Program evaluation with remotely sensed outcomes

Ashesh Rambachan
MIT

Rahul Singh
Harvard

Davide Viviano*
Harvard

First draft: November 2024
This draft: April 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 to use its predictions as the outcomes in the experiment. We show that this method is biased whenever the RSV is *post*-outcome, i.e. if variation in the economic outcome causes variation in the RSV. In program evaluation, 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 that underlies common practice: the conditional distribution of the RSV given the outcome and treatment is stable across the samples. Based on our identifying formula, we find that the efficient representation of RSVs for causal inference requires three predictions rather than one. Valid inference does not require any rate conditions on RSV predictions, justifying the use of complex deep learning algorithms with unknown statistical properties. We re-analyze the effect of an anti-poverty program in India using satellite images.

Keywords: Causal inference, data fusion, treatment-experiments, satellite images

*We thank Isaiah Andrews, Josh Angrist, Arun Chandrasekhar, Melissa Dell, Namrata Kala, Sylvia Klosin, Ben Olken, Pritham Raja, Jonathan Roth, and Jesse Shapiro, as well seminar audiences at Stanford and Harvard/MIT 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.

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. 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); and satellite images as a measure of air 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; Aiken et al., 2022).[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 the papers in general interest economics journals from 2015-2024 that use remotely sensed outcomes.[2] The common practice is to predict the economic outcome from the RSV, and then to use 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.

As motivation, we prove that this intuitive method can lead to positive or negative bias in the estimated treatment effect 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 between the treatment and the outcome. 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 reversal; it is present even without machine learning.

As an example, consider a binary outcome indicating whether a plot of land has been burned

[1] Figure 8 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 analysis. 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% use a similar logic, but without an explicit formula for data combination.

(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 post-outcome. A common practice is to predict fires in the experimental sample, using a machine learning algorithm trained on labeled satellite images from an observational sample, and then to compute the difference of 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 at zero, regardless 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. Of independent interest, we propose a diagnostic to evaluate whether an RSV is relevant enough for an economic outcome to use it for program evaluation.

Our secondary contribution is to characterize the optimal representation of the RSV for inference on the treatment effect. Given our main result, the techniques are standard and hence straightforward to implement; our contribution is to point out the connection between modern remote sensing and classical conditional moments (Chamberlain, 1987; Newey, 1993), and to interpret its consequences. We find that three predictions are necessary for semiparametrically 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.

Because modern remote sensing typically involves unstructured data and complex machine learning, we derive valid $n^{-1/2}$ 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; (ii) the limit predicts the outcome of interest, and for this we provide a diagnostic. More precise predictions improve efficiency of program evaluation.

Finally, we conduct a semi-synthetic exercise, calibrated to an existing randomized control trial 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 embeddings of satellite images. Compared to an unbiased benchmark, estimated from true outcomes in the experiment, our method recovers similar treatment effects with similar precision, despite using remotely sensed variables to compensate for missing experimental outcomes.[3] By contrast, common practice can have positive or negative bias for the treatment effect. Conservative cost calculations suggest that using remotely sensed outcomes, instead of directly surveyed outcomes, can recover the treatment effect of interest, while saving about $3 million dollars in survey costs.

1.1 Related work

Compared to various auxiliary variable models in causal inference, we study a different auxiliary variable. Whereas a surrogate is a mediator between the treatment and outcome (Athey et al., 2024; Kallus and Mao, 2024), the RSV is a *post*-outcome variable in our framework. We show that misusing a post-outcome RSV as a surrogate leads to arbitrary biases for treatment effects. The negative control literature extends the surrogacy framework to deal with unobserved confounding (Ghassami et al., 2022; Imbens et al., 2024), yet such extensions face the same limitation. See Remark 1 for details.

Compared to a 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

[3]Experimental outcomes may be missing for a random half of villages, or for all treated villages.

the variable of interest, is stable across samples. This leads 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). By contrast, we allow the researcher to observe no outcomes for any experimental units.

Our results do not require a correctly specified generative model of how treatments and outcomes affect RSVs, which may be prone to mis-specification when the RSV is a satellite image. Several previous works propose methods based on generative modeling (possibly in combination with PPI) that do require correct specification (Gentzkow et al., 2019; Alix-Garcia and Millimet, 2023; Proctor et al., 2023; Battaglia et al., 2024). Similarly, methods for causal inference on outcomes that are latent concepts require a generative model (Egami et al., 2022; Knox et al., 2022; Stoetzer et al., 2024).

Section 2 formalizes our main assumption: stability of the RSV. Section 3 proves our main result: nonparametric identification. Section 4 characterizes the optimal representation, and provides inference without rate or complexity restrictions. Section 5 shows that our method outperforms common practice in evaluating an anti-poverty program from satellite images.

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.[4] 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 can be low dimensional (e.g., the output of some pre-trained machine learning algorithm). The researcher would like

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

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 define our object of interest.

