estimate (which is equivalent to the SCQE CATE estimate when using the δ-adjusted outcome $\tilde{Y}$), under an assumption of

[3] We may define a sufficiently strong instrument through an F-statistic cutoff of 10, though this particular value may be anti-conservative (Lee et al., 2020).

homoskedasticity. Given an instrument Z, a model

$$\tilde{Y} = \beta_0 + \beta_{\text{SCQE}} D + u$$

with fitted residuals

$$\hat{u}_i = \tilde{Y}_i - \hat{\beta}_0 - \hat{\beta}_{\text{SCQE}} D_i = Y_i - \delta Z_i - \hat{\beta}_0 - \hat{\beta}_{\text{SCQE}} D_i,$$

and assuming $\text{Var}(u) = \sigma^2 I_n$, the asymptotic variance estimator of the SCQE estimate is:

$$\widehat{\text{Var}}\left(\hat{\beta}_{\text{SCQE}}\right) = \frac{\hat{\sigma}_u^2}{N\,\hat{\rho}_{D,Z}^2\,\hat{\sigma}_D^2} \tag{A.1}$$

If we have unit level data, we can estimate each quantity in this variance directly from the sample, plugging in $\frac{1}{N-2}\sum \hat{u}_i^2$ for $\hat{\sigma}_u^2$; the sample variance of D for $\hat{\sigma}_D^2$; and the R^2 from an OLS regression of D on Z for $\hat{\rho}_{D,Z}^2$.

Without unit level data however, we can only estimate the variance if the Y, Z, and D are all binary. Doing so requires the same 8 counts we used in Section 3.3, and we derive this count-based re-expression of A.1 quantity by quantity below.

Count-based variance derivation

Starting with $\hat{\sigma}_u^2$,

$$\begin{aligned}\hat{\sigma}_u^2 &= \widehat{\text{Var}}(Y - \delta Z - \hat{\beta}_0 - \hat{\beta}_{\text{SCQE}} D) \\ &= \frac{1}{N-2}\sum\nolimits_i (Y_i - \delta Z_i - \hat{\beta}_0 - \hat{\beta}_{\text{SCQE}} D_i)^2\end{aligned}$$

The term in this sum can take on one of 8 values, depending on the value of Y_i, Z_i, and D_i, and we know how often each take place from the supplied counts (and differences between them, e.g. $N_{Y=0,D=0,Z=1} = N_{D=0,Z=1} - N_{Y=1,D=0,Z=1}$). Therefore, we can split up the sum as follows:

$$\begin{aligned}
\hat{\sigma}_u^2 &= \frac{1}{N-2}\sum\nolimits_i (Y_i - \delta Z_i - \hat{\beta}_0 - \hat{\beta}_{\text{SCQE}} D_i)^2 \\
&= \frac{1}{N-2}\Big[N_{Y=1,D=1,Z=1}(1 - \delta - \hat{\beta}_0 - \hat{\beta}_{\text{SCQE}})^2 \\
&\qquad + N_{Y=1,D=0,Z=1}(1 - \delta - \hat{\beta}_0)^2 \\
&\qquad + N_{Y=0,D=1,Z=1}(-\delta - \hat{\beta}_0 - \hat{\beta}_{\text{SCQE}})^2 \\
&\qquad + N_{Y=0,D=0,Z=1}(-\delta - \hat{\beta}_0)^2 \\
&\qquad + N_{Y=1,D=1,Z=0}(1 - \hat{\beta}_0 - \hat{\beta}_{\text{SCQE}})^2 \\
&\qquad + N_{Y=1,D=0,Z=0}(1 - \hat{\beta}_0)^2 \\
&\qquad + N_{Y=0,D=1,Z=0}(-\hat{\beta}_0 - \hat{\beta}_{\text{SCQE}})^2 \\
&\qquad + N_{Y=0,D=0,Z=0}(-\hat{\beta}_0)^2\Big]
\end{aligned}$$

