Program Evaluation with Remotely Sensed Outcomes

Ashesh Rambachan
MIT

Rahul Singh
Harvard

Davide Viviano*
Harvard

First draft: November 2024
This draft: October 2025

Abstract

Economists often estimate treatment effects in experiments using remotely sensed variables (RSVs), e.g., satellite images or mobile phone activity, in place of directly measured economic outcomes. A common practice is to use an observational sample to train a predictor of the economic outcome from the RSV, and then use these predictions as the outcomes in the experiment. We show that this method is biased whenever the RSV is a *post*-outcome variable, meaning that variation in the economic outcome causes variation in the RSV. For example, changes in poverty or environmental quality cause changes in satellite images, but not vice versa. As our main result, we nonparametrically identify the treatment effect by formalizing the intuition underlying common practice: the conditional distribution of the RSV given the outcome and treatment is stable across samples. Our identifying formula reveals that efficient inference requires predictions of three quantities from the RSV—the outcome, treatment, and sample indicator—whereas common practice only predicts the outcome. Valid inference does not require any rate conditions on RSV predictions, justifying the use of complex deep learning algorithms with unknown statistical properties. We reanalyze the effect of an anti-poverty program in India using satellite images.

Keywords: Causal inference, data fusion, experiments, satellite images, machine learning.

*Email: asheshr@mit.edu, rahul_singh@fas.harvard.edu, dviviano@fas.harvard.edu. We thank Isaiah Andrews, Josh Angrist, Arun Chandrasekhar, Raj Chetty, Kevin Chen, Ben Deaner, Melissa Dell, Seema Jayachandran, Namrata Kala, Sylvia Klosin, Ben Olken, Pritham Raja, Jonathan Roth, and Jesse Shapiro, as well as audiences at Bocconi, CEMFI, CREST/Sciences Po, Harvard/MIT, the NBER Summer Institute, the Online Causal Inference Seminar, Opportunity Insights, Stanford, University of Bologna, University of Chicago, University of Southern California, University of Toronto, and Yale for helpful discussions. Haya Alsharif, Peter Chen, Marvin Lob, Leonard Mushunje, Miriam Nelson, Kevin Wang, and Sammi Zhu provided excellent research assistance. Davide Viviano gratefully acknowledges funding from the Harvard Griffin Fund in Economics and NSF Grant SES 2447088. Replication code is available here. We provide the `remoteoutcome` R package to implement our method.

1 Introduction

While traditional program evaluations rely on surveys to measure impact, important economic outcomes such as living standards and environmental quality may be costly or infeasible to collect at scale. As a consequence, researchers increasingly estimate treatment effects on economic outcomes using remotely sensed variables (RSVs). Examples include night lights as a measure of local economic activity (Chen and Nordhaus, 2011; Henderson et al., 2012; Asher et al., 2021), roofing material as a measure of housing quality (Marx et al., 2019; Michaels et al., 2021; Huang et al., 2021), mobile phone transactions as a measure of local wealth or consumption (Blumenstock et al., 2015; Aiken et al., 2025), and satellite images as a measure of pollution (Currie et al., 2023), deforestation (Jayachandran et al., 2017; Assuncao et al., 2023), fires (Jack et al., 2025; Balboni et al., 2024), flooding (Chen et al., 2017; Patel, 2024), and local poverty (Jean et al., 2016).[1] Our research question is how researchers should rigorously estimate treatment effects from remotely sensed outcomes.

A recurring empirical practice appears in about 50% of papers in general interest economics journals from 2015-2024 that use remotely sensed outcomes.[2] This common practice predicts the economic outcome from the RSV and then uses the predicted outcome in lieu of a true outcome measurement in an experimental sample. Researchers often form such predictions using an auxiliary, observational sample collected in some other context, which contains the RSV and linked outcome measurements. The predictor is typically complex, e.g., a deep learning algorithm with unknown statistical properties.

We show that this intuitive method can produce arbitrarily biased treatment effect estimates—including flipped signs—when the RSV is a *post*-outcome variable. Using the predicted outcome in lieu of the true outcome implicitly uses the RSV as a surrogate that mediates between the treatment and the outcome (Prentice, 1989; Athey et al., 2025). However, in many empirical applications using RSVs, the opposite is more plausible: the treatment affects the outcome, and both may affect the RSV. The bias is fundamentally due to this

[1] Figure A.1 illustrates the rapid rise of papers published in economics journals and general interest science journals using RSVs. Of those published in economics journals, we find that 90% use high-dimensional satellite images as RSVs, and 40% use the RSVs as the main outcome in their empirical analyses. See also Burke et al. (2021) and Jack and Walker (2023) for recent reviews on the empirical uses of RSVs.

[2] We survey AEA journals, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics*, and *Review of Economic Studies*. The other 50% of papers use a similar logic, but without an explicit formula for data combination.

reversal; it is present even without machine learning.

As an example, consider a binary outcome indicating whether a plot of land has been burned (Balboni et al., 2024; Jack et al., 2025) and an RSV summarizing the color saturation in a satellite image. Fires cause changes in satellite images, but not vice versa; the color saturation of satellite images is a post-outcome variable. Common practice predicts fires in the experimental sample using a machine learning algorithm trained on labeled satellite images from an observational sample, and then computes the difference in predicted outcomes between treated and control units in the experimental sample. Because the RSV is post-outcome, this method's estimand combines two quantities: the desired effect of the treatment on the outcome, and the correlation between the RSV and the outcome. In the extreme case where the RSV fails to predict the outcome—when an ideal method should report infinite standard errors—this popular method will instead report a precise estimate of zero, regardless of the true treatment effect. In general, this method may flip the sign of the true treatment effect.

Our main contribution is a novel formula to nonparametrically identify treatment effects using RSVs by combining (i) an experimental sample where the outcome is missing, and (ii) an observational sample in which the treatment is non-randomized and possibly missing. Our key assumption formalizes the logic underlying the examples above: the conditional distribution of the RSV given the outcome and the treatment is *stable* across both samples. Consequently, the relationship between the RSV and the outcome can be learned from the observational sample and transported to the experimental sample. If the treatment is missing in the observational sample, then we need an additional assumption to restore identification: the treatment only affects the RSV through the outcome.

Our main identifying assumptions are jointly testable and lend themselves to simple diagnostics, allowing researchers to assess their plausibility in applications. We propose a diagnostic to evaluate whether an RSV is sufficiently relevant for an economic outcome to justify its use in program evaluation. Our framework also naturally extends to quasi-experiments, including instrumental variables and difference-in-differences designs, which we present in Appendix J.

Our secondary contribution is to characterize the representation of the RSV that achieves valid, precise, and robust $n^{-1/2}$ inference on the treatment effect. Because modern remote sensing typically involves unstructured data and complex machine learning, we derive valid inference without rate conditions and without complexity restrictions on RSV-based predictions. Valid inference only requires that (i) a learned RSV representation has some limit, and (ii) the limit predicts the outcome of interest, for which we provide a diagnostic. This

enables researchers to use complex deep learning algorithms and still conduct valid inference on treatment effects.

Specifically, we build on our main identification result to establish the connection between modern remote sensing and classical conditional moments (Chamberlain, 1987; Newey, 1993). This allows us to derive an expression for a simple representation of the RSV that maximizes precision from the conditional moments we derived. We find that three predictions are necessary for efficient downstream causal inference based on RSVs: predictions of the outcome, the treatment, and the sample indicator given the RSV. By contrast, common practice only predicts the outcome given the RSV. Moreover, $n^{-1/2}$ inference remains valid even if all three predictions are misspecified, which is a strong form of robustness. We provide the `remoteoutcome` R package to implement our method.

Finally, we conduct a semi-synthetic exercise, calibrated to an existing field experiment in India, which we merge with existing satellite images. Following Muralidharan et al. (2016, 2023), we study the effect of Smartcards, a biometrically authenticated payments infrastructure, on village-level poverty measures. We use the geographic coordinates of each village to extract nighttime luminosity and high-dimensional, pre-trained embeddings of satellite images. Despite using outcomes for only half of the units, we recover the treatment effect of interest with the same precision as an unbiased regression method that has access to outcomes for all units. By contrast, common practice can have positive or negative bias for the treatment effect. Our method enables large savings in survey costs and therefore opens up possibilities for program evaluation in cost-constrained environments.

1.1 Related Work

Our framework differs from existing approaches to auxiliary variables and data combination along two key dimensions: (i) the causal direction between the auxiliary variable and the outcome, and (ii) the data requirements.

In the surrogacy framework (Prentice, 1989; Athey et al., 2025; Kallus and Mao, 2024), the auxiliary variable (surrogate) is a mediator between the treatment and the outcome, whereas the RSV is a *post*-outcome variable in our framework. We show that misusing a post-outcome RSV as a surrogate leads to arbitrary bias in treatment effects. The negative control literature extends the surrogacy framework to address unobserved confounding (Ghassami et al., 2022; Imbens et al., 2024), yet such extensions face the same limitation. See Remark 2 for details.

Compared to the vast literature on data combination (e.g., Cross and Manski, 2002; Ridder and Moffitt, 2007; Bareinboim and Pearl, 2016; D'Haultfœuille et al., 2025) and nonclassical measurement error (e.g., Chen et al., 2011; Schennach, 2020), we place what appears to be a different key assumption. Several influential works handle measurement error in moment condition models by using auxiliary data and assuming that the conditional distribution of the variable of interest, given the imperfect measurement, is stable across samples (Chen et al., 2005, 2008; Graham et al., 2016), akin to the surrogacy framework. Our key identifying assumption is the opposite: the conditional distribution of the imperfect measurement, given the variable of interest, is stable across samples. This is natural when the imperfect measurement is caused by the outcome (as with satellite images of poverty or environmental outcomes) rather than merely correlated with it. Other works focus on mismeasured or missing covariates in a sample where the outcome is observed (e.g., Fan et al., 2014; Battaglia et al., 2024; D'Haultfœuille et al., 2025, 2024). By contrast, in our setting, the outcome is missing yet a post-outcome variable is present. Both of these distinctions lead to a novel identifying formula.

The main difference between our RSV framework and the prediction-powered inference (PPI) framework (Angelopoulos et al., 2023; Lu et al., 2025; Kluger et al., 2025) is also along these lines: the PPI framework uses machine learning predictions as surrogates (Ji et al., 2025). Another difference concerns data availability. In our terminology, the PPI approach would require the researcher to observe the treatment, outcome, and RSV for a random subsample of experimental units. The data requirements in other works are similar to those in the PPI literature (Fong and Tyler, 2021; Allon et al., 2023; Gordon et al., 2023; Egami et al., 2024; Carlson and Dell, 2025). By contrast, we allow the researcher to observe no outcomes for any experimental units, enabling inference in settings where PPI-style techniques cannot be applied.

Our results do not require a correctly specified generative model of how treatments and outcomes affect RSVs, which may be prone to misspecification when the RSV is a satellite image or some other type of unstructured data. Several previous works propose methods based on generative modeling that do require correct specification (Gentzkow et al., 2019; Alix-Garcia and Millimet, 2023; Proctor et al., 2023). Similarly, methods for causal inference on outcomes that are latent concepts require a correctly specified generative model (Egami et al., 2022; Knox et al., 2022; Stoetzer et al., 2025).

2 Model and Assumptions

2.1 Goal: Identification using Remotely Sensed Outcomes

The researcher observes units in two samples, indicated by the variable $S \in \{e, o\}$: an experimental sample ($S = e$) and an observational sample ($S = o$).

Within the experimental sample ($S = e$), we observe pre-treatment covariates $X \in \mathcal{X}$ and a binary treatment $D \in \{0,1\}$. However, the outcome $Y \in \mathcal{Y}$ is missing.[3] In its place, we have access to a remotely sensed outcome variable $R \in \mathcal{R}$. We typically think of R as high-dimensional (e.g., unstructured data such as satellite images), but it could be low-dimensional (e.g., the output of some pre-trained machine learning algorithm). The researcher would like to use the remotely sensed variable (RSV) as an imperfect measurement of the outcome in the experimental sample, without placing parametric assumptions on their relationship.

The causal parameter of interest is the effect of the treatment D on the outcome Y in the experimental sample. Though the outcome Y is unobserved in the experimental sample, we may still define its potential outcomes $Y(d)$ and our parameter of interest.[4]

Definition 1 (Causal parameter)**.** The average treatment effect (ATE) in the experimental sample is $\theta := \mu(1) - \mu(0)$, where $\mu(d) := \mathbb{E}\{Y(d) \mid S = e\}$.

Without further assumptions, point identification for this causal parameter is impossible (Horowitz and Manski, 1995). Even if R can predict Y with great accuracy, if the prediction is at all imperfect, then an assumption is necessary. Therefore, we place additional structure on the problem, inspired by recent empirical work in environmental and development economics.

A popular practice in environmental and development economics is to use an auxiliary dataset of outcomes and RSVs, e.g., of labeled satellite images. We refer to the auxiliary dataset as the observational sample ($S = o$). For these units, we observe baseline covariates X, the outcome Y, and the remotely sensed variable R. We may or may not observe the treatment D. If we do, we denote it by $D \in \{0,1\}$ and refer to this scenario as having "complete" cases. If we do not, or if treatment is deterministic in the observational study, we set $D = 0$ for all units in the observational study and refer to this latter scenario as having "incomplete"

[3] For ease of exposition, we focus on the case where the outcome Y is completely missing in the experimental sample. Remark 4 gives the extension where Y is only partially missing in the experimental sample.

[4] This definition of potential outcomes has no spillovers. We discuss spillovers in Appendix I.1.

Table 1: Summary of the data environment.

Sample S	Covariate X	Treatment D	Outcome Y	Remotely sensed R
Experimental	✓	✓	Missing	✓
Observational: Complete	✓	✓	✓	✓
Observational: Incomplete	✓	Missing or deterministic	✓	✓

Notes: Here, ✓ denotes that the variable is observed. When treatment is missing or deterministic in the observational sample, we encode it as $D=0$ in the observational sample.

cases.[5] As we discuss below, incomplete cases will require stronger assumptions.

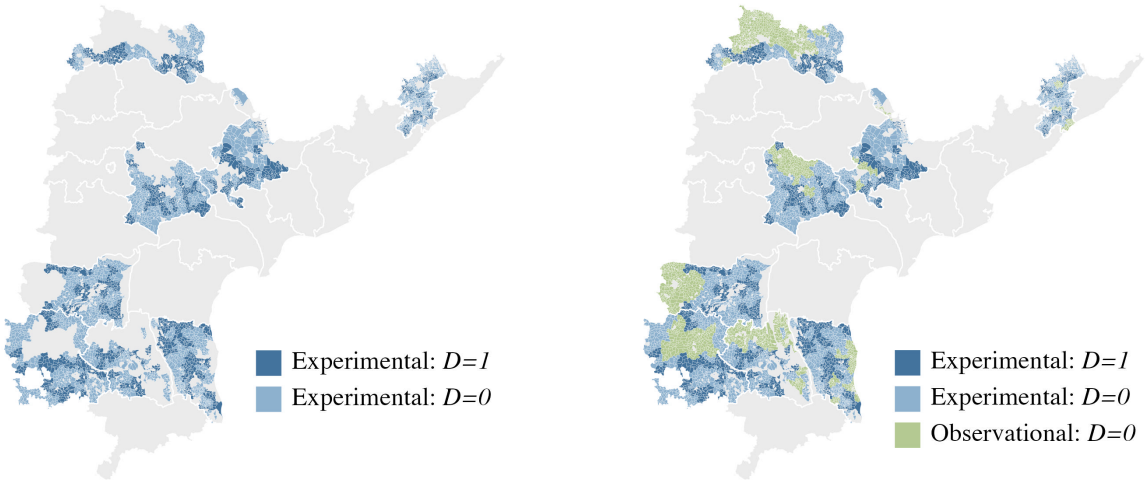
Table 1 summarizes the setting. Each unit is characterized by the random vector $(S,X,D,Y(0),Y(1),R)$, which we assume to be independent and identically distributed.[6] For units in the experimental sample ($S=e$), we observe (X,D,R); for units in the observational sample ($S=o$), we observe (X,D,Y,R) in complete cases or (X,Y,R) in incomplete cases.

Example 1 (Environmental impacts)**.** Consider a randomized experiment that offers cash payments to households in order to incentivize environmental conservation (i.e., "payments for ecosystem services" or PES). Access to PES contracts is often randomized at the village level. We would like to measure whether access to PES contracts D reduces harmful environmental behaviors Y, such as deforestation (Jayachandran et al., 2017) or crop burning (Jack et al., 2025). In a separate observational sample, we link satellite images R to direct measurements Y of deforestation or crop burning (e.g., Hansen et al., 2013; Walker et al., 2022). While it is expensive to hire surveyors to record measurements of tree cover or crop management practices in rural areas, it is inexpensive to collect satellite images. We investigate how to combine these data sources and thereby identify the effect of the PES contracts in the experimental sample. ▲

Example 2 (Household poverty)**.** Consider a randomized experiment evaluating an anti-poverty program, such as an unconditional cash transfer (Egger et al., 2022) or biometrically authenticated payment (Muralidharan et al., 2023). Treatment is often randomized at the village level. We would like to study the effect of the anti-poverty program D on village-level poverty Y. In a separate observational sample, we link satellite images R to census statistics

[5] Incomplete cases in the observational sample may refer to three scenarios. First, the treatment status may be missing. Second, the treatment status may be present, and all observational units are untreated, hence $D=0$. Third, the treatment status may be present, and all observational units are treated. Relabeling treatment values gives $1-D=0$.

[6] Independence is not used to derive our main identification argument.



(a) Experimental units only.

(b) Auxiliary sample: Observational units.

Figure 1: We illustrate the two samples that we will use to evaluate an anti-poverty program in Andhra Pradesh, India (Muralidharan et al., 2023). With experimental units alone and completely missing outcomes, point identification is impossible. Therefore we introduce an auxiliary sample of observational units. See Section 5 for further details.

on village-level poverty Y. It is well documented that poverty can be predicted from satellite images, with some error (e.g., Jean et al., 2016; Rolf et al., 2021). For example, Huang et al. (2021) use deep learning methods to predict household wealth in Kenya from roof quality. While it is expensive to collect poverty measures through in-person surveys in the experimental sample, it is inexpensive to collect satellite images. We identify the effect of the anti-poverty program in the experimental sample.

Figure 1 illustrates an example of incomplete cases, using data from an evaluation of an anti-poverty program in India, where we apply our method in Section 5. ▲

2.2 Main Assumption: Stability

We formalize this causal setting via three assumptions. Our identifying assumptions allow (X,Y,R) to be discrete or continuous. For readability, we slightly abuse notation: for a random variable W, we use the symbol $f_W(\cdot\,|\,...)$ to refer to its (conditional) probability mass function if W is discrete, or its (conditional) density if W is continuous.[7]

Assumption 1 (Experimental unconfoundedness)**.** Suppose the following:

[7]Formally, $f_W(\cdot\,|\,...)$ is our symbol for the Radon-Nikodym derivative.

i. SUTVA: $Y = DY(1) + (1-D)Y(0)$ almost surely.

ii. Randomization: $D \perp\!\!\!\perp \{Y(0), Y(1)\} \mid X, S = e$.

iii. Overlap: $\Pr(D = 1 \mid X, S = e)$ is bounded away from zero and one almost surely.

In many empirical applications involving RSVs, such as Examples 1 and 2, Assumption 1 is satisfied by design: experimental units are chosen as aggregates without spillovers, e.g., villages, and the treatment is randomly assigned for these experimental units. We defer to Appendix J the study of quasi-experimental settings, such as instrumental variables and difference-in-differences.

Under Assumption 1, if we were to observe the outcome in the experimental sample, the ATE could be identified using standard arguments. However, the outcome is not observed in the experiment; instead, we have an RSV.

We resolve this measurement issue by leveraging the observational sample. Intuitively, the idea is to learn the relationship between the RSV and the outcome of interest in the observational sample and to "transport" it to the experimental sample.

Assumption 2 (Stability of the remotely sensed variable)**.** Suppose the following:

i. Stability: $S \perp\!\!\!\perp R \mid X, D, Y$.

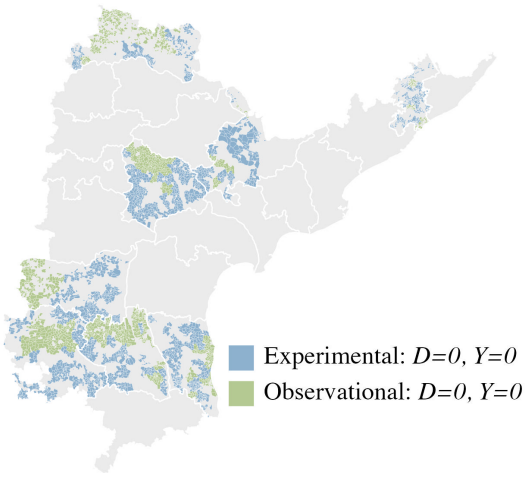
ii. Common support: for some outcome support $\mathcal{Y}$, $\Pr(Y \in \mathcal{Y} \mid S = e, X) = 1$ almost surely, and $f_Y(y \mid S = o, X)$ is bounded away from zero almost surely for all $y \in \mathcal{Y}$.

iii. Coverage: $f_R(r \mid S, X, D)$ is bounded away from zero almost surely for all $r \in \mathcal{R}$.

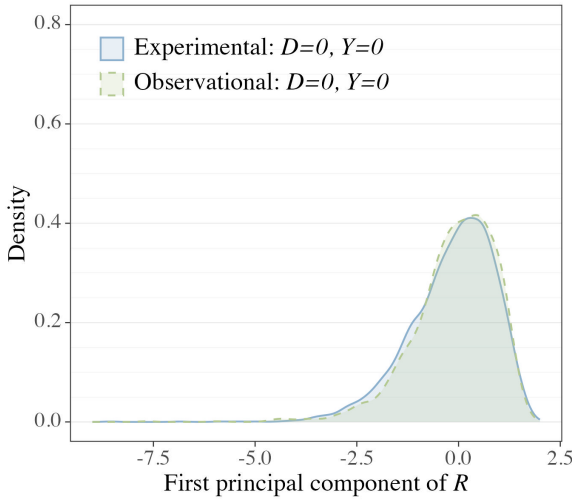
iv. Two samples: $\Pr(S = e|X)$ is bounded away from zero and one almost surely.

Assumption 2(i) is the main assumption of our framework: the conditional distribution of the remotely sensed variable R, given (X, D, Y), is *stable* across the experimental and observational samples. This allows us to "transport" the measurement error distribution from the observational sample to the experimental sample. Importantly, this condition does not require stability of the underlying treatment effects, which may differ across samples. See Remark 2 for comparisons between Assumption 2(i) in the RSV model and assumptions in alternative models.

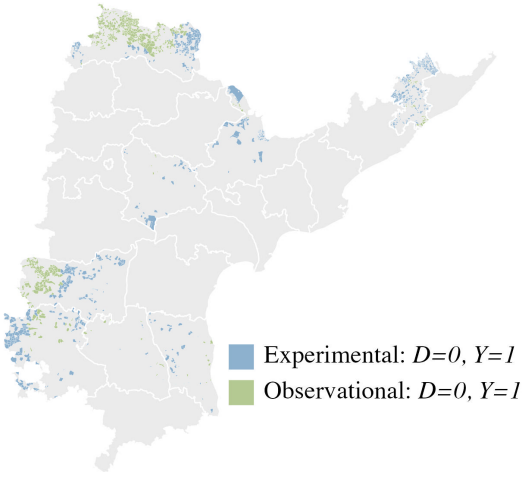
Returning to our two leading examples, Assumption 2(i) requires that the conditional distribution of tree cover pixels R, given environmental outcomes Y and interventions D (as well as other pre-treatment covariates), is stable across the experimental and observational



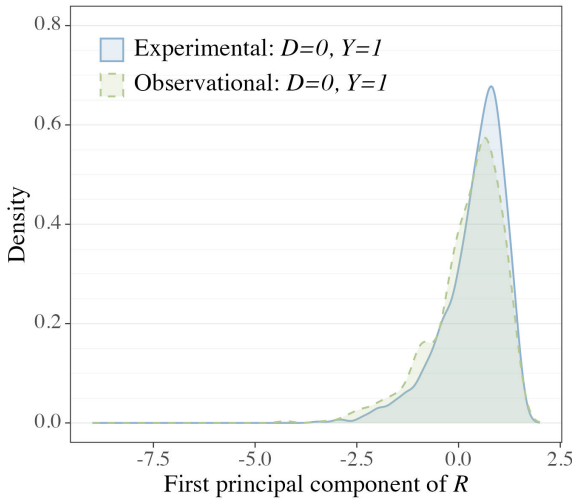
(a) Units with $D=0$ and $Y=0$.



(b) Densities of $R \mid S$, $D=0$, $Y=0$.



(c) Units with $D=0$ and $Y=1$.



(d) Densities of $R \mid S$, $D=0$, $Y=1$.

Figure 2: Our main assumption (Assumption 2(i)) is plausible in real data. We compare $f_R(R \mid S=e, D=0, Y=0)$ with $f_R(R \mid S=o, D=0, Y=0)$ in Figure 2b, and $f_R(R \mid S=e, D=0, Y=1)$ with $f_R(R \mid S=o, D=0, Y=1)$ in Figure 2d, using data from the Smartcard experiment conducted by Muralidharan et al. (2016) that we analyze in Section 5. Because the satellite image $R \in \mathbb{R}^{4000}$ is high-dimensional, we visualize the density of its standardized first principal component on the right hand side, for units highlighted on the left hand side.

samples. Analogously, it requires that the conditional distribution of the satellite image R, given village-level poverty Y and the anti-poverty program D (as well as other pre-treatment covariates), is stable across the experimental and observational samples.

Our main assumption is empirically plausible, as illustrated by Figure 2. We use data from an anti-poverty program in India, where outcomes are observed. If Assumption 2 holds, then

the conditional densities of the RSV given the outcome and treatment should be the same across the experimental and observational samples. They appear to coincide in this empirical setting.[8]

The remaining aspects of Assumption 2 are weak regularity conditions. Assumption 2(ii) requires that the outcome in the observational sample has a common (or larger) support than the outcome in the experiment. Assumption 2(iii) ensures that the RSV distribution does not degenerate for any stratum. Assumption 2(iv) requires that we observe some data from both the experimental and observational samples.

Assumptions 1 and 2 imply identification when the observational sample has complete cases. When the observational sample has incomplete cases, we require a further assumption. In other words, if the treatment is missing or deterministic in the observational sample, then a further restriction is necessary for point identification.

Assumption 3 (Observational completeness)**.** Suppose that *either* condition holds:

i. Complete cases: $\Pr(D=1 \mid S=o,X)$ is bounded away from zero and one almost surely;

ii. No direct effect: $D \perp\!\!\!\perp R \mid X,Y$.

Assumption 3 imposes only one of two conditions.

Assumption 3(i) implies that we have access to complete cases, i.e., some observations of (X,D,Y,R) where D has variation. Within the observational sample, the treatment is observed and varies, although it does not need to be randomized and may suffer from *unobserved confounding*. Whenever we have complete cases, under Assumption 3(i), no further causal assumptions are needed. In particular, the treatment D may have a direct effect on the remotely sensed variable R. In Example 1, this would allow the environmental program to affect satellite images both indirectly, i.e., via crop burning, and directly, e.g., via visible investments in farm equipment.

If Assumption 3(i) is violated, then we have no complete cases, i.e., no observations of (X,D,Y,R) where D is variable. Without joint observations of the outcome and treatment, a further restriction is needed. Assumption 3(ii) fills this gap, requiring that the treatment

[8]When $X=\varnothing$, Assumption 2 imposes four equalities: $f_R(R \mid S=e,D=d,Y=y)=f_R(R \mid S=o,D=d,Y=y)$ for $d \in \{0,1\}$ and $y \in \{0,1\}$. Each equality can be evaluated with a diagnostic plot if outcome data are available. For example, using the experimental and observational units satisfying $D=0$ and $Y=0$ in Figure 2a, we can visualize whether the density of $R \mid S=e,D=0,Y=0$ aligns with the density of $R \mid S=o,D=0,Y=0$ in Figure 2b. Since R is high-dimensional, we simplify the visualization by comparing the densities of its first principal component.

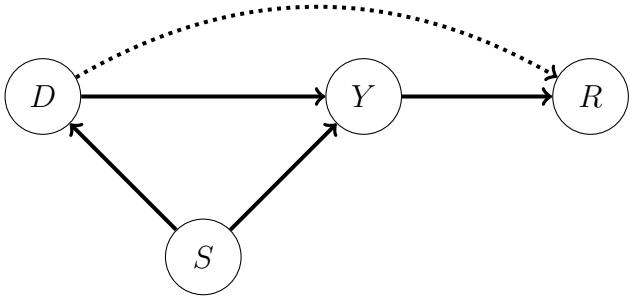


Figure 3: Causal graph for remotely sensed variables under Assumptions 3(i) versus 3(ii). Complete cases allow the dotted line. Assumption 2 rules out the line from S to R.

D affects the remotely sensed variable R only via its effect on the outcome Y. In Example 1, we may be comfortable assuming that the PES contract has no direct effect on the specific infrared band used to measure charred soil in satellite images. Assumption 3(ii) may also become more plausible when Y is a vector of outcomes. Several outcomes may approximate all mechanisms through which the treatment affects the RSV. For exposition, we focus on scalar outcomes in the main text and we generalize to vector outcomes in Appendix E.

Together, Assumption 2(i) and Assumption 3(ii) imply that $(S,D) \perp\!\!\!\perp R \mid X,Y$. In the next section, we will show that Assumptions 2(i) and 3(ii) are jointly testable, even when no outcome is observed from the experimental sample (Remark 3).

Figure 3 illustrates our identifying assumptions as a causal graph. The treatment affects the outcome, which in turn affects the RSV. By Assumption 2, the sample indicator does not change the conditional distribution of the RSV given the treatment and outcome. Depending on which version of Assumption 3 is imposed, the treatment may have a direct effect on the RSV, as illustrated by the dotted line. Table A.1 summarizes the implications of our main assumptions.

3 Main Result: Identification

In this causal setting, a commonly used procedure may lead to causal estimates with arbitrary bias. Motivated by this negative result, we prove a positive one: we nonparametrically identify the causal parameter by combining the experimental and observational samples differently.

To streamline notation, we initially focus on the setting where Assumption 3(ii) holds, then return to the setting where Assumption 3(i) holds at the end of this section. For exposition, we will also assume that the outcome is binary, i.e., $\mathcal{Y}=\{0,1\}$. Finally, we assume that the support of R is at least as large as the support of Y (it may be binary, discrete, or continuous).

Appendices E and F extend our results to discrete and continuous outcomes, respectively.

3.1 Current Practice may have Positive or Negative Bias

In empirical research, it is common to use RSVs in two steps: (i) researchers train a predictor of the outcome Y from the remotely sensed variable R in the observational sample, and (ii) the predictor is applied to the experimental sample, where its predictions are used as surrogate outcomes to estimate treatment effects. While intuitive, this empirical strategy can lead to arbitrary bias for the ATE in the experimental sample.

Suppose there are no pre-treatment covariates for simplicity. The widely used two-step estimation procedure implicitly targets the estimand $\widetilde{\theta} = \widetilde{\mu}(1) - \widetilde{\mu}(0)$, where $\widetilde{\mu}(d) := \mathbb{E}\{\mathbb{E}(Y \mid R, S = o) \mid D = d, S = e\}$ for $d \in \{0,1\}$. Within this expression, the first step estimates the conditional expectation function $\mathbb{E}(Y \mid R = r, S = o)$. The second step evaluates and averages this function over the treated and untreated subgroups in the experimental sample.

As a starting point, if the RSV fails to predict the outcome, i.e., if $\mathbb{E}(Y|R,S=o) = \mathbb{E}(Y|S=o)$, then the implicit target $\widetilde{\theta}$ is zero regardless of the true treatment effect. Common practice would return a precise estimate of zero, even though the RSV provides no information about treatment effects in this case.

More importantly, even if the RSV does predict the outcome, the implicit target $\widetilde{\theta}$ can incur bias with arbitrary sign for the ATE in the experimental sample.