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. We therefore 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 data set 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" 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

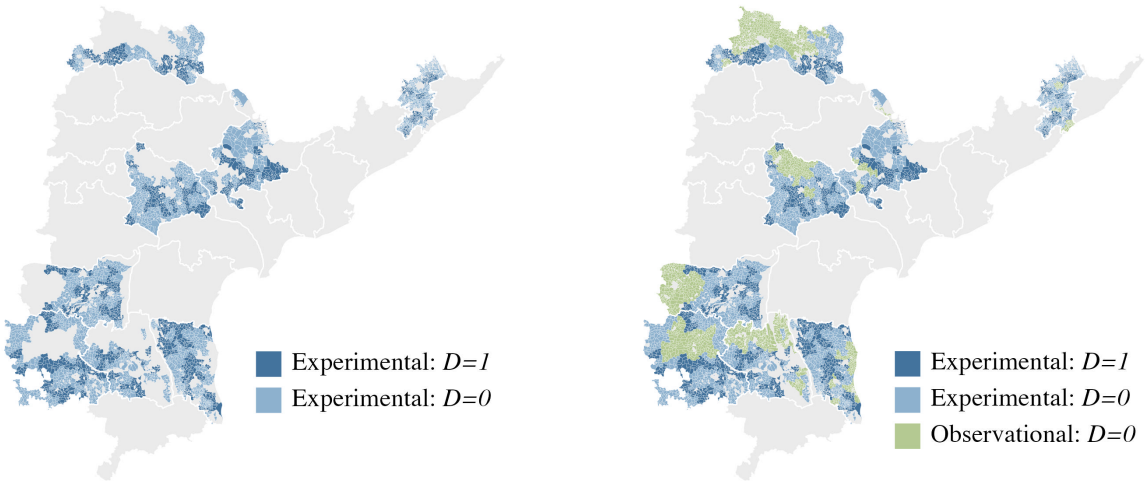
[5] Incomplete cases in the observational sample may refer to three scenarios. First, the treatment status may be missing. Second, the treatment status may be present, and all observational units are untreated, hence $D = 0$. Third, the treatment status may be present, and all observational units are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

[6] Independence is not used to derive our main identification argument, and our framework directly extends to weakly dependent data, as long as a central limit theorem applies.

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 the variable is observed. When treatment is missing or deterministic in the observational sample, we encode it as $D=0$ in the observational sample.



(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.

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 cheap 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 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 cheap 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. The first is standard.

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

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.

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, Assumption 2 allows us 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 $\Pr(Y = y \mid S = o, X)$ is bounded away from zero almost surely for all $y \in \mathcal{Y}$.

iii. Coverage: $\Pr(R=r \mid S,X,D)$ 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 on the underlying treatment effects, which may differ across samples. See Remark 1 for comparisons between Assumption 2(i) in the RSV model and alternative 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 the observational 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.

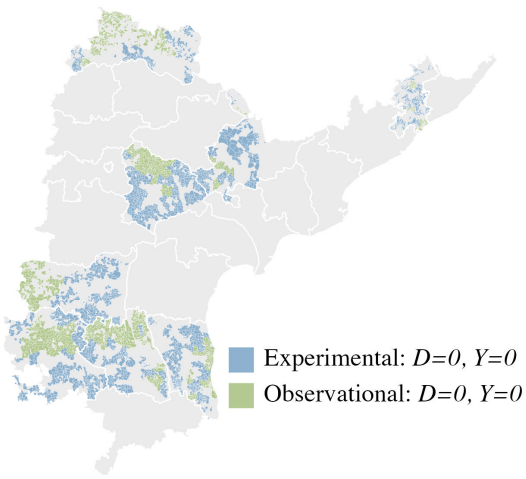
Our main assumption is empirically plausible, as illustrated by Figure 2. We use data from an anti-poverty program in India, where outcomes are observed.

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. [7]

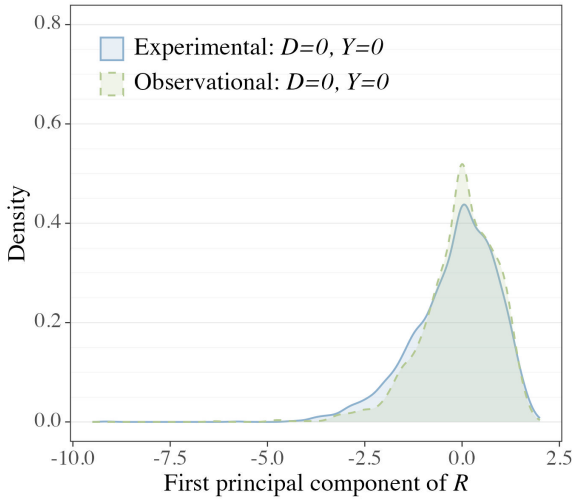
The remaining aspects of Assumption 2 are weak regularity conditions. Assumption 2(ii) requires that the outcome in the observational sample has a common (or larger) support than the outcome in the experiment. Assumption 2(iii) ensures that the RSV distribution does not degenerate for any stratum. Assumption 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.

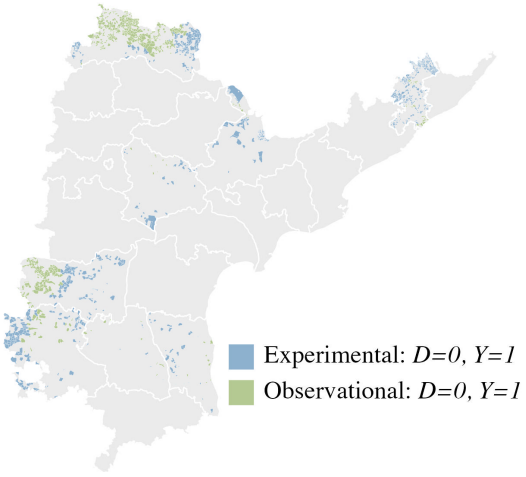
[7] When $X=\varnothing$, Assumption 2 imposes four equalities: $\Pr(R \mid S=e,D=d,Y=y)=\Pr(R \mid S=o,D=d,Y=y)$ for $d \in \{0,1\}$ and $y \in \{0,1\}$. Each equality can be evaluated with a diagnostic plot if outcome data are available. For example, using the experimental and observational units satisfying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



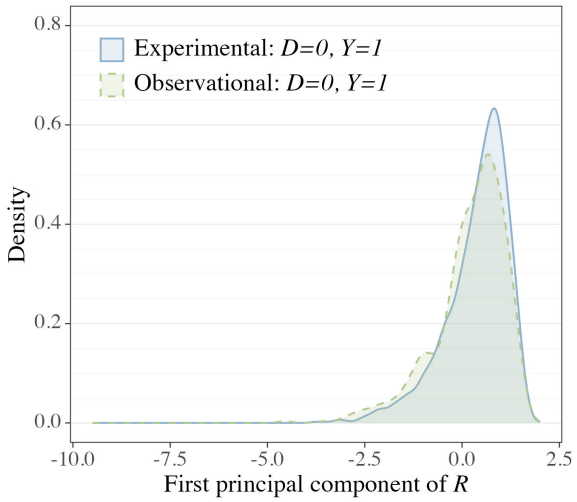
(a) Units with $D=0$ and $Y=0$.