And we estimate $\hat{\beta}_{\text{SCQE}}$ and $\hat{\beta}_0$ with:

$$\begin{aligned}
\hat{\beta}_{\text{SCQE}} &= \frac{\mathbb{E}[\tilde{Y}|Z=1] - \mathbb{E}[\tilde{Y}|Z=0]}{\mathbb{E}[\tilde{D}|Z=1] - \mathbb{E}[\tilde{D}|Z=0]} \\
\hat{\beta}_0 &= \mathbb{E}[\tilde{Y}] - \hat{\beta}_{\text{SCQE}}\bar{D}
\end{aligned}$$

Moving on to $\hat{\rho}_{D,Z}^2$, we want to calculate the square of the sample Pearson correlation coefficient, $\hat{r}_{D,Z}$:

$$\hat{r}_{D,Z}^2 = \frac{\left(\sum_i (D_i - \bar{D})(Z_i - \bar{Z})\right)^2}{\sum_i (D_i - \bar{D})^2 \ \sum_i (Z_i - \bar{Z})^2}$$

The sum in the numerator breaks down in the same way:

$$\begin{aligned}
\Big(\sum\nolimits_i (D_i - \bar{D})(Z_i - \bar{Z})\Big)^2 = \Big[&N_{D=1,Z=1}(1 - \bar{D})(1 - \bar{Z}) \\
&+ N_{D=1,Z=0}(1 - \bar{D})(-\bar{Z}) \\
&+ N_{D=0,Z=1}(-\bar{D})(1 - \bar{Z}) \\
&+ N_{D=0,Z=0}(-\bar{D})(-\bar{Z})\Big]^2
\end{aligned}$$

while both quantities in the denominator simplify:

$$\begin{aligned}\sum\nolimits_i (D_i - \bar{D})^2 &= \Big[N_{D=1}(1-\bar{D})^2 + N_{D=0}(-\bar{D})^2\Big] \\ &= N\Big[\bar{D}(1-\bar{D})^2 + (1-\bar{D})(\bar{D})^2\Big] \\ &= N\Big[\bar{D}(1-\bar{D})\left((1-\bar{D})+\bar{D}\right)\Big] \\ &= N\bar{D}(1-\bar{D})\end{aligned}$$

and, similarly,

$$\sum\nolimits_i (Z_i - \bar{Z})^2 = N\bar{Z}(1-\bar{Z})$$

The last quantity, $\hat{\sigma}_D^2$, simplifies to $\frac{1}{N-1}\overline{D}(1-\overline{D})$.

Putting all of these quantities together produces a count-based variance estimator for the SCQE estimate:

$$\begin{aligned}\widehat{\text{Var}}\left(\hat{\beta}_{\text{SCQE}}\right) =& \frac{N-1}{N-2}\bar{Z}(1-\bar{Z}) \\ &\times\Big[N_{Y=1,D=1,Z=1}(1-\delta-\hat{\beta}_0-\hat{\beta}_{\text{SCQE}})^2 \\ &\quad + N_{Y=1,D=0,Z=1}(1-\delta-\hat{\beta}_0)^2 \\ &\quad + N_{Y=0,D=1,Z=1}(-\delta-\hat{\beta}_0-\hat{\beta}_{\text{SCQE}})^2 \\ &\quad + N_{Y=0,D=0,Z=1}(-\delta-\hat{\beta}_0)^2 \\ &\quad + N_{Y=1,D=1,Z=0}(1-\hat{\beta}_0-\hat{\beta}_{\text{SCQE}})^2 \\ &\quad + N_{Y=1,D=0,Z=0}(1-\hat{\beta}_0)^2 \\ &\quad + N_{Y=0,D=1,Z=0}(-\hat{\beta}_0-\hat{\beta}_{\text{SCQE}})^2 \\ &\quad + N_{Y=0,D=0,Z=0}(-\hat{\beta}_0)^2\Big] \\ &\times\Big[N_{D=1,Z=1}(1-\bar{D})(1-\bar{Z}) \\ &\quad + N_{D=1,Z=0}(1-\bar{D})(-\bar{Z}) \\ &\quad + N_{D=0,Z=1}(-\bar{D})(1-\bar{Z}) \\ &\quad + N_{D=0,Z=0}(-\bar{D})(-\bar{Z})\Big]^{-2}\end{aligned}$$