Proposition 1 (Bias of common practice)**.** Suppose Assumptions 1, 2, and 3(ii) hold with $X = \varnothing$. Suppose $\Pr(D = 1 \mid S = o) = 0$ and $S \perp\!\!\!\perp (Y,R) \mid D$. Then the following hold.

i. The bias is $\widetilde{\theta} - \theta = \mu(1) \int \{w(r) - 1\} f_R(r|Y = 1, S = e) \mathrm{d}r$, where $w(r) = \frac{\Pr\{Y(0)=1|S=e\} f_R(r|D=1)}{\Pr\{Y(1)=1|S=e\} f_R(r|D=0)}$.

ii. There exists a data-generating process satisfying the above restrictions with $\widetilde{\theta} - \theta > 0$, and a different data-generating process satisfying the above restrictions with $\widetilde{\theta} - \theta < 0$.

Proposition 1 derives the bias of current empirical practice, which uses the RSV as a surrogate outcome. Under Assumptions 2 and 3(ii), the conditional distribution of the RSV $f_R(R \mid Y, S = o)$ is stable across the experimental and observational samples. Existing practice attempts to transport the predictions $f_Y(Y \mid R, S = o)$ into the experimental sample. By Bayes' rule, this induces a bias due to differences in the marginal distributions of the outcomes and RSVs across the samples. Importantly, the bias of common practice arises even if units are randomly allocated into the experimental and observational samples.[9]

[9] See Appendix B for a formal verification of this claim.

Remark 1 (Bias summary)**.** To gain insight into the source of bias, we consider three nested cases. (i) In the extreme "irrelevance" case, the RSV fails to predict the outcome. We would like an estimator whose variance is infinity. However, $\mathbb{E}(Y|R,S=o)$ is a constant function of R, so $\widetilde{\theta}=0$; the common practice always reports zero. (ii) In the simple linear case, something similar occurs. Suppose the data-generating process satisfies $\mathbb{E}(Y|R,S=o)=\tilde{\beta}_0+\tilde{\beta}R$ and $\mathbb{E}(R|Y,D,S=e)=\beta_0+\beta Y$. Combining these expressions, $\widetilde{\theta}=\tilde{\beta}\beta\theta_0$; the common practice suffers from attenuation bias (see Appendix B for further details). (iii) In the general case, where the relationships among variables are nonlinear, Proposition 1 shows that the bias of common practice can be positive or negative.

Remark 2 (Comparison to the surrogacy framework)**.** Proposition 1 provides a direct comparison to the surrogacy framework. Within the surrogacy framework, the implicit target of empirical practice $\widetilde{\theta}$ recovers the causal parameter θ if $(D,S)\perp\!\!\!\perp Y \mid R,X$, i.e., if surrogacy and surrogate compatibility are satisfied (Prentice, 1989; Athey et al., 2025). These assumptions contrast our Assumptions 2 and 3. The surrogacy assumptions state that the surrogate fully mediates the effect of the treatment on the outcome, and so the surrogate is *pre*-outcome. Assumptions 2 and 3 imply the opposite: the outcome mediates the effect of the treatment on the RSV, either partially (Assumption 3(i)) or fully (Assumption 3(ii)); the RSV is *post*-outcome. Figure A.2 illustrates the difference via a causal graph. Below, we argue that the RSV model is more plausible in environmental and development applications with satellite images.

The surrogacy framework can be viewed as a version of the moment condition model with auxiliary data studied by e.g., Chen et al. (2005, 2008); Graham et al. (2016). In our terminology, those works require that the conditional distribution of the outcome given the RSV and treatment is stable across samples. By contrast, we require that the conditional distribution of the RSV given the outcome and treatment is stable across samples. Figure 2 suggests that our assumption is plausible in a real development application with satellite images.

The surrogacy model, and related models which lead to the estimand $\widetilde{\theta}$, have been extended to include negative controls (e.g., Ghassami et al., 2022; Imbens et al., 2024). As generalizations of the surrogacy model, they suffer from the same drawback formalized in Proposition 1.

Similar to the RSV framework, the single negative control framework (Park et al., 2024) has one auxiliary variable. The frameworks have two key differences. Unlike the single negative control model, we allow the treatment D to affect both the outcome Y and the auxiliary variable R when Assumption 3(i) holds. When no complete cases exist, i.e. the setting covered by Assumption 3(ii), the single negative control model provides no guidance of how to proceed.

3.2 Example: Current Practice may Underestimate Environmental Impacts

We revisit an experiment conducted by Jack et al. (2025) that studied whether payments for ecosystem services (PES) contracts can incentivize farmers to reduce crop burning. We measure the average treatment effect θ of being offered a PES contract $D \in \{0,1\}$ on the likelihood that a farmer burns their fields $Y \in \{0,1\}$. Here, $Y = 1$ means not burning, so a positive θ means environmental benefit.

It is costly to measure whether the crop residue on a particular field has been burned, requiring a surveyor to make frequent visits to rural fields. Therefore, it is natural to turn to a remotely sensed variable R for the outcome Y. To construct such a remotely sensed variable, Jack et al. (2025) link surveyor-collected measurements of crop burning to satellite-based spectral indices and then train a supervised learning algorithm to predict whether these fields have been burned. Here, $R \in \{0,1\}$ is a classifier for whether a field has not been burned, which applies a threshold rule to a machine learning prediction of the probability that a field has not been burned.[10]

Our causal assumptions are plausible in this setting. Because burning crops would alter satellite images, but altering satellite images would not burn crops, R is a *post*-outcome variable rather than a *pre*-outcome variable. Since access to PES contracts was randomized at the village level, Assumption 1 is satisfied by design. Since the authors conducted randomized spot checks, surveying the outcome Y for randomly selected fields, Assumption 2 is also satisfied by design if we define the observational sample $S = o$ as the fields in which these random spot checks occurred. Finally, since specific infrared bands are used to form R, it is plausible that PES contracts affect these specific infrared bands only via crop burning, so Assumption 3(ii) is reasonable. Figure G.1 illustrates the two samples used in our re-analysis.

We implement the common empirical practice: we use the RSV as a surrogate for crop burning. Column (1) of Table 2 estimates $\widetilde{\theta}$. Offering any PES contract to farmers appears to reduce crop burning by 7.9%.[11]

However, by Proposition 1, $\widetilde{\theta}$ is typically biased for θ. To quantify the magnitude of its

[10] Jack et al. (2025) construct two binary RSVs for crop burning by applying two alternative threshold rules to the estimated probability a field has not been burned. We use their "max accuracy" RSV in the main text, and we report analogous results in Table G.1 using the authors' "balanced accuracy" RSV.

[11] We modify the specification of Jack et al. (2025) in two ways. (i) While they distinguish between two types of PES contracts, we define the treatment as whether any PES contract was offered. (ii) They analyze the effects of PES contracts by defining a farmer-level outcome, whereas we analyze effects at the field level.

Table 2: Underestimation of treatment effects in crop burning experiment.

Estimand	Common practice $\widetilde{\theta}$	Bias β	Causal parameter θ
Estimate	0.079	0.530	0.148
	(0.041)	(0.072)	(0.084)

Notes: The RSV is the field-level "maximum accuracy" label defined by Jack et al. (2025), which applies a threshold rule to the predicted probability of not being burned. The "observational sample" has fields that received a random spot check, and the "experimental sample" has other fields. For illustration, we conduct linear estimation, controlling for stratum fixed effects. Standard errors are based on 5000 bootstrap replications clustered at the village level.

bias, we estimate $\beta := \mathbb{E}(R \mid Y=1) - \mathbb{E}(R \mid Y=0)$ using information collected through random spot checks of the fields. Under Assumption 3(ii), an algebraic argument in Appendix B yields $\widetilde{\theta} = \beta\theta$ in this setting.[12] Column (2) of Table 2 reports the estimate of β. It suggests that $\widetilde{\theta}$ understates the treatment effect of being offered any PES contract by approximately 47%. This motivates us to ask: what should empirical researchers do instead?

3.3 Main Result: An Identification Formula

Our main result nonparametrically identifies the ATE in the experimental sample without a parametric model that restricts the distribution or dimensionality of the RSV. We derive a novel formula for combining the experimental and observational samples.

To begin, we express the RSV distribution in the experimental sample as an affine transformation of the potential outcomes we wish to identify. The slope and intercept depend on how the RSV relates to the outcome, and they are identified by the observational sample.

Lemma 1 (Identification as generative model)**.** Suppose Assumptions 1, 2, and 3(ii) hold. Then, for any $d \in \{0,1\}$, $x \in \mathcal{X}$, and $r \in \mathcal{R}$, $\delta^e_d(r,x) = \{\delta^o_1(r,x) - \delta^o_0(r,x)\}\mu(d,x) + \delta^o_0(r,x)$, where $\mu(d,x) := \mathbb{E}\{Y(d) \mid S=e, X=x\}$ is the conditional average potential outcome in the experimental sample. Here, $\delta^e_d(r,x) := f_R(r \mid S=e, X=x, D=d)$ and $\delta^o_y(r,x) := f_R(r \mid S=o, X=x, Y=y)$ are RSV conditional densities in the experimental and observational samples, respectively.

By Lemma 1, we recover the ATE in the experimental sample by combining (i) how the RSV varies with the treatment in the experimental sample with (ii) how the RSV varies with

[12] Compared to Remark 1(ii), here we have $\mathbb{E}(Y \mid R, S=o) = R$, which implies $\tilde{\beta} = 1$ and hence $\widetilde{\theta} = \beta\theta_0$.

the outcome in the observational sample. This combination leverages our key assumption: stability of the RSV across samples.

While Lemma 1 identifies the ATE in the experimental sample, it suggests a challenging estimation problem: it involves the conditional distribution of the high-dimensional RSV. One path forward is to develop a complex parametric model, e.g., a generative model of the satellite image distribution conditional upon poverty measurements or environmental outcomes. Even if such a generative model could be developed, it would be prone to misspecification.

We follow a different path that does not require a generative model for R. We use Bayes' rule to rewrite Lemma 1 as a conditional moment equation. This transformation avoids estimation of the RSV's conditional distribution and recovers a classic econometric estimation problem.

Let $\theta(x) := \mathbb{E}\{Y(1) - Y(0) \,|\, S = e, X = x\} = \mu(1,x) - \mu(0,x)$ denote the conditional average treatment effect in the experimental sample.

Theorem 1 (Identification as conditional moment)**.** Under the conditions of Lemma 1,

$$\text{for any } x \in \mathcal{X}, \quad \mathbb{E}\{\Delta^e(x) - \Delta^o(x)\theta(x) \,|\, X = x, R\} = 0 \quad \text{almost surely, where}$$

$\Delta^e(x) := \frac{1\{D=1,S=e\}}{\Pr(D=1,S=e|X=x)} - \frac{1\{D=0,S=e\}}{\Pr(D=0,S=e|X=x)}$ and $\Delta^o(x) := \frac{1\{Y=1,S=o\}}{\Pr(Y=1,S=o|X=x)} - \frac{1\{Y=0,S=o\}}{\Pr(Y=0,S=o|X=x)}$.

Theorem 1 identifies the treatment effect as the solution to a set of conditional moment equalities. The conditional average treatment effect $\theta(x)$ balances treatment variation from the experimental sample $\Delta^e(x)$ and outcome variation from the observational sample $\Delta^o(x)$, so that their projections onto the remotely sensed variable R match. Consequently, we can leverage a celebrated literature on conditional moment equalities (e.g., Chamberlain, 1987; Newey, 1993) for estimation and inference.

For intuition on Theorem 1, consider the case without covariates. Then Δ^e is simply a scalar computed from the experimental sample, where the treatment is observed. Similarly, Δ^o is simply a scalar computed from the observational sample, where the outcome is observed. Theorem 1 shows that $\theta_0 = \frac{\mathbb{E}(\Delta^e|R)}{\mathbb{E}(\Delta^o|R)}$. This ratio appears to be a new formula for data combination, where the numerator and denominator are from different samples. The numerator reflects how both the treatment and the outcome affect the RSV, so we must divide it by the effect of the outcome on the RSV, in order to isolate the desired causal parameter.

For identification, Theorem 1 implies that we may introduce representations of the RSV. Such representations can be arbitrary, as long as they predict outcome variation.

Corollary 1 (Identification as representation)**.** Under Lemma 1's conditions, $\theta(x) = \frac{\mathbb{E}\{H(x,R)\Delta^e(x)|X=x\}}{\mathbb{E}\{H(x,R)\Delta^o(x)|X=x\}}$ for any representation $H(x,R)$ with $\mathbb{E}\{H(x,R)\Delta^o(x) \,|\, X = x\} \neq 0$.

Remark 3 (Testable implication)**.** Our main identifying assumptions are jointly testable. By Corollary 1, any predictive representation $H(x,R)$ identifies $\theta(x)$. If different representations yield significantly different estimates, then we can reject our identifying assumptions. Future work may develop bounds under violations of our identifying assumptions.

Once again, for intuition on Corollary 1, consider the scenario without covariates. In this case, $H(R)$ is simply a scalar representation of the RSV. Corollary 1 shows that $\theta_0 = \frac{\mathbb{E}\{\Delta^e H(R)\}}{\mathbb{E}\{\Delta^o H(R)\}}$. Any representation $H(R)$ is valid as long as the representation is predictive: $\mathbb{E}\{\Delta^o H(R)\} \neq 0$. A weak RSV test asks whether $\mathbb{E}\{\Delta^o H(R)\} \approx 0$, which can be tested using existing frameworks for weak instruments. In the same spirit, a joint test of our identifying assumptions asks whether two representations $H(R)$ and $H'(R)$ give similar estimates. Dissimilar estimates would be evidence against our assumptions.

By Corollary 1, many representations of the RSV provide identification. However, naive choices may be inefficient, producing needlessly large standard errors. Section 4 asks: what representation of the RSV should be chosen for precise, downstream causal inference? Our answer develops a connection between "representation learning" (e.g., Johannemann et al., 2019; Vafa et al., 2025) and classical results for conditional moment equalities.

So far, we have focused on the setting of Assumption 3(ii), which allows incomplete cases but disallows direct effects of the treatment on the RSV. Our main result also holds for the setting of Assumption 3(i), which requires complete cases yet allows direct effects of the treatment on the RSV. Recall that $\mu(d,x) := \mathbb{E}\{Y(d) \mid S=e, X=x\}$ is the conditional average potential outcome in the experimental sample.

Theorem 2 (Identification as conditional moment with direct effects)**.** Suppose Assumptions 1, 2, and 3(i) hold. Then, for any $d \in \{0,1\}$ and any $x \in \mathcal{X}$,

$$\mathbb{E}\{\widetilde{\Delta}^e(d,x) - \widetilde{\Delta}^o(d,x)\mu(d,x) \mid R, X=x\} = 0 \quad \text{almost surely, where}$$

$\widetilde{\Delta}^e(d,x) := \frac{1\{D=d,S=e\}}{\Pr(D=d,S=e|X=x)} - \frac{1\{Y=0,D=d,S=o\}}{\Pr(Y=0,D=d,S=o|X=x)}$, and $\widetilde{\Delta}^o(d,x) := \frac{1\{Y=1,D=d,S=o\}}{\Pr(Y=1,D=d,S=o|X=x)} - \frac{1\{Y=0,D=d,S=o\}}{\Pr(Y=0,D=d,S=o|X=x)}$.

Once again, the causal estimand reconciles treatment variation from the experimental sample with outcome variation from the observational sample, so that their projections onto the remotely sensed variable R match. Theorem 2 allows direct effects; the key assumption in our framework is stability in Assumption 2. Corollary 1 and Remark 3 extend accordingly.

Remark 4 (Some experimental outcomes)**.** In some empirical applications, researchers may additionally collect outcomes for a small subsample of the experimental sample. This information can be directly incorporated into our procedure for estimation and inference. For this extension, define the extended sampling indicator as $\tilde{S} \in \{\{e,o\},e,o\}$. Here, $\tilde{S}=\{e,o\}$ indicates that a unit is experimental, and we have Y for this unit. $\tilde{S}=e$ indicates that a unit is experimental but we do not have Y. Finally, $\tilde{S}=o$ indicates that a unit is observational. To apply our results, replace the expression $S=e$ with $e \in \tilde{S}$, and replace the expression $S=o$ with $o \in \tilde{S}$.

Remark 5 (Multi-valued outcomes)**.** Our identification, estimation, and inference results generalize to discrete and continuous outcomes. Appendix E extends our identification result to discrete outcomes. We generalize the conditional moment equations. Estimation and inference remain essentially the same, under a minimum rank condition that requires the RSV to predict each outcome value well. Appendix F extends our results to continuous outcomes.

4 Estimation and Inference

As a secondary contribution, we demonstrate that our identification result provides guidance on how to choose a representation of the RSV for valid, precise, and robust inference on the causal parameter. We point out and interpret the connection between modern remote sensing and classical conditional moment analysis, which allows us to use standard techniques. We derive valid $n^{-1/2}$ inference without rate conditions and without complexity restrictions on the researcher's RSV-based predictions, justifying the use of complex deep learning algorithms that may be misspecified.

For clarity, we focus on the simple case in which the outcome is binary, there are no pretreatment covariates, and Assumption 3(ii) holds. We discuss the more general case with discrete outcomes and discrete or continuous covariates in Appendix E.

4.1 Choice of Representation for Program Evaluation

Without covariates, our main identification result in Theorem 1 simplifies. The treatment and outcome variation are simply scalars:

$$\Delta^e = \frac{1\{D=1,S=e\}}{\Pr(D=1,S=e)} - \frac{1\{D=0,S=e\}}{\Pr(D=0,S=e)}, \quad \Delta^o = \frac{1\{Y=1,S=o\}}{\Pr(Y=1,S=o)} - \frac{1\{Y=0,S=o\}}{\Pr(Y=0,S=o)}. \quad (1)$$

Since (S,D,Y) are binary, the denominators can be estimated by simple counts. The treatment effect is identified as $\theta = \frac{\mathbb{E}(\Delta^e|R)}{\mathbb{E}(\Delta^o|R)}$. Therefore, $\theta = \frac{\mathbb{E}\{\Delta^e H(R)\}}{\mathbb{E}\{\Delta^o H(R)\}}$ for any representation $H(R)$ that

is predictive of outcome variation in the sense that $\mathbb{E}\{\Delta^o H(R)\} \neq 0$. While any predictive representation can be used for inference, different choices have different efficiency properties. We now ask, which representation should researchers use in practice?

By rearranging our identifying formula, we highlight the connection to classical ideas in econometrics and thereby derive an answer. Specifically, the treatment effect θ may be viewed as the coefficient in the regression model

$$\Delta^e = \theta \Delta^o + \epsilon, \quad \mathbb{E}(\epsilon | R) = 0 \tag{2}$$

where the residual $\epsilon := \Delta^e - \theta \Delta^o$ is mean zero conditional on R by Theorem 1. Results in Chamberlain (1987) and Newey (1993) demonstrate that the representation that achieves efficiency within a class of models satisfying (2) is $H^*(R) = \frac{\mathbb{E}(\Delta^o | R)}{\sigma^2(\theta, R)}$, where $\sigma^2(\theta, R) := \mathbb{E}\{(\Delta^e - \Delta^o \theta)^2 \,|\, R\}$.

For downstream causal inference, we find that the efficient representation of the high-dimensional remotely sensed variable $R \in \mathcal{R}$ is a simple, continuous scalar $H^*(R) \in \mathbb{R}$. The formula for $H^*(R)$ is a conditional expectation divided by a conditional variance.

This connection has a direct consequence for empirical practice: to improve efficiency with the RSV, we should not only predict the outcome Y using observational data, but also predict the treatment D using experimental data, and predict the sample indicator S using all data. The first prediction is part of common practice, but the second and third are not. By using all three predictions, we use the RSV more efficiently.

We confirm this connection by further developing the expression for $H^*(R)$: the numerator $\mathbb{E}(\Delta^o \,|\, R)$ contains $\Pr(Y = 1, S = o | R)$ while the denominator $\sigma^2(\theta, R)$ contains $\Pr(D = 1, S = e | R)$; see Lemma 2 below. Finally, by the definition of conditional probability, $\Pr(Y = 1, S = o | R) = \Pr(Y = 1 | S = o, R) \Pr(S = o | R)$ and $\Pr(D = 1, S = e | R) = \Pr(D = 1 | S = e, R) \Pr(S = e | R)$.

4.2 Inference with Learned Representations

For simplicity, we describe our inferential procedure using sample splitting with TRAIN and TEST folds, though our proofs in Appendix D allow for cross-fitting with any fixed number of folds. We first state our inferential procedure at a high level before filling in the details:

1. Divide the sample into TRAIN and TEST folds.
2. Learn the representation on TRAIN: $\widehat{H}(R)$.

- Train predictors of Y, D, and S: $\text{PRED}_Y(R)$, $\text{PRED}_D(R)$, $\text{PRED}_S(R)$.

- Construct an initial estimate, by regressing $\widehat{\mathbb{E}}(\Delta^e|R)$ on $\widehat{\mathbb{E}}(\Delta^o|R)$: $\widehat{\theta}_{\text{INIT}}$.
- Combine these into the representation: $\widehat{H}(R)$.

3. Construct a causal estimate on TEST: $\widehat{\theta}$.

The overall structure is familiar (Angrist et al., 1999; Chernozhukov et al., 2018), although here we will show that no rate condition is required on $\widehat{H}(R)$. To simplify the algorithm statement, we introduce some notation for simple events and an algebraic identity.

Let $\mathbb{E}_{\text{TRAIN}}(\cdot) = \frac{1}{|\text{TRAIN}|}\sum_{i\in\text{TRAIN}}(\cdot)$ and $\mathbb{E}_{\text{TEST}}(\cdot) = \frac{1}{|\text{TEST}|}\sum_{i\in\text{TEST}}(\cdot)$. Slightly abusing notation, the counting operation, $\text{COUNT}_{\text{EVENT}}$, can be $\mathbb{E}_{\text{TRAIN}}(1_{\text{EVENT}})$ in the second step or $\mathbb{E}_{\text{TEST}}(1_{\text{EVENT}})$ in the third step, which will be clear from the context.

Lemma 2. $(\Delta^e - \Delta^o\theta)^2 = \frac{1\{D=1,S=e\}}{\Pr(D=1,S=e)^2} + \frac{1\{D=0,S=e\}}{\Pr(D=0,S=e)^2} + \theta^2\left[\frac{1\{Y=1,S=o\}}{\Pr(Y=1,S=o)^2} + \frac{1\{Y=0,S=o\}}{\Pr(Y=0,S=o)^2}\right]$.

Algorithm 1 (Inference)**.** Given $\{S_i, 1\{S_i=e\}D_i, 1\{S_i=o\}Y_i, R_i\}$:

1. Divide the sample into TRAIN and TEST folds.

2. Learn the representation on TRAIN: $\widehat{H}(R)$.

(a) Count marginals: $\text{COUNT}_{Y=1,S=o}$, $\text{COUNT}_{Y=0,S=o}$, $\text{COUNT}_{D=1,S=e}$, $\text{COUNT}_{D=0,S=e}$.

(b) Train predictors: $\text{PRED}_Y(R)$ estimates $\Pr(Y=1|S=o,R)$, $\text{PRED}_D(R)$ estimates $\Pr(D=1|S=e,R)$, and $\text{PRED}_S(R)$ estimates $\Pr(S=e|R)$, using machine learning.

(c) Initially estimate $\widehat{\theta}_{\text{INIT}} = \text{argmin}_\theta \mathbb{E}_{\text{TRAIN}}[\{\widehat{\mathbb{E}}(\Delta^e|R) - \widehat{\mathbb{E}}(\Delta^o|R)\theta\}^2]$, where $\widehat{\mathbb{E}}(\Delta^e|R)$ and $\widehat{\mathbb{E}}(\Delta^o|R)$ are constructed from the marginal probabilities and predictors according to (1).

(d) Learn the representation: $\widehat{H}(R) = \frac{\widehat{\mathbb{E}}(\Delta^o|R)}{\widehat{\sigma}^2(\widehat{\theta}_{\text{INIT}},R)}$ where $\widehat{\sigma}^2(\widehat{\theta}_{\text{INIT}},R)$ is constructed from the marginal probabilities, predictors, and initial estimate via Lemma 2.

3. Construct a causal estimate on TEST: $\widehat{\theta}$.

(a) Count marginals: $\text{COUNT}_{Y=1,S=o}$, $\text{COUNT}_{Y=0,S=o}$, $\text{COUNT}_{D=1,S=e}$, $\text{COUNT}_{D=0,S=e}$.

(b) Construct a causal estimate: $\widehat{\theta} = \frac{\mathbb{E}_{\text{TEST}}\{\widehat{\Delta}^e\widehat{H}(R)\}}{\mathbb{E}_{\text{TEST}}\{\widehat{\Delta}^o\widehat{H}(R)\}}$ where $\widehat{\Delta}^e$ and $\widehat{\Delta}^o$ are constructed from marginal probabilities according to (1).

(c) Bootstrap its confidence interval: $\widehat{\theta} \pm c_\alpha \hat{v} n^{-1/2}$, where c_α is the $1-\alpha/2$ quantile of the standard Gaussian and $\hat{v}n^{-1/2}$ is the bootstrap standard error of $\widehat{\theta}$ while fixing $\widehat{H}(R)$.

See Appendix D for explicit computations of $\widehat{\mathbb{E}}(\Delta^e|R)$, $\widehat{\mathbb{E}}(\Delta^o|R)$, and $\widehat{\sigma}^2(\widehat{\theta}_{\text{INIT}},R)$ in terms of the marginal probabilities, predictors, and initial estimate.

Sample splitting may be eliminated under complexity restrictions that tolerate simple machine learning procedures. See e.g., Chernozhukov et al. (2020) for a recent summary.

4.3 Robustness to Misspecification

In empirical research, the RSV is typically unstructured and high-dimensional, e.g., a satellite image. The prediction is often conducted by a complex machine learning algorithm, e.g., a deep convolutional neural network. For this realistic setting, rates of convergence are often unknown. In other words, we have no reason to believe that $\text{PRED}_Y(R)$ converges to $\Pr(Y=1|S=o,R)$, nor that $\widehat{H}(R)$ converges to $H^*(R)$. Even carefully crafted architectures positing a generative model would typically be misspecified.

For this reason, we place a weaker regularity condition: the predictions, and hence the representation estimator, have *some* probability limit; they may be misspecified.

Assumption 4 (Limit)**.** The learned representation has some mean square limit: $\mathbb{E}_R[\{\widehat{H}(R)-\widetilde{H}(R)\}^2]=o_p(1)$, where $\mathbb{E}\{\widetilde{H}(R)^2\}$ is finite, and possibly $\widetilde{H}(R)\neq H^*(R)$. This limit is correlated with outcome variation: $\mathbb{E}\{\widetilde{H}(R)\Delta^o\}$ is bounded away from zero.

Assumption 4 does not require any complexity restriction, nor any rate of convergence. The moment restriction in (2) is infinite-order Neyman orthogonal (Mackey et al., 2018; Chen et al., 2020), so Algorithm 1 enjoys the best of both worlds: no complexity restriction (Chernozhukov et al., 2018, 2023), and no rate requirement (Chamberlain, 1987; Newey, 1993). Even if all three of its predictions are misspecified, Algorithm 1 still delivers valid $n^{-1/2}$ inference, which is a strong form of robustness to misspecification.

In the following statements, we refer to $\Pr(D=1,S=e)$, $\Pr(D=0,S=e)$, $\Pr(Y=1,S=o)$, and $\Pr(Y=0,S=o)$ as the marginal probabilities.

Proposition 2 (Inference with known counts)**.** Suppose Theorem 1's conditions and Assumption 4 hold. If the marginal probabilities are known and bounded away from zero, then $n^{1/2}(\hat{\theta}-\theta)\rightsquigarrow\mathcal{N}\left(0,\frac{\mathbb{E}\{(\Delta^e-\theta\Delta^o)^2\widetilde{H}(R)^2\}}{[\mathbb{E}\{\Delta^o\widetilde{H}(R)\}]^2}\right)$. Moreover, if $\widetilde{H}(R)=H^*(R)$, then $\hat{\theta}$ is semiparametrically efficient for θ satisfying (2) with marginal probabilities known to the researcher.

Proposition 3 (Inference with unknown counts)**.** Suppose Theorem 1's conditions and Assumption 4 hold. If the marginal probabilities and their counting estimators are bounded away from zero, then $n^{1/2}(\hat{\theta}-\theta)\rightsquigarrow\mathcal{N}\left(0,\frac{V}{[\mathbb{E}\{\Delta^o\widetilde{H}(R)\}]^2}\right)$, where V is defined in Lemma D.6.

When marginal probabilities are known, the asymptotic variance is standard (Proposition 2). In this case, (2) matches the conditional moment of Chamberlain (1987), so the semiparametric efficiency bound is known.

When marginal probabilities are unknown, the asymptotic variance is weighted by them (Proposition 3). The asymptotic variance in Proposition 3 differs from the one in Proposition 2 because estimation of the marginal probabilities introduces an additional estimation error of order $n^{-1/2}$ that we must take into account for asymptotic inference. In this case, it is simpler to use a bootstrap than an analytic variance estimator.

Theorem 2 allows direct effects of the treatment on the remotely sensed variable, as long as treatment is non-missing and non-deterministic in the observational sample. We have already derived the conditional moment equations. Estimation and inference are similar.

Remark 6 (Inference with covariates)**.** For clarity of exposition, this section has focused on the setting without covariates X.

When X is discrete with a finite support, estimation and inference are straightforward: apply Algorithm 1 within each covariate stratum, then average over X. See Appendix E.3.

When X is continuous with a low dimension, there are two main modifications: (i) estimate the conditional probabilities $\Pr(D=d,S=e\,|\,X=x)$ and $\Pr(Y=y,S=o\,|\,X=x)$ as smooth functions of x; (ii) estimate the conditional average treatment effect $\theta(x)$ via local regression or series methods, before averaging over X. Appendix E.4 discusses this setting.

Crucially, the main robustness property—that $n^{-1/2}$ inference requires no rate condition on the learned representation $\widehat{H}(x,R)$—continues to hold in both settings. This justifies the use of complex machine learning algorithms for RSV representations.

5 Program Evaluation using Satellite Images

To empirically validate our method, we conduct three semi-synthetic exercises that are increasingly realistic. We use real RSV distributions, together with

- synthetic treatment effects and synthetic sample definitions;
- real treatment effects and synthetic sample definitions; or
- real treatment effects and real sample definitions.

Across the exercises, we use data from an experiment analyzed in Muralidharan et al. (2016, 2023), illustrated in Figure 1. The authors collect data in an experimental sample and

an observational sample of villages in Andhra Pradesh, India. The treatment D is the early introduction of Smartcards, which are a biometrically authenticated payments infrastructure. The outcome Y is a measure of village-level poverty in administrative data from 2012 to 2013 (after the experiment was deployed). See Appendix H for details on the experiment.

We use the geographic coordinates of each village to extract a remotely sensed variable R, from free, open-access databases. The RSV concatenates nighttime luminosity measures from 2012 to 2020 (a vector in $\mathbb{R}^{50}$) (Asher et al., 2021) and satellite images from 2019 (a high-dimensional, pre-trained embedding vector in $\mathbb{R}^{4000}$) (Rolf et al., 2021). These variables have been extensively validated as predictors of poverty (Henderson et al., 2011; Jean et al., 2016; Stoeffler et al., 2016; Michaels et al., 2021; Huang et al., 2021; Sherman et al., 2023). Our framework allows us to test their relevance for program evaluation, similar to a "first stage" exercise in instrumental variable analysis. In our semi-synthetic exercises, we will evaluate how well we can conduct program evaluation using this real RSV.