(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 $\Pr(R \mid S=e, D=0, Y=0)$ with $\Pr(R \mid S=o, D=0, Y=0)$ in Figure 2b, and $\Pr(R \mid S=e, D=0, Y=1)$ with $\Pr(R \mid S=o, D=0, Y=1)$ in Figure 2d, using data from Muralidharan et al. (2023) 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.

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;

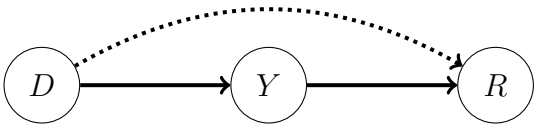


Figure 3: Causal graph of Assumptions 3(i) versus 3(ii). Complete cases allow the dotted line.

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

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 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 D only affects the remotely sensed variable R 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 become more plausible when Y is a vector of outcomes. Several outcomes may approximate all mechanisms though which the treatment affects the RSV. For readability, we focus on scalar outcomes in the main text, and vector outcomes in Appendix E.

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

Figure 3 illustrates our identifying assumptions as a causal graph. The treatment affects the outcome, which in turn affects the RSV. Depending on which version of Assumption 3 is imposed, the treatment may also have a direct effect on the RSV, as illustrated by the dotted line. Table 3 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. We will also assume the outcome is binary, i.e. $\mathcal{Y}=\{0,1\}$. Appendices E and F extend our results to (multivalued) discrete and continuous outcomes, respectively.

3.1 Current practice may have positive or negative bias

In empirical research, it is common for researchers 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; (ii) the predictor is applied on the experimental sample, and its predictions are used as surrogate outcomes to estimate treatment effects. While intuitive, this empirical strategy can lead to arbitrary bias in the downstream causal estimate, i.e. 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 $\widetilde{\mu}(d):=\mathbb{E}\{\mathbb{E}(Y\,|\,R,S=o)\,|\,D=d,S=e\}$ 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 on the treated and untreated subgroups in the experimental sample.

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 biases with arbitrary signs for the ATE in the experimental sample, even if units are randomly allocated between the experimental and observational samples.

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

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

2. There exists a data-generating process satisfying the above restrictions with $\widetilde{\theta}-\theta>0$, and a different data-generating process 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 $\Pr(R\,|\,Y,S=o)$ is stable across the experimental and the observational samples. Existing practice attempts to transport the predictions $\Pr(Y\,|\,R,S=o)$ 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.

Remark 1 (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 $(D,S)\perp\!\!\!\perp Y\,|\,R,X$, i.e. if surrogacy and surrogate compatibility are satisfied (Prentice, 1989; Athey et al., 2024). These assumptions are non-nested with 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. Below, we argue that the RSV model is more plausible in environmental and development applications with satellite images. Figure 9 illustrates the difference via a causal graph.

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), Chen et al. (2005), and 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 exists, 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 ecosystems 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 professional surveyor. 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.[8]

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 only affect these specific infrared bands via crop burning, so Assumption 3(ii) is reasonable. Figure 10 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%.[9]

[8] 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 4 using the authors' "balanced accuracy" RSV.

[9] We modify the main specification of Jack et al. (2025) in two ways. First, while they distinguish between two types of PES contracts, we define the treatment as whether any PES contract was offered. Second, the authors' analyze the effects of the 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.

However, by Proposition 1, $\widetilde{\theta}$ is typically biased for θ. To quantify the magnitude of its 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.1 yields $\widetilde{\theta} = \beta\theta$ in linear settings. 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%.

3.3 Main result: A novel 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 experimental and observational data.

To begin, we express the RSV distribution in the experimental sample as a mixture of potential outcomes. The mixture weights 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 experiment, while $\delta^e_d(r,x) := \Pr(R=r \mid S=e, X=x, D=d)$ and $\delta^o_y(r,x) := \Pr(R=r \mid S=o, X=x, Y=y)$ are RSV distributions.

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 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 possible way 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 mis-specification.

We follow a different path that does not require a generative model on 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 $\theta(x) := \mathbb{E}\{Y(1)-Y(0) \mid S=e, X=x\} = \mu(1,x)-\mu(0,x)$ denote the conditional average treatment effect in the experimental sample.

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

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

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

Theorem 1 identifies the treatment effect as the solution to a set of conditional moment equalities. Intuitively, 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 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 $\mathbb{E}\{H(x,R)\Delta^o(x) \mid X=x\} \neq 0$.

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 is the optimal representation of the RSV for downstream causal inference? Our answer develops a connection between "representation learning" e.g. Johannemann et al. (2019); Vafa et al. (2024) and classical results for conditional moment equalities.

Remark 2 (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.

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 $\mu(d,x) := \mathbb{E}\{Y(d) \,|\, S=e, X=x\}$ 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 $x \in \mathcal{X}$,

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

$\widetilde{\Delta}^e(d,x) := \frac{1\{D=d,S=e\}}{\Pr(D=d,S=e|X=x)} - \frac{1\{Y=0,D=d,S=o\}}{\Pr(Y=0,D=d,S=o|X=x)}$, 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 2 extend accordingly.