A.3.2 One-cohort variance

In the one-cohort setting, we begin with the ratio estimator for the ATT (equation 4.4), but instead of using the Fieller approach (which we effectively do in Section 4.2.1 by transforming the one-cohort inference question into a faux-two-cohort question), we use the delta method to derive the approximate variance of that ratio (Oehlert, 1992). That approach gives us:

$$\begin{aligned}
\widehat{\text{Var}}(\widehat{\text{ATT}}_{1C}) &= \widehat{\text{Var}}\left(\frac{\overline{Y} - \mu_{Y0}}{\pi}\right) \\
&\approx \frac{(\overline{Y} - \mu_{Y0})^2}{\pi^2}\left[\frac{\widehat{\text{Var}}(\overline{Y} - \mu_{Y0})}{(\overline{Y} - \mu_{Y0})^2} + \frac{\widehat{\text{Var}}(\pi)}{\pi^2} - 2\frac{\widehat{\text{Cov}}(\overline{Y} - \mu_{Y0}, \pi)}{(\overline{Y} - \mu_{Y0})\pi}\right]
\end{aligned}$$

We are able to simplify this expression by focusing its final numerator:

$$\begin{aligned}
\widehat{\text{Cov}}(\overline{Y} - \mu_{Y0}, \pi) &= \widehat{\text{Cov}}(\overline{Y}, \pi) \\
&= \widehat{\text{Cov}}\left(\frac{1}{N}\sum(Y_i(0) + \widehat{\text{ATT}}_{1C} \cdot D_i)\ ,\ \pi\right) \\
&= \widehat{\text{Cov}}\left(\frac{1}{N}\sum Y_i(0)\ ,\ \pi\right) + \widehat{\text{Cov}}\left(\frac{1}{N}\sum \widehat{\text{ATT}}_{1C} \cdot D_i\ ,\ \pi\right) \\
&= \widehat{\text{Cov}}(\overline{Y(0)}, \pi) + \widehat{\text{Cov}}(\widehat{\text{ATT}}_{1C} \cdot \pi, \pi) \\
&= \widehat{\text{ATT}}_{1C} \cdot \widehat{\text{Var}}(\pi),
\end{aligned}$$

in which we assume the covariance of $\overline{Y(0)}$ and π is 0.

We can then continue:

$$\begin{aligned}
\widehat{\mathrm{Var}}(\mathrm{ATT}_{1C}) &\approx \frac{(\overline{Y}-\mu_{Y0})^2}{\pi^2}\left[\frac{\widehat{\mathrm{Var}}(\overline{Y}-\mu_{Y0})}{(\overline{Y}-\mu_{Y0})^2}+\frac{\widehat{\mathrm{Var}}(\pi)}{\pi^2}-2\frac{\widehat{\mathrm{Cov}}(\overline{Y}-\mu_{Y0},\pi)}{(\overline{Y}-\mu_{Y0})\pi}\right]\\
&= \frac{(\overline{Y}-\mu_{Y0})^2}{\pi^2}\left[\frac{\widehat{\mathrm{Var}}(\overline{Y})}{(\overline{Y}-\mu_{Y0})^2}+\frac{\widehat{\mathrm{Var}}(\pi)}{\pi^2}-2\frac{\widehat{\mathrm{ATT}_{1C}}\cdot\widehat{\mathrm{Var}}(\pi)}{(\overline{Y}-\mu_{Y0})\pi}\right]\\
&= \frac{(\overline{Y}-\mu_{Y0})^2}{\pi^2}\left[\frac{\widehat{\mathrm{Var}}(\overline{Y})}{(\overline{Y}-\mu_{Y0})^2}+\frac{\widehat{\mathrm{Var}}(\pi)}{\pi^2}-2\frac{\widehat{\mathrm{Var}}(\pi)}{\pi^2}\right]\\
&= \frac{(\overline{Y}-\mu_{Y0})^2}{\pi^2}\left[\frac{\widehat{\mathrm{Var}}(\overline{Y})}{(\overline{Y}-\mu_{Y0})^2}-\frac{\widehat{\mathrm{Var}}(\pi)}{\pi^2}\right]\\
&= \frac{\widehat{\mathrm{Var}}(\overline{Y})}{\pi^2}-\frac{(\overline{Y}-\mu_{Y0})^2\widehat{\mathrm{Var}}(\pi)}{\pi^4}
\end{aligned}$$