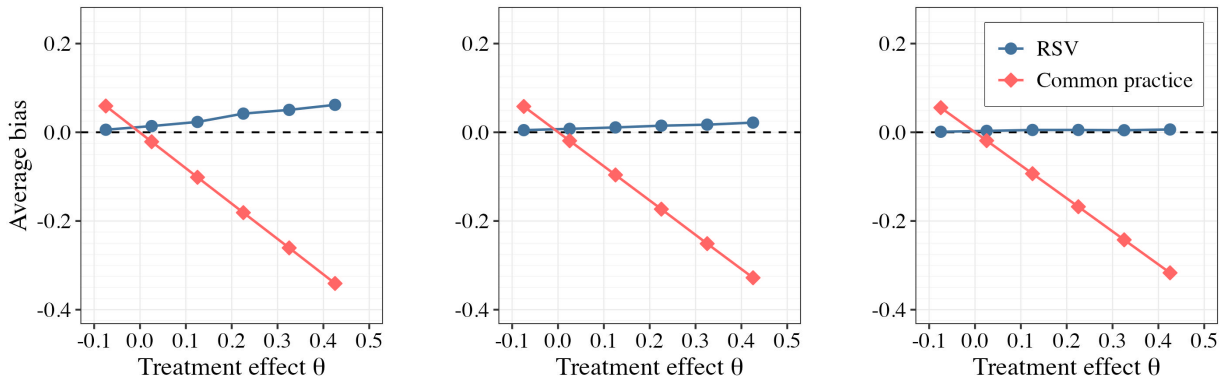
In terms of our identification framework, the semi-synthetic settings have incomplete cases (Assumption 3(ii)): the treatment is missing in the observational sample. The settings also have some experimental outcomes (Remark 4). We will be explicit about sample definitions below. Appendix H contains a complete description of the data.

5.1 Our Method Outperforms Current Practice across Effect Sizes and Sample Sizes

We impose all of our assumptions in a calibrated, synthetic data-generating process (DGP). First, we simulate the binary treatment D as a fair coin toss (Assumption 1). If $D=0$, we simulate the binary outcome Y as a weighted coin toss, calibrated to the empirical probability of $Y=1$ among untreated experimental units in the real data. If $D=1$, we simulate Y as a different weighted coin toss, calibrated to the empirical probability of $Y=1$ among treated experimental units in the real data.

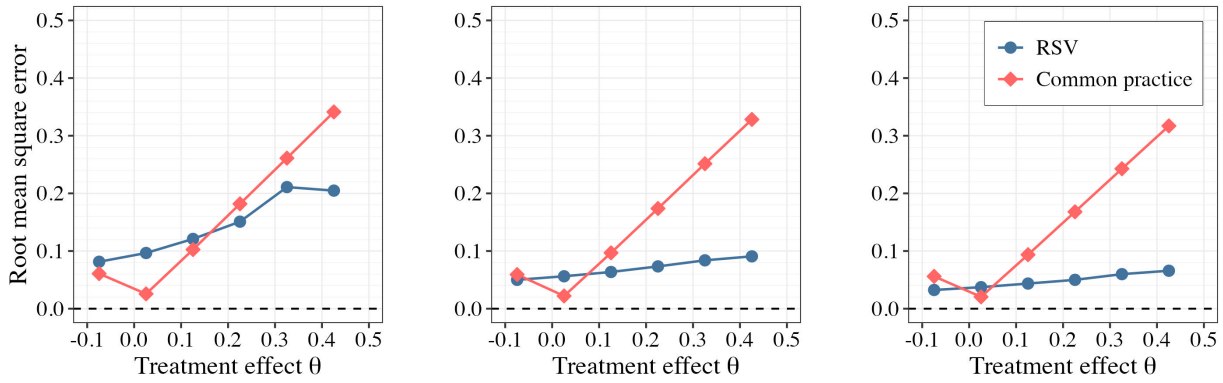
In this baseline version of the DGP, the synthetic treatment effect is calibrated to the real one: -0.07. We consider additional synthetic treatment effect values $\theta=-0.07+\tau$ by augmenting the probability of $Y=1$ when $D=1$ for alternative values of τ.

Next, we draw RSV values from the real data. If $Y=0$, we draw R from the empirical distribution of $R|Y=0$ in the real experimental data. Likewise, we draw R when $Y=1$. This imposes no direct effects (Assumption 3(ii)). For computational feasibility in this exercise,



(a) $n=1000$. (b) $n=2000$. (c) $n=3000$.

Figure 4: In the first exercise, our method outperforms common practice in terms of average bias. For each value of the synthetic treatment effect θ and each sample size n, we conduct 500 replications.



(a) $n=1000$. (b) $n=2000$. (c) $n=3000$.

Figure 5: In the first exercise, our method outperforms common practice in terms of root mean square error. For each value of the synthetic treatment effect θ and each sample size n, we conduct 500 replications.

we use luminosity and only the initial 1,000 satellite image features as our RSV.

Finally, to mimic a setting with missing outcomes, we delete Y if $D=1$. In terms of Remark 4, we set $\tilde{S}=e$ when $D=1$ and $\tilde{S}=\{e,o\}$ when $D=0$.[13] This imposes stability across samples (Assumption 2). Recent methods that use machine learning predictions as outcomes, e.g., prediction-powered inference (Angelopoulos et al., 2023), provide no guidance for this DGP, because no treated unit has an observed outcome.

[13]For simplicity, we do not have $\tilde{S}=o$. This could be easily done: if $D=0$, toss a coin to determine whether $\tilde{S}=\{e,o\}$ or $\tilde{S}=o$. With or without $\tilde{S}=o$, a direct extension of Theorem 1 applies: within the identifying formula given there, simply replace $S=e$ with $e\in\tilde{S}$ in Δ^e, and similarly replace $S=o$ with $o\in\tilde{S}$ in Δ^o.

Figures 4 and 5 demonstrate that our method outperforms common practice across treatment effect values and sample sizes. We consider treatment effect values $\theta = -0.07 + \tau$ with $\tau \in \{0, 0.1, ..., 0.5\}$ and sample sizes $n \in \{1000, 2000, 3000\}$. By varying these aspects of the synthetic DGP, we evaluate relative performance for different signal-to-noise ratios.

We compare the bias and root mean square error of the two methods. Whereas the bias of our method is always small and vanishing with sample size, the bias of the common practice is typically very large and constant across sample sizes. Common practice has positive or negative bias for the treatment effect, consistent with Proposition 1. While the variance of our method is similar to the variance of the common practice, our method's large improvement in bias translates into similar improvement in mean square error when the sample size is large enough.

This exercise has clear consequences for empirical practice. If the goal is to conduct valid inference on the treatment effect, common practice can be misleading due to large bias that does not vanish as the sample size increases. It can return the wrong sign, as characterized by Proposition 1. By contrast, our method reports unbiased estimates, with valid asymptotic inference.

5.2 Our Method Recovers the True Effect with Randomly Missing Outcomes

While our first exercise involves synthetic treatment effects, our second exercise involves real treatment effects. We now ask: compared to an unbiased benchmark, how does our method perform? The benchmark is the difference-in-means estimate and confidence interval an economist would obtain if they could observe all treatments and outcomes in the experiment. Our method gives an estimate and confidence interval using treatments and RSVs in an experiment, and outcomes and RSVs in an observational sample.[14]

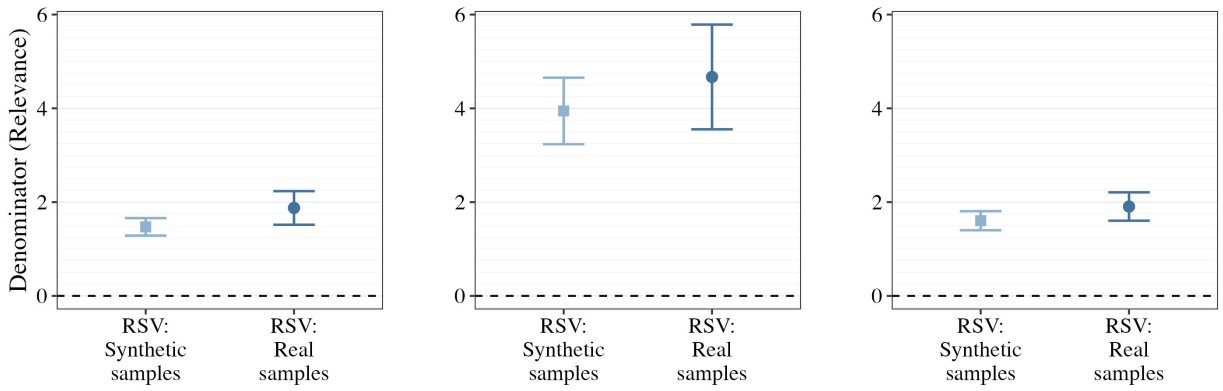
We conduct this exercise with three possible poverty measurements: (i) is the village in the bottom quartile of villages for per capita consumption, (ii) does a village have only low income households, and (iii) does a village have only low and middle income households. While (i) is a measure of average consumption, (ii) and (iii) describe the income distribution using formal categories from Indian administrative data; see Appendix H for details.[15]

In this exercise, we now use the full RSV: the 50 luminosity measures, and the full 4000-dimensional embedding of satellite images.

In a semi-synthetic way, we classify villages into the samples described in Assumption 3(ii)

[14] Following Remark 4, we also observe some outcomes in the experiment, as described below.

[15] We used the low consumption outcome (i) in the first semi-synthetic exercise.



(a) Low consumption. (b) Low income. (c) Low and middle income.

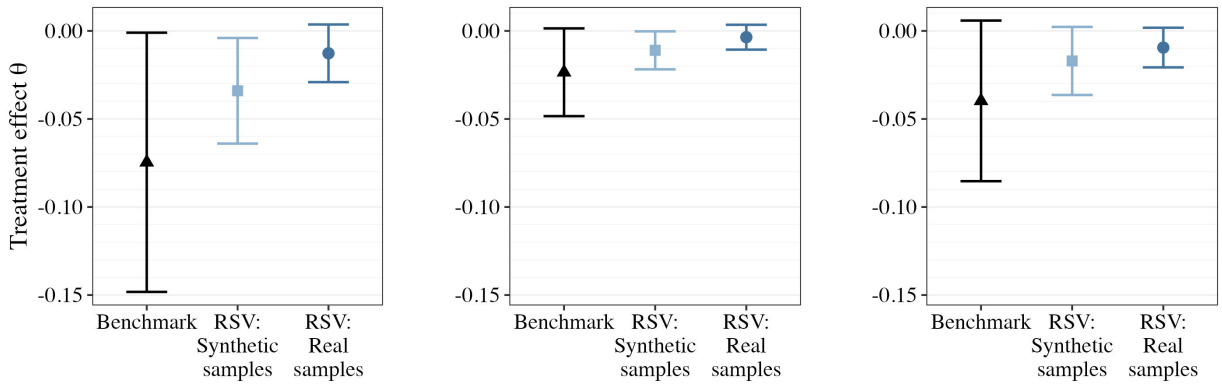
Figure 6: Satellite images are relevant to the poverty outcomes. Each plot is for a poverty outcome. Within each plot, we report $\mathbb{E}_n\{\widehat{H}(R)\widehat{\Delta}^o\}$ and its 90% confidence interval using our learned representation. In light blue, we visualize the relevance of our learned representation in the second exercise ("synthetic samples"). In dark blue, we visualize the relevance of our learned representation in the third exercise ("real samples"). Bootstrap standard errors, based on 1000 replications, are clustered at the sub-district level.

and Remark 4. We classify the real observational villages as $\tilde{S}=o$, for which we observe (Y,R). Among the real experimental villages, we randomly classify some as $\tilde{S}=e$, for which we observe (D,R), and some as $\tilde{S}=\{e,o\}$, for which we observe (D,Y,R). In other words, for real experimental villages, we randomly delete half of their outcomes.

This exercise satisfies the key assumptions of our framework. The real treatment variable was randomized in the real experiment (Assumption 1). Stability of the RSV conditional distribution is plausible, as demonstrated in Figure 2 (Assumption 2). Finally, because the real data have incomplete cases, we must argue that the treatment only affects the RSV via the outcome (Assumption 3(ii)). As supporting evidence, Figures H.1 and H.2 show that the distribution of $R|Y=y,D=1$ visually matches the distribution of $R|Y=y,D=0$.

Another requirement in our framework is that the RSV is relevant: $\mathbb{E}\{H(R)\Delta^o\}\neq 0$ in Corollary 1. Intuitively, the representation of the RSV $H(R)$ used for inference should be correlated with outcome variation Δ^o. This requirement is testable; for our learned representation in Algorithm 1, we can test whether the empirical analogue $\mathbb{E}_n\{\widehat{H}(R)\widehat{\Delta}^o\}$ is significantly nonzero. Figure 6 confirms that our RSV is relevant.

We further interpret and compare our learned representation of the satellite image in Appendix H. Our optimal representation is based on three predictions: the outcome, treatment,



(a) Low consumption. (b) Low income. (c) Low and middle income.

Figure 7: Our method recovers the unbiased benchmark estimate and its 90% confidence interval. Each plot is for a particular poverty outcome. Within each plot, we visualize the benchmark in black versus our method in blue. The benchmark is the difference-in-means an economist would obtain if they could observe treatments and outcomes in the experiment. Our method uses treatments and RSVs in an experiment, and outcomes and RSVs in an observational sample. In light blue, we visualize our method in the second exercise, where we observe outcomes for a random subset of experimental villages ("synthetic samples"). In dark blue, we visualize our method in the third exercise, where we observe outcomes for only the untreated experimental villages ("real samples"). Bootstrap standard errors, based on 1000 replications, are clustered at the sub-district level.

and sample indicator given the RSV. By contrast, common practice only uses the prediction of the outcome given the RSV. Interestingly, our optimal representation allows for extrapolation, i.e. negative weights on villages, whereas common practice does not.

Figure 7 ("synthetic samples") shows that our method recovers the treatment effect estimated by the unbiased benchmark in this exercise. For each poverty outcome, our treatment effects have the same signs and similar magnitudes as the benchmark, i.e. as if the economist could observe all treatments and outcomes in the experiment. These effects are consistent in sign with the findings in Muralidharan et al. (2023): early adoption of Smartcards reduces poverty. Compared to the benchmark, our confidence intervals are similar, and sometimes shorter; efficiently using the additional information in RSVs can boost statistical power.[16]

These findings have a practical implication: our method may allow economists to significantly reduce survey costs. In this exercise, we mimic what would happen if an economist paid

[16]Assumption 3(ii) is an additional restriction that may improve asymptotic precision. Relative magnitudes of standard errors may also reflect finite sample estimation error.

surveyors to collect outcomes for half of the experimental villages, and relied on free satellite images for the other half. We can calculate the savings, compared to paying surveyors to collect outcomes for all of the experimental villages. If surveying each individual in a village conservatively costs \$0.50, our results suggest savings of \$3.6 million.[17]

5.3 Our Method Recovers the True Effect in a Realistic Setting

The third exercise is identical to the second exercise, except that we classify villages in a more realistic way. As before, we classify the real observational villages as $\tilde{S}=o$, for which we observe (Y,R). Among the real experimental villages, we now classify the treated ones as $\tilde{S}=e$, for which we observe (D,R), and the untreated ones as $\tilde{S}=\{e,o\}$, for which we observe (D,Y,R). In other words, for real experimental villages, we systematically delete the outcomes of all treated villages. Methods based on missingness-at-random provide no guidance in this setting.

As before, this exercise appears to satisfy our identifying assumptions. The real treatment variable was randomized in the real experiment (Assumption 1). The conditional distribution of the RSV appears stable in Figure 2 (Assumption 2). There do not appear to be direct effects in Figures H.1 and H.2 (Assumption 3(ii)). The RSV is relevant in Figure 6 (Corollary 1).

In this more challenging exercise, we find qualitatively similar results. Our method recovers comparable estimates and confidence intervals to the benchmark in Figure 7 ("real samples"). Our learned representations allow efficient extrapolation in Figure H.3.

We conclude this discussion with two final remarks on the interpretation of our semi-synthetic exercises.

Remark 7 (Spillovers)**.** In the first exercise, we compare our method to the true value of the treatment effect in Definition 1. There are no spillovers by design.

In the second and third exercises, we compare our method to a natural benchmark: a difference-in-means estimate an economist would obtain if they could observe all treatments and outcomes in the experiment. If there are no spillovers, then we are comparing our method to an unbiased estimate of the treatment effect in Definition 1. If there are spillovers, then we are comparing to an unbiased estimate of a reasonable estimand: a difference-in-means of expected outcomes, indexing by each unit's treatment status and averaging over other all other units.

[17] Table H.1 summarizes the number of experimental villages and the populations in experimental villages, which we use to calculate savings. This calculation is highly conservative; Viviano and Rudder (2024) find that phone surveys in Pakistan cost \$7 per individual, rather than \$0.50.

In Appendix I, we consider spillovers more carefully, by introducing additional notation. We discuss how our method can be adapted to estimate the global average treatment effect, a common parameter of interest in models with spillovers.

Remark 8 (Timing)**.** Outcomes were measured at the end of the experiment in 2013, so we study the effect on poverty in 2013. Meanwhile, RSVs were measured through 2020. Appendix I confirms that it is valid to use RSVs measured after the outcomes. The auxiliary sample links outcomes from 2013 with RSVs through 2020, so the estimand and estimator remain unchanged despite these differences in timing.

6 Recommendations for Practice

Common empirical practice can be highly biased when the remotely sensed variable is *post*-outcome, i.e., when the RSV is caused by the outcome of interest. For example, changes in satellite images are caused by fires or variation in local income, but not vice versa. Theoretically, we demonstrate that this bias can have arbitrary sign. Empirically, we find that common practice may attenuate the treatment effect in a real environmental application, and it may have positive or negative bias in a semi-synthetic development application.

However, the core intuition underlying empirical work is powerful: the conditional distribution of the RSV given the outcome and treatment should be stable across samples. By formalizing this intuition, we develop a framework that nonparametrically identifies treatment effects and efficiently uses RSVs for program evaluation.

We conclude with some concrete recommendations for researchers conducting program evaluation with remotely sensed outcomes. We organize our recommendations into (i) implementation steps and (ii) diagnostic tests.

6.1 Three Steps for Robust Inference with RSVs

To apply our framework, researchers should follow three steps when designing studies with remotely sensed outcomes.

Step 1: Construct a credible observational sample. Build an auxiliary, observational sample that contains RSVs with linked outcomes. The key identifying assumption is stability: the conditional distribution of the RSV given the covariates, treatment, and outcome should be stable across the experimental and observational samples. Consequently,

the researcher's most important design decision is how to construct an observational sample in a way that makes the stability assumption defensible.

Stability is most credible if the researcher randomly selects units for the experimental sample and observational sample prior to conducting the experiment. When random selection is infeasible, researchers should construct an observational sample that is similar to the experimental sample along a few dimensions. (i) Select units that are geographically close to the experimental sample. (ii) Select units with similar baseline covariates, e.g. rural-urban compositions, to the experimental sample. (iii) Select units that align with the timing of the experimental sample. That is, in the observational sample, collect outcomes measured in the same period as the experiment's endline, and collect RSVs measured at the same time that RSVs will be collected in the experiment. If exact alignment is not possible, then match the time difference so that the temporal relationship might be preserved.

Step 2: Estimate three predictions from RSVs. Efficient use of RSVs for downstream causal inference requires three predictions: the outcome, treatment, and sample indicator given the RSV. Common practice uses only the first prediction, in a way that sacrifices not only accuracy but also precision. In other words, by using all three predictions, our method improves not only bias but also variance. Researchers can construct these predictions using complex deep learning algorithms with unknown statistical properties.

Step 3: Apply robust inference algorithm. Algorithm 1 aggregates the three predictions into an efficient scalar representation of the high-dimensional RSV (where efficiency is within the class of models satisfying the conditional moments we derive). This aggregation reduces the dimension while preserving information relevant for causal inference. Critically, Algorithm 1 is robust to misspecification; it provides valid, $n^{-1/2}$ inference even if all three predictions are misspecified. It only requires that the learned representation converges at any rate to any limit that predicts outcome variation. This robustness property justifies the use of deep neural networks, which are widely implemented by empirical researchers in this literature, while providing the safeguard of a principled inferential framework.

6.2 Three Diagnostics for RSVs

In addition to the point estimate and confidence interval of Algorithm 1, researchers should report the following diagnostics to defend the credibility of their RSV.

Diagnostic 1: RSV relevance test. Just as weak instruments threaten inference in

instrumental variable models, weak RSVs—those that poorly predict the outcome—threaten inference in our framework. Researchers should assess whether the correlation between their estimated representation and the outcome is significantly different than zero. Figure 6 illustrates this diagnostic, showing that satellite images are strongly related to village-level poverty. When relevance is weak, researchers should either improve RSV measurements (e.g., by using higher resolution images or additional data sources) or acknowledge that inference is inconclusive.

Diagnostic 2: Joint test of identifying assumptions. Our identifying assumptions are jointly testable. If two different RSV representations yield significantly different treatment effect estimates, then reject the null hypothesis that our identifying assumptions jointly hold: stability (Assumption 2) and no direct effects (Assumption 3(ii)). This is essentially a specification test. Researchers should verify that their results are robust to alternative ways of representing the RSV, such as using different machine learning architectures or different subsets of RSV features.

Diagnostic 3: RSV density plots if possible. When outcomes are observed for some experimental units, researchers can directly visualize whether stability holds by comparing conditional densities. Specifically, plot the RSV density in the experimental sample against the RSV density in the observational sample, for each combination of treatment and outcome values. If these densities align, then stability is plausible, as in Figure 2. Similarly, researchers can assess the no-direct-effects assumption (Assumption 3(ii)) by comparing the RSV density of treated units against the RSV density of untreated units, for each outcome value. If treatment only affects the RSV via the outcome, these densities should coincide, as in Figures H.1 and H.2.

6.3 Broader Implications

By following these steps, researchers can conduct rigorous program evaluation while substantially reducing costs. In our semi-synthetic exercise, we recover treatment effects and confidence intervals comparable to those obtained with direct outcome measurements, despite using satellite images for half the sample.

Our framework generalizes beyond simple randomized experiments. Appendix J extends our identification results to quasi-experiments (e.g., instrumental variables and difference-in-differences). The key insight remains the same: stability in how the RSV relates to other variables across samples allows us to transport this relationship to the sample where the outcome is missing.

Our framework poses new questions. Future research may study a new experimental design problem: how to simultaneously design treatment assignment, outcome collection, and sensor deployment to maximize statistical power subject to budget constraints. Our main identification result provides the foundation to set up this problem. Our main result may inform not only the use of satellite images, but also the design of cheap, noisy surveys. Finally, future work may extend our inference guarantees to accommodate high-dimensional outcomes.

As remote sensing expands the frontier of data availability in economics, our framework offers practical principles for its use in program evaluation. This paper corrects a widespread but biased practice and provides concrete recommendations on how to incorporate satellite images and other remotely sensed data into causal inference. These new data sources and new methods broaden access to rigorous impact evaluation in cost-constrained environments.

References

Ai, C. and X. Chen (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica 71*(6), 1795–1843.

Aiken, E., S. Bellue, J. E. Blumenstock, D. Karlan, and C. Udry (2025). Estimating impact with surveys versus digital traces: Evidence from randomized cash transfers in Togo. *Journal of Development Economics 175*, 103477.

Alix-Garcia, J. and D. L. Millimet (2023). Remotely incorrect? Accounting for nonclassical measurement error in satellite data on deforestation. *Journal of the Association of Environmental and Resource Economists 10*(5), 1335–1367.

Allon, G., D. Chen, Z. Jiang, and D. Zhang (2023). Machine learning and prediction errors in causal inference. *The Wharton School Research Paper*.

Angelopoulos, A. N., S. Bates, C. Fannjiang, M. I. Jordan, and T. Zrnic (2023). Prediction-powered inference. *Science 382*(6671), 669–674.

Angrist, J. D., G. W. Imbens, and A. B. Krueger (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics 14*(1), 57–67.

Asher, S., T. Lunt, R. Matsuura, and P. Novosad (2021). Development research at high geographic resolution: An analysis of night-lights, firms, and poverty in India using the SHRUG open data platform. *The World Bank Economic Review 35*(4), 845–871.

Asher, S. and P. Novosad (2020). Rural roads and local economic development. *American Economic Review 110*(3), 797–823.

Assuncao, J., R. McMillan, J. Murphy, and E. Souza-Rodrigues (2023). Optimal environmental targeting in the Amazon rainforest. *The Review of Economic Studies 90*(4), 1608–1641.

Athey, S., R. Chetty, G. W. Imbens, and H. Kang (2025, 09). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. *The Review of Economic Studies*, rdaf087.

Balboni, C., R. Burgess, and B. A. Olken (2024). The origins and control of forest fires in the tropics. *NBER Working Paper Series*.

Bareinboim, E. and J. Pearl (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences 113*(27), 7345–7352.

Battaglia, L., T. Christensen, S. Hansen, and S. Sacher (2024). Inference for regression with variables generated from unstructured data. *arXiv:2402.15585*.

Blumenstock, J., G. Cadamuro, and R. On (2015). Predicting poverty and wealth from mobile phone metadata. *Science 350*(6264), 1073–1076.

Burke, M., A. Driscoll, D. B. Lobell, and S. Ermon (2021). Using satellite imagery to understand and promote sustainable development. *Science 371*(6535), eabe8628.

Carlson, J. and M. Dell (2025). A unifying framework for robust and efficient inference with unstructured data. *arXiv:2505.00282*.

Carrasco, M., J.-P. Florens, and E. Renault (2007). Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of Econometrics 6*, 5633–5751.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics 34*(3), 305–334.

Chen, J., D. L. Chen, and G. Lewis (2020). Mostly harmless machine learning: Learning optimal instruments in linear IV models. *arXiv:2011.06158*.

Chen, J. J., V. Mueller, Y. Jia, and S. K.-H. Tseng (2017). Validating migration responses to flooding using satellite and vital registration data. *American Economic Review 107*(5), 441–445.

Chen, X., H. Hong, and D. Nekipelov (2011). Nonlinear models of measurement errors. *Journal of Economic Literature 49*(4), 901–37.

Chen, X., H. Hong, and E. Tamer (2005). Measurement error models with auxiliary data. *The Review of Economic Studies 72*(2), 343–366.

Chen, X., H. Hong, and A. Tarozzi (2008). Semiparametric efficiency in GMM models with auxiliary data. *The Annals of Statistics 36*(2), 808.

Chen, X. and W. D. Nordhaus (2011). Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences 108*(21), 8589–8594.

Chen, X. and M. Reiss (2011). On rate optimality for ill-posed inverse problems in econometrics. *Econometric Theory 27*(3), 497–521.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. K. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Chernozhukov, V., W. Newey, R. Singh, and V. Syrgkanis (2020). Adversarial estimation of Riesz representers. *arXiv:2101.00009*.

Chernozhukov, V., W. K. Newey, and R. Singh (2023). A simple and general debiased machine learning theorem with finite-sample guarantees. *Biometrika 110*(1), 257–264.

Cross, P. J. and C. F. Manski (2002). Regressions, short and long. *Econometrica 70*(1), 357–368.

Currie, J., J. Voorheis, and R. Walker (2023). What caused racial disparities in particulate exposure to fall? New evidence from the Clean Air Act and satellite-based measures of air quality. *American Economic Review 113*(1), 71–97.

D'Haultfœuille, X., C. Gaillac, and A. Maurel (2024). Linear regressions with combined data. *arXiv:2412.04816*.

D'Haultfœuille, X., C. Gaillac, and A. Maurel (2025). Partially linear models under data combination. *The Review of Economic Studies 92*(1), 238–267.

Egami, N., C. J. Fong, J. Grimmer, M. E. Roberts, and B. M. Stewart (2022). How to make causal inferences using texts. *Science Advances 8*(42), eabg2652.

Egami, N., M. Hinck, B. Stewart, and H. Wei (2024). Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. *Advances in Neural Information Processing Systems 36*.

Egger, D., J. Haushofer, E. Miguel, P. Niehaus, and M. Walker (2022). General equilibrium effects of cash transfers: Experimental evidence from Kenya. *Econometrica 90*(6), 2603–2643.

Fan, Y., R. Sherman, and M. Shum (2014). Identifying treatment effects under data combination. *Econometrica 82*(2), 811–822.

Fong, C. and M. Tyler (2021). Machine learning predictions as regression covariates. *Political Analysis 29*(4), 467–484.

Gentzkow, M., J. M. Shapiro, and M. Taddy (2019). Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica 87*(4), 1307–1340.

Ghassami, A., A. Yang, D. Richardson, I. Shpitser, and E. T. Tchetgen (2022). Combining experimental and observational data for identification and estimation of long-term causal effects. *arXiv:2201.10743*.

Gordon, M., M. Ayers, E. Stone, and L. C. Sanford (2023). Remote control: Debiasing remote sensing predictions for causal inference. In *ICLR 2023 Workshop on Tackling Climate Change with Machine Learning*.

Graham, B. S., C. C. de Xavier Pinto, and D. Egel (2016). Efficient estimation of data combination models by the method of auxiliary-to-study tilting (AST). *Journal of Business & Economic Statistics 34*(2), 288–301.

Hansen, M. C., P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, et al. (2013). High-resolution global maps of 21st-century forest cover change. *Science 342*(6160), 850–853.

Henderson, V., A. Storeygard, and D. N. Weil (2011, May). A bright idea for measuring economic growth. *American Economic Review 101*(3), 194–99.

Henderson, V., A. Storeygard, and D. N. Weil (2012). Measuring economic growth from outer space. *American Economic Review 102*(2), 994–1028.

Horowitz, J. L. and C. F. Manski (1995). Identification and robustness with contaminated and corrupted data. *Econometrica 63*(2), 281–302.

Huang, L. Y., S. M. Hsiang, and M. Gonzalez-Navarro (2021). Using satellite imagery and deep learning to evaluate the impact of anti-poverty programs. *National Bureau of Economic Research*.

Imbens, G., N. Kallus, X. Mao, and Y. Wang (2024, 10). Long-term causal inference under persistent confounding via data combination. *Journal of the Royal Statistical Society Series B: Statistical Methodology 87*(2), 362–388.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Jack, B. K., S. Jayachandran, N. Kala, and R. Pande (2025). Money (not) to burn: Payments for ecosystem services to reduce crop residue burning. *American Economic Review: Insights 7*(1), 39–55.

Jack, B. K. and K. Walker (2023). Integrating remote sensing and randomized controlled trials: Challenges, opportunities, and practical guidance. Technical report, Haas School of Business, UC Berkeley.

Jayachandran, S., J. De Laat, E. F. Lambin, C. Y. Stanton, R. Audy, and N. E. Thomas (2017). Cash for carbon: A randomized trial of payments for ecosystem services to reduce deforestation. *Science 357*(6348), 267–273.

Jean, N., M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon (2016). Combining satellite imagery and machine learning to predict poverty. *Science 353*(6301), 790–794.

Ji, W., L. Lei, and T. Zrnic (2025). Predictions as surrogates: Revisiting surrogate outcomes in the age of AI. *arXiv:2501.09731*.

Johannemann, J., V. Hadad, S. Athey, and S. Wager (2019). Sufficient representations for categorical variables. *arXiv:1908.09874*.

Kallus, N. and X. Mao (2024, 10). On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *Journal of the Royal Statistical Society Series B: Statistical Methodology 87*(2), 480–509.

Kluger, D. M., K. Lu, T. Zrnic, S. Wang, and S. Bates (2025). Prediction-powered inference with imputed covariates and nonuniform sampling. *arXiv:2501.18577*.

Knox, D., C. Lucas, and W. K. T. Cho (2022). Testing causal theories with learned proxies. *Annual Review of Political Science 25*(1), 419–441.

Kress, R. (1989). *Linear Integral Equations*, Volume 3. Springer.

Lu, K., D. M. Kluger, S. Bates, and S. Wang (2025). Regression coefficient estimation from remote sensing maps. *Remote Sensing of Environment 330*, 114949.

Mackey, L., V. Syrgkanis, and I. Zadik (2018). Orthogonal machine learning: Power and limitations. In *International Conference on Machine Learning*, pp. 3375–3383. PMLR.

Marx, B., T. M. Stoker, and T. Suri (2019). There is no free house: Ethnic patronage in a Kenyan slum. *American Economic Journal: Applied Economics 11*(4), 36–70.

Michaels, G., D. Nigmatulina, F. Rauch, T. Regan, N. Baruah, and A. Dahlstrand (2021). Planning ahead for better neighborhoods: Long-run evidence from Tanzania. *Journal of Political Economy 129*(7), 2112–2156.

Muralidharan, K., P. Niehaus, and S. Sukhtankar (2016). Building state capacity: Evidence from biometric smartcards in India. *American Economic Review 106*(10), 2895–2929.

Muralidharan, K., P. Niehaus, and S. Sukhtankar (2023). General equilibrium effects of (improving) public employment programs: Experimental evidence from India. *Econometrica 91*(4), 1261–1295.