Remark 3 (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 4 (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. We describe how researchers can conduct estimation and inference using a discrete approximation, and characterize the

worst-case bias of that discrete approximation. Intuitively, more complexity of $\mathcal{Y}$ must be compensated by additional regularity elsewhere.

4 Estimation and inference

As a secondary contribution, we demonstrate that our main contribution (Theorem 1) implies an optimal representation of the RSV for downstream causal estimation and inference. The techniques are standard; we simply point out and interpret the connection between modern remote sensing and classical conditional moment analysis. Using well known 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.

For clarity, we focus on the case in which the outcome is binary, there are no pretreatment covariates, and Assumption 3(ii) holds. A more general case is straightforward to implement by incorporating additional moment conditions, as discussed in Appendix E.

4.1 Optimal representation for program evaluation

In this setting, Theorem 1 and Corollary 1 simplify. The causal parameter satisfies $\mathbb{E}(\Delta^e - \Delta^o\theta \mid R) = 0$ and hence $\mathbb{E}\{(\Delta^e - \Delta^o\theta)H(R)\} = 0$ for any predictive representation $H(R)$. The treatment and outcome variation simplify to scalars:

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

Since (S,D,Y) are binary, the denominators can be estimated by simple counts.

While any predictive representation H gives inference, different choices have different efficiency properties. We appeal to classical ideas in econometrics to derive the optimal representation H^*, which achieves the semiparametric efficiency bound. Write the conditional moment as the regression $\Delta^e = \Delta^o\theta + \epsilon$ with $\mathbb{E}(\epsilon \mid R) = 0$. The optimal representation is then $H^*(R) = \frac{\mathbb{E}(\Delta^o|R)}{\sigma^2(\theta,R)}$, where $\sigma^2(\theta,R) := \mathbb{E}\{(\Delta^e - \Delta^o\theta)^2 \mid R\}$ (Chamberlain, 1987; Newey, 1993).

For downstream causal inference, the optimal representation of the high dimensional remotely sensed variable $R \in \mathcal{R}$ is a scalar $H^*(R) \in \mathbb{R}$. Concretely, $H^*(R)$ is the optimal compression of a satellite image for the task of program evaluation. Interestingly, different causal estimands imply different optimal compressions.

This connection has another consequence for empirical practice: to optimally use 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 most 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 optimal 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 causal estimate, by regressing $\widehat{\mathbb{E}}(\Delta^e|R)$ on $\widehat{\mathbb{E}}(\Delta^o|R)$: $\widehat{\theta}_{\text{INIT}}$.
   - Combine these into the optimal representation: $\widehat{H}(R)$.
3. Construct an efficient causal estimate on TEST: $\widehat{\theta}$.

The overall structure is familiar (Angrist et al., 1999; Chernozhukov et al., 2018). To simplify the algorithm statement, we introduce some notation for simple events and an algebraic identity.

Let $\mathbb{E}_{\text{TRAIN}}=\frac{1}{|\text{TRAIN}|}\sum_{i\in\text{TRAIN}}(\cdot)$ and $\mathbb{E}_{\text{TEST}}=\frac{1}{|\text{TEST}|}\sum_{i\in\text{TRAIN}}(\cdot)$. 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. $(\Delta^e-\Delta^o\theta)^2=\frac{1\{D=1,S=e\}}{\Pr(D=1,S=e)^2}+\frac{1\{D=0,S=e\}}{\Pr(D=0,S=e)^2}+\theta^2\left[\frac{1\{Y=1,S=o\}}{\Pr(D=Y,S=o)^2}+\frac{1\{Y=0,S=o\}}{\Pr(Y=0,S=o)^2}\right]$.

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 optimal 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 optimal 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 an efficient 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 an efficient 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 mis-specification

In empirical research, the RSV is typically unstructured and high dimensional, e.g. a satellite image. The prediction is typically 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 mis-specified.

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

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 Theorem 1 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).

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 the conditional moment in Theorem 1.

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). When marginal probabilities are unknown, the asymptotic variance is weighted by them (Proposition 3). For the latter, 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 remain essentially the same.

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 Andra 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. See Appendix G 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 luminosity (a scalar) (Asher et al., 2021) and satellite images (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 3). We will be explicit about sample definitions below.

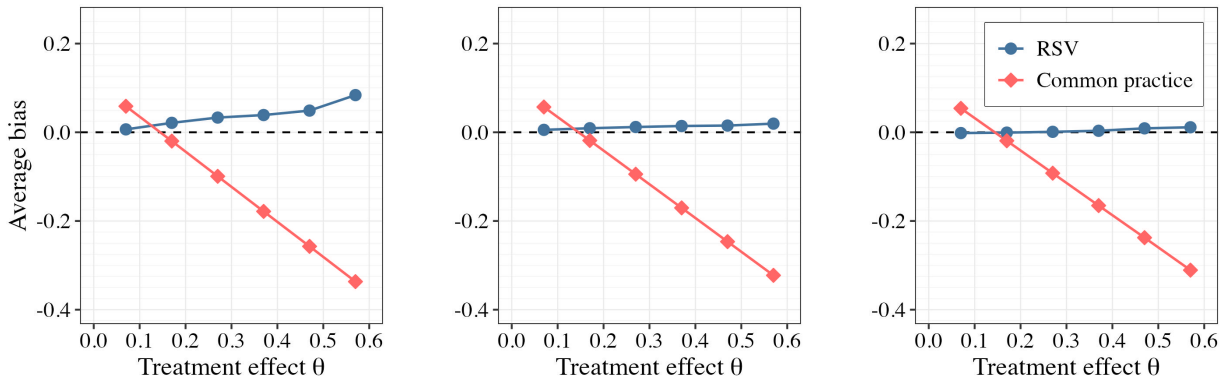
5.1 Our method outperforms current practice across effect sizes and sample sizes

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.