This variance estimator can also be expressed using only counts when the outcome is binary:

$$\begin{aligned}
\widehat{\mathrm{Var}}(\mathrm{ATT}_{1C}) &\approx \frac{\widehat{\mathrm{Var}}(\overline{Y})}{\pi^2}-\frac{(\overline{Y}-\mu_{Y0})^2\widehat{\mathrm{Var}}(\pi)}{\pi^4}\\
&= \frac{\widehat{\mathrm{Var}}(Y)}{N^2\pi^2}-\frac{(\overline{Y}-\mu_{Y0})^2\widehat{\mathrm{Var}}(D)}{N^2\pi^4}\\
&= \frac{\overline{Y}(1-\overline{Y})}{N(N-1)\pi^2}-\frac{(\overline{Y}-\mu_{Y0})^2\overline{D}(1-\overline{D})}{N(N-1)\pi^4}\\
&= \frac{\overline{Y}(1-\overline{Y})}{N(N-1)\overline{D}^2}-\frac{(\overline{Y}-\mu_{Y0})^2(1-\overline{D})}{N(N-1)\overline{D}^3}
\end{aligned}$$

Bibliography

Akolo, C., Adetifa, I., Shepperd, S., and Volmink, J. (2010). Treatment of latent tuberculosis infection in hiv infected persons. *Cochrane database of systematic reviews*.

Alam, N., Hobbelink, E. L., van Tienhoven, A.-J., van de Ven, P. M., Jansma, E. P., and Nanayakkara, P. W. (2014). The impact of the use of the early warning score (ews) on patient outcomes: a systematic review. *Resuscitation*, 85(5):587–594.

Anderson, T. W., Rubin, H., et al. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical statistics*, 20(1):46–63.

Andrews, I., Stock, J. H., and Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11:727–753.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

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

Assebe, L. F., Reda, H. L., Wubeneh, A. D., Lerebo, W. T., and Lambert, S. M. (2015). The effect of isoniazid preventive therapy on incidence of tuberculosis among hiv-infected clients under pre-art care, jimma, ethiopia: a retrospective cohort study. *BMC Public Health*, 15(1):346.

Baiocchi, M., Cheng, J., and Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in medicine*, 33(13):2297–2340.

Bell, B., Blundell, R., and Van Reenen, J. (1999). Getting the unemployed back to work: the role of targeted wage subsidies. *International tax and public finance*, 6(3):339–360.

Block, B. L., Martin, T. M., Boscardin, W. J., Covinsky, K. E., Mourad, M., Hu, L. L., and Smith, A. K. (2021). Variation in covid-19 mortality across 117 us hospitals in high-and low-burden settings. *Journal of Hospital Medicine*.

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

Broderick, T., Giordano, R., and Meager, R. (2020). An automatic finite-sample robustness metric: Can dropping a little data change conclusions?

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

Buist, M. D., Moore, G. E., Bernard, S. A., Waxman, B. P., Anderson, J. N., and Nguyen, T. V. (2002). Effects of a medical emergency team on reduction of incidence of and mortality from unexpected cardiac arrests in hospital: preliminary study. *Bmj*, 324(7334):387–390.

Cain, L. E., Cole, S. R., Greenland, S., Brown, T. T., Chmiel, J. S., Kingsley, L., and Detels, R. (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. C. and Trivedi, P. K. (2005). *Microeconometrics: methods and applications*. Cambridge university press.