Newey, W. K. (1993). Efficient estimation of models with conditional moment restrictions. In R. F. Engle and D. L. McFadden (Eds.), *Handbook of Statistics*, Volume 11, Chapter 16, pp. 419–454. Elsevier.

Newey, W. K. and J. L. Powell (2003). Instrumental variable estimation of nonparametric models. *Econometrica 71*(5), 1565–1578.

Park, C., D. B. Richardson, and E. J. Tchetgen Tchetgen (2024). Single proxy control. *Biometrics 80*(2), ujae027.

Patel, D. (2024). Floods. Technical report, Department of Economics, Harvard University.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine 8*(4), 431–440.

Proctor, J., T. Carleton, and S. Sum (2023). Parameter recovery using remotely sensed variables. *National Bureau of Economic Research*.

Ridder, G. and R. Moffitt (2007). The econometrics of data combination. *Handbook of Econometrics 6*, 5469–5547.

Rolf, E., J. Proctor, T. Carleton, I. Bolliger, V. Shankar, M. Ishihara, B. Recht, and S. Hsiang (2021). A generalizable and accessible approach to machine learning with global satellite imagery. *Nature Communications 12*(1), 4392.

Roth, J. and P. H. C. Sant'Anna (2023). When is parallel trends sensitive to functional form? *Econometrica 91*(2), 737–747.

Schennach, S. M. (2020). Mismeasured and unobserved variables. In *Handbook of Econometrics*, Volume 7, pp. 487–565. Elsevier.

Sherman, L., J. Proctor, H. Druckenmiller, H. Tapia, and S. M. Hsiang (2023). Global high-resolution estimates of the United Nations Human Development Index using satellite imagery and machine-learning. *National Bureau of Economic Research*.

Stoeffler, Q., B. Mills, and C. Del Ninno (2016). Reaching the poor: Cash transfer program targeting in Cameroon. *World Development 83*, 244–263.

Stoetzer, L. F., X. Zhou, and M. Steenbergen (2025). Causal inference with latent outcomes. *American Journal of Political Science 69*(2), 624–640.

Vafa, K., S. Athey, and D. M. Blei (2025). Estimating wage disparities using foundation models. *Proceedings of the National Academy of Sciences 122*(22), e2427298122.

Viviano, D. and J. Rudder (2024). Policy design in experiments with unknown interference. *arXiv:2011.08174*.

Walker, K., B. Moscona, K. Jack, S. Jayachandran, N. Kala, R. Pande, J. Xue, and M. Burke (2022). Detecting crop burning in India using satellite data. *arXiv:2209.10148*.

Program Evaluation with Remotely Sensed Outcomes

Appendix Materials

Ashesh Rambachan, Rahul Singh, Davide Viviano

Appendix A provides additional figures and tables that motivate our model. Appendix B derives the bias of common practice. Appendices C and D prove our main identification and inference results, respectively. Appendices E and F extend these results from binary outcomes to discrete and continuous outcomes, respectively. Appendix G provides details for the crop burning illustration. Appendices H and I provide details and extensions, respectively, for the Smartcard illustration. Finally, Appendix J extends our identification framework to quasi-experimental settings.

A Related Work and Additional Model Details

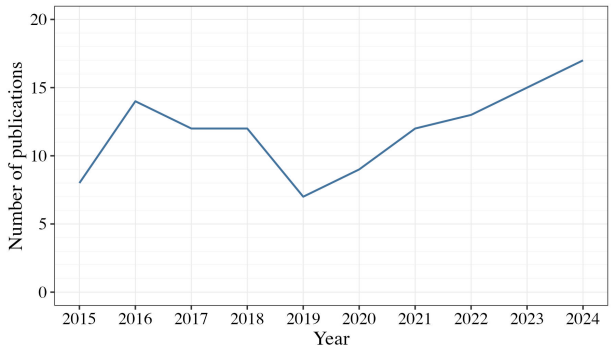


Figure A.1: Remotely sensed variables are increasingly popular in published papers.

Figure A.1 illustrates the increasing popularity of RSVs in empirical research. First, we collected papers published in the AEA journals, *Econometrica*, *Quarterly Journal of Economics*, *Review of Economic Studies* and *Journal of Political Economy* using a keyword search of "remotely sensed variables", "mobile phone", "satellite", "machine learning", and "drones / aerial" on their websites. Next, we collected papers published in *Nature* and *Science* using a keyword search of "remotely sensed variables" on their websites. We subsetted to the papers with RSVs in their main empirical analysis.

Figure A.2 illustrate the causal graph associated with the surrogacy identifying assumptions (Prentice, 1989; Athey et al., 2025): the surrogate R fully mediates the effect of the

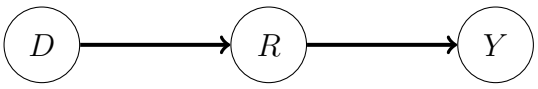


Figure A.2: Causal graph of surrogacy model.

treatment D on the outcome Y. Our RSV identifying assumptions are the opposite, as illustrated by Figure 3.

Table A.1: Implications of RSV identifying assumptions.

Assumption	Quantity	Experimental	Observational	Description
1	$\Pr\{Y(d)\|S,X,D\}$	$\Pr\{Y(d)\|S=e,X\}$	$\Pr\{Y(d)\|S=o,X,D\}$	Differs across samples
2	$\Pr(R\|S,X,D,Y)$	$\Pr(R\|X,D,Y)$	$\Pr(R\|X,D,Y)$	Stable across samples
3(i)	$\Pr(R\|S,X,D,Y)$	$\Pr(R\|X,D,Y)$	$\Pr(R\|X,D,Y)$	Differs across treatments if complete cases
3(ii)	$\Pr(R\|S,X,D,Y)$	$\Pr(R\|X,Y)$	$\Pr(R\|X,Y)$	Stable across treatments if incomplete cases

Table A.1 summarizes the implications of our identifying assumptions. Our identifying assumptions allow the propensity score $\Pr(D=1\,|\,S,X)$ to differ across samples.

B Bias of Common Practice Proofs

B.1 Bias with Binary Variables

Assumption B.1 (Binary setting)**.** Suppose $X=\varnothing$, and $D,Y,R\in\{0,1\}$.

Lemma B.1 (Linear representations)**.** Suppose Assumptions 1, 2, and 3(ii) hold, as well as Assumption B.1. Then, without loss of generality, $\mathbb{E}(Y|R,S=o)=\tilde{\beta}_0+\tilde{\beta}R$, $\mathbb{E}(R|Y,D,S=e)=\beta_0+\beta Y$, and $\theta_0=\mathbb{E}(Y|D=1,S=e)-\mathbb{E}(Y|D=0,S=e)$ for some scalars $(\tilde{\beta}_0,\tilde{\beta},\beta_0,\beta)$.

Proof. The conditional distribution of binary variables is summarized by their correlation, with appropriate scaling by the variance. □

Proposition B.1 (Bias of common practice: Binary)**.** Under the conditions of Lemma B.1, $\widetilde{\theta}=\tilde{\beta}\beta\theta_0$.

Proof. By the surrogate formula, Lemma B.1, and law of iterated expectations,

$$\begin{aligned}
\widetilde{\theta}&=\mathbb{E}\{\mathbb{E}(Y|R,S=o)|D=1,S=e\}-\mathbb{E}\{\mathbb{E}(Y|R,S=o)|D=0,S=e\}\\
&=\tilde{\beta}\{\mathbb{E}(R|D=1,S=e)-\mathbb{E}(R|D=0,S=e)\}\\
&=\tilde{\beta}[\mathbb{E}\{\mathbb{E}(R|Y,D=1,S=e)|D=1,S=e\}-\mathbb{E}\{\mathbb{E}(R|Y,D=0,S=e)|D=0,S=e\}]\\
&=\tilde{\beta}\beta\{\mathbb{E}(Y|D=1,S=e)-\mathbb{E}(Y|D=0,S=e)\}=\tilde{\beta}\beta\theta_0.
\end{aligned}$$

□

Corollary B.1 (Bias of common practice: Binary constrained)**.** Under the conditions of Lemma B.1, if $\tilde{\beta}=1$, then $\widetilde{\theta}=\beta\theta_0$.

Proof. The result is immediate from Proposition B.1. □

Proposition B.2 (Identification with binary variables)**.** Under the conditions of Lemma B.1,

$$\theta_0=\beta^{-1}\{\mathbb{E}(R|D=1,S=e)-\mathbb{E}(R|D=0,S=e)\}.$$

Proof. Within the proof of Proposition B.1, we have shown

$$\tilde{\beta}\{\mathbb{E}(R|D=1,S=e)-\mathbb{E}(R|D=0,S=e)\}=\tilde{\beta}\beta\theta_0.$$

□

B.2 Bias with Random Sample Selection

As in Proposition 1, consider the setting with binary outcomes, no covariates, and incomplete cases. We now verify that the common practice is biased, even when the researcher randomly assigns units to either the experimental or the observational sample.

In this thought experiment, the researcher implements the following study: (i) randomization of units into the experimental or observational sample; (ii) randomization of experimental units to treatment or control; (iii) the common practice described in Section 3.1. No observational units are treated.

To prove this claim, it is helpful to explicitly write the potential outcome as $Y(d)$ and the potential RSV as $R(y)$. The potential RSV is not indexed by the treatment under Assumption 3(ii). Each unit is characterized by the random vector $\{S,D,Y(0),Y(1),R(0),R(1)\}$.

Proposition B.3 (Bias of common practice with random sample selection)**.** Suppose Assumptions 1, 2, and 3(ii) hold with $X=\varnothing$. Further assume

i. SUTVA for RSV: $R=YR(1)+(1-Y)R(0)$ almost surely.

ii. Randomized sample selection: $S \perp\!\!\!\perp \{Y(0),Y(1),R(0),R(1)\}$.

iii. Experimental sample has randomized treatment: $D \perp\!\!\!\perp \{Y(0),Y(1),R(0),R(1)\}|S=e$.

iv. Observational sample is untreated: $\Pr(D=0|S=o)=1$.

Despite randomized sample selection, the bias of common practice persists. In particular, the conclusions of Proposition 1 continue to hold.

Proof. We demonstrate that Proposition B.3's hypotheses imply Proposition 1's hypotheses. Therefore, the conclusions of Proposition 1 continue to hold. Formally, we verify $S \perp\!\!\!\perp (Y,R)|D=1$ and $S \perp\!\!\!\perp (Y,R)|D=0$.

1. By the hypothesis that the observational sample is untreated, $\Pr(S=e|D=1)=1$. Therefore, S becomes a degenerate random variable conditional upon $D=1$ and we conclude that $S \perp\!\!\!\perp (Y,R)|D=1$ trivially.

2. Fix any bounded, measurable function $h(Y,R)$. We show that $\mathbb{E}\{h(Y,R)|D=0,S=e\} = \mathbb{E}\{h(Y,R)|D=0,S=o\}$, which implies $S \perp\!\!\!\perp (Y,R)|D=0$ as desired. By SUTVA, randomized treatment assignment, and randomized sample selection,

$$\begin{aligned}
\mathbb{E}\{h(Y,R)|D=0,S=e\} &= \mathbb{E}[h\{Y,YR(1)+(1-Y)R(0)\}|D=0,S=e] \\
&= \mathbb{E}(h[Y(0),Y(0)R(1)+\{1-Y(0)\}R(0)]|D=0,S=e) \\
&= \mathbb{E}(h[Y(0),Y(0)R(1)+\{1-Y(0)\}R(0)]|S=e) \\
&= \mathbb{E}(h[Y(0),Y(0)R(1)+\{1-Y(0)\}R(0)]).
\end{aligned}$$

By SUTVA, untreated observational units, and randomized sample selection,

$$\begin{aligned}
\mathbb{E}\{h(Y,R)|D=0,S=o\} &= \mathbb{E}[h\{Y,YR(1)+(1-Y)R(0)\}|D=0,S=o] \\
&= \mathbb{E}(h[Y(0),Y(0)R(1)+\{1-Y(0)\}R(0)]|D=0,S=o) \\
&= \mathbb{E}(h[Y(0),Y(0)R(1)+\{1-Y(0)\}R(0)]|S=o) \\
&= \mathbb{E}(h[Y(0),Y(0)R(1)+\{1-Y(0)\}R(0)]).
\end{aligned}$$

□

B.3 Proof of Proposition 1

We characterize the bias. For simplicity, we take $X=\varnothing$. We allow R to be discrete or continuous, and impose that Y is binary for the sake of exposition. When constructing adversarial data-generating processes, we will choose simple examples with binary R, again for the sake of exposition. We proceed in steps.

1. To begin, we write the average potential outcome in the experimental sample as

$$\begin{aligned}\mu(1) &:= \Pr\{Y(1)=1 \mid S=e\} \\ &\overset{(1)}{=} \Pr\{Y(1)=1 \mid D=1, S=e\} \\ &= \Pr(Y=1 \mid D=1, S=e) \\ &= \int \Pr(Y=1 \mid R=r, D=1, S=e) f_R(r \mid D=1, S=e) \mathrm{d}r,\end{aligned}$$

where (1) follows by Assumption 1.

2. Next we rewrite the implicit target

$$\widetilde{\mu}(1) := \int \Pr(Y=1 \mid R=r, S=o) f_R(r \mid D=1, S=e) \mathrm{d}r.$$

Within this expression,

$$\begin{aligned}\Pr(Y=1 \mid R, S=o) &\overset{(1)}{=} \frac{f_R(R \mid Y=1, S=o) \Pr(Y=1 \mid S=o)}{f_R(R \mid S=o)} \\ &\overset{(2)}{=} \frac{f_R(R \mid Y=1, D=1, S=e) P(Y=1 \mid S=o)}{f_R(R \mid S=o)} \\ &\overset{(3)}{=} \Pr(Y=1 \mid R, D=1, S=e) \frac{f_R(R \mid D=1, S=e) \Pr(Y=1 \mid S=o)}{\Pr(Y=1 \mid D=1, S=e) f_R(R \mid S=o)} \\ &\overset{(4)}{=} \Pr(Y=1 \mid R, D=1, S=e) \frac{\Pr(Y=1 \mid S=o)}{\Pr\{Y(1)=1 \mid S=e\}} \frac{f_R(R \mid D=1, S=e)}{f_R(R \mid S=o)}\end{aligned}$$

where (1) follows by Bayes' rule; (2) follows since the contraction axiom gives $(S,D) \perp\!\!\!\perp R \mid Y$ under Assumptions 2 and 3(ii); (3) follows by Bayes' rule; and (4) applies Assumption 1.

We introduce additional structure to simplify the expression. First, we further assume $S \perp\!\!\!\perp (R,Y) \mid D$, which in turn implies $S \perp\!\!\!\perp Y \mid R, D$ by the weak union axiom. Second, since we further assume $\Pr(D=1 \mid S=o) = 0$, we have that

$$\Pr(Y=1 \mid S=o) = \Pr(Y=1 \mid D=0, S=o) = \Pr(Y=1 \mid D=0, S=e) = \Pr\{Y(0)=1 \mid S=e\}$$

where the first equality uses the law of total probability, the second uses $S \perp\!\!\!\perp Y \mid D$, and the third uses Assumption 1. Similarly, $f_R(R \mid D=1, S=e) = f_R(R \mid D=1)$ using $R \perp\!\!\!\perp S \mid D$ and

$$f_R(R \mid S=o) = f_R(R \mid D=0, S=o) = f_R(R \mid D=0)$$

using $\Pr(D=1 \mid S=o) = 0$ and $S \perp\!\!\!\perp R \mid D$.

Therefore, we have that $\widetilde{\mu}(1)$ equals

$$\frac{\Pr\{Y(0)=1 \mid S=e\}}{\Pr\{Y(1)=1 \mid S=e\}} \int \frac{f_R(r \mid D=1)}{f_R(r \mid D=0)} \Pr(Y=1 \mid R=r, D=1, S=e) f_R(r \mid D=1, S=e) \mathrm{d}r.$$

3. Combining results, the bias for the treated potential outcome, $\widetilde{\mu}(1)-\mu(1)$, equals

$$\int\left[\frac{\Pr\{Y(0)=1\,|\,S=e\}}{\Pr\{Y(1)=1\,|\,S=e\}}\frac{f_R(r\,|\,D=1)}{f_R(r\,|\,D=0)}-1\right]\Pr(Y=1\,|\,R=r,D=1,S=e)f_R(r\,|\,D=1,S=e)\mathrm{d}r.$$

Within this expression,

$$\begin{aligned}&\Pr(Y=1\,|\,R,D=1,S=e)f_R(R\,|\,D=1,S=e)=f_{Y,R}(Y=1,R\,|\,D=1,S=e)\\&=f_R(R\,|\,Y=1,D=1,S=e)\Pr(Y=1\,|\,D=1,S=e\}\\&=f_R(R\,|\,Y=1,D=1,S=e)\Pr\{Y(1)=1\,|\,S=e\}\\&=f_R(R\,|\,Y=1,D=1,S=e)\mu(1).\end{aligned}$$

by the law of total probability and Assumption 1.

Consequently, under the additional structure introduced, we have that $\widetilde{\mu}(1)-\mu(1)$ equals

$$\mu(1)\int\left[\frac{\Pr\{Y(0)=1\,|\,S=e\}}{\Pr\{Y(1)=1\,|\,S=e\}}\frac{f_R(r\,|\,D=1)}{f_R(r\,|\,D=0)}-1\right]f_R(r\,|\,Y=1,D=1,S=e)\mathrm{d}r.$$

Finally, since the contraction axiom gives $(S,D)\perp\!\!\!\perp R\,|\,Y$ under Assumptions 2 and 3(ii), the weak union axiom gives $R\perp\!\!\!\perp D|Y,S$ and therefore $f_R(R\,|\,Y=1,D=1,S=e)=f_R(R\,|\,Y=1,S=e)$.

4. We can follow a similar argument for $\widetilde{\mu}(0)$. As in Step 2,

$$\begin{aligned}&\widetilde{\mu}(0)=\int\Pr(Y=1\,|\,R=r,S=o)f_R(r\,|\,D=0,S=e)\mathrm{d}r\\&=\int\Pr(Y=1\,|\,R=r,D=0,S=e)\frac{\Pr(Y=1\,|\,S=o)}{\Pr\{Y(0)=1\,|\,S=e\}}\frac{f_R(r\,|\,D=0,S=e)}{f_R(r\,|\,S=o)}f_R(r\,|\,D=0,S=e)\mathrm{d}r\end{aligned}$$

by identical arguments as before, replacing $D=1$ with $D=0$. We have previously shown $\Pr(Y=1\,|\,S=o)=\Pr\{Y(0)=1\,|\,S=e\}$. Moreover, $f_R(R\,|\,D=0,S=e)=f_R(R\,|\,D=0)$ since $R\perp\!\!\!\perp S|D$, and we have shown $f_R(R\,|\,D=0)=f_R(R|S=o)$. Therefore,

$$\widetilde{\mu}(0)=\int\Pr(Y=1\,|\,R=r,D=0,S=e)f_R(r\,|\,D=0,S=e)\mathrm{d}r=\mu(0)$$

where the second equality uses the same argument as Step 1.

5. Since $\widetilde{\mu}(0)=\mu(0)$, we conclude that $\widetilde{\theta}-\theta=\widetilde{\mu}(1)-\mu(1)$ under the stated conditions.

6. Next we demonstrate the bias can be positive or negative by constructing the claimed DGPs. Suppose that R is binary with

$$R\,|\,Y,D,S=\begin{cases}Y\text{ with probability }1/2\\1\text{ otherwise.}\end{cases}$$

This process implies $R \perp\!\!\!\perp (D,S) | Y$, which satisfies $R \perp\!\!\!\perp S \,|\, D,Y$ (Assumption 2) by the weak union axiom. It also directly satisfies $R \perp\!\!\!\perp D \,|\, Y$ (Assumption 3(ii)).

Under this DGP,

$$\Pr(R=1 \,|\, Y=1,D=1,S=e)=1, \quad \Pr(R=0 \,|\, Y=1,D=1,S=e)=0.$$

Consequently, the bias of the implicit target simplifies to

$$\left[\frac{\Pr\{Y(0)=1 \,|\, S=e\}}{\Pr\{Y(1)=1 \,|\, S=e\}} \frac{\Pr(R=1 \,|\, D=1)}{\Pr(R=1 \,|\, D=0)} - 1\right] \mu(1).$$

Furthermore, under this DGP,

$$\Pr(R=1 \,|\, Y=1,D=d,S=e)=1, \quad \Pr(R=1 \,|\, Y=0,D=d,S=e)=\frac{1}{2}.$$

By similar arguments to those above, we therefore have

$$\begin{aligned}
\Pr(R=1 \,|\, D=d) &= \Pr(R=1 \,|\, D=d,S=e) \\
&= \int \Pr(R=1,Y=y \,|\, D=d,S=e) \mathrm{d}y \\
&= \int \Pr(R=1 \,|\, Y=y,D=d,S=e) \Pr(Y=y \,|\, D=d,S=e) \mathrm{d}y \\
&= \Pr(Y=1 \,|\, D=d,S=e) + \frac{1}{2} \Pr(Y=0 \,|\, D=d,S=e) \\
&= \Pr\{Y(d)=1 \,|\, S=e\} + \frac{1}{2} \Pr\{Y(d)=0 \,|\, S=e\} \\
&= \frac{1}{2}[1 + \Pr\{Y(d)=1 \,|\, S=e\}]
\end{aligned}$$

using the extra structure $S \perp\!\!\!\perp R \mid D$, the law of total probability, the specific DGP, and Assumption 1. Therefore,

$$\frac{\Pr(R=1 \,|\, D=1)}{\Pr(R=1 \,|\, D=0)} = \frac{1+\Pr\{Y(1)=1 \,|\, S=e\}}{1+\Pr\{Y(0)=1 \,|\, S=e\}}$$

and so we have shown that $\widetilde{\mu}(1) - \mu(1)$ equals

$$\left[\frac{\Pr\{Y(0)=1 \,|\, S=e\}}{\Pr\{Y(1)=1 \,|\, S=e\}} \cdot \frac{1+\Pr\{Y(1)=1 \,|\, S=e\}}{1+\Pr\{Y(0)=1 \,|\, S=e\}} - 1\right] \Pr\{Y(1)=1 \,|\, S=e\}.$$

Lightening notation by writing $a := \Pr\{Y(0)=1 \,|\, S=e\}$ and $b := \Pr\{Y(1)=1 \,|\, S=e\}$,

$$\widetilde{\mu}(1) - \mu(1) = \left(\frac{a}{b} \cdot \frac{1+b}{1+a} - 1\right) b = \frac{a-b}{a+1},$$

which is positive when $a > b$ and negative otherwise. □

C Identification Proofs

C.1 Proof of Lemma 1

To begin, we consider the general case where (X,Y,R) may be discrete or continuous. Then, we specialize the result to the case where Y is binary. Recall that $f_W(\cdot|...)$ is the Radon-Nikodym derivative.

By the law of total probability,

$$\begin{aligned}\delta_d^e(r,x) &= f_R(r\,|\,S=e,X=x,D=d)\\ &= \int f_{R,Y}(r,y\,|\,S=e,X=x,D=d)\mathrm{d}y\\ &= \int f_R(r\,|\,S=e,X=x,D=d,Y=y)f_Y(y\,|\,S=e,X=x,D=d)\mathrm{d}y.\end{aligned}$$

By Assumption 1, $f_Y(y\,|\,S=e,X=x,D=d)=f_{Y(d)}(y\,|\,S=e,X=x)$. Next, notice that

$$\begin{aligned}f_R(r\,|\,S=e,X=x,D=d,Y=y) &= f_R(r\,|\,S=o,X=x,D=d,Y=y)\\ &= f_R(r\,|\,S=o,X=x,Y=y)\\ &= \delta_y^o(r,x),\end{aligned}$$

where the first equality applies Assumption 2 and the second equality applies Assumption 3(ii).

Combining the previous displays, we arrive at the general result

$$\delta_d^e(r,x) = \int \delta_y^o(r,x) f_{Y(d)}(y\,|\,S=e,X=x)\mathrm{d}y. \tag{3}$$

When Y is binary,

$$\begin{aligned}f_{Y(d)}(1\,|\,S=e,X=x) &= \mathbb{E}\{Y^{(d)}\,|\,S=e,X=x\} = \mu(d,x)\\ f_{Y(d)}(0\,|\,S=e,X=x) &= 1-\mathbb{E}\{Y^{(d)}\,|\,S=e,X=x\} = 1-\mu(d,x).\end{aligned}$$

Therefore, the general result specializes to

$$\delta_d^e(r,x) = \delta_1^o(r,x)\mu(d,x)+\delta_0^o(r,x)\{1-\mu(d,x)\} = \delta_0^o(r,x)+\{\delta_1^o(r,x)-\delta_0^o(r,x)\}\mu(d,x). \qquad \square$$

C.2 Proof of Theorem 1

Recall the notation from Lemma 1. To prove this result, we apply Bayes' rule to rewrite

$$\begin{aligned}\delta_y^o(r,x) &:= f_R(r\,|\,Y=y,S=o,X=x) = \frac{\Pr(Y=y,S=o\,|\,R=r,X=x)f_R(r|X=x)}{\Pr(Y=y,S=o|X=x)}, \qquad (4)\\ \delta_d^e(r,x) &:= f_R(r\,|\,D=d,S=e,X=x) = \frac{\Pr(D=d,S=e\,|\,R=r,X=x)f_R(r|X=x)}{\Pr(D=d,S=e|X=x)}.\end{aligned}$$

Applying Lemma 1 and canceling $f_R(r|X=x)$, we have for $d\in\{0,1\}$, $x\in\mathcal{X}$, and $r\in\mathcal{R}$,

$$\begin{aligned}&\frac{\Pr(D=d,S=e\,|\,R=r,X=x)}{\Pr(D=d,S=e|X=x)}-\frac{\Pr(Y=0,S=o\,|\,R=r,X=x)}{\Pr(Y=0,S=o|X=x)}\\&=\left\{\frac{\Pr(Y=1,S=o\,|\,R=r,X=x)}{\Pr(Y=1,S=o|X=x)}-\frac{\Pr(Y=0,S=o\,|\,R=r,X=x)}{\Pr(Y=0,S=o|X=x)}\right\}\mu(d,x).\end{aligned}$$

By iterated expectations, this can be further rewritten as

$$\begin{aligned}&\mathbb{E}\left[\frac{1\{D=d,S=e\}}{\Pr(D=d,S=e|X=x)}-\frac{1\{Y=0,S=o\}}{\Pr(Y=0,S=o|X=x)}\,|\,R=r,X=x\right]\\&=\mathbb{E}\left[\frac{1\{Y=1,S=o\}}{\Pr(Y=1,S=o|X=x)}-\frac{1\{Y=0,S=o\}}{\Pr(Y=1,S=o|X=x)}\,|\,R=r,X=x\right]\mu(d,x).\end{aligned}$$

Consequently, it immediately follows that

$$\begin{aligned}&\mathbb{E}\left[\frac{1\{D=1,S=e\}}{\Pr(D=1,S=e|X=x)}-\frac{1\{D=0,S=e\}}{\Pr(D=0,S=e|X=x)}\,|\,R=r,X=x\right]\\&=\mathbb{E}\left[\frac{1\{Y=1,S=o\}}{\Pr(Y=1,S=o|X=x)}-\frac{1\{Y=0,S=o\}}{\Pr(Y=1,S=o|X=x)}\,|\,R=r,X=x\right]\theta(x),\end{aligned}$$

proving the result as desired. □

C.3 Proof of Theorem 2

The proof follows the same steps as the proofs of Lemma 1 and Theorem 1.

1. As in Lemma 1, for general (X,Y,R) and for $\delta^e_d(r,x):=f_R(r\,|\,S=e,X=x,D=d)$,

$$\delta^e_d(r,x)=\int f_R(r\,|\,S=e,X=x,D=d,Y=y)f_Y(y\,|\,S=e,X=x,D=d)\mathrm{d}y.$$

By Assumption 1, $f_Y(y\,|\,S=e,X=x,D=d)=f_{Y(d)}(y\,|\,S=e,X=x)$. Now, however,

$$f_R(r\,|\,S=e,X=x,D=d,Y=y)=f_R(r\,|\,S=o,X=x,D=d,Y=y)=:\delta^o_{y,d}(r,x),$$

where the first equality applies Assumption 2 and the second equality applies Assumption 3(i).

Combining the previous displays, we arrive at the general result

$$\delta^e_d(r,x)=\int\delta^o_{y,d}(r,x)f_{Y(d)}(y\,|\,S=e,X=x)\mathrm{d}y. \tag{5}$$

When Y is binary, as before

$$f_{Y(d)}(1\,|\,S=e,X=x)=\mu(d,x),\quad f_{Y(d)}(0\,|\,S=e,X=x)=1-\mu(d,x).$$

Therefore, the general result specializes to

$$\delta^e_d(r,x)=\delta^o_{1,d}(r,x)\mu(d,x)+\delta^o_{0,d}(r,x)\{1-\mu(d,x)\}=\delta^o_{0,d}(r,x)+\mu(d,x)\{\delta^o_{1,d}(r,x)-\delta^o_{0,d}(r,x)\}.$$

2. As in Theorem 1, we next apply Bayes' rule to rewrite

$$\delta^o_{y,d}(r,x) = \frac{\Pr(Y=y,D=d,S=o\,|\,R=r,X=x)f_R(r|X=x)}{\Pr(Y=y,D=d,S=o|X=x)} \tag{6}$$
$$\delta^e_d(r,x) = \frac{\Pr(D=d,S=e\,|\,R=r,X=x)f_R(r|X=x)}{\Pr(D=d,S=e|X=x)}.$$

It therefore follows by canceling $f_R(r|X=x)$ that, for $d\in\{0,1\}$ and $r\in\mathcal{R}$,

$$\begin{aligned}&\frac{\Pr(D=d,S=e\,|\,R=r,X=x)}{\Pr(D=d,S=e|X=x)} - \frac{\Pr(Y=0,D=d,S=o\,|\,R=r,X=x)}{\Pr(Y=0,D=d,S=o|X=x)}\\ &= \left\{\frac{\Pr(Y=1,D=d,S=o\,|\,R=r,X=x)}{\Pr(Y=1,D=d,S=o|X=x)} - \frac{\Pr(Y=0,D=d,S=o\,|\,R=r,X=x)}{\Pr(Y=0,D=d,S=o|X=x)}\right\}\mu(d,x).\end{aligned}$$

By iterated expectations, this can be further rewritten as

$$\begin{aligned}&\mathbb{E}\left[\frac{1\{D=d,S=e\}}{\Pr(D=d,S=e|X=x)} - \frac{1\{Y=0,D=d,S=o\}}{\Pr(Y=0,D=d,S=o|X=x)}\,|\,R=r,X=x\right]\\ &= \mathbb{E}\left[\frac{1\{Y=1,D=d,S=o\}}{\Pr(Y=1,D=d,S=o|X=x)} - \frac{1\{Y=0,D=d,S=o\}}{\Pr(Y=0,D=d,S=o|X=x)}\,|\,R=r,X=x\right]\mu(d,x).\end{aligned}$$

The result then follows immediately. □

D Estimation and Inference Proofs

D.1 Algorithm Details

Algorithm D.1 (Inference: Details)**.** Given $\{S_i, 1\{S_i=e\}D_i, 1\{S_i=o\}Y_i, R_i\}$:

1. Divide the sample into TRAIN and TEST folds.

2. Learn the representation on TRAIN: $\widehat{H}(R)$.

(a) Count marginals: $\text{COUNT}_{Y=1,S=o}$, $\text{COUNT}_{Y=0,S=o}$, $\text{COUNT}_{D=1,S=e}$, $\text{COUNT}_{D=0,S=e}$.

(b) Train predictors: $\text{PRED}_Y(R)$ estimates $\Pr(Y=1|S=o,R)$, $\text{PRED}_D(R)$ estimates $\Pr(D=1|S=e,R)$, and $\text{PRED}_S(R)$ estimates $\Pr(S=e|R)$, using machine learning.