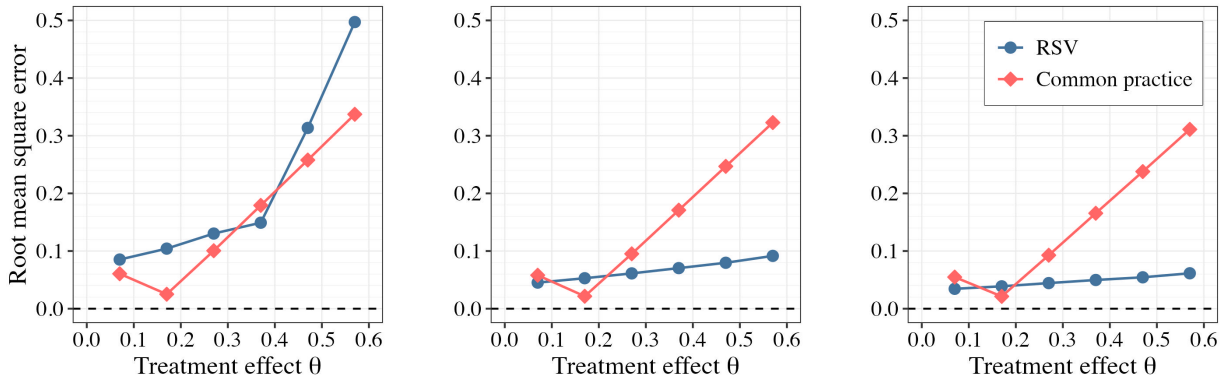
Next, we draw RSV values from the real data. If $Y=0$, we draw R from the empirical distribution of $R|Y=0$ in the real experimental data. Likewise, we draw R when $Y=1$. This imposes no direct effects (Assumption 3(ii)). For computational feasibility, we use luminosity and only the initial 1,000 satellite image features as our RSV.

Finally, to mimic a setting with missing outcomes, we delete Y if $D=1$. In terms of



(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.

Remark 3, we set $\tilde{S}=e$ when $D=1$ and $\tilde{S}=\{e,o\}$ when $D=0$.[10] This imposes stability across samples (Assumption 2). Recent methods with 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.

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

[10]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 $\tilde{S}=o$. Either way, Theorem 1 applies: use $e\in\tilde{S}$ instead of $S=e$ in Δ^e, and $o\in\tilde{S}$ instead of $S=o$ in Δ^o.

$\tau \in \{0,0.1,...,0.5\}$ 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. 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 foreshadowed 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 both 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.[11]

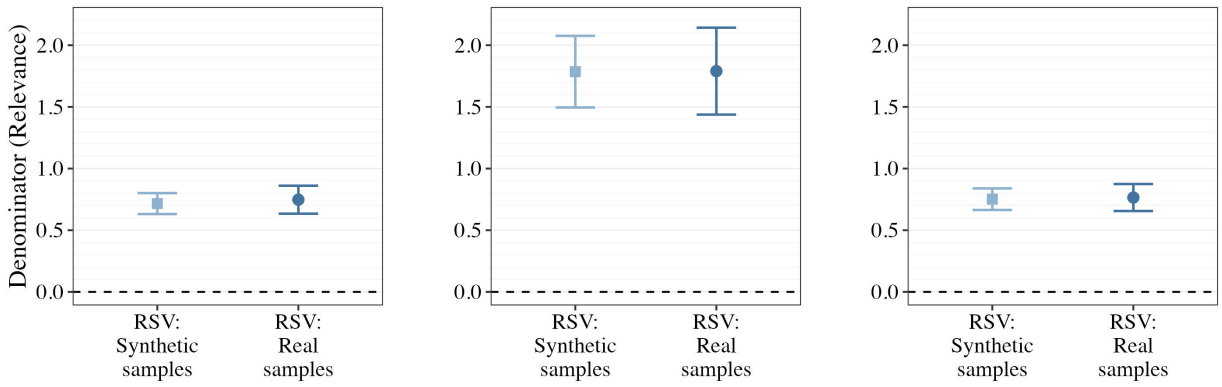
We conduct this exercise with three possible poverty measurements: (i) is the village in the bottom quartile of villages for per capita consumption, (ii) does a village have only low income households, and (iii) does a village have only low and middle income households. While (i) is a measure of average consumption, (ii) and (iii) describe the income distribution using formal categories from Indian administrative data; see Appendix G for details.[12]

In this exercise, we now use the full RSV: luminosity, 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) and Remark 3. 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

[11] Following Remark 3, we also observe some outcomes in the experiment, as described below.