Cathey, L. (2020). Timeline: Tracking trump alongside scientific developments on hydroxychloroquine. *ABC News*, 2020-08-08.

Cavalcanti, A. B., Zampieri, F. G., Rosa, R. G., Azevedo, L. C., Veiga, V. C., Avezum, A., Damiani, L. P., Marcadenti, A., Kawano-Dourado, L., Lisboa, T., et al. (2020). Hydroxychloroquine with or without azithromycin in mild-to-moderate covid-19. *New England Journal of Medicine*, 383(21):2041–2052.

Chan, P. S., Jain, R., Nallmothu, B. K., Berg, R. A., and Sasson, C. (2010). Rapid response teams: a systematic review and meta-analysis. *Archives of internal medicine*, 170(1):18–26.

Chen, T. and Guestrin, C. (2016). Xgboost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*.

Cinelli, C. and Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):39–67.

Cinelli, C. and Hazlett, C. (2021). An omitted variable bias framework for sensitivity analysis of instrumental variables. Unpublished Manuscript.

Conley, T. G., Hansen, C. B., and Rossi, P. E. (2012). Plausibly exogenous. *Review of Economics and Statistics*, 94(1):260–272.

Dixon, P. M., Ellison, A. M., and Gotelli, N. J. (2005). Improving the precision of estimates of the frequency of rare events. *Ecology*, 86(5):1114–1123.

Egami, N. and Hartman, E. (2020). Elements of external validity: Framework, design, and analysis. *SSRN*.

Escobar, G. J., Liu, V. X., Schuler, A., Lawson, B., Greene, J. D., and Kipnis, P. (2020). Automated identification of adults at risk for in-hospital clinical deterioration. *New England Journal of Medicine*, 383(20):1951–1960.

FDA Press Release (2020). *Coronavirus (COVID-19) Update: FDA Revokes Emergency Use Authorization for Chloroquine and Hydroxychloroquine.* [Press Release] https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-fda-revokes-emergency-use-authorization-chloroquine-and.

Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 16(2):175–185.

Fox, W., Ellard, G. A., and Mitchison, D. A. (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.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58(1):21–29.

Geleris, J., Sun, Y., Platt, J., Zucker, J., Baldwin, M., Hripcsak, G., Labella, A., Manson, D. K., Kubin, C., Barr, R. G., et al. (2020). Observational study of hydroxychloroquine in hospitalized patients with covid-19. *New England Journal of Medicine*.

Genest, C., Zidek, J. V., et al. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–135.

Geremew, D., Endalamaw, A., Negash, M., Eshetie, S., and Tessema, B. (2019). The protective effect of isoniazid preventive therapy on tuberculosis incidence among hiv positive patients receiving art in ethiopian settings: a meta-analysis. *BMC infectious diseases*, 19(1):405.

Gokhale, M., Buse, J. B., DeFilippo Mack, C., Jonsson Funk, M., Lund, J., Simpson, R. J., and Stürmer, T. (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, C. and Rapson, D. S. (2018). Regression discontinuity in time: Considerations for empirical applications. *Annual Review of Resource Economics*, 10:533–552.

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

Hazlett, C., Maokola, W., and Wulf, D. A. (2020). Inference without randomization or ignorability: A stability controlled quasi-experiment on the prevention of tuberculosis. *Statistics in Medicine*, 39:4169–4186.

Howell, M. D., Ngo, L., Folcarelli, P., Yang, J., Mottley, L., Marcantonio, E. R., Sands, K. E., Moorman, D., and Aronson, M. D. (2012). Sustained effectiveness of a primary-team–based rapid response system. *Critical care medicine*, 40(9):2562.

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

Ji, X., Small, D. S., Leonard, C. E., and Hennessy, S. (2017). The trend-in-trend research design for causal inference. *Epidemiology (Cambridge, Mass.)*, 28(4):529.

Johnston, K., Gustafson, P., Levy, A., and Grootendorst, P. (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.