(c) Initially estimate $\widehat{\theta}_{\text{INIT}} = \text{argmin}_\theta \mathbb{E}_{\text{TRAIN}}[\{\widehat{\mathbb{E}}(\Delta^e|R) - \widehat{\mathbb{E}}(\Delta^o|R)\theta\}^2]$, where $\widehat{\mathbb{E}}(\Delta^e|R)$ and $\widehat{\mathbb{E}}(\Delta^o|R)$ are constructed from the marginal probabilities and predictors according to (1):

$$\begin{aligned}\widehat{\mathbb{E}}(\Delta^e|R) &= \left[\frac{\text{PRED}_D(R)}{\text{COUNT}_{D=1,S=e}} - \frac{1-\text{PRED}_D(R)}{\text{COUNT}_{D=0,S=e}}\right]\text{PRED}_S(R)\\ \widehat{\mathbb{E}}(\Delta^o|R) &= \left[\frac{\text{PRED}_Y(R)}{\text{COUNT}_{Y=1,S=o}} - \frac{1-\text{PRED}_Y(R)}{\text{COUNT}_{Y=0,S=o}}\right]\{1-\text{PRED}_S(R)\}.\end{aligned}$$

(d) Learn the representation: $\widehat{H}(R) = \frac{\widehat{\mathbb{E}}(\Delta^o|R)}{\widehat{\sigma}^2(\widehat{\theta}_{\text{INIT}},R)}$ where $\widehat{\sigma}^2(\widehat{\theta}_{\text{INIT}},R)$ is constructed from the marginal probabilities, predictors, and initial estimate via Lemma 2:

$$\widehat{\sigma}^2(\widehat{\theta}_{\text{INIT}},R) = \left[\frac{\text{PRED}_D(R)}{\text{COUNT}^2_{D=1,S=e}} + \frac{1-\text{PRED}_D(R)}{\text{COUNT}^2_{D=0,S=e}}\right]\text{PRED}_S(R) + \widehat{\theta}^2_{\text{INIT}}\left[\frac{\text{PRED}_Y(R)}{\text{COUNT}^2_{Y=1,S=o}} + \frac{1-\text{PRED}_Y(R)}{\text{COUNT}^2_{Y=0,S=o}}\right]\{1-\text{PRED}_S(R)\}.$$

3. Construct a causal estimate on TEST: $\widehat{\theta}$.

(a) Count marginals: $\text{COUNT}_{Y=1,S=o}$, $\text{COUNT}_{Y=0,S=o}$, $\text{COUNT}_{D=1,S=e}$, $\text{COUNT}_{D=0,S=e}$.

(b) Construct a causal estimate: $\widehat{\theta} = \frac{\mathbb{E}_{\text{TEST}}\{\widehat{\Delta}^e\widehat{H}(R)\}}{\mathbb{E}_{\text{TEST}}\{\widehat{\Delta}^o\widehat{H}(R)\}}$ where $\widehat{\Delta}^e$ and $\widehat{\Delta}^o$ are constructed from marginal probabilities according to (1):

$$\widehat{\Delta}^e = \frac{1\{D=1,S=e\}}{\text{COUNT}_{D=1,S=e}} - \frac{1\{D=0,S=e\}}{\text{COUNT}_{D=0,S=e}}, \quad \widehat{\Delta}^o = \frac{1\{Y=1,S=o\}}{\text{COUNT}_{Y=1,S=o}} - \frac{1\{Y=0,S=o\}}{\text{COUNT}_{Y=0,S=o}}.$$

(c) Bootstrap its confidence interval: $\widehat{\theta} \pm c_\alpha \hat{v} n^{-1/2}$, where c_α is the $1-\alpha/2$ quantile of the standard Gaussian and $\hat{v}n^{-1/2}$ is the bootstrap standard error of $\widehat{\theta}$ while fixing $\widehat{H}(R)$.

D.2 Proof of Lemma 2

Since the events in Δ^e are exclusive of the events in Δ^o,

$$(\Delta^e - \Delta^o\theta)^2 = (\Delta^e)^2 - (2\theta)\Delta^e\Delta^o + \theta^2(\Delta^o)^2 = (\Delta^e)^2 + \theta^2(\Delta^o)^2.$$

Similarly, since the events within Δ^e are exclusive of each other,

$$(\Delta^e)^2 = \left[\frac{1\{D=1,S=e\}}{\Pr(D=1,S=e)}\right]^2 + \left[\frac{1\{D=0,S=e\}}{\Pr(D=0,S=e)}\right]^2 = \frac{1\{D=1,S=e\}}{\Pr(D=1,S=e)^2} + \frac{1\{D=0,S=e\}}{\Pr(D=0,S=e)^2}.$$

The argument for $(\Delta^o)^2$ is similar. □

D.3 Proof of Proposition 2

To lighten notation, let $Y = \Delta^e$, $X = \Delta^o$, $U = Y - X\theta$, $Z = \widetilde{H}(R)$, and $\hat{Z} = \widehat{H}(R)$. By construction, $\mathbb{E}(U^2|Z) \leq \bar{\sigma}^2_U$ for some constant $\bar{\sigma}^2_U < \infty$ due to Lemma 2 and the assumption that marginal probabilities are bounded away from zero. Moreover, $|X| \leq \bar{X}$ almost surely for some constant $\bar{X} < \infty$ due to the assumption that marginal probabilities are bounded away from zero.

With known marginal probabilities, the argument uses standard techniques, similar to Mackey et al. (2018); Chen et al. (2020). In this lighter notation,

$$n^{1/2}(\hat{\theta} - \theta) = n^{1/2}\left\{\frac{\mathbb{E}_n(Y\hat{Z})}{\mathbb{E}_n(X\hat{Z})} - \theta\right\} = \frac{n^{1/2}\mathbb{E}_n(U\hat{Z})}{\mathbb{E}_n(X\hat{Z})}.$$

Focusing on the numerator, if $n^{1/2}\mathbb{E}_n(U\hat{Z}) = n^{1/2}\mathbb{E}_n(UZ) + o_p(1)$ and $n^{1/2}\mathbb{E}_n(UZ) \rightsquigarrow \mathcal{N}\{0,\mathbb{E}(U^2Z^2)\}$, then $n^{1/2}\mathbb{E}_n(U\hat{Z}) \rightsquigarrow \mathcal{N}\{0,\mathbb{E}(U^2Z^2)\}$ by Slutsky's theorem .

Focusing on the denominator, if $\mathbb{E}_n(X\hat{Z}) = \mathbb{E}_n(XZ) + o_p(1)$ and $\mathbb{E}_n(XZ) = \mathbb{E}(XZ) + o_p(1)$ then $\mathbb{E}_n(X\hat{Z}) = \mathbb{E}(XZ) + o_p(1)$ by the continuous mapping theorem.

Overall, we conclude that $n^{1/2}(\hat{\theta} - \theta) \rightsquigarrow \mathcal{N}\left[0, \frac{\mathbb{E}(U^2Z^2)}{\{\mathbb{E}(XZ)\}^2}\right]$ by Slutsky's theorem. In the following lemmas, we prove these high probability statements. □

While proving the results, we use standard notation for cross-fitting. Let there be L folds, each denoted by I_ℓ with $\ell \in [L]$. Each fold contains $n_\ell = n/L$ observations. The complement of I_ℓ is $I_{-\ell}$. If $i \in I_\ell$, then $\hat{Z}_i = \widehat{H}_\ell(R_i)$ is constructed from $\widehat{H}_\ell$ estimated on the remaining folds $I_{-\ell}$.

Lemma D.1. Under Proposition 2's conditions, $n^{1/2}\mathbb{E}_n(U\hat{Z}) = n^{1/2}\mathbb{E}_n(UZ) + o_p(1)$.

Proof. We express the difference as

$$n^{1/2}\mathbb{E}_n\{U(\hat{Z} - Z)\} = n^{1/2}\frac{1}{L}\frac{1}{n_\ell}\sum_{\ell=1}^{L}\sum_{i\in I_\ell} U_i(\hat{Z}_i - Z_i) = L^{1/2}\frac{1}{L}\sum_{\ell=1}^{L} n_\ell^{1/2}\frac{1}{n_\ell}\sum_{i\in I_\ell} U_i(\hat{Z}_i - Z_i).$$

Focusing on the foldwise quantity, it suffices to control

$$\mathbb{E}\left[\left\{n_\ell^{1/2}\frac{1}{n_\ell}\sum_{i\in I_\ell} U_i(\hat{Z}_i - Z_i)\right\}^2\right] = \mathbb{E}\left(\mathbb{E}\left[\left\{n_\ell^{1/2}\frac{1}{n_\ell}\sum_{i\in I_\ell} U_i(\hat{Z}_i - Z_i)\right\}^2 \mid I_{-\ell}\right]\right).$$

Due to cross-fitting, the inner expectation is

$$\begin{aligned}
&\mathbb{E}\left[\left\{n_\ell^{1/2}\frac{1}{n_\ell}\sum_{i\in I_\ell} U_i(\hat{Z}_i - Z_i)\right\}^2 \mid I_{-\ell}\right] = \frac{1}{n_\ell}\mathbb{E}\left\{\sum_{i,j\in I_\ell} U_i(\hat{Z}_i - Z_i)U_j(\hat{Z}_j - Z_j) \mid I_{-\ell}\right\} \\
&= \frac{1}{n_\ell}\mathbb{E}\left\{\sum_{i\in I_\ell} U_i^2(\hat{Z}_i - Z_i)^2 \mid I_{-\ell}\right\} = \mathbb{E}\{U_i^2(\hat{Z}_i - Z_i)^2 \mid I_{-\ell}\} = \mathbb{E}\{\mathbb{E}(U_i^2 \mid Z_i, I_{-\ell})(\hat{Z}_i - Z_i)^2 \mid I_{-\ell}\} \\
&\leq \bar{\sigma}_U^2\mathbb{E}\{(\hat{Z}_i - Z_i)^2 \mid I_{-\ell}\} = \bar{\sigma}_U^2\mathcal{R}(\hat{Z}) = o_p(1).
\end{aligned}$$

In the inequality, we use $\mathbb{E}(U_i^2 \mid Z_i, I_{-\ell}) = \mathbb{E}(U_i^2 \mid Z_i) \leq \bar{\sigma}_U^2$, where $\bar{\sigma}_U^2 < \infty$ under our assumptions. In the final equality, we write $\mathcal{R}(\hat{Z}) = \mathbb{E}\{(\hat{Z}_i - Z_i)^2 \mid I_{-\ell}\} = o_p(1)$ for the mean square limit. □

Lemma D.2. Under Proposition 2's conditions, $n^{1/2}\mathbb{E}_n(UZ) \rightsquigarrow \mathcal{N}\{0,\mathbb{E}(U^2Z^2)\}$.

Proof. By Theorem 1, $\mathbb{E}(U_iZ_i) = \mathbb{E}\{\mathbb{E}(U_i|Z_i)Z_i\} = 0$. Moreover, $\mathbb{E}(U_i^2Z_i^2) = \mathbb{E}\{\mathbb{E}(U_i^2|Z_i)Z_i^2\} \leq \bar{\sigma}_U^2\mathbb{E}(Z_i^2)$ since $\mathbb{E}(U_i^2|Z_i) \leq \bar{\sigma}_U^2$ under our assumptions. Here, $\mathbb{E}(Z_i^2)$ is finite by Assumption 4, so we apply the Lindeberg-Levy central limit theorem. □

Lemma D.3. Under Proposition 2's conditions, $\mathbb{E}_n(X\hat{Z}) = \mathbb{E}_n(XZ) + o_p(1)$.

Proof. We express the difference as

$$\mathbb{E}_n\{X(\hat{Z}-Z)\}=\frac{1}{L}\frac{1}{n_\ell}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}X_i(\hat{Z}_i-Z_i)=\frac{1}{L}\sum_{\ell=1}^{L}\frac{1}{n_\ell}\sum_{i\in I_\ell}X_i(\hat{Z}_i-Z_i).$$

Focusing on the foldwise quantity, it suffices to control

$$\mathbb{E}\left\{\left|\frac{1}{n_\ell}\sum_{i\in I_\ell}X_i(\hat{Z}_i-Z_i)\right|\right\}=\mathbb{E}\left[\mathbb{E}\left\{\left|\frac{1}{n_\ell}\sum_{i\in I_\ell}X_i(\hat{Z}_i-Z_i)\right||I_{-\ell}\right\}\right].$$

The inner expectation is

$$\begin{aligned}&\mathbb{E}\left\{\left|\frac{1}{n_\ell}\sum_{i\in I_\ell}X_i(\hat{Z}_i-Z_i)\right||I_{-\ell}\right\}\leq\mathbb{E}\left\{\frac{1}{n_\ell}\sum_{i\in I_\ell}\left|X_i(\hat{Z}_i-Z_i)\right||I_{-\ell}\right\}\leq\mathbb{E}\left\{\frac{\bar{X}}{n_\ell}\cdot\sum_{i\in I_\ell}\left|\hat{Z}_i-Z_i\right||I_{-\ell}\right\}\\&=\bar{X}\mathbb{E}\left\{\left|\hat{Z}_i-Z_i\right||I_{-\ell}\right\}\leq\bar{X}[\mathbb{E}\{(\hat{Z}_i-Z_i)^2|I_{-\ell}\}]^{1/2}=\bar{X}\mathcal{R}(\hat{Z})^{1/2}=o_p(1).\end{aligned}$$

We use $|X_i|\leq\bar{X}$ almost surely, since marginal probabilities are bounded away from zero. □

Lemma D.4. Under Proposition 2's conditions, $\mathbb{E}_n(XZ)=\mathbb{E}(XZ)+o_p(1)$.

Proof. By Chebyshev's inequality, it suffices to bound $\mathbb{V}(X_iZ_i)\leq\mathbb{E}(X_i^2Z_i^2)\leq\bar{X}^2\mathbb{E}(Z_i^2)$. In summary, we use $|X_i|\leq\bar{X}$ almost surely since marginal probabilities are bounded away from zero and $\mathbb{E}(Z_i^2)<\infty$ due to Assumption 4. □

D.4 Proof of Proposition 3

With unknown marginal probabilities, some extra care is required. Extending the notation from the proof of Proposition 2, if $\hat{Y}=\widehat{\Delta}^e$, $\hat{X}=\widehat{\Delta}^o$, and $\hat{U}=\hat{Y}-\hat{X}\theta$, then

$$n^{1/2}(\hat{\theta}-\theta)=n^{1/2}\left\{\frac{\mathbb{E}_n(\hat{Y}\hat{Z})}{\mathbb{E}_n(\hat{X}\hat{Z})}-\theta\right\}=\frac{n^{1/2}\mathbb{E}_n(\hat{U}\hat{Z})}{\mathbb{E}_n(\hat{X}\hat{Z})}.$$

Focusing on the numerator, if $n^{1/2}\mathbb{E}_n(\hat{U}\hat{Z})=n^{1/2}\mathbb{E}_n(\hat{U}Z)+o_p(1)$ and $n^{1/2}\mathbb{E}_n(\hat{U}Z)\rightsquigarrow\mathcal{N}(0,V)$, then $n^{1/2}\mathbb{E}_n(U\hat{Z})\rightsquigarrow\mathcal{N}(0,V)$ by Slutsky's theorem.

Focusing on the denominator, if $\mathbb{E}_n(\hat{X}\hat{Z})=\mathbb{E}_n(\hat{X}Z)+o_p(1)$, $\mathbb{E}_n(\hat{X}Z)=\mathbb{E}_n(XZ)+o_p(1)$, and $\mathbb{E}_n(XZ)=\mathbb{E}(XZ)+o_p(1)$, then $\mathbb{E}_n(\hat{X}\hat{Z})=\mathbb{E}(XZ)+o_p(1)$ by the continuous mapping theorem.

Overall, we conclude that $n^{1/2}(\hat{\theta}-\theta)\rightsquigarrow\mathcal{N}\left[0,\frac{V}{\{\mathbb{E}(XZ)\}^2}\right]$ by Slutsky's theorem. In the following lemmas, we prove the remaining high probability statements and derive V. □

Lemma D.5. Under Proposition 3's conditions, $n^{1/2}\mathbb{E}_n(\hat{U}\hat{Z})=n^{1/2}\mathbb{E}_n(\hat{U}Z)+o_p(1)$.

Proof. The argument is similar to Lemma D.1, using $|\hat{U}_i| \leq \bar{U}'$ almost surely, which follows from our assumptions. □

Lemma D.6. Under Proposition 3's conditions, $n^{1/2}\mathbb{E}_n(\hat{U}Z) \rightsquigarrow \mathcal{N}(0,V)$ where $V = v^\top \Sigma v$, $\Sigma_{ij} = cov(B_i, B_j)$, and

$$v = \begin{pmatrix} \mathbb{E}(B_5)^{-1} \\ -\mathbb{E}(B_6)^{-1} \\ -\theta\mathbb{E}(B_7)^{-1} \\ \theta\mathbb{E}(B_8)^{-1} \\ -\mathbb{E}(B_1)\mathbb{E}(B_5)^{-2} \\ \mathbb{E}(B_2)\mathbb{E}(B_6)^{-2} \\ \theta\mathbb{E}(B_3)\mathbb{E}(B_7)^{-2} \\ -\theta\mathbb{E}(B_4)\mathbb{E}(B_8)^{-2} \end{pmatrix}, \quad B = \begin{pmatrix} 1_{D=1,S=e}Z \\ 1_{D=0,S=e}Z \\ 1_{Y=1,S=o}Z \\ 1_{Y=0,S=o}Z \\ 1_{D=1,S=e} \\ 1_{D=0,S=e} \\ 1_{Y=1,S=o} \\ \mathbb{1}_{Y=0,S=o} \end{pmatrix}.$$

Proof. We unpack the definition of $\hat{U}$:

$$\hat{U} = \hat{Y} - \hat{X}\theta = \widehat{\Delta}^e - \widehat{\Delta}^o\theta = \frac{1_{D=1,S=e}}{\mathbb{E}_n(1_{D=1,S=e})} - \frac{1_{D=0,S=e}}{\mathbb{E}_n(1_{D=0,S=e})} - \left\{ \frac{1_{Y=1,S=o}}{\mathbb{E}_n(1_{Y=1,S=o})} - \frac{1_{Y=0,S=o}}{\mathbb{E}_n(1_{Y=0,S=o})} \right\}\theta.$$

Therefore, for $h(B) = \frac{B_1}{B_5} - \frac{B_2}{B_6} - \theta\frac{B_3}{B_7} + \theta\frac{B_4}{B_8}$,

$$\begin{aligned} n^{1/2}\mathbb{E}_n(\hat{U}Z) &= n^{1/2}\left[\frac{\mathbb{E}_n(1_{D=1,S=e}Z)}{\mathbb{E}_n(1_{D=1,S=e})} - \frac{\mathbb{E}_n(1_{D=0,S=e}Z)}{\mathbb{E}_n(1_{D=0,S=e})} - \left\{ \frac{\mathbb{E}_n(1_{Y=1,S=o}Z)}{\mathbb{E}_n(1_{Y=1,S=o})} - \frac{\mathbb{E}_n(1_{Y=0,S=o}Z)}{\mathbb{E}_n(1_{Y=0,S=o})} \right\}\theta \right] \\ &= n^{1/2} h\{\mathbb{E}_n(B)\}. \end{aligned}$$

We make three observations. First, $\mathbb{E}(Z^2) < \infty$ by Assumption 4, so by the Lindeberg-Levy central limit theorem, $n^{1/2}\{\mathbb{E}_n(B) - \mathbb{E}(B)\} \rightsquigarrow \mathcal{N}(0,\Sigma)$.

Second, by the conditional moment equation,

$$\begin{aligned} h\{\mathbb{E}(B)\} &= \frac{\mathbb{E}(1_{D=1,S=e}Z)}{\mathbb{E}(1_{D=1,S=e})} - \frac{\mathbb{E}(1_{D=0,S=e}Z)}{\mathbb{E}(1_{D=0,S=e})} - \theta\frac{\mathbb{E}(1_{Y=1,S=o}Z)}{\mathbb{E}(1_{Y=1,S=o})} + \theta\frac{\mathbb{E}(1_{Y=0,S=o}Z)}{\mathbb{E}(1_{Y=0,S=o})} \\ &= \mathbb{E}\{(\Delta^e - \theta\Delta^o)Z\} = \mathbb{E}[\mathbb{E}\{(\Delta^e - \theta\Delta^o)|Z\}Z] = 0. \end{aligned}$$

Third, the derivative is

$$\{\nabla h(B)\}^\top = \left(B_5^{-1}, \ -B_6^{-1}, \ -\theta B_7^{-1}, \ \theta B_8^{-1}, \ -B_1 B_5^{-2}, \ B_2 B_6^{-2}, \ \theta B_3 B_7^{-2}, \ -\theta B_4 B_8^{-2} \right).$$

Therefore, by the delta method,

$$n^{1/2}\mathbb{E}_n(\hat{U}Z) = n^{1/2}[h\{\mathbb{E}_n(B)\} - h\{\mathbb{E}(B)\}] \rightsquigarrow \mathcal{N}(0, [\nabla h\{\mathbb{E}(B)\}]^\top \Sigma [\nabla h\{\mathbb{E}(B)\}]).$$

□

Lemma D.7. Under Proposition 3's conditions, $\mathbb{E}_n(\hat{X}\hat{Z}) = \mathbb{E}_n(\hat{X}Z) + o_p(1)$.

Proof. The argument is similar to Lemma D.3, using $|\hat{X}_i| \leq \bar{X}'$ almost surely, which follows from our assumptions. □

Lemma D.8. Under Proposition 3's conditions, $\mathbb{E}_n(\hat{X}Z) = \mathbb{E}_n(XZ) + o_p(1)$.

Proof. We express the difference as

$$\mathbb{E}_n(\hat{X}Z) - \mathbb{E}_n(XZ) = \mathbb{E}_n\{(\hat{X}-X)Z\} \leq [\mathbb{E}_n\{(\hat{X}-X)^2\}]^{1/2}\{\mathbb{E}_n(Z^2)\}^{1/2}.$$

Since $\mathbb{E}_n(Z^2) = \mathbb{E}(Z^2) + o_p(1)$ when $\mathbb{E}(Z^2) < \infty$ by the weak law of large numbers, it suffices to study

$$\mathbb{E}_n\{(\hat{X}-X)^2\} = \frac{1}{L}\sum_{\ell=1}^{L}\frac{1}{n_\ell}\sum_{i\in I_\ell}(\hat{X}_i - X_i)^2 = \frac{1}{L}\sum_{\ell=1}^{L}\mathbb{E}_\ell\{(\hat{X}_i - X_i)^2\}, \quad \mathbb{E}_\ell(\cdot) = \frac{1}{n_\ell}\sum_{i\in I_\ell}(\cdot).$$

Unpacking the notation of $\mathbb{E}_\ell\{(\hat{X}_i - X_i)^2\}$,

$$\begin{aligned}
\hat{X} - X = \hat{\Delta}^o - \Delta^o &= \frac{1_{Y=1,S=o}}{\mathbb{E}_\ell(1_{Y=1,S=o})} - \frac{1_{Y=0,S=o}}{\mathbb{E}_\ell(1_{Y=0,S=o})} - \left\{\frac{1_{Y=1,S=o}}{\mathbb{E}(1_{Y=1,S=o})} - \frac{1_{Y=0,S=o}}{\mathbb{E}(1_{Y=0,S=o})}\right\} \\
&= 1_{Y=1,S=o}\left\{\frac{1}{\mathbb{E}_\ell(1_{Y=1,S=o})} - \frac{1}{\mathbb{E}(1_{Y=1,S=o})}\right\} - 1_{Y=0,S=o}\left\{\frac{1}{\mathbb{E}_\ell(1_{Y=0,S=o})} - \frac{1}{\mathbb{E}(1_{Y=0,S=o})}\right\} \\
&= 1_{Y=1,S=o}\left\{\frac{\mathbb{E}(1_{Y=1,S=o}) - \mathbb{E}_\ell(1_{Y=1,S=o})}{\mathbb{E}_\ell(1_{Y=1,S=o})\mathbb{E}(1_{Y=1,S=o})}\right\} - 1_{Y=0,S=o}\left\{\frac{\mathbb{E}(1_{Y=0,S=o}) - \mathbb{E}_\ell(1_{Y=0,S=o})}{\mathbb{E}_\ell(1_{Y=0,S=o})\mathbb{E}(1_{Y=0,S=o})}\right\}.
\end{aligned}$$

With population and empirical counts bounded away from zero,

$$|\hat{X} - X| \lesssim |\mathbb{E}(1_{Y=1,S=o}) - \mathbb{E}_\ell(1_{Y=1,S=o})| + |\mathbb{E}(1_{Y=0,S=o}) - \mathbb{E}_\ell(1_{Y=0,S=o})|.$$

By Hoeffding's inequality and the union bound, with probability $1-2\delta$,

$$|\mathbb{E}(1_{Y=1,S=o}) - \mathbb{E}_\ell(1_{Y=1,S=o})| \leq \left\{\frac{\ln(2/\delta)}{2n_\ell}\right\}^{1/2}, \quad |\mathbb{E}(1_{Y=0,S=o}) - \mathbb{E}_\ell(1_{Y=0,S=o})| \leq \left\{\frac{\ln(2/\delta)}{2n_\ell}\right\}^{1/2}.$$

Therefore with probability $1-\delta$ for all $i \in [n]$ simultaneously,

$$|\hat{X}_i - X_i| \lesssim 2\left\{\frac{\ln(4/\delta)}{2n_\ell}\right\}^{1/2} \lesssim \frac{\ln(4/\delta)^{1/2}}{n_\ell^{1/2}}.$$

We conclude that, under this union event that holds with probability $1-\delta$,

$$\mathbb{E}_\ell\{(\hat{X}-X)^2\} \lesssim \mathbb{E}_\ell\left\{\frac{\ln(4/\delta)}{n_\ell}\right\} = \frac{\ln(4/\delta)}{n_\ell}.$$

For all $\delta > 0$, this quantities vanishes for large $n_\ell = n/L$, so $\mathbb{E}_\ell\{(\hat{X}-X)^2\} = o_p(1)$. The continuous mapping theorem implies the desired result. □

Lemma D.9. Under Proposition 3's conditions, $\mathbb{E}_n(XZ) = \mathbb{E}(XZ) + o_p(1)$.

Proof. The argument is identical to Lemma D.4. □

E Discrete Outcomes

In this section, we extend our identification results and estimation strategy from $\mathcal{Y}=\{0,1\}$ to $\mathcal{Y}=\{y_1,...,y_K\}$ for $K>2$. Our results in the main text follow from the special case with $K=2$ with $y_1=1$ and $y_2=0$.

E.1 Identification

As in the main text, our identification result allows for the covariates X to be either discrete or continuous. We will also assume that the support of R has at least K elements (i.e., it is either multi-valued or continuous).

Lemma E.1 (Identification as generative model)**.** Suppose Assumptions 1 and 2 hold.

i. If Assumption 3(i) holds, then, for any $d\in\{0,1\}$, $x\in\mathcal{X}$ and $r\in\mathcal{R}$,

$$f_R(r\,|\,S=e,X,D=d)=\sum_{y\in\mathcal{Y}}f_R(r\,|\,S=o,X,D=d,Y=y)\Pr\{Y(d)=y\,|\,S=e,X\}.$$

ii. If Assumption 3(ii) holds, then, for any $d\in\{0,1\}$, $x\in\mathcal{X}$ and $r\in\mathcal{R}$,

$$f_R(r\,|\,S=e,X,D=d)=\sum_{y\in\mathcal{Y}}f_R(r\,|\,S=o,X,Y=y)\Pr\{Y(d)=y\,|\,S=e,X\}.$$

Proof. The former result specializes the general equation (5). The latter result specializes the general equation (3). □

The next step is Bayes' rule. Define the treatment weights in the experimental sample as

$$\pi_d(X,R):=\frac{\Pr(D=d,S=e\,|\,X,R)}{\Pr(D=d,S=e\,|\,X)},$$

and the outcome weights in the observational sample as

$$\gamma_{y,d}(X,R)=\frac{\Pr(Y=y,D=d,S=o\,|\,X,R)}{\Pr(Y=y,D=d,S=o\,|\,X)}\text{ or }\gamma_y(X,R)=\frac{\Pr(Y=y,S=o\,|\,X,R)}{\Pr(Y=y,S=o\,|\,X)}.$$

Lemma E.2 (Bayes' rule)**.** Suppose Assumptions 1 and 2 hold.

i. If Assumption 3(i) holds, $\pi_d(X,R)=\sum_{y\in\mathcal{Y}}\gamma_{y,d}(X,R)\Pr\{Y(d)=y\,|\,S=e,X\}$ almost surely.

ii. If Assumption 3(ii) holds, $\pi_d(X,R)=\sum_{y\in\mathcal{Y}}\gamma_y(X,R)\Pr\{Y(d)=y\,|\,S=e,X\}$ almost surely.

Proof. Consider the first statement. By (6),

$$f_R(R\,|\,S=o,X,D=d,Y=y)=\frac{\Pr(Y=y,D=d,S=o\,|\,R,X)f_R(R|X)}{\Pr(Y=y,D=d,S=o|X)}=\gamma_{y,d}(X,R)f_R(R|X)$$
$$f_R(R\,|\,D=d,S=e,X)=\frac{\Pr(D=d,S=e\,|\,R,X)f_R(R|X)}{\Pr(D=d,S=e|X)}=\pi_d(X,R)f_R(R|X).$$

Substituting these expressions into Lemma E.1(i) and canceling $f_R(R\,|\,X)$ yields the result.

Consider the second statement. By (4),

$$f_R(R\,|\,Y=y,S=o,X)=\frac{\Pr(Y=y,S=o\,|\,R,X)f_R(R|X)}{\Pr(Y=y,S=o|X)}=\gamma_y(X,R)f_R(R|X),$$
$$f_R(R\,|\,D=d,S=e,X)=\frac{\Pr(D=d,S=e\,|\,R,X)f_R(R|X)}{\Pr(D=d,S=e|X)}=\pi_d(X,R)f_R(R|X).$$

Substituting these expressions into Lemma E.1(ii) and canceling $f_R(R\,|\,X)$ yields the result. □

The previous two lemmas give the generalization of our main result, expressing the causal parameter as a conditional moment and unconditional moment. Generalizing the notation of the main text, let $\mu(d,X):=[\Pr\{Y(d)=y_1\,|\,S=e,X\},...,\Pr\{Y(d)=y_{K-1}\,|\,S=e,X\}]^\top\in\mathbb{R}^{K-1}$.[18]

Theorem E.1 (Identification as conditional moment)**.** Suppose Assumptions 1 and 2 hold.

i. If Assumption 3(i) holds, then, for $d\in\{0,1\}$,

$$\mathbb{E}\Big\{\widetilde{\Delta}^e(d,X)-\widetilde{\Delta}^o(d,X)^\top\mu(d,X)\,|\,X,R\Big\}=0 \text{ almost surely,}$$

where $\widetilde{\Delta}^e(d,x):=\frac{1\{D=d,S=e\}}{\Pr(D=d,S=e|X=x)}-\frac{1\{Y=y_K,D=d,S=o\}}{\Pr(Y=y_K,D=d,S=o|X=x)}$ and

$$\widetilde{\Delta}^o(d,x):=\begin{bmatrix}\frac{1\{Y=y_1,D=d,S=o\}}{\Pr(Y=y_1,D=d,S=o|X=x)}-\frac{1\{Y=y_K,D=d,S=o\}}{\Pr(Y=y_K,D=d,S=o|X=x)}\\ \vdots\\ \frac{1\{Y=y_{K-1},D=d,S=o\}}{\Pr(Y=y_{K-1},D=d,S=o|X=x)}-\frac{1\{Y=y_K,D=d,S=o\}}{\Pr(Y=y_K,D=d,S=o|X=x)}\end{bmatrix}\in\mathbb{R}^{K-1}$$

ii. If Assumption 3(ii) holds, then, for $d\in\{0,1\}$,

$$\mathbb{E}\big\{\Delta^e(d,X)-\Delta^o(X)^\top\mu(d,X)\,|\,X,R\big\}=0 \text{ almost surely,}$$

where $\Delta^e(d,x):=\frac{1\{D=d,S=e\}}{\Pr(D=d,S=e|X=x)}-\frac{1\{Y=y_K,S=o\}}{\Pr(Y=y_K,S=o|X=x)}$ and

$$\Delta^o(x):=\begin{bmatrix}\frac{1\{Y=y_1,S=o\}}{\Pr(Y=y_1,S=o|X=x)}-\frac{1\{Y=y_K,S=o\}}{\Pr(Y=y_K,S=o|X=x)}\\ \vdots\\ \frac{1\{Y=y_{K-1},S=o\}}{\Pr(Y=y_{K-1},S=o|X=x)}-\frac{1\{Y=y_K,S=o\}}{\Pr(Y=y_K,S=o|X=x)}\end{bmatrix}\in\mathbb{R}^{K-1}.$$

[18] In particular, the main text takes $K=2$, $y_1=1$, and $y_2=0$.