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



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

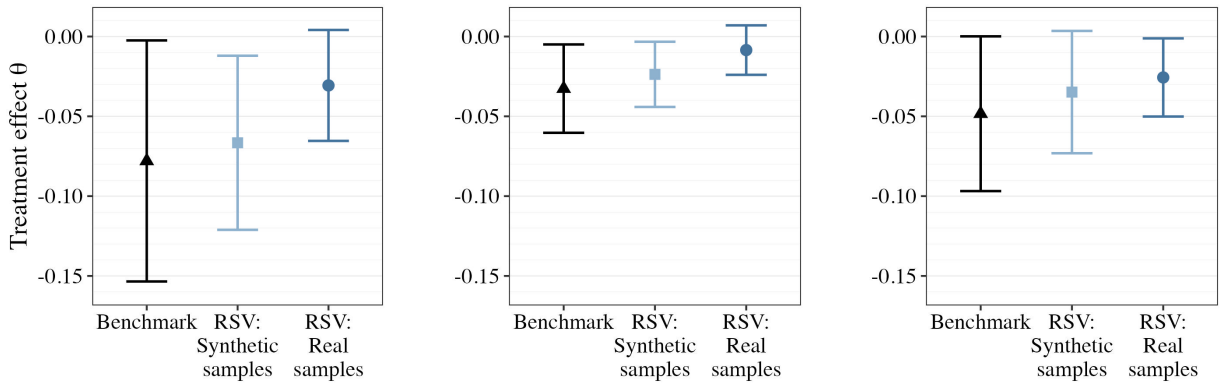
Figure 6: Satellite images are relevant to the poverty outcomes. Each plot is for a poverty outcome. Within each plot, we report $\mathbb{E}_n\{\widehat{H}(R)\widehat{\Delta}^o\}$ and its 90% confidence interval using our learned representation. In light blue, we visualize the relevance of our learned 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.

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 11 and 12 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 G. Our optimal representation is based on three predictions: the outcome, treatment, 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,



(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.

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 magnitudes as the benchmark, i.e. as if the economist could observe treatments and outcomes in the experiment. These effects are consistent 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.[13]

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 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

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

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.[14]

5.3 Our method recovers the true effect in a realistic setting

Finally, as our third exercise, we use the data of Muralidharan et al. (2016, 2023) as realistically as possible. 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 11 and 12 (Assumption 3(ii)). Moreover, the RSV is relevant in Figure 6 (Corollary 1).

In this more challenging exercise, we find qualitatively similar results. Our learned representations allow efficient extrapolation in Figure 13. Our method recovers similar estimates and confidence intervals to the benchmark in Figure 7 ("real samples"), consistent with prior work.

6 Recommendations for practice

Common empirical practice can be highly biased if the remotely sensed variable is *post*-outcome, i.e. 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 the resulting bias can be positive or negative. Empirically, we demonstrate that common practice may attenuate the treatment effect in a real environmental application, and it may have positive or negative bias for the treatment effect in a semi-synthetic development application.

However, the intuition underlying empirical work is powerful: the conditional distribution of the RSV given the outcome and treatment is stable across samples. We use such an

[14]Table 5 summarizes the number of experimental villages and the populations in experimental villages, which we use to calculate savings. This calculation is extremely conservative; Viviano and Rudder (2024) find that phone surveys in Pakistan cost \$7 per individual, rather than \$0.50.

assumption to nonparametrically identify treatment effects, and derive a novel formula. We demonstrate that it is empirically plausible in a real development application. Finally, we characterize the optimal way to use remotely sensed outcomes for program evaluation.

Our findings inform how to conduct program evaluation with RSVs in practice.

- **Auxiliary sample.** Collect RSVs with linked outcomes in an auxiliary, observational sample. In this sample, the treatments may be missing or deterministic. It is helpful, but unnecessary, to collect some outcomes in the primary, experimental sample. The distribution of RSVs given outcomes and treatments should be stable across samples.

- **Three predictions.** Use machine learning to predict not only the outcome, but also the treatment and sample indicator, given the RSV. Researchers can use complex machine learning methods with unknown statistical properties, while remaining agnostic about how the RSV is generated.

- **Robust inference.** Use Algorithm 1 to aggregate the three predictions into an efficient representation of the RSV for program evaluation. Algorithm 1 is robust to misspecification; it provides valid inference as long as the learned representation converges at any rate to any limit that is correlated with outcome variation.

Our framework clarifies the key diagnostics that empirical researchers should assess.

- **RSV relevance test.** Evaluate whether the learned RSV representation is correlated with outcome variation. Tests for weak instruments can detect weak RSVs.

- **Stability and no direct effect joint test.** If the treatment effects estimated with two different RSV representations are significantly different, then reject the null hypothesis that both conditions hold: stability (Assumption 2), and no direct effects (Assumption 3(ii)).

- **RSV density plots.** If possible, collect some experimental outcomes. If the conditional density of the RSV given the outcome and treatment is similar across experimental and observational units, then stability (Assumption 2) is plausible. If it is similar across treated and untreated experimental units, then no direct effects (Assumption 3(ii)) is plausible.

In summary, researchers may substantially reduce research costs by using post-outcome RSVs instead of directly measured outcomes, without compromising valid inference.

Our framework poses new questions, e.g. how to jointly design experiments (treatment assignment), surveys (outcome collection), and sensors (RSV extraction) to optimally estimate treatment effects, subject to cost constraints. Our main result (Theorem 1) may inform not only the use of satellite images, but also the design of cheap, noisy surveys. Future work may also extend our inference guarantees (Propositions 2 and 3) to accommodate high dimensional outcomes.

As remote sensing expands the frontier of data availability in economics, our framework provides practical principles for its use in program evaluation.

References

Aiken, E., Bellue, S., Karlan, D., Udry, C., and Blumenstock, J. E. (2022). Machine learning and phone data can improve targeting of humanitarian aid. *Nature*, 603(7903):864–870.

Alix-Garcia, J. and Millimet, D. L. (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., Chen, D., Jiang, Z., and Zhang, D. (2023). Machine learning and prediction errors in causal inference. *The Wharton School Research Paper*.

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

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

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

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

Athey, S., Chetty, R., Imbens, G. W., and Kang, H. (2024). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. *NBER Working Paper Series*.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Currie, J., Voorheis, J., and Walker, R. (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., Gaillac, C., and Maurel, A. (2025). Partially linear models under data combination. *The Review of Economic Studies*, 92(1):238–267.