Johnston, L. G., McLaughlin, K. R., El Rhilani, H., Latifi, A., Toufik, A., Bennani, A., Alami, K., Elomari, B., and Handcock, M. S. (2015). Estimating the size of hidden populations using respondent-driven sampling data: case examples from morocco. *Epidemiology (Cambridge, Mass.)*, 26(6):846.

Knox, D., Yamamoto, T., Baum, M. A., and Berinsky, A. J. (2019). Design, identification, and sensitivity analysis for patient preference trials. *Journal of the American Statistical Association*, pages 1–27.

Lee, D. L., McCrary, J., Moreira, M. J., and Porter, J. (2020). Valid t-ratio inference for iv. *arXiv preprint arXiv:2010.05058*.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies*, 76(3):1071–1102.

Lee, D. S. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of economic literature*, 48(2):281–355.

Lyons, P. G., Edelson, D. P., and Churpek, M. M. (2018). Rapid response systems. *Resuscitation*, 128:191–197.

Mack, C. D., Brookhart, M. A., Glynn, R. J., Meyer, A. M., Carpenter, W. R., Sandler, R. S., and Stürmer, T. (2015). Comparative effectiveness of oxaliplatin vs. 5-flourouricil in older adults: an instrumental variable analysis. *Epidemiology (Cambridge, Mass.)*, 26(5):690.

Magagnoli, J., Narendran, S., Pereira, F., Cummings, T., Hardin, J. W., Sutton, S. S., and Ambati, J. (2020). Outcomes of hydroxychloroquine usage in united states veterans hospitalized with covid-19. *medRxiv*.

Maharaj, R., Raffaele, I., and Wendon, J. (2015). Rapid response systems: a systematic review and meta-analysis. *Critical Care*, 19(1):1–15.

Mahévas, M., Tran, V.-T., Roumier, M., Chabrol, A., Paule, R., Guillaud, C., Fois, E., Lepeule, R., Szwebel, T.-A., Lescure, F.-X., et al. (2020). Clinical efficacy of hydroxychloroquine in patients with covid-19 pneumonia who require oxygen: observational comparative study using routine care data. *Bmj*, 369.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323.

Manski, C. F. (2009). *Identification for prediction and decision.* Harvard University Press.

MERIT Study Investigators (2005). Introduction of the medical emergency team (met) system: a cluster-randomised controlled trial. *The Lancet*, 365(9477):2091–2097.

Meyer, B. D. (1995). Natural and quasi-experiments in economics. *Journal of business & economic statistics*, 13(2):151–161.

National AIDS Control Program (2009). *National Guidelines for the Management of HIV and AIDS.* 3 edition.

Neyman, J. 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.

Oakley, J. (2020). *SHELF: Tools to Support the Sheffield Elicitation Framework.* R package version 1.7.0.

Oehlert, G. W. (1992). A note on the delta method. *The American Statistician*, 46(1):27–29.

O'Hagan, A., Buck, C., Daneshkhah, A., Eiser, J., Garthwaite, P., Jenkinson, D., Oakley, J., and Rakow, T. (2006). *Uncertain Judgements: Eliciting Expert Probabilities.* John Wiley & Sons.

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

Pearl, J. (2009). *Causality: Models, Reasoning and Inference.* Cambridge University Press, 2nd edition.

RECOVERY Collaborative Group (2020). Effect of hydroxychloroquine in hospitalized patients with covid-19. *New England Journal of Medicine*, 383(21):2030–2040.

Rosenberg, E. S., Dufort, E. M., Udo, T., Wilberschied, L. A., Kumar, J., Tesoriero, J., Weinberg, P., Kirkwood, J., Muse, A., DeHovitz, J., et al. (2020). Association of treatment with hydroxychloroquine or azithromycin with in-hospital mortality in patients with covid-19 in new york state. *Jama*.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Rubin, D. B. (1990). [on the application of probability theory to agricultural experiments. essay on principles. section 9.] comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, pages 472–480.