Proof. We prove the second statement; the proof of the first statement is analogous. By Lemma E.2(ii), for any $r \in \mathcal{R}$ and $x \in \mathcal{X}$:

$$\pi_d(x,r) = \sum_{j=1}^{K} \gamma_{y_j}(x,r) \Pr\{Y(d) = y_j \mid S = e, X = x\}$$

where

$$\pi_d(x,r) = \frac{\Pr(D=d, S=e \mid X=x, R=r)}{\Pr(D=d, S=e \mid X=x)} \text{ and } \gamma_{y_j}(x,r) = \frac{\Pr(Y=y_j, S=o \mid X=x, R=r)}{\Pr(Y=y_j, S=o \mid X=x)}.$$

Since $\sum_{j=1}^{K} \Pr\{Y(d) = y_j \mid S = e, X = x\} = 1$, we can rewrite the sum as

$$\pi_d(x,r) = \sum_{j=1}^{K-1} \gamma_{y_j}(x,r) \Pr\{Y(d) = y_j \mid S = e, X = x\} + \gamma_{y_K}(x,r) \left[1 - \sum_{j=1}^{K-1} \Pr\{Y(d) = y_j \mid S = e, X = x\} \right].$$

Further re-arranging then yields:

$$\pi_d(x,r) - \gamma_{y_K}(x,r) = \sum_{j=1}^{K-1} \left\{ \gamma_{y_j}(x,r) - \gamma_{y_K}(x,r) \right\} \Pr\{Y(d) = y_j \mid S = e, X = x\}.$$

We then define

$$\tilde{\gamma}(x,r) := \begin{bmatrix} \gamma_{y_1}(x,r) - \gamma_{y_K}(x,r) \\ \vdots \\ \gamma_{y_{K-1}}(x,r) - \gamma_{y_K}(x,r) \end{bmatrix}, \quad \mu(d,x) := \begin{bmatrix} \Pr\{Y(d) = y_1 \mid S = e, X = x\} \\ \vdots \\ \Pr\{Y(d) = y_{K-1} \mid S = e, X = x\} \end{bmatrix}$$

and rewrite the previous display as

$$\pi_d(x,r) - \gamma_{y_K}(x,r) = \tilde{\gamma}(x,r)^\top \mu(d,x).$$

The result then follows immediately by iterated expectations. □

Corollary E.1 (Identification as representation)**.** Suppose Assumptions 1 and 2 hold. Consider any measurable function $H_d : \mathcal{X} \times \mathcal{R} \to \mathbb{R}^J$ with $J \geq K-1$.

i. If Assumption 3(i) holds, then, for $d \in \{0,1\}$,

$$\mathbb{E}\{H_d(X,R) \widetilde{\Delta}^e(d,X) \mid X\} = \mathbb{E}\{H_d(X,R) \widetilde{\Delta}^o(d,X)^\top \mid X\} \mu(d,X) \text{ almost surely.}$$

If $\mathbb{E}\{H_d(X,R) \widetilde{\Delta}^o(d,X)^\top \mid X\}$ has full column rank almost surely, then $\mu(d,X)$ is identified.

ii. If Assumption 3(ii) holds, then, for $d \in \{0,1\}$,

$$\mathbb{E}\{H_d(X,R) \Delta^e(d,X) \mid X\} = \mathbb{E}\left\{H_d(X,R) \Delta^o(X)^\top \mid X\right\} \mu(d,X) \text{ almost surely.}$$

If $\mathbb{E}\{H_d(X,R) \Delta^o(X)^\top \mid X\}$ has full column rank almost surely, then $\mu(d,X)$ is identified.

E.2 Choice of Representation

As in the main text, Theorem E.1 and Corollary E.1 imply that we can use existing results on conditional moment restrictions for identification and inference on the average potential outcomes in the experimental sample. Importantly, any choice of representation satisfying the full rank condition suffices for identification as well as inference on the average potential outcomes in the experimental sample, as we will discuss next.

Furthermore, we can again apply results in Chamberlain (1987) and Newey (1993) to derive the efficient representation for downstream causal inference with a discrete outcome. For exposition, consider the case in which Assumption 3(ii) holds. Theorem E.1(ii) yields the stacked conditional moment restriction:

$$\mathbb{E}\left\{\begin{array}{l}\Delta^e(0,X)-\Delta^o(X)^\top\mu(0,X)\\ \Delta^e(1,X)-\Delta^o(X)^\top\mu(1,X)\end{array}\middle|\, X,R\right\}=0.$$

We take the difference of the treated and untreated moment restrictions. Generalizing the notation of the main text, let $\Delta^e(X)=\Delta^e(1,X)-\Delta^e(0,X)$ and $\theta(X)=\mu(1,X)-\mu(0,X)$. Then the differenced moment restriction is simply

$$\mathbb{E}\big\{\Delta^e(X)-\Delta^o(X)^\top\theta(X)|X,R\big\}=0, \tag{7}$$

generalizing (2) from the main text. We conclude that the efficient representation in the class of models satisfying (7) is

$$H^*(X,R)=\frac{\mathbb{E}\{\Delta^o(X)\,|\,X,R\}}{\sigma^2(\theta,X,R)},\quad \sigma^2(\theta,X,R):=\mathbb{E}[\{\Delta^e(X)-\Delta^o(X)^\top\theta(X)\}^2\,|\,X,R]$$

where $H^*(X,R)\in\mathbb{R}^{K-1}$, $\mathbb{E}\{\Delta^o(X)\,|\,X,R\}\in\mathbb{R}^{K-1}$, and $\sigma^2(\theta,X,R)\in\mathbb{R}$. Its jth component is the scalar $H^*_j(X,R)=\frac{\mathbb{E}\{\Delta^o_j(X)|X,R\}}{\sigma^2(\theta,X,R)}$, where $\Delta^o_j(X)=\frac{1\{Y=y_j,S=o\}}{\Pr(Y=y_j,S=o|X)}$.

As before, this formula reveals that the efficient representation of the RSV requires three types of predictions: (i) prediction of each outcome category given the RSV, which appears in the numerator $\mathbb{E}\{\Delta^o(X)\,|\,X,R\}$; (ii) prediction of the treatment given the RSV, which appears in the denominator $\sigma^2(\theta,X,R)$; and (iii) prediction of the sample indicator given the RSV, which appears in both the numerator and the denominator.

Finally, Theorem E.1 and Corollary E.1 provide testable implications of the identifying assumptions through over-identifying restrictions: we can test whether alternative choices of the representation $H(X,R)$ and $H'(X,R)$ yield significantly different estimates of the average potential outcomes in the experimental sample.

E.3 Estimation and Inference with Discrete Covariates

In light of these identification results, we extend our inferential procedure to allow for discrete outcomes and discrete covariates. We continue to focus on the case of Assumption 3(ii) for exposition.

As in the main text, for clarity, we describe our inferential procedure using sample splitting with TRAIN and TEST folds; the generalization to cross-fitting is straightforward. The extension for discrete covariates is conceptually simple: for each $x \in \mathcal{X}$, we estimate the generalized conditional average treatment effect in the experimental sample $\theta(x) \in \mathbb{R}^{K-1}$. Then, we average over X to estimate the generalized average treatment effect in the experimental sample $\theta \in \mathbb{R}^{K-1}$. This may be further reduced to the scalar $\theta_0 \in \mathbb{R}$, which is the standard average treatment effect. Importantly, the inference results from Section 4 apply directly for each covariate value $x \in \mathcal{X}$.

Algorithm E.1 (Inference with discrete outcome and discrete covariates)**.** Given $\{S_i, 1\{S_i = e\}D_i, 1\{S_i = o\}Y_i, X_i, R_i\}$, where $Y \in \mathcal{Y} = \{y_1, ..., y_K\}$ and $X \in \mathcal{X} = \{x_1, ..., x_{|\mathcal{X}|}\}$:

1. Divide the sample into TRAIN and TEST folds.

2. Learn the representations on TRAIN: $\widehat{H}(R,x)$. For each $x \in \mathcal{X}$,

(a) Count conditional probabilities:

$$\text{COUNT}(D=d, S=e|X=x) = \frac{\sum_{i \in \text{TRAIN}} 1(D_i=d, S_i=e, X_i=x)}{\sum_{i \in \text{TRAIN}} 1(X_i=x)} \text{ for each } d \in \{0,1\},$$

$$\text{COUNT}(Y=y, S=o|X=x) = \frac{\sum_{i \in \text{TRAIN}} 1(Y_i=y, S_i=o, X_i=x)}{\sum_{i \in \text{TRAIN}} 1(X_i=x)} \text{ for each } y \in \mathcal{Y}.$$

(b) Train predictors on the subsample with $X = x$:

$$\text{PRED}_Y(R,x) : \mathcal{R} \to [0,1]^K \text{ using } \{(R_i, Y_i) : i \in \text{TRAIN}, X_i = x, S_i = o\}$$

$$\text{PRED}_D(R,x) : \mathcal{R} \to [0,1] \text{ using } \{(R_i, D_i) : i \in \text{TRAIN}, X_i = x, S_i = e\}$$

$$\text{PRED}_S(R,x) : \mathcal{R} \to [0,1] \text{ using } \{(R_i, S_i) : i \in \text{TRAIN}, X_i = x\}.$$

(c) Construct conditional expectations for this subsample:

$$\widehat{\mathbb{E}}\{\Delta^e(x)|R,x\} = \left\{ \frac{\text{PRED}_D(R,x)}{\text{COUNT}(D=1, S=e|X=x)} - \frac{1-\text{PRED}_D(R,x)}{\text{COUNT}(D=0, S=e|X=x)} \right\} \text{PRED}_S(R,x)$$

$$\widehat{\mathbb{E}}\{\Delta^o_j(x)|R,x\} = \left\{ \frac{\text{PRED}_{Y,j}(R,x)}{\text{COUNT}(Y=y_j, S=o|X=x)} - \frac{\text{PRED}_{Y,K}(R,x)}{\text{COUNT}(Y=y_K, S=o\,|\,X=x)} \right\} \{1-\text{PRED}_S(R,x)\}$$

(d) Initially estimate $\widehat{\theta}_{\text{INIT}}(x)$ with this subsample:

$$\widehat{\theta}_{\text{INIT}}(x) = \underset{\theta}{\text{argmin}} \sum_{i \in \text{TRAIN}: X_i = x} \left[\widehat{\mathbb{E}}\{\Delta^e(x)|R_i,x\} - \widehat{\mathbb{E}}\{\Delta^o(x)|R_i,x\}^\top \theta \right]^2.$$

(e) Learn the representation with this subsample: $\widehat{H}(R,x) = \frac{\widehat{\mathbb{E}}\{\Delta^o(x)|R,x\}}{\widehat{\sigma}^2(\widehat{\theta}_{\text{INIT}},R,x)}$ where

$$\widehat{\sigma}^2(\widehat{\theta}_{\text{INIT}},R,x) = \left[\frac{\text{PRED}_D(R,x)}{\text{COUNT}(D=1,S=e|X=x)^2} + \frac{1-\text{PRED}_D(R,x)}{\text{COUNT}(D=0,S=e|X=x)^2}\right]\text{PRED}_S(R,x) + \widehat{\theta}^2_{\text{INIT}}\left[\frac{\text{PRED}_{Y,j}(R,x)}{\text{COUNT}(Y=1,S=o|X=x)^2} + \frac{\text{PRED}_{Y,K}(R,x)}{\text{COUNT}(Y=0,S=o|X=x)^2}\right]\{1-\text{PRED}_S(R,x)\}.$$

3. Construct generalized CATE estimates on TEST: $\widehat{\theta}(x)$. For each $x \in \mathcal{X}$,

(a) Count conditional probabilities:

$$\text{COUNT}(D=d,S=e|X=x) = \frac{\sum_{i\in\text{TEST}} 1(D_i=d,S_i=e,X_i=x)}{\sum_{i\in\text{TEST}} 1(X_i=x)} \text{ for each } d\in\{0,1\},$$
$$\text{COUNT}(Y=y,S=o|X=x) = \frac{\sum_{i\in\text{TEST}} 1(Y_i=y,S_i=o,X_i=x)}{\sum_{i\in\text{TEST}} 1(X_i=x)} \text{ for each } y\in\mathcal{Y}.$$

(b) Compute treatment and outcome variation:

$$\widehat{\Delta}^e(x) = \frac{1(D=1,S=e)}{\text{COUNT}(D=d,S=e|X=x)} - \frac{1(D=0,S=e)}{\text{COUNT}(D=d,S=e|X=x)}$$
$$\widehat{\Delta}^o_j(x) = \frac{1(Y=y_j,S=o)}{\text{COUNT}(Y=y_j,S=o|X=x)} - \frac{1(Y=y_K,S=o)}{\text{COUNT}(Y=y_K,S=o\,|\,X=x)} \text{ for } j=1,...,K-1.$$

(c) Compute the generalized CATE estimate:

$$\widehat{\theta}(x) = \left\{\sum_{i\in\text{TEST}:X_i=x} \widehat{H}(R_i,x)\widehat{\Delta}^o_i(x)^\top\right\}^{-1}\left\{\sum_{i\in\text{TEST}:X_i=x} \widehat{H}(R_i,x)\widehat{\Delta}^e_i(x)\right\}.$$

4. Aggregate into ATE estimate on TEST: $\widehat{\theta}_0$.

(a) Compute the generalized ATE estimate: $\widehat{\theta} = \mathbb{E}_{\text{TEST}}\{\widehat{\theta}(X)\}$

(b) Compute the ATE estimate: $\widehat{\theta}_0 = \sum_{j=1}^{K-1} y_j\widehat{\theta}_j + y_K\left(1-\sum_{j=1}^{K-1}\widehat{\theta}_j\right)$.

5. Bootstrap its confidence interval: $\widehat{\theta}_0 \pm c_\alpha \widehat{v} n^{-1/2}$, where c_α is the $1-\alpha/2$ quantile of the standard Gaussian and $\widehat{v}n^{-1/2}$ is the bootstrap standard error of $\widehat{\theta}_0$ fixing the estimated representations.

Our final estimator $\widehat{\theta}_0 \in \mathbb{R}$, for the ATE in the experimental sample, is asymptotically normal by direct extensions of the arguments in Section 4 and Appendix D. The key insight is that, for each fixed $x \in \mathcal{X}$, estimation of $\theta(x)$ has the same structure as the covariate-free case analyzed in the main text. Critically, we do not require the learned representation $\widehat{H}(R,x)$ to converge to the efficient representation $H^*(R,x)$.

Assumption E.1 (Limit with discrete covariates)**.** For each $x \in \mathcal{X}$ and each $j \in \{1,...,K-1\}$, the learned representation has some mean square limit: $\mathbb{E}_R[\{\widehat{H}_j(R,x) - \widetilde{H}_j(R,x)\}^2] = o_p(1)$, where $\mathbb{E}\{\widetilde{H}_j(R,x)^2\} < \infty$, and possibly $\widetilde{H}_j(R,x) \neq H_j^*(R,x)$. Moreover, the limit is evenly correlated with outcome variation: $\mathbb{E}\left\{\widetilde{H}(R,x)\Delta^o(x)^\top\right\}$ has full rank.

This assumption does not impose any complexity restriction on the machine learning algorithms used to construct the predictors $\text{PRED}_Y(R,x)$, $\text{PRED}_D(R,x)$, and $\text{PRED}_S(R,x)$. Nor does it require any rate of convergence for these predictors. The representation $\widehat{H}(R,x)$ may be misspecified, as long as it has some limit that predicts outcome variation. This robustness justifies the use of complex deep learning algorithms with unknown statistical properties.

E.4 Estimation and Inference with Continuous Covariates

As a final extension, we discuss how to modify our inferential procedure to accommodate continuous covariates. We continue to impose Assumption 3(ii). Here, we focus on low-dimensional covariates. High-dimensional covariates are an exciting direction for future econometric theory.

Our identification results already apply to continuous covariates as written. The conditional moment equation from Theorem 1 applies for each x, as does its generalization in (7) above:

$$\mathbb{E}\{\Delta^e(x) - \Delta^o(x)^\top \theta(x) \,|\, X = x, R\} = 0,$$

where $\theta_j(x) = \Pr\{Y(1) = y_j \,|\, S = e, X = x\} - \Pr\{Y(0) = y_j \,|\, S = e, X = x\}$ is the generalized CATE in the experimental sample. When Y is binary, this simplifies to the familiar CATE in the experimental sample $\theta(x) = \theta_1(x) = \mathbb{E}\{Y(1) - Y(0) | S = e, X = x\}$.[19]

The key modification is that the conditional probabilities $\Pr(D = d, S = e \,|\, X = x)$ and $\Pr(Y = y, S = o \,|\, X = x)$ must now be estimated as smooth functions of x using nonparametric methods (e.g., kernel smoothing, local polynomials, or series). The overall procedure is similar to Algorithm E.1, with the following modifications:

- In Steps 2 and 3, replace the conditional probability estimators. Instead of the counting estimators $\text{COUNT}(D = d, S = e | X = x)$ and $\text{COUNT}(Y = y, S = o | X = x)$, use smooth nonparametric estimators $\text{EST}(D = d, S = e | X = x)$ and $\text{EST}(Y = y, S = o | X = x)$.

- In Steps 2 and 3, replace the generalized CATE estimators. The objectives of $\widehat{\theta}_{\text{INIT}}(x)$ and $\widehat{\theta}(x)$ implicitly use the discrete kernel $K(X,x) = 1(X = x)$.[20] Now, instead use a continuous kernel such as $K_h(X,x) = 1(|X - x| \leq h)$, where $h > 0$ is a vanishing bandwidth.

[19] Take $K = 2$, $y_1 = 1$, and $y_2 = 0$ to return to the simpler setting.

[20] For example, $\sum_{i \in \text{TRAIN}: X_i = x}(\cdot) = \sum_{i \in \text{TRAIN}} 1(X_i = x)(\cdot) = \sum_{i \in \text{TRAIN}} K(X_i, x)(\cdot)$.

Valid inference would require: (i) standard smoothness conditions on the conditional probabilities as functions of x; (ii) undersmoothing or bias correction for their nonparametric estimates, to ensure that they contribute only second-order terms; (iii) the representation $\widehat{H}(R,x)$ has some mean square limit $\widetilde{H}(R,x)$ correlated with outcome variation.

The asymptotic variance must account for estimation of the "first stage" conditional probabilities, and it can be computed via the bootstrap. The proof strategy with continuous covariates would largely follow our analysis with discrete covariates, combined with standard arguments for Z-estimation with nonparametric first stages (e.g., Ai and Chen, 2003).

As before, no rate condition will be required on the estimated representation due to infinite order Neyman orthogonality and sample splitting. The key technical insight—that the RSV moment condition permits misspecified representation learning—continues to hold. This permits complex machine learning for RSV representations even with continuous (low-dimensional) covariates.

F Continuous Outcomes

We extend our results to outcomes that are continuous and bounded. Recall that $f_W(\cdot|...)$ is our general symbol for the Radon-Nikodym derivative. Let $f_{Y(d)}(y \mid S,X,R)$ denote the conditional density of $Y(d)$ given S, X, and R.

Assumption F.1 (Bounded outcome)**.** Suppose that $Y(d) \in [-U,U]^p$ for some $U < \infty$. Suppose that for $d \in \{0,1\}$ and $s \in \{e,o\}$, $f_{Y(d)}(y \mid S=s,X,D,R) > 0$ for $y \in [-U,U]^p$ almost surely.

We consider two complementary approaches: discretization and deconvolution. The former is easier to implement, but requires more restrictive conditions than the latter.

F.1 Discretization

The first approach discretizes the continuous outcome Y into a discrete approximation $\tilde{Y}_\varepsilon$ by binning up its support. Specifically, we bin up the continuous support $\mathcal{Y}$ using the grid $\mathcal{Y}_\varepsilon = \{y_1,\cdots,y_{|\mathcal{Y}_\varepsilon|}\}$, where each grid value is the center of a bin of radius ε. The grid $\mathcal{Y}_\varepsilon$ forms an ε-cover of $[-U,U]^p$. The discretized random variable can be written as $\tilde{Y}_\varepsilon = \operatorname{argmin}_{y \in \mathcal{Y}_\varepsilon} \|y - Y\|_\infty$. Given a continuous random variable Y, $\tilde{Y}_\varepsilon$ is a discrete random variable that takes the value of the nearest grid location.

We introduce some additional notation for binning. Let $B_y(\varepsilon)$ defining an l_∞-ball of radius $\varepsilon>0$ around the value y, i.e. $B_y(\varepsilon)=\{y'\in\mathcal{Y}:\|y-y'\|_\infty\leq\varepsilon\}$. Furthermore, define the discretized outcome weights

$$\gamma^\varepsilon_{y,d}(X,R)=\frac{\Pr\{Y\in B_y(\varepsilon),D=d,S=o\,|\,X,R\}}{\Pr\{Y\in B_y(\varepsilon),D=d,S=o\,|\,X\}} \text{ and } \gamma^\varepsilon_{y}(X,R)=\frac{\Pr\{Y\in B_y(\varepsilon),S=o\,|\,X,R\}}{\Pr\{Y\in B_y(\varepsilon),S=o\,|\,X\}}.$$

These discretized outcome weights closely resemble the outcome weights in Appendix E.

Finally, recall the previously defined treatment weights

$$\pi_d(X,R):=\frac{\Pr(D=d,S=e\,|\,X,R)}{\Pr(D=d,S=e\,|\,X)}.$$

Proposition F.1 (Binning)**.** Suppose Assumptions 1, 2 with $\tilde{Y}_\varepsilon$ in lieu of Y, and F.1 hold.

i. If Assumption 3(i) holds, then $\pi_d(X,R)=\sum_{y\in\mathcal{Y}_\varepsilon}\gamma^\varepsilon_{y,d}(X,R)\Pr\{Y(d)\in B_y(\varepsilon)\,|\,S=e,X\}$.

ii. If Assumption 3(ii) holds with $\tilde{Y}_\varepsilon$ in lieu of Y, then $\pi_d(X,R)=\sum_{y\in\mathcal{Y}_\varepsilon}\gamma^\varepsilon_{y}(X,R)\Pr\{Y(d)\in B_y(\varepsilon)\,|\,S=e,X\}$.

Proof. We prove the second statement. The argument for the first statement is similar. The steps mirror the derivations of equations (3) and (4), respectively. Throughout, Assumption F.1 and $\varepsilon>0$ guarantee that various events happen with positive probability.

1. By the law of total probability and the definition of $\tilde{Y}_\varepsilon$,

$$\begin{aligned}f_R(r\,|\,S=e,X=x,D=d)&=\int f_{R,\tilde{Y}_\varepsilon}(r,y\,|\,S=e,X=x,D=d)\mathrm{d}y\\&=\int f_R(r\,|\,S=e,X=x,D=d,\tilde{Y}_\varepsilon=y)f_{\tilde{Y}_\varepsilon}(y\,|\,S=e,X=x,D=d)\mathrm{d}y\\&=\sum_{y\in\mathcal{Y}_\varepsilon}f_R(r\,|\,S=e,X=x,D=d,\tilde{Y}_\varepsilon=y)\Pr(\tilde{Y}_\varepsilon=y\,|\,S=e,X=x,D=d).\end{aligned}$$

By the definition of $\tilde{Y}_\varepsilon$ and Assumption 1,

$$\begin{aligned}\Pr(\tilde{Y}_\varepsilon=y\,|\,S=e,X=x,D=d)&=\int_{y'\in B_y(\varepsilon)}f_Y(y'\,|\,S=e,X=x,D=d)\mathrm{d}y'\\&=\int_{y'\in B_y(\varepsilon)}f_{Y(d)}(y'\,|\,S=e,X=x)\mathrm{d}y'\\&=\Pr\{Y(d)\in B_y(\varepsilon)\,|\,S=e,X=x\}.\end{aligned}$$

Next, notice that

$$\begin{aligned}f_R(r\,|\,S=e,X=x,D=d,\tilde{Y}_\varepsilon=y)&=f_R(r\,|\,S=o,X=x,D=d,\tilde{Y}_\varepsilon=y)\\&=f_R(r\,|\,S=o,X=x,\tilde{Y}_\varepsilon=y)\end{aligned}$$

where the first equality applies Assumption 2 with $\tilde{Y}_\varepsilon$ in lieu of Y, and the second equality applies Assumption 3(ii) with $\tilde{Y}_\varepsilon$ in lieu of Y.

Combining the previous displays, we arrive at

$$f_R(r\,|\,S=e,X=x,D=d)=\sum_{y\in\mathcal{Y}_\varepsilon} f_R(r\,|\,S=o,X=x,\tilde{Y}_\varepsilon=y)\Pr\{Y(d)\in B_y(\varepsilon)\,|\,S=e,X=x\}.$$

2. We apply Bayes' rule to rewrite

$$f_R(r\,|\,S=o,X=x,\tilde{Y}_\varepsilon=y)=\frac{\Pr(\tilde{Y}_\varepsilon=y,S=o\,|\,R=r,X=x)f_R(r|X=x)}{\Pr(\tilde{Y}_\varepsilon=y,S=o|X=x)}=\gamma_y^\varepsilon(x,r)f_R(r|X=x),$$
$$f_R(r\,|\,S=e,X=x,D=d)=\frac{\Pr(D=d,S=e\,|\,R=r,X=x)f_R(r|X=x)}{\Pr(D=d,S=e|X=x)}=\pi_d(x,r)f_R(r|X=x).$$

Substituting these expressions into the previous step and canceling $f_R(R\,|\,X)$ yields the result.

□

Consequently, for any fixed $\varepsilon>0$, we can perform estimation and inference for the estimand

$$\theta(\varepsilon):=\sum_{y\in\mathcal{Y}_\varepsilon} y[\Pr\{Y(1)\in B_y(\varepsilon)\,|\,S=e\}-\Pr\{Y(0)\in B_y(\varepsilon)\,|\,S=e\}]$$

by directly applying the arguments given in Appendix E. The estimand $\theta(\varepsilon)$ can be interpreted as a discrete approximation to the ATE on the experimental sample over the ε-cover. Of course, since $\varepsilon>0$, this estimand will be biased for the ATE in the experimental sample

$$\theta:=\int_{[-U,U]^p} y\big\{f_{Y(1)}(y\,|\,S=e)-f_{Y(0)}(y\,|\,S=e)\big\}\mathrm{d}y.$$

Nonetheless, we expect that these two estimands will be close to each other for low-dimensional outcomes. When $p=1$, we show that the worst-case bias of $\theta(\varepsilon)$ is no greater than 2ε. In other words, the bias is at most the width of each bin.[21]

Proposition F.2 (Discretization bias)**.** Under the conditions of Proposition F.1, if $p=1$ then $|\theta(\varepsilon)-\theta|\leq 2\varepsilon$.

Proof. To prove this result, let $f_{Y(d)}(y|S)$ denote the conditional density of $Y(d)$ given S. By the definition of $\theta(\varepsilon)$, we write

$$\begin{aligned}\theta(\varepsilon)&:=\sum_{y\in\mathcal{Y}_\varepsilon} y[\Pr\{Y(1)\in B_y(\varepsilon)\,|\,S=e\}-\Pr\{Y(0)\in B_y(\varepsilon)\,|\,S=e\}]\\&=\sum_{y\in\mathcal{Y}_\varepsilon} y\int_{y'\in B_y(\varepsilon)}\{f_{Y(1)}(y'\,|\,S=e)-f_{Y(0)}(y'\,|\,S=e)\}\mathrm{d}y'\\&=\sum_{y\in\mathcal{Y}_\varepsilon}\int_{y'\in B_y(\varepsilon)} y\{f_{Y(1)}(y'\,|\,S=e)-f_{Y(0)}(y'\,|\,S=e)\}\mathrm{d}y'.\end{aligned}$$

[21] Since the radius of one ball is ε, the width of each bin is 2ε.

Since $\mathcal{Y}_\varepsilon$ is an ε cover of $\mathcal{Y}$,

$$\theta := \int_{[-U,U]} y' \{ f_{Y(1)}(y' \mid S=e) - f_{Y(0)}(y' \mid S=e) \} \mathrm{d}y'$$
$$\theta = \sum_{y \in \mathcal{Y}_\varepsilon} \int_{y' \in B_y(\varepsilon)} y' \{ f_{Y(1)}(y' \mid S=e) - f_{Y(0)}(y' \mid S=e) \} \mathrm{d}y'.$$

Consequently, by the triangle inequality

$$\begin{aligned} |\theta(\varepsilon) - \theta| &= \left| \sum_{y \in \mathcal{Y}_\varepsilon} \int_{y' \in B_y(\varepsilon)} (y-y') \{ f_{Y(1)}(y' \mid S=e) - f_{Y(0)}(y' \mid S=e) \} \mathrm{d}y' \right| \\ &\leq \left| \sum_{y \in \mathcal{Y}_\varepsilon} \int_{y' \in B_y(\varepsilon)} (y-y') f_{Y(1)}(y' \mid S=e) \mathrm{d}y' \right| + \left| \sum_{y \in \mathcal{Y}_\varepsilon} \int_{y' \in B_y(\varepsilon)} (y-y') \{ f_{Y(0)}(y' \mid S=e) \mathrm{d}y' \right|. \end{aligned}$$

Focusing on the former term,

$$\begin{aligned} \left| \sum_{y \in \mathcal{Y}_\varepsilon} \int_{y' \in B_y(\varepsilon)} (y-y') f_{Y(1)}(y' \mid S=e) \mathrm{d}y' \right| &\leq \sum_{y \in \mathcal{Y}_\varepsilon} \left| \int_{y' \in B_y(\varepsilon)} (y-y') f_{Y(1)}(y' \mid S=e) \mathrm{d}y' \right| \\ &\leq \sum_{y \in \mathcal{Y}_\varepsilon} \int_{y' \in B_y(\varepsilon)} |y-y'| f_{Y(1)}(y' \mid S=e) \mathrm{d}y' \\ &\leq \varepsilon \sum_{y \in \mathcal{Y}_\varepsilon} \int_{y' \in B_y(\varepsilon)} f_{Y(1)}(y' \mid S=e) \mathrm{d}y' \\ &= \varepsilon \int_{y' \in [-U,U]} f_{Y(1)}(y' \mid S=e) \mathrm{d}y' \\ &= \varepsilon. \end{aligned}$$

by the triangle inequality, Hölder inequality, definition of the ε-cover, and definition of a density function. Applying the same argument to the other term yields the result.

□

Consequently, for any fixed $\varepsilon > 0$, researchers can apply our results to conduct inference on $\theta(\varepsilon)$, which is a discrete approximation to the ATE in the experimental sample. Furthermore, the bias of $\theta(\varepsilon)$ for θ, which is introduced by our discrete approximation, can be bounded.

F.2 Deconvolution

A complementary strategy is based on deconvolution. This approach allows us to use the original versions of Assumption 2 and 3, which are weaker. However, the procedure departs from the one described in Appendix E.

Defined the generalized outcome weights

$$\gamma^0_{y,d}(X,R) = \frac{f_{Y,D,S}(y,d,o|X,R)}{f_{Y,D,S}(y,d,o|X)} \text{ and } \gamma^0_y(X,R) = \frac{f_{Y,S}(y,o|X,R)}{f_{Y,S}(y,o|X)}.$$

Here, $f_{Y,D,S}(\cdot|...)$ denotes the (conditional) Radon-Nikodym derivative, where Y is continuous, D is binary, and S indicates the sample. Recall the previously defined treatment weights

$$\pi_d(X,R) := \frac{\Pr(D=d,S=e\,|\,X,R)}{\Pr(D=d,S=e\,|\,X)}.$$

Proposition F.3 (Identification with continuous outcome)**.** Suppose Assumptions 1, 2, and F.1 hold.

i. If Assumption 3(i) holds, $\pi_d(X,R) = \int_{y\in\mathcal{Y}} \gamma^0_{y,d}(X,R) f_{Y(d)}(y|S=e,X)\mathrm{d}y$.

ii. If Assumption 3(ii) holds, $\pi_d(X,R) = \int_{y\in\mathcal{Y}} \gamma^0_{y}(X,R) f_{Y(d)}(y|S=e,X)\mathrm{d}y$.