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

Egami, N., Hinck, M., Stewart, B., and Wei, H. (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., Haushofer, J., Miguel, E., Niehaus, P., and Walker, M. (2022). General equilibrium effects of cash transfers: Experimental evidence from kenya. *Econometrica*, 90(6):2603–2643.

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

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

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

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

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

Henderson, J. V., Storeygard, A., and Weil, D. N. (2011). A bright idea for measuring economic growth. *American Economic Review*.

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

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

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

Imbens, G., Kallus, N., Mao, X., and Wang, Y. (2024). Long-term causal inference under persistent confounding via data combination. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae095.

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

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

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

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

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

Kallus, N. and Mao, X. (2024). 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*, page qkae099.

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

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

Lu, K., Kluger, D. M., Bates, S., and Wang, S. (2025). Regression coefficient estimation from remote sensing maps. *arXiv:2407.13659*.

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

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

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

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

Muralidharan, K., Niehaus, P., and Sukhtankar, S. (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 Engle, R. F. and McFadden, D. L., editors, *Handbook of Statistics*, volume 11, chapter 16, pages 419–454. Elsevier.

Park, C., Richardson, D. B., and Tchetgen Tchetgen, E. J. (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., Carleton, T., and Sum, S. (2023). Parameter recovery using remotely sensed variables. *National Bureau of Economic Research*.

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

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

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

Sherman, L., Proctor, J., Druckenmiller, H., Tapia, H., and Hsiang, S. M. (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., Mills, B., and Del Ninno, C. (2016). Reaching the poor: Cash transfer program targeting in cameroon. *World Development*, 83:244–263.

Stoetzer, L. F., Zhou, X., and Steenbergen, M. (2024). Causal inference with latent outcomes. *American Journal of Political Science*.

Vafa, K., Athey, S., and Blei, D. M. (2024). Estimating wage disparities using foundation models. *arXiv:2409.09894*.

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

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

A Related work and model details

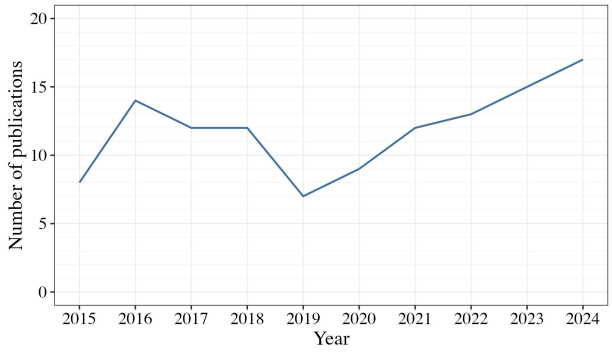


Figure 8: Remotely sensed variables are increasingly popular in published papers.

Figure 8 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.

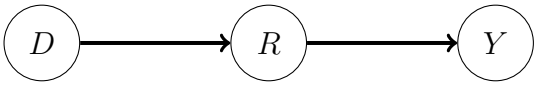


Figure 9: Causal graph of surrogacy model.

Figure 9 illustrate the causal graph associated with the surrogacy identifying assumptions (Prentice, 1989; Athey et al., 2024): the surrogate R fully mediates the effect of the treatment D on the outcome Y. Our RSV identifying assumptions are the opposite, as illustrated by Figure 3.

Table 3: 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 3 summarizes the implications of our identifying assumptions. Our identifying assumptions allow the propensity score $\Pr(D=1\,|\,S,X)$ to differ across samples.

B Crop burning illustration

B.1 Characterizing 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,

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

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 (Binary)**.** 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

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

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.

Table 4: Underestimation of treatment effects in crop burning experiment: Revisited.

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

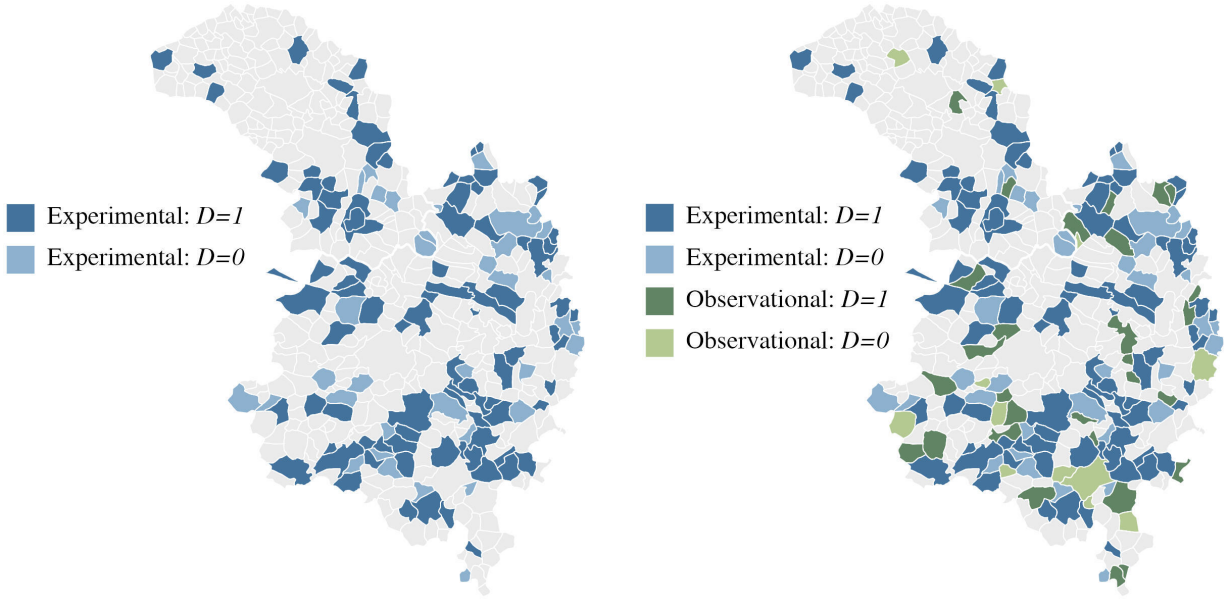
Notes: The RSV is the field-level "balanced accuracy" label defined by Jack et al. (2025), which applies a threshold rule to the predicted probability of not being burned. The "observational sample" has fields that received a random spot check, and the "experimental sample" has other fields. For illustration,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

B.2 Implementation details

We present details for the empirical example in Section 3.2, based on Jack et al. (2025). 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 4, 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 10a, we classify a village as experimental if less than 50% of its fields received a random spot check. Similarly, in Figure 10b, we classify a village as observational if more than 50% of its fields received a random spot check.