Rubin, E. J., Harrington, D. P., Hogan, J. W., Gatsonis, C., Baden, L. R., and Hamel, M. B. (2020). The urgency of care during the covid-19 pandemic—learning as we go.

Sabasaba, A., Mwambi, H., Somi, G., Ramadhani, A., and Mahande, M. J. (2019). Effect of isoniazid preventive therapy on tuberculosis incidence and associated risk factors among hiv infected adults in tanzania: a retrospective cohort study. *BMC infectious diseases*, 19(1):62.

Self, W. H., Semler, M. W., Leither, L. M., Casey, J. D., Angus, D. C., Brower, R. G., Chang, S. Y., Collins, S. P., Eppensteiner, J. C., Filbin, M. R., et al. (2020). Effect of hydroxychloroquine on clinical status at 14 days in hospitalized patients with covid-19: a randomized clinical trial. *JAMA*, 324(21):2165–2176.

Shetty, K. D., Vogt, W. B., and Bhattacharya, J. (2009). Hormone replacement therapy and cardiovascular health in the united states. *Medical Care*, pages 600–605.

Small, D. S. and Rosenbaum, P. R. (2008). War and wages: the strength of instrumental variables and their sensitivity to unobserved biases. *Journal of the American Statistical Association*, 103(483):924–933.

Smieja, M., Marchetti, C., Cook, D., and Smaill, F. M. (1999). Isoniazid for preventing tuberculosis in non-HIV infected persons. *Cochrane Database of Systematic Reviews*.

Stefan, A. M., Evans, N. J., and Wagenmakers, E.-J. (2020). Practical challenges and methodological flexibility in prior elicitation. *Psychological Methods*.

Stock, J. H. and Yogo, M. (2002). Testing for weak instruments in linear iv regression.

Streeter, A. J., Lin, N. X., Crathorne, L., Haasova, M., Hyde, C., Melzer, D., and Henley, W. E. (2017). Adjusting for unmeasured confounding in nonrandomized longitudinal studies: a methodological review. *Journal of clinical epidemiology*, 87:23–34.

Temesgen, B., Kibret, G. D., Alamirew, N. M., Melkamu, M. W., Hibstie, Y. T., Petrucka, P., and Alebel, A. (2019). Incidence and predictors of tuberculosis among hiv-positive adults on antiretroviral therapy at debre markos referral hospital, northwest ethiopia: a retrospective record review. *BMC public health*, 19(1):1566.

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

Uddin, M. J., Groenwold, R. H., Ali, M. S., de Boer, A., Roes, K. C., Chowdhury, M. A., and Klungel, O. H. (2016). Methods to control for unmeasured confounding in pharmacoepidemiology: an overview. *International journal of clinical pharmacy*, 38(3):714–723.

van Kippersluis, H. and Rietveld, C. A. (2018). Beyond plausibly exogenous. *The Econometrics Journal*, 21(3):316–331.

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

Vansteelandt, S., Bekaert, M., and Claeskens, G. (2012). On model selection and model misspecification in causal inference. *Statistical methods in medical research*, 21(1):7–30.

Wagner, A. K., Soumerai, S. B., Zhang, F., and Ross-Degnan, D. (2002). Segmented regression analysis of interrupted time series studies in medication use research. *Journal of clinical pharmacy and therapeutics*, 27(4):299–309.

WHO (2008). Isoniazid preventive therapy. In *Implementing the WHO Stop TB Strategy: A Handbook for National Tuberculosis Control Programmes*, chapter 7.

Winkler, R. L. (1967). The assessment of prior distributions in bayesian analysis. *Journal of the American Statistical Association*, 62(319):776–800.

Winkler, R. L., Smith, J. E., and Fryback, D. G. (2002). The role of informative priors in zero-numerator problems: being conservative versus being candid. *The American Statistician*, 56(1):1–4.

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

Zhang, X., Faries, D. E., Li, H., Stamey, J. D., and Imbens, G. W. (2018). Addressing unmeasured confounding in comparative observational research. *Pharmacoepidemiology and drug safety*, 27(4):373–382.