Proof. We prove the second statement. The argument for the first statement is similar.

1. By equation (3),

$$f_R(r\,|\,S=e,X=x,D=d) = \int f_R(r\,|\,S=o,X=x,Y=y) f_{Y(d)}(y\,|\,S=e,X=x)\mathrm{d}y.$$

2. By Bayes' rule,

$$f_R(r\,|\,Y=y,S=o,X=x) = \frac{f_{Y,S}(y,o\,|\,R=r,X=x) f_R(r|X=x)}{f_{Y,S}(y,o|X=x)} = \gamma^0_y(X,R) f_R(r|X=x),$$
$$f_R(r\,|\,D=d,S=e,X=x) = \frac{\Pr(D=d,S=e\,|\,R=r,X=x) f_R(r|X=x)}{\Pr(D=d,S=e|X=x)} = \pi_d(X,R) f_R(r|X=x).$$

Substituting these expressions into the previous step and canceling $f_R(r\,|\,X=x)$ yields the result.

□

Proposition F.3 identifies the conditional counterfactual density $f_{Y(d)}(y|S=e,X)$ as the solution to an integral equation. For example, consider the integral equation of Proposition F.3(ii):

$$\pi_d(X,R) = \int_{y\in\mathcal{Y}} \gamma^0_y(X,R) f_{Y(d)}(y|S=e,X)\mathrm{d}y.$$

The left hand side is a ratio of the regression functions $\Pr(D=d,S=e\,|\,X,R)$ and $\Pr(D=d,S=e\,|\,R)$, each of which is identified from the experimental sample. The right hand side is the composition of the integral operator $\tilde{f} \mapsto \int_{y\in\mathcal{Y}} \gamma^0_y(X,R)\tilde{f}(y|X)\mathrm{d}y$ and the target density $f_{Y(d)}(y|S=e,X)$. Clearly the integral operator is identified from the observational sample.

Recovering $f_{Y(d)}(y|S=e,X)$ is a classic statistical problem called a Fredholm integral equation of the first kind (Kress, 1989). In general, it is an ill posed inverse problem because

to isolate $f_{Y(d)}(y|S=e,X)$ we must invert the integral operator. This problem is closely related to nonparametric instrumental variable regression and deconvolution, for which several nonparametric estimators are available in the literature; see e.g. Carrasco et al. (2007) for a review. Existence of the solution $f_{Y(d)}(y|S=e|X)$ will be guaranteed by Picard's criterion (Kress, 1989). Uniqueness of the solution $f_{Y(d)}(y|S=e|X)$ will be guaranteed by a completeness condition on the integral operator (Newey and Powell, 2003). The rate of estimation for $f_{Y(d)}(y|S=e|X)$ will be the rate of estimation for $\pi_d(X,R)$, as throttled by the ill posedness of inverting the integral operator (e.g. Chen and Reiss, 2011). Articulating the technical details of such an approach is an exciting direction for future econometric theory.

G Crop Burning Illustration: Additional Details

We present details for the empirical example in Section 3.2, based on Jack et al. (2025).

The bias formula is derived in Appendix B. Within the main text, we take $\tilde{\beta}=1$ so that $\widetilde{\theta}=\beta\theta_0$ by Corollary B.1 since Jack et al. (2025) directly plug-in the RSV R as the outcome in the experiment.

We classify a field as experimental ($S=e$) if it did not receive a random spot check. We classify a field as observational ($S=o$) if it received a random spot check. This example has complete cases (Assumption 3(ii)); the treatment varies in the observational sample.

In Section 3.2, we define $R\in\{0,1\}$ as the authors' "maximum accuracy" classifier for whether a field has not been burned. Jack et al. (2025) construct another "balanced accuracy" classifier based on an alternative threshold rule. We find similar results in Table G.1, now defining $R\in\{0,1\}$ as the authors' "balanced accuracy" classifier.

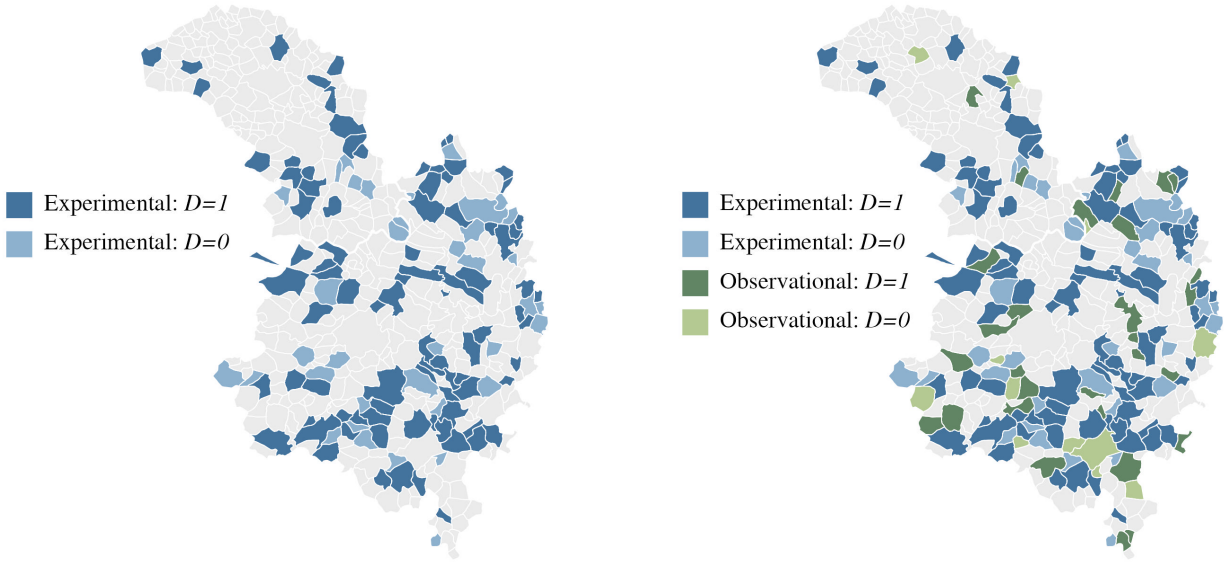
Finally, we visualize the experimental and observational fields, summarized at the village level. Due to privacy concerns, longitude and latitude coordinates for individual fields in the experiment are unavailable. Therefore, in Figure G.1a, we classify a village as experimental if less than 50% of its fields received a random spot check. Similarly, in Figure G.1b, we classify a village as observational if more than 50% of its fields received a random spot check.

Figure G.1 is provided purely for the sake of placing our empirical exercise of Section 3.2 on a map, keeping in mind that our analysis is at the field level.

Table G.1: Underestimation of treatment effects in crop burning experiment: Revisited.

Estimand	Common practice $\widetilde{\theta}$	Bias β	Causal parameter θ
Estimate	0.074	0.601	0.123
	(0.049)	(0.068)	(0.086)

Notes: The RSV is the field-level "balanced accuracy" label defined by Jack et al. (2025), which applies a threshold rule to the predicted probability of not being burned. The "observational sample" has fields that received a random spot check, and the "experimental sample" has other fields. For illustration, we conduct linear estimation, controlling for stratum fixed effects. Standard errors are based on 5000 bootstrap replications clustered at the village level.



(a) Experimental units only.

(b) Auxiliary sample: Observational units.

Figure G.1: We illustrate the samples in our re-analysis of the crop burning experiment in Jack et al. (2025), plotting the map of villages in Bathinda and Faridkot, which are two districts in Punjab, India. Experimental villages are those in which less than 50% of fields received a spot check. Observational villages are those in which more than 50% of fields received a spot check.

H Smartcard Illustration: Additional Details

H.1 Real Data from a Randomized Control Trial

Randomization took place at the level of the mandal, i.e. subdistrict, of Andhra Pradesh, India. Muralidharan et al. (2023) partition 396 mandals into the following subgroups:

- treated mandals (111), randomly assigned to receive Smartcards in 2010;

- buffer mandals (136), randomly assigned to receive Smartcards in 2011;
- untreated mandals (44), randomly assigned to receive Smartcards in 2012;
- non-study mandals (105), which were excluded from the experiment.

We study villages within mandals as the units of analysis, where villages are defined by Asher et al. (2021). For each village, our treatment D indicates whether the village received Smartcards in 2010. We interpret villages within the treated mandals as treated experimental units; villages within the buffer and untreated mandals as untreated experimental units; and villages within the non-study mandals as observational units with missing treatment. Finally, we drop villages with fewer than 100 individuals. This removes about 2% of villages.

Figure 1 illustrates our village classification. The causal parameter is the effect of early adoption (2010 Smartcards) for villages in the experiment.

Table H.1 summarizes village characteristics. We study 8,320 villages, with an average population exceeding 2,000 individuals per village. Villages are located in rural areas. The populated area within a typical village is geographically concentrated.

Table H.1: Village summary statistics.

Sample	Observational	Experimental: Untreated	Experimental: Untreated	Experimental: Treated
Smartcards	N/A	2012	2011	2010
Number of villages	2257	852	2929	2274
Average population	1985	2255	2105	2116
Average fraction female	0.491	0.492	0.496	0.497

For each village, we collect poverty measurements to serve as the outcome. Following Asher and Novosad (2020), the data sources are the 2012-2013 Socio-Economic and Caste Census (SECC), and the 2013 Indian Economic Census. Our main outcome variable, used in all three semi-synthetic exercises, indicates whether a village's per capita consumption is in the bottom quartile. We consider two additional outcome variables in the second and third semi-synthetic exercises: does a village have only low income households, i.e. no earner making above 5,000 rupees; and does a village only low and middle income households, i.e. no earner making above 10,000 rupees. The definitions of low and middle income households is from the SECC.

H.2 Real Satellite Images

For each village, we extract satellite images to serve as the RSV. First, we extract coordinates for the perimeter of the village (Asher et al., 2021). Then, we extract luminosity measures, a vector in $\mathbb{R}^{50}$, which includes the minimum, maximum, mean, and sum of night light within each polygon, along with the total number of pixels in the polygon, from 2012 to 2020 (Asher et al., 2021). Finally, we extract satellite images from 2019, summarized as a high-dimensional, pre-trained embedding vector in $\mathbb{R}^{4000}$ (Rolf et al., 2021). The concatenation of these objects is our remotely sensed variable R.

In the first exercise, we truncate the RSV from $\mathbb{R}^{4050}$ to $\mathbb{R}^{1050}$ for computational tractability. In particular, we use the initial 1,000 features of the pre-trained embedding vector. In the second and third exercise, we use the full RSV in $\mathbb{R}^{4050}$.

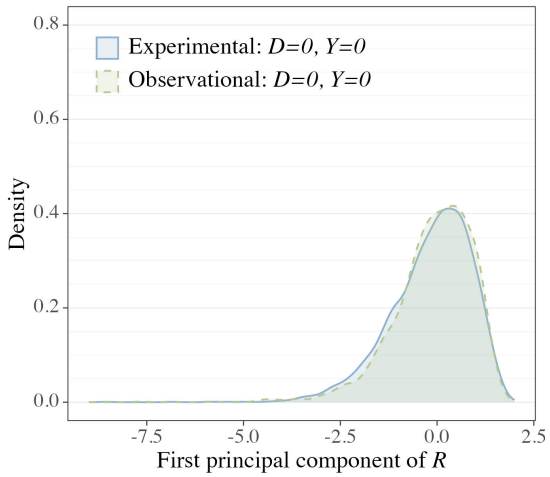
While no direct effects (Assumption 3(ii)) holds by design in the first exercise, it must be defended in the second and third exercises. Figure H.1 provides such evidence for the low consumption outcome. We compare $f_R(R \mid S=s, D=0, Y=y)$ in the top row with $f_R(R \mid S=s, D=1, Y=y)$ in the bottom row. The columns subset observations by $y \in \mathcal{Y}$, and the colors subset observations by $s \in \{e, o\}$. If, in a given column, the same-colored densities in the top and bottom row look similar, that is evidence that treatment only affects the RSV via the outcome. For example, focusing on the first column and the blue densities, we visually compare $f_R(R \mid S=e, D=0, Y=0)$ in the top row with $f_R(R \mid S=e, D=1, Y=0)$ in the bottom row. The densities are similar, as desired. Figure H.2 provides similar evidence for the two other poverty outcomes.

H.3 Implementation Details

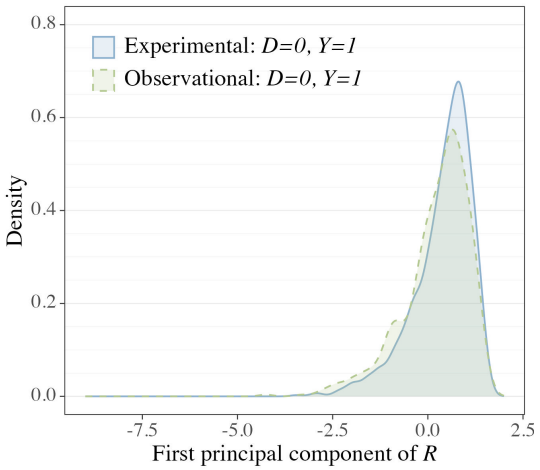
Across empirical exercises, we use random forest predictions: $\text{PRED}_Y(R)$ estimates $\Pr(Y=1 \mid S=o, R)$, $\text{PRED}_D(R)$ estimates $\Pr(D=1 \mid S=e, R)$, and $\text{PRED}_S(R)$ estimates $\Pr(S=e \mid R)$. The first prediction appears in the common practice, while all three appear in the efficient representation.

In the first exercise, we use luminosity as well as the initial 1,000 features of the satellite image embedding, for computational tractability of the 500 replications. Using this subset of features makes cross-fitting computationally feasible. This procedure is justified by Proposition 3.

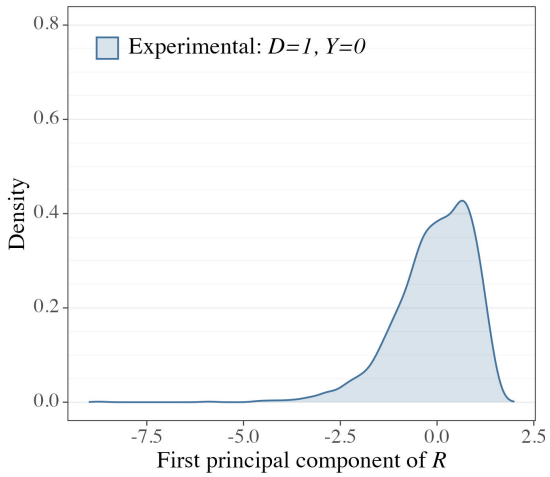
In the second and third exercises, we use luminosity measures as well as all 4,000 features of the satellite image embedding, since we are only running one replication. Random forests satisfy stability conditions, which allow us to eliminate cross-fitting; the argument



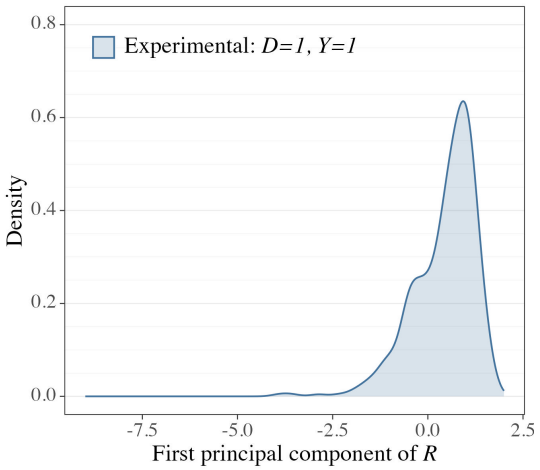
(a) Densities of $R \mid S$, $D=0$, $Y=0$.



(b) Densities of $R \mid S$, $D=0$, $Y=1$.



(c) Densities of $R \mid S=e, D=1, Y=0$.

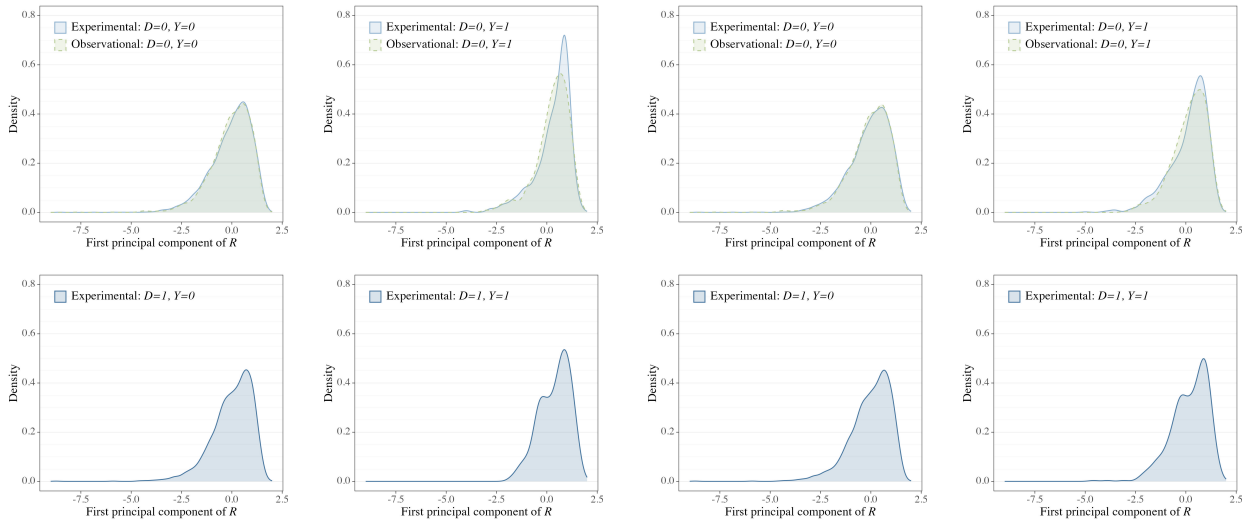


(d) Densities of $R \mid S=e$, $D=1$, $Y=1$.

Figure H.1: No direct effects (Assumption 3(ii)) is plausible in the second and third exercises for the low consumption outcome. Within each plot, we compare $f_R(R \mid S=s, D=0, Y=y)$ in the top row with $f_R(R \mid S=s, D=1, Y=y)$ in the bottom row, using data from Muralidharan et al. (2023). Because the RSV R is high-dimensional, we visualize the density of its standardized first principal component.

is a straightforward extension of Proposition 3, using stability in place of independence to handle the stochastic equicontinuity terms. See e.g. Chernozhukov et al. (2020, Theorem 2). Alternatively, we could use the limited complexity of random forests, along the lines of Chernozhukov et al. (2020, Theorem 3).

The technical details of the random forest implementation are as follows. We use the `R` package `randomForest` with 100 trees and the package default options. When predicting the outcomes, we set the class weights ten-to-one because one outcome value is much more



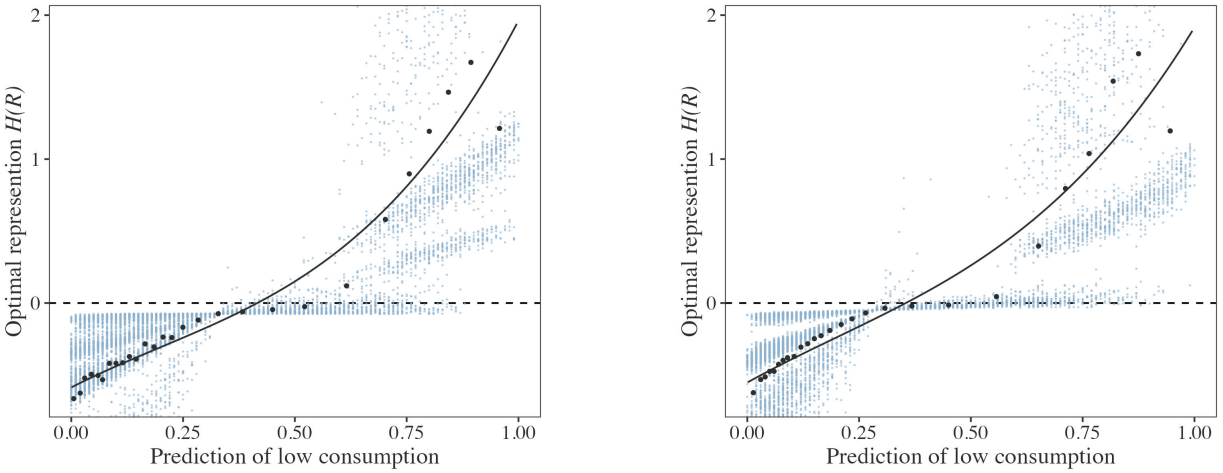
(a) Low income.

(b) Low and middle income.

Figure H.2: No direct effects (Assumption 3(ii)) is plausible in the second and third exercises for the additional poverty outcomes. Within each plot, we compare $f_R(R \mid S=s, D=0, Y=y)$ in the top row with $f_R(R \mid S=s, D=1, Y=y)$ in the bottom row, using data from Muralidharan et al. (2023). Because the RSV R is high-dimensional, we visualize the density of its standardized first principal component.

frequent than the other.

H.4 Comparing Alternative Representations



(a) RSV: Synthetic samples

(b) RSV: Real samples

Figure H.3: We contrast efficient versus simple representations of the RSV. The efficient representation combines three predictions. The simple representation uses only one prediction.

Figure H.3 visualizes how our efficient representation $\widehat{H}(R)$ compares to the implicit representation of common practice $\text{PRED}_Y(R)$, using the first outcome variable. Even though the RSV is high-dimensional, these representations are scalars. Therefore, we can visualize each village as a point in a plot with $\widehat{H}(R)$ on the vertical axis and $\text{PRED}_Y(R)$ on the horizontal axis. For the sake of visualization, we also plot the binscatter and the best fitting curve. We make such a plot for the second exercise (Figure H.3a; "synthetic samples") and for the third exercise (Figure H.3b; "real samples").

There are some notable differences. The efficient representation $\widehat{H}(R)$ varies over $(-2,4)$, while the simple representation associated with common practice $\text{PRED}_Y(R)$ is bounded in the unit interval $(0,1)$. It is efficient to extrapolate beyond observed villages. Using $\text{PRED}_Y(R)$ in place of $\widehat{H}(R)$ in Algorithm 1 would only interpolate among observed villages. The representations $\widehat{H}(R)$ and $\text{PRED}_Y(R)$ are generally correlated, but their relationship is nonlinear.

I Smartcard Illustration: Extensions

We expand our discussion of spillovers (Remark 7) and timing (Remark 8). These extensions allow a more nuanced interpretation of the second and third semi-synthetic exercises. To ease exposition, we consider a binary outcome, omit pre-treatment covariates, and focus on "incomplete" cases (Assumption 3(ii)) throughout this discussion.

I.1 Spillovers

I.1.1 Generalized Potential Outcomes

With spillover effects, we write the potential outcome for unit i as $Y_i(d_i,\mathbf{d}_{-i})$. The potential outcome is indexed by the unit of interest's treatment assignment $d_i \in \{0,1\}$, as well as the vector of other units' treatment assignments $\mathbf{d}_{-i} \in \{0,1\}^{n-1}$.

From this potential outcome, we define the direct potential outcome by taking the expectation over other units' treatment assignments: $Y_i^{\text{dir}}(d_i) = \mathbb{E}_{\mathbf{D}_{-i}}\{Y_i(d_i,\mathbf{D}_{-i}\}|D_i=d_i,S_i=e\}$. Note that it remains random due to unobserved heterogeneity.[22] The direct potential outcome may be viewed as design-based, since the expectation is over the design-induced distribution of $\mathbf{D}_{-i}$ conditional on $D_i=d$ in the experiment.

[22]Formally, we may denote unobserved heterogeneity by η_i and use the nonseparable model notation $Y_i = Y_i(D_i,\mathbf{D}_{-i},\eta_i)$. Then the potential outcome is $Y_i(d_i,\mathbf{d}_{-i},\eta_i)$ and the direct potential outcome is $\int Y_i(d_i,\mathbf{d}_{-i},\eta_i)\}\text{dPr}(\mathbf{d}_{-i}|D_i=d_i,S_i=e,\eta_i)$.

From direct potential outcomes, we define the average direct treatment effect. We relax the assumption of identical distribution across observations that is maintained throughout the main text. As such, it is a sample average direct treatment effect, though we omit "sample" for brevity.

Definition I.1 (Average direct treatment effect)**.** The average direct treatment effect in the experimental sample is $\theta_n = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\{Y_i^{\text{dir}}(1) - Y_i^{\text{dir}}(0) | S_i = e\}$.

The parameter θ_n may be viewed as an average direct effect, because it involves counterfactuals for unit i under a "direct" intervention on unit i's treatment assignment.

We now generalize Assumption 1 for the setting with spillovers.

Assumption I.1 (Spillovers)**.** Suppose the following:

i. Spillovers: If $D_i = d_i$ and $\mathbf{D}_{-i} = \mathbf{d}_{-i}$ then $Y_i = Y_i(d_i, \mathbf{d}_{-i})$.

ii. Randomization: $D_i \perp\!\!\!\perp \{Y_i(d_i, \mathbf{d}_{-i})\} \,|\, S_i = e$.[23]

iii. Overlap: $\Pr(D_i = 1 \,|\, S_i = e)$ is bounded away from zero and one almost surely.

Lemma I.1. Under Assumption I.1, the average direct treatment effect equals a difference of outcome means. Formally, $\theta_n = \frac{1}{n}\sum_{i=1}^{n}\{\mathbb{E}(Y_i | D_i = 1, S_i = e) - \mathbb{E}(Y_i | D_i = 0, S_i = e)\}$.

Proof. Using nonseparable model notation with unobserved heterogeneity η_i, as defined in Footnote 22,

$$\begin{aligned}
\mathbb{E}(Y_i | D_i = d, S_i = e) &= \mathbb{E}\{Y_i(d, \mathbf{D}_{-i}, \eta_i) | D_i = d, S_i = e\} \\
&= \mathbb{E}[\mathbb{E}\{Y_i(d, \mathbf{D}_{-i}, \eta_i) | D_i = d, S_i = e, \eta_i)\} | D_i = d, S_i = e] \\
&= \mathbb{E}\{Y_i^{\text{dir}}(d, \eta_i) | D_i = d, S_i = e\} \\
&= \mathbb{E}\{Y_i^{\text{dir}}(d, \eta_i) | S_i = e\}
\end{aligned}$$

where the final line appeals to randomization. □

If the outcomes were perfectly observed in the experimental sample, then θ_n would be identified by Lemma I.1.

Our benchmark estimator in the semi-synthetic exercise is the empirical analogue to this expression. It is the benchmark that an economist would obtain if they could fully observe the treatments and outcomes in the experiment. By Lemma I.1, the benchmark is an unbiased estimator of a reasonable estimand, even in the presence of spillovers.

[23] In nonseparable model notation, $D_i \perp\!\!\!\perp \eta_i | S_i = e$.

As before, the crux of our problem is that the outcomes are not perfectly observed in the experimental sample. Instead, we have access to an auxiliary, observational sample. Our method recovers θ_n under appropriate modifications of our identifying assumptions. We now extend Assumptions 2 and 3 accordingly.

Assumption I.2 (Stability with spillovers)**.** Suppose Assumption 2 holds for each unit $i \in \{1,...,n\}$.

Assumption I.3 (No direct effects with spillovers)**.** Suppose Assumption 3(ii) holds for each unit $i \in \{1,...,n\}$.

Most importantly, we modify stability (Assumption 2). In the presence of spillovers, stability requires $S_i \perp R_i \,|\, (Y_i, D_i, X_i)$: conditional on a unit's *own* outcome, treatment, and covariates, the distribution of the unit's RSV R_i would be the same had it belonged to the experimental or observational samples. In particular, the unit's RSV distribution depends on its own outcome, but does not depend on other units' outcomes.

In our semi-synthetic exercise, unit i is a village. Treatments are randomized at the mandal level. A mandal is a sub-district containing villages, roughly comparable to a U.S. county. Substantively, this modified stability assumption means that the satellite image of a village depends only on that village's poverty level and on the treatment status of the mandal in which the village lies. It does not depend on the poverty levels of other villages, nor on the treatment statuses of other mandals. We think of this as a plausible approximation; intuitively the satellite image of a village should only depends on variables observed in that village.

In summary, our method extends as long as spillovers are in treatment effects, not in RSV distributions. We formalize this claim in what follows.

Theorem I.1 (Identification with spillovers)**.** Suppose Assumptions I.1, I.2, and I.3 hold. Then

$$\theta_n = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{E}(\Delta_i^e | R_i)}{\mathbb{E}(\Delta_i^o | R_i)} \quad \text{almost surely,}$$

where $\Delta_i^e = \frac{1(D_i=1,S_i=e)}{\Pr(D_i=1,S_i=e)} - \frac{1(D_i=0,S_i=e)}{\Pr(D_i=0,S_i=e)}$ and $\Delta_i^o = \frac{1(Y_i=1,S_i=o)}{\Pr(Y_i=1,S_i=o)} - \frac{1(Y_i=0,S_i=o)}{\Pr(Y_i=0,S_i=o)}$.

Proof. The argument mirrors the proof of Lemma 1 and Theorem 1.

1. By the law of total probability,

$$\begin{aligned}
\delta_{d,i}^e(r) &:= f_{R_i}(r \,|\, S_i = e, D_i = d) \\
&= \int f_{R_i,Y_i}(r, y \,|\, S_i = e, D_i = d) \mathrm{d}y \\
&= \int f_{R_i}(r \,|\, S_i = e, D_i = d, Y_i = y) f_{Y_i}(y \,|\, S_i = e, D_i = d) \mathrm{d}y.
\end{aligned}$$

Next, notice that

$$\begin{aligned} f_{R_i}(r \mid S_i=e,D_i=d,Y_i=y) &= f_{R_i}(r \mid S_i=o,D_i=d,Y_i=y) \\ &= f_{R_i}(r \mid S_i=o,Y_i=y) \\ &=: \delta^o_{y,i}(r), \end{aligned}$$

where the first equality applies Assumption I.2 and the second equality applies Assumption I.3.

Combining the previous displays, we arrive at the general result

$$\delta^e_{d,i}(r) = \int \delta^o_{y,i}(r) f_{Y_i}(y \mid S_i=e,D_i=d)\mathrm{d}y.$$

When Y is binary,

$$\begin{aligned} f_{Y_i}(1 \mid S_i=e,D_i=d) &= \mathbb{E}(Y_i \mid S_i=e,D_i=d) =: \mu_i(d) \\ f_{Y_i}(0 \mid S_i=e,D_i=d) &= 1-\mathbb{E}(Y_i \mid S_i=e,D_i=d) = 1-\mu_i(d). \end{aligned}$$

By the proof of Lemma I.1, $\mu_i(d) = \mathbb{E}\{Y_i^{\mathrm{dir}}(d) | S_i=e\}$. Therefore, the general result specializes to

$$\delta^e_{d,i}(r) = \delta^o_{1,i}(r)\mu_i(d) + \delta^o_{0,i}(r)\{1-\mu_i(d)\} = \delta^o_{0,i}(r) + \{\delta^o_{1,i}(r) - \delta^o_{0,i}(r)\}\mathbb{E}\{Y_i^{\mathrm{dir}}(d) | S_i=e\}.$$

2. We apply Bayes' rule to rewrite

$$\begin{aligned} \delta^o_{y,i}(r) &:= f_{R_i}(r \mid S_i=o,Y_i=y) = \frac{\Pr(Y_i=y,S_i=o \mid R_i=r) f_{R_i}(r)}{\Pr(Y_i=y,S_i=o)}, \\ \delta^e_{d,i}(r) &:= f_{R_i}(r \mid S_i=e,D_i=d) = \frac{\Pr(D_i=d,S_i=e \mid R_i=r) f_{R_i}(r)}{\Pr(D_i=d,S_i=e)}. \end{aligned}$$

Substituting these expressions into the previous step and canceling $f_{R_i}(r)$ gives

$$\mathbb{E}(\Delta_i^e | R_i) = \mathbb{E}(\Delta_i^o | R_i)\theta_i, \quad \theta_i = \mathbb{E}\{Y_i^{\mathrm{dir}}(1) - Y_i^{\mathrm{dir}}(0) | S_i=e\}.$$

We conclude that $\theta_n = \frac{1}{n}\sum_{i=1}^n \theta_i = \frac{1}{n}\sum_{i=1}^n \frac{\mathbb{E}(\Delta_i^e|R_i)}{\mathbb{E}(\Delta_i^o|R_i)}$. □

I.1.2 Average Global Treatment Effect

Theorem I.1 identifies the average direct treatment effect, which is a reasonable estimand. We now ask: when does this reasonable estimand coincide with the standard estimand for models with spillovers, namely the average global treatment effect? We describe two plausible and empirically relevant scenarios. In these two scenarios, our method estimates the average global treatment effect in the presence of spillovers.

Definition I.2 (Average global treatment effect)**.** The average global treatment effect in the experimental sample is $\tilde{\theta}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\{Y_i(1,\mathbf{1}_{n-1}) - Y_i(0,\mathbf{0}_{n-1})|S_i = e\}$, where $\mathbf{1}_{n-1}$ and $\mathbf{0}_{n-1}$ are vectors of repeated entries in $\mathbb{R}^{n-1}$.

The parameter $\tilde{\theta}_n$ may be viewed as an average global effect, because it involves counterfactuals for unit i under a "global" intervention on all units' treatment assignments.

Corollary I.1 (Spillovers within but not across mandals)**.** Suppose the assumptions of Theorem I.1 hold. Suppose that each unit i is a village, and that treatment assignment is randomized at the mandal level, as in the real experiment we study. Now, further assume that spillovers only occur within mandals. Then $\tilde{\theta}_n = \theta_n = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{E}(\Delta_i^e|R_i)}{\mathbb{E}(\Delta_i^o|R_i)}$.

Proof. Because spillovers only occur within mandals, the potential outcome of village i only depends on the treatment assignments of other villages within its mandal; it does not depend on treatment assignments of other villages in other mandals. Because treatment assignment is at the mandal level, the treatment of villages i matches the treatment of all other villages in its mandal. These statements imply that $\mathbb{E}\{Y_i^{\text{dir}}(1)|S_i = e\} = \mathbb{E}\{Y_i(1,\mathbf{1}_{n-1})|S_i = e\}$. The same is true for $d_i = 0$. Therefore, $\theta_n = \tilde{\theta}_n$ and we are done.

We formally prove this statement for village $i = 1$ and $d_1 = 1$. Suppose that, among villages $\{1,...,n\}$, the initial m belong to the same mandal. As argued in the proof of Lemma I.1, with nonseparable model notation defined in Footnote 22,

$$\begin{aligned}
\mathbb{E}\{Y_1^{\text{dir}}(1,\eta_1)|S_1 = e\} &= \mathbb{E}\{Y_1(1,D_2,...,D_n,\eta_1)|D_1 = 1, S_1 = e\} \\
&= \mathbb{E}\{Y_1(1,D_2,...,D_m,\eta_1)|D_1 = 1, S_1 = e\} \\
&= \mathbb{E}\{Y_1(1,\mathbf{1}_{m-1},\eta_1)|D_1 = 1, S_1 = e\} \\
&= \mathbb{E}\{Y_1(1,\mathbf{1}_{n-1},\eta_1)|D_1 = 1, S_1 = e\} \\
&= \mathbb{E}\{Y_1(1,\mathbf{1}_{n-1},\eta_1)|S_1 = e\}
\end{aligned}$$

by spillovers only within mandals, mandal-level treatment assignment, spillovers only within mandals, and randomization. □

By assuming no spillovers across mandals, our method estimates the average global treatment effect from the full sample. Therefore, under the conditions of Corollary I.1, Figures 6 and 7 report average global treatment effect estimates in the presence of spillovers. For simplicity, Corollary I.1 assumes that there are no spillovers across mandals. In fact, our argument will go through as long as spillovers across mandals are asymptotically negligible.

Next, we consider an alternative restriction on spillovers: that they only occur within a fixed geographic radius. This alternative restriction is also widely adopted in the empirical literature. For example, Muralidharan et al. (2023) posit a radius of 20 km. After appropriately subsetting villages, our method once again estimates the average global treatment effect.

Corollary I.2 (Spillovers within a geographic radius)**.** Suppose the assumptions of Theorem I.1 hold. Suppose that each unit i is a village, and that treatment assignment is randomized at the mandal level, as in the real experiment we study. Now, further assume that spillovers only occur within a geographic radius of 20 km. Then $\tilde{\theta}_n$ coincides with θ_n after dropping any village that is within 20 km of another village with the opposite treatment assignment.

Proof. The argument is similar to the proof of Corollary I.1. We prove the key equality.

As before, consider village $i=1$ and $d_1=1$. Suppose that, among villages $\{1,...,n\}$, the initial m are within a 20 km radius. Each of these m villages must have been treated due to the subetting rule.

As argued in the proof of Lemma I.1, with nonseparable model notation defined in Footnote 22,

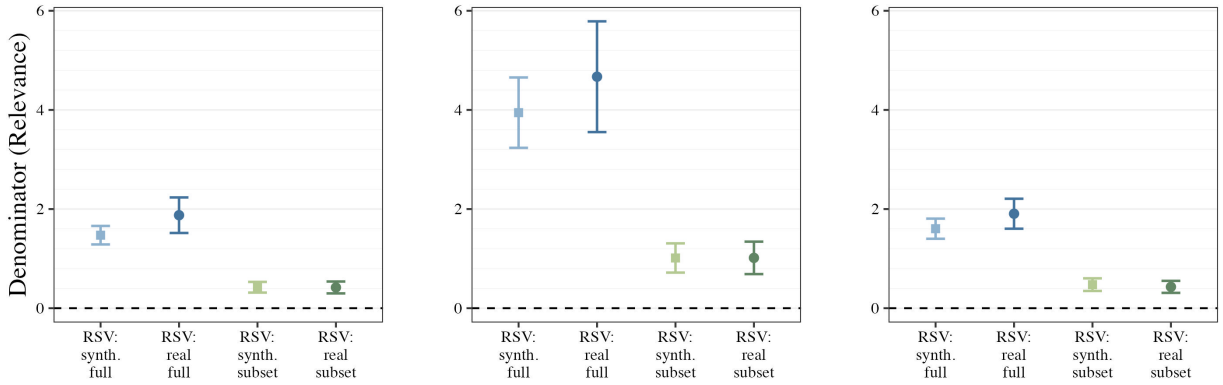
$$\begin{aligned}
\mathbb{E}\{Y_1^{\text{dir}}(1,\eta_1)|S_1=e\} &= \mathbb{E}\{Y_1(1,D_2,...,D_n,\eta_1)|D_1=1,S_1=e\} \\
&= \mathbb{E}\{Y_1(1,D_2,...,D_m,\eta_1)|D_1=1,S_1=e\} \\
&= \mathbb{E}\{Y_1(1,\mathbf{1}_{m-1},\eta_1)|D_1=1,S_1=e\} \\
&= \mathbb{E}\{Y_1(1,\mathbf{1}_{n-1},\eta_1)|D_1=1,S_1=e\} \\
&= \mathbb{E}\{Y_1(1,\mathbf{1}_{n-1},\eta_1)|S_1=e\}
\end{aligned}$$

by spillovers only within the radius, the subsetting rule, spillovers only within the radius, and randomization. □

By assuming no spillovers beyond a 20 km radius, our method estimates the average global treatment effect from a subset of the sample. On the one hand, this assumption is closer to that in the empirical literature. On the other hand, it comes at the cost of a smaller effective sample size.

As a robustness check, we now repeat our second and third semi-synthetic exercises after subsetting the villages in the manner described above. The subsetting step eliminates about 27% of the total number villages. In particular, about 27% of villages in the full sample have at least one village within a 20 km radius exposed to opposite treatment status.[24].

[24]The distance between two villages is measured using their centroids, following Muralidharan et al. (2023)



(a) Low consumption. (b) Low income. (c) Low and middle income.

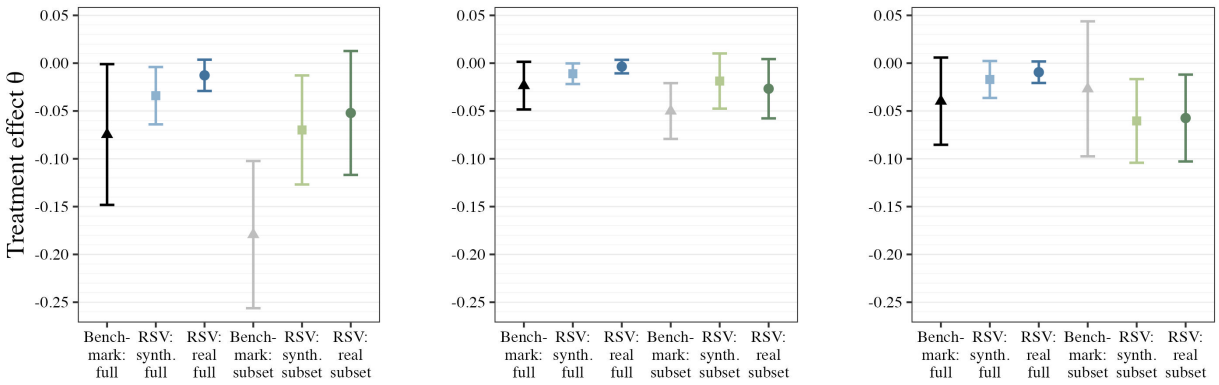
Figure I.1: Satellite images are relevant to the poverty outcomes in the presence of spillovers. Each plot is for a poverty outcome. Within each plot, we report $\mathbb{E}_n\{\widehat{H}(R)\widehat{\Delta}^o\}$ and its 90% confidence interval using our learned representation. The blue estimates correspond to Corollary I.1. In light blue, we visualize the relevance of our learned representation in the second exercise ("synthetic samples") while using the full sample ("full"). In dark blue, we visualize the relevance of our learned representation in the third exercise ("real samples") using the full sample ("full"). The green estimates correspond to Corollary I.2. In light green, we visualize the relevance of our learned representation in the second exercise ("synthetic samples") using the subsetted sample ("subset"). In dark green, we visualize the relevance of our learned representation in the third exercise ("real samples") using the subsetted sample ("subset"). Bootstrap standard errors, based on 1000 replications, are clustered at the sub-district level.

Under the conditions of Corollary I.2, Figures I.1 and I.2 report global average treatment effect estimates in the presence of spillovers. We find qualitatively similar results as before. Here, the benchmark uses the same subset of villages as our method.

I.2 Timing of Outcomes and RSVs

In the main text, we avoided time indexing and viewed the RSV as a post-outcome variable. In real data, such as the Smartcard illustration, the data have time indices. In particular, the RSV and the outcome may be measured at different times. Moreover, the treatment variable may be defined as early adoption, raising the question of whether our method applies to such a setting. We now confirm that it does, and clarify the interpretation of our reported estimate.

Suppose each unit is independent and identically distributed. Suppose early adoption of the treatment is randomly assigned at time t, the outcome is collected at time $t' \geq t$, and the



(a) Low consumption. (b) Low income. (c) Low and middle income.

Figure I.2: Our method recovers the unbiased benchmark estimate and its 90% confidence interval in the presence of spillovers. Each plot is for a particular poverty outcome. Within each plot, we visualize the benchmark versus our method. The black and blue estimates correspond to Corollary I.1, i.e. using the full sample of villages ("full"). The black benchmark is the difference-in-means an economist would obtain if they could observe treatments and outcomes in the experiment. Our method uses treatments and RSVs in an experiment, and outcomes and RSVs in an observational sample. In light blue, we visualize our method in the second exercise, where we observe outcomes for a random subset of experimental villages ("synthetic samples"). In dark blue, we visualize our method in the third exercise, where we observe outcomes for only the untreated experimental villages ("real samples"). The gray and green estimates correspond to Corollary I.2, i.e. using the subsetted sample ("subset"). The interpretations are analogous. Bootstrap standard errors, based on 1000 replications, are clustered at the sub-district level.

RSV is collected at time $t'' \geq t'$. In particular, only the treated experimental units receive treatment at time t, while all remaining units receive treatment at time t'. Overall, each unit is now characterized by the random vector

$$\{S, D_t, Y_{t'}(0), Y_{t'}(1), R_{t''}\}$$

where $Y_{t'}(d_t)$ are potential outcomes. For units in the experimental sample ($S = e$), we observe $(D_t, R_{t''})$. For units in the observational sample ($S = o$), we observe $(Y_{t'}, R_{t''})$. In this setting, $D_{t'} = 1$ and $D_{t''} = 1$ for all units, so we omit them.

The Smartcard illustration takes this form. Simplifying some of the details in Appendix H, only treated units in the experimental sample received Smartcards at time $t = 2010$. Untreated units in the experimental sample received Smartcards by the time the outcomes were collected

in $t' = 2013$. All remaining units, i.e. the observational sample, received Smartcards in $t' = 2013$ as well. Satellite images were collected at the later time $t'' = 2019$.

The parameter of interest is defined from the time t' potential outcomes. In the Smartcard illustration, it is the effect of 2010 Smartcard adoption on 2013 poverty levels.

Definition I.3 (Average time-specific treatment effect)**.** The average time-specific treatment effect in the experimental sample is $\theta_{t'} = \mathbb{E}\{Y_{t'}(1) - Y_{t'}(0) | S = e\}$.

We modify our assumption about the experimental sample to reflect the time variation in the measurement of the outcome and RSV.

Assumption I.4 (Experiment with time variation)**.** Suppose the following:

i. SUTVA: $Y_{t'} = D_t Y_{t'}(1) + (1 - D_t) Y_{t'}(0)$ almost surely.

ii. Randomization: $D_t \perp\!\!\!\perp \{Y_{t'}(0), Y_{t'}(1)\} \,|\, S = e$.

iii. Overlap: $\Pr(D_t = 1 \,|\, S = e)$ is bounded away from zero and one almost surely.

iv. Ultimate adoption: $D_{t'} = 1$ and $D_{t''} = 1$ almost surely.

Next, we modify our assumptions on stability and on the observational sample to reflect the time variation in the measurement of the outcome and RSV.

Assumption I.5 (Stability with time variation)**.** Suppose Assumption 2 holds replacing (D, Y, R) with $(D_t, Y_{t'}, R_{t''})$.

Assumption I.5 remains the main assumption of our framework: $S \perp\!\!\!\perp R_{t''} | D_t, Y_{t'}$. In an experiment with time variation, stability now requires that the conditional distribution of the RSV $R_{t''}$ given $(D_t, Y_{t'})$ is stable across the experimental and observational samples. Importantly, the RSV may be collected at a later time than the outcome. Figure 2 provides empirical evidence supporting this assumption.

Assumption I.6 (No direct effects with time variation)**.** Suppose Assumption 3(ii) holds replacing (D, Y, R) with $(D_t, Y_{t'}, R_{t''})$.

Assumption I.6 becomes $D_t \perp\!\!\!\perp R_{t''} | Y_{t'}$. Given the time variation in the measurement of the outcome and the RSV, this amounts to a Markov property: the effect of the earlier treatment D_t on the later RSV $R_{t''}$ must operate only though the its effect on the intermediate outcome

$Y_{t'}$. Importantly, $R_{t''}$ may depend on later outcomes $Y_{t''}$, but any effect from the earlier treatment D_t must be via the intermediate outcome $Y_{t'}$.

Concretely, 2010 Smartcards may affect 2013 consumption; 2013 consumption may affect 2019 consumption; and 2013 and 2019 consumption may affect 2019 satellite images. However, in this example, 2010 Smartcards cannot affect 2019 consumption via in any channel besides 2013 consumption. In this sense there is a Markov restriction. Figures H.1 and H.2 provide empirical evidence supporting this assumption.

Theorem I.2 (Identification with time variation)**.** Suppose Assumptions I.4, I.5, and I.6 hold. Then

$$\theta_{t'} = \frac{\mathbb{E}(\Delta_t^e | R_{t''})}{\mathbb{E}(\Delta_{t'}^o | R_{t''})} \quad \text{almost surely,}$$

where $\Delta_t^e = \frac{1(D_t=1,S=e)}{\Pr(D_t=1,S=e)} - \frac{1(D_t=0,S=e)}{\Pr(D_t=0,S=e)}$ and $\Delta_{t'}^o = \frac{1(Y_{t'}=1,S=o)}{\Pr(Y_{t'}=1,S=o)} - \frac{1(Y_{t'}=0,S=o)}{\Pr(Y_{t'}=0,S=o)}$.

Proof. The argument is identical to the proof of Lemma 1 and Theorem 1, replacing (D,Y,R) with $(D_t,Y_{t'},R_{t''})$. □

In summary, our results directly apply when D_t is an early adoption treatment, and when $Y_{t'}$ and $R_{t''}$ are an outcome and an RSV collected later on. In the main text, we report the effect of 2010 Smartcard adoption on 2013 consumption. With incomplete cases, there is an implicit Markov restriction that researchers must asses. However, with complete cases, this additional restriction may be relaxed.

J Quasi-Experiments with Remotely Sensed Outcomes

In the main text, we focused on the case in which the treatment is unconfounded in the experimental sample (Assumption 1) and the causal parameter of interest is the average treatment effect in the experimental sample (Definition 1). In this section, we describe how our main identifying assumption (Assumption 2(i)) can be used to identify treatment effects in quasi-experimental samples via data combination with the observational sample. We discuss two quasi-experimental strategies: instrumental variables and difference-in-differences. To ease exposition, we consider a binary outcome, omit pre-treatment covariates, and focus on "incomplete" cases (Assumption 3(ii)) throughout this discussion.

J.1 Instrumental Variables

Each unit is now characterized by the random vector

$$\{S,Z,D(0),D(1),Y(0,0),Y(0,1)Y(1,0),Y(1,1),R\},$$

where $Z \in \{0,1\}$ is a binary instrument, $D(z)$ are potential treatments and $Y(d,z)$ are potential outcomes, following Imbens and Angrist (1994). For units in the quasi-experimental sample ($S=e$), we observe (Z,D,R). For units in the observational sample ($S=o$), we observe (Y,R) since we focus on the incomplete case.

We modify our previous assumption about the experimental sample to accommodate the instrument. This sample may be viewed as a quasi-experiment, or as an experiment with imperfect compliance.

Assumption J.1 (Instrumental variable)**.** Suppose the following:

i. Instrument exclusion: for all $z \in \{0,1\}$ and $d \in \{0,1\}$, $Y(d,z) := Y(d)$ almost surely.

ii. SUTVA: $D = ZD(1)+(1-Z)D(0)$ and $Y = DY(1)+(1-D)Y(0)$ almost surely.

iii. Instrument randomization: $Z \perp\!\!\!\perp \{D(0),D(1),Y(0),Y(1)\} \mid S=e$.

iv. Instrument overlap: $\Pr(Z=1 \mid S=e)$ is bounded away from zero and one almost surely.

v. Monotonicity: $\Pr\{D(1) \geq D(0) | S=e\} = 1$ and $\Pr\{D(1) > D(0) | S=e\} > 0$.

Under Assumption J.1, if we were to observe the outcome in the quasi-experimental sample, then the local average treatment effect (LATE) $\theta^{LATE} := \mathbb{E}\{Y(1)-Y(0) \mid D(1) > D(0), S=e\}$ could be identified using standard arguments. Since the outcome is unobserved in the quasi-experimental sample, we will use the observational sample as in the main text.

We next modify our stability and observational completeness assumptions to accommodate the instrument.

Assumption J.2 (Stability with an instrument)**.** Suppose Assumption 2 holds replacing D with (Z,D).

Assumption J.3 (No direct effects with an instrument)**.** Suppose Assumption 3(ii) holds replacing D with (Z,D).

Assumption J.2 remains the main assumption of our framework: $S \perp\!\!\!\perp R | Z,D,Y$. In the presence of an instrument, stability now requires that the conditional distribution of the remotely sensed variable R given (Z,D,Y) is stable across the quasi-experimental and observational samples. Analogously, Assumption J.3 becomes $Z,D \perp\!\!\!\perp R|Y$. It now requires that the instrument Z and the treatment D only effect the remotely sensed variable R via their effect on the outcome Y. Together Assumption J.2 and Assumption J.3 imply that $(S,D,Z) \perp\!\!\!\perp R \,|\, Y$, which is a testable implication as before.

These conditions allow us to identify θ^{LATE} by combining the quasi-experimental and observational samples. As notation, let $\alpha(z) = \mathbb{E}(Y \,|\, S=e, Z=z)$, and $\beta(z) := \mathbb{E}(D \,|\, S=e, Z=z)$. By standard arguments, under Assumption J.1, $\theta^{LATE} = \frac{\alpha(1)-\alpha(0)}{\beta(1)-\beta(0)}$. Of course, $\beta(0)$ and $\beta(1)$ are identified from the quasi-experimental sample since (Z,D) are observed. We will therefore identify $\alpha(0)$ and $\alpha(1)$ by combining the quasi-experimental and observational samples under Assumption J.2 and Assumption J.3.

As a stepping stone, we first identify $\alpha(d,z) := \mathbb{E}(Y \,|\, S=e, D=d, Z=z)$.

Theorem J.1 (Identification with an instrument)**.** Suppose Assumption J.1, J.2 and J.3 hold. Then, for any $z \in \{0,1\}$ and $d \in \{0,1\}$,

$$\alpha(d,z) = \frac{\mathbb{E}\{\Delta^e(d,z) | R\}}{\mathbb{E}(\Delta^o | R)} \text{ almost surely,}$$

where $\Delta^e(d,z) := \frac{1\{D=d,Z=z,S=e\}}{\Pr(D=d,Z=z,S=e)} - \frac{1\{Y=0,S=o\}}{\Pr(Y=0,S=o)}$ and $\Delta^o := \frac{1\{Y=1,S=o\}}{\Pr(Y=1,S=o)} - \frac{1\{Y=0,S=o\}}{\Pr(Y=0,S=o)}$.

Proof. The argument mirrors the proof of Lemma 1 and Theorem 1.

1. By the law of total probability,

$$\begin{aligned}
\delta^e_{d,z}(r) &:= f_R(r \,|\, S=e, D=d, Z=z) \\
&= \int f_{R,Y}(r,y \,|\, S=e, D=d, Z=z) \mathrm{d}y \\
&= \int f_R(r \,|\, S=e, D=d, Z=z, Y=y) f_Y(y \,|\, S=e, D=d, Z=z) \mathrm{d}y.
\end{aligned}$$

Next, notice that

$$\begin{aligned}
f_R(r \,|\, S=e, D=d, Z=z, Y=y) &= f_R(r \,|\, S=o, D=d, Z=z, Y=y) \\
&= f_R(r \,|\, S=o, Y=y) \\
&=: \delta^o_y(r),
\end{aligned}$$

where the first equality applies Assumption J.2 and the second equality applies Assumption J.3.

Combining the previous displays, we arrive at the general result

$$\delta^e_{d,z}(r) = \int \delta^o_y(r) f_Y(y \mid S=e, D=d, Z=z) \mathrm{d}y.$$

When Y is binary,

$$f_Y(1 \mid S=e, D=d, Z=z) = \mathbb{E}(Y \mid S=e, D=d, Z=z) = \alpha(d,z)$$
$$f_Y(0 \mid S=e, D=d, Z=z) = 1 - \mathbb{E}(Y \mid S=e, D=d, Z=z) = 1 - \alpha(d,z).$$

Therefore, the general result specializes to

$$\delta^e_{d,z}(r) = \delta^o_1(r)\alpha(d,z) + \delta^o_0(r)\{1-\alpha(d,z)\} = \delta^o_0(r) + \{\delta^o_1(r) - \delta^o_0(r)\}\alpha(d,z).$$

2. We apply Bayes' rule to rewrite

$$\delta^o_y(r) := f_R(r \mid S=o, Y=y) = \frac{\Pr(Y=y, S=o \mid R=r) f_R(r)}{\Pr(Y=y, S=o)},$$
$$\delta^e_{d,z}(r) := f_R(r \mid S=e, D=d, Z=z) = \frac{\Pr(D=d, Z=z, S=e \mid R=r) f_R(r)}{\Pr(D=d, Z=z, S=e)}.$$

Substituting these expressions into the previous step and canceling $f_R(r)$ yields

$$\mathbb{E}\{\Delta^e(d,z) - \Delta^o \alpha(d,z) \mid R\} = 0.$$

□

Corollary J.1 (Representation with an instrument)**.** Under Theorem J.1's conditions, for any $z \in \{0,1\}$, $d \in \{0,1\}$ and any representation $H(R)$ with $\mathbb{E}\{H(R)\Delta^o\} \neq 0$,

$$\alpha(d,z) = \frac{\mathbb{E}\{H(R)\Delta^e(d,z)\}}{\mathbb{E}\{H(R)\Delta^o\}}.$$

Theorem J.1 and Corollary J.1 immediately imply that $\alpha(0)$ and $\alpha(1)$ are identified; for $z \in \{0,1\}$, the law of iterated expectations gives $\alpha(z) = \alpha(1,z)\beta(z) + \alpha(0,z)\{1-\beta(z)\}$. Therefore, θ^{LATE} is also identified. Estimation and inference can then follow by suitably stacking moments and extending the discussion provided in Appendix E.

J.2 Two-Period Difference-in-Differences

Next, we consider a setting with two periods $t \in \{1,2\}$. Treated units ($D=1$) receive a treatment between period $t=1$ and $t=2$, and untreated units ($D=0$) remain untreated in both periods. Each unit is characterized by the random vector

$$\{S, D, Y_1(0), Y_1(1), Y_2(0), Y_2(1), R_1, R_2\}.$$

For units in the quasi-experimental sample ($S=e$), we observe (D,R_1,R_2), For units in the observational sample ($S=o$), we observe (Y_1,Y_2,R_1,R_2).

We again modify our previous assumption about the experimental sample, this time to accommodate the panel structure.

Assumption J.4 (Difference-in-differences)**.** Suppose the following:

i. SUTVA: For $t\in\{1,2\}$, $Y_t=DY_t(1)+(1-D)Y_t(0)$.

ii. Overlap: $\Pr(D=1)$ is bounded away from zero and one.

iii. Parallel trends: $\mathbb{E}\{Y_2(0)-Y_1(0)\,|\,S=e,D=1\}=\mathbb{E}\{Y_2(0)-Y_1(0)\,|\,S=e,D=0\}$.

iv. No anticipation: $\Pr\{Y_1(0)=Y_1(1)|S=e,D=1\}=1$.

Under Assumption J.4, if we were to observe the outcome in the quasi-experimental sample, then the average treatment effect on the treated $\theta^{ATT}:=\mathbb{E}\{Y_2(1)-Y_2(0)\,|\,S=e,D=1\}$ could be identified using standard arguments. Assumption J.4 is akin to a parallel trends-type assumption on the cumulative distribution functions for untreated potential outcomes (Roth and Sant'Anna, 2023).

We modify our stability and observational completeness assumptions to accommodate the panel setting.

Assumption J.5 (Stability for difference-in-differences)**.** Suppose Assumption 2 holds for each time period $t\in\{1,2\}$.

Assumption J.6 (No direct effects for difference-in-differences)**.** Suppose Assumption 3(ii) holds for each time period $t\in\{1,2\}$.

In other words, we impose the same assumptions as in the main text, period by period. The main conditions are $S\perp\!\!\!\perp R_t|D,Y_t$ and $D\perp\!\!\!\perp R_t|Y_t$. Together, Assumption J.5 and Assumption J.6 imply that, for each $t\in\{1,2\}$, $(S,D)\perp\!\!\!\perp R_t\,|\,Y_t$, which is a testable implication as before.

These conditions allow us to identify θ^{ATT} by combining the quasi-experimental and observational samples. As notation, let $\alpha_t(d)=\mathbb{E}(Y_t\,|\,S=e,D=d)$. By standard arguments, under Assumption J.4, $\theta^{ATT}=\{\alpha_2(1)-\alpha_1(1)\}-\{\alpha_2(0)-\alpha_1(0)\}$. We will identify $\alpha_t(d)$ for $t\in\{1,2\}$ and $d\in\{0,1\}$ by combining the quasi-experimental and observational samples, under Assumption J.5 and Assumption J.6.

Theorem J.2 (Identification for difference-in-differences)**.** Suppose Assumption J.4, J.5 and J.6 hold. Then, for any $t \in \{1,2\}$ and $d \in \{0,1\}$,

$$\alpha_t(d) = \frac{\mathbb{E}\{\Delta_t^e(d) | R_t\}}{\mathbb{E}(\Delta_t^o | R_t)} \text{ almost surely,}$$

where $\Delta_t^e(d) := \frac{1\{D=d,S=e\}}{\Pr(D=d,S=e)} - \frac{1\{Y_t=0,S=o\}}{\Pr(Y_t=0,S=o)}$ and $\Delta_t^o := \frac{1\{Y_t=1,S=o\}}{\Pr(Y_t=1,S=o)} - \frac{1\{Y_t=0,S=o\}}{\Pr(Y_t=0,S=o)}$.

Proof. The argument mirrors the proof of Lemma 1 and Theorem 1.

1. By the law of total probability,

$$\begin{aligned}
\delta_{d,t}^e(r) &:= f_{R_t}(r \,|\, S=e, D=d) \\
&= \int f_{R_t,Y_t}(r,y \,|\, S=e, D=d) \mathrm{d}y \\
&= \int f_{R_t}(r \,|\, S=e, D=d, Y_t=y) f_{Y_t}(y \,|\, S=e, D=d) \mathrm{d}y.
\end{aligned}$$

Next, notice that

$$\begin{aligned}
f_{R_t}(r \,|\, S=e, D=d, Y_t=y) &= f_{R_t}(r \,|\, S=o, D=d, Y_t=y) \\
&= f_{R_t}(r \,|\, S=o, Y_t=y) \\
&=: \delta_{y,t}^o(r),
\end{aligned}$$

where the first equality applies Assumption J.5 and the second equality applies Assumption J.6.

Combining the previous displays, we arrive at the general result

$$\delta_{d,t}^e(r) = \int \delta_{y,t}^o(r) f_{Y_t}(y \,|\, S=e, D=d) \mathrm{d}y.$$

When Y is binary,

$$\begin{aligned}
f_{Y_t}(1 \,|\, S=e, D=d) &= \mathbb{E}(Y_t \,|\, S=e, D=d) = \alpha_t(d) \\
f_{Y_t}(0 \,|\, S=e, D=d) &= 1 - \mathbb{E}(Y_t \,|\, S=e, D=d) = 1 - \alpha_t(d).
\end{aligned}$$

Therefore, the general result specializes to

$$\delta_{d,t}^e(r) = \delta_{1,t}^o(r)\alpha_t(d) + \delta_{0,t}^o(r)\{1-\alpha_t(d)\} = \delta_{0,t}^o(r) + \{\delta_{1,t}^o(r) - \delta_{0,t}^o(r)\}\alpha_t(d).$$

2. We apply Bayes' rule to rewrite

$$\begin{aligned}
\delta_y^o(r) &:= f_{R_t}(r \,|\, Y_t=y, S=o) = \frac{\Pr(Y_t=y, S=o \,|\, R_t=r) f_{R_t}(r)}{\Pr(Y_t=y, S=o)}, \\
\delta_d^e(r) &:= f_{R_t}(r \,|\, D=d, S=e) = \frac{\Pr(D=d, S=e \,|\, R_t=r) f_{R_t}(r)}{\Pr(D=d, S=e)}.
\end{aligned}$$

Substituting these expressions into the previous step and canceling $f_{R_t}(r)$ yields

$$\mathbb{E}\{\Delta_t^e(d) - \Delta_t^o \alpha_t(d) \,|\, R_t\} = 0.$$

□

Corollary J.2. Under Theorem J.2's conditions, for any $t \in \{1,2\}$, $d \in \{0,1\}$ and representation $H_t(R_t)$ with $\mathbb{E}\{H_t(R_t)\Delta_t^o\} \neq 0$,

$$\alpha_t(d) = \frac{\mathbb{E}\{H_t(R_t)\Delta_t^e(d)\}}{\mathbb{E}\{H_t(R_t)\Delta_t^o\}}.$$

Theorem J.2 and Corollary J.2 imply that $\alpha_t(d)$ are identified. Therefore, θ^{ATT} is also identified. Estimation and inference can again follow by suitably stacking moments and extending the discussion provided in Appendix E.