(a) Experimental units only.

(b) Auxiliary sample: Observational units.

Figure 10: 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.

C Identification proofs

C.1 Proof of Proposition 1

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

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

where (1) follows by Assumption 1. Next we rewrite the implicit target

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

In particular,

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

where (1) follows by Bayes' rule; (2) follows since $(S,D)\perp\!\!\!\perp R\,|\,Y$ 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 $S\perp\!\!\!\perp(R,Y)\,|\,D$, which in turn implies $S\perp\!\!\!\perp Y\,|\,R,D$. Second, since we further assume $\Pr(D=1\,|\,S=o)=0$, we have that

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

where the second equality uses $S\perp\!\!\!\perp Y\,|\,D$ and the third equality uses Assumption 1. Similarly, $\Pr(R\,|\,D=1,S=e)=\Pr(R\,|\,D=1)$ using $R\perp\!\!\!\perp S\,|\,D$ and

$$\Pr(R\,|\,S=o)=\Pr(R\,|\,D=0,S=o)=\Pr(R\,|\,D=0)$$

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

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

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

and $\widetilde{\mu}(1)-\mu(1)$ equals

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

Further rearranging, we can write

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

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

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

We can follow a similar argument for $\widetilde{\mu}(0)$. In particular, $\widetilde{\mu}(0)$ equals

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

The first expression uses identical arguments as before, replacing $D=1$ with $D=0$. The equality uses Assumption 1, SUTVA, $S \perp\!\!\!\perp Y|D$, and $D=0$ when $S=o$ to simplify

$$\begin{aligned}\Pr\{Y(0)=1\,|\,S=e\} &= \Pr\{Y(0)=1\,|\,D=0,S=e\} = \Pr(Y=1\,|\,D=0,S=e) \\ &= \Pr(Y=1\,|\,D=0,S=o) = \Pr(Y=1\,|\,S=o),\end{aligned}$$

as well as $\Pr(R=r\,|\,D=0,S=e)=\Pr(R=r\,|\,D=0)$ as before.

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

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

$$R\,|\,Y,D,S = \begin{cases} Y \text{ with probability } 1/2 \\ 0 \text{ otherwise.} \end{cases}$$

This satisfies both $S \perp\!\!\!\perp (R,Y)\,|\,D$ and $D \perp\!\!\!\perp R\,|\,Y$ (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,

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

By similar arguments to those above, we therefore have

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

using the extra structure $S \perp\!\!\!\perp R\,|\,D$, 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.2 Proof of Lemma 1

Let $\delta^e_{y,d}(r,x) := \Pr(R=r \,|\, Y=y, D=d, S=e, X=x)$. By the law of total probability,

$$\begin{aligned}\delta^e_d(r,x) &= \delta^e_{0,d}(r,x) + \left\{\delta^e_{1,d}(r,x) - \delta^e_{0,d}(r,x)\right\} \Pr\{Y(d)=1 \,|\, D=d, S=e, X=x\} \\ &= \delta^e_{0,d}(r,x) + \left\{\delta^e_{1,d}(r,x) - \delta^e_{0,d}(r,x)\right\} \Pr\{Y(d)=1 \,|\, S=e, X=x\}\end{aligned}$$

where the second equality applies Assumption 1. Next, notice that

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

where the first equality applies Assumption 2 and the second equality applies Assumption 3(ii). Combining the previous two displays,

$$\delta^e_d(r,x) = \delta^o_0(r,x) + \{\delta^o_1(r,x) - \delta^o_0(r,x)\} \Pr\{Y(d)=1 \,|\, S=e, X=x\}. \quad \square$$

C.3 Proof of Theorem 1

To prove this result, we apply Bayes' rule to rewrite

$$\begin{aligned}\delta^o_y(r,x) &:= \Pr(R=r \,|\, Y=y, S=o, X=x) = \frac{\Pr(Y=y, S=o \,|\, R=r, X=x) \Pr(R=r|X=x)}{\Pr(Y=y, S=o|X=x)}, \\ \delta^e_d(r,x) &:= \Pr(R=r \,|\, D=d, S=e, X=x) = \frac{\Pr(D=d, S=e \,|\, R=r, X=x) \Pr(R=r|X=x)}{\Pr(D=d, S=e|X=x)}.\end{aligned}$$

Applying Lemma 1, we have for $d \in \{0,1\}, x \in \mathcal{X}$ and $r \in \mathcal{R}$,

$$\frac{\Pr(D=d, S=e \,|\, R=r, X=x)}{\Pr(D=d, S=e|X=x)} - \frac{\Pr(Y=0, S=o \,|\, R=r, X=x)}{\Pr(Y=0, S=o|X=x)} =$$

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

By iterated expectations, this can be further rewritten as

$$\mathbb{E}\left[\frac{1\{D=d, S=e\}}{\Pr(D=d, S=e|X=x)} - \frac{1\{Y=y, S=o\}}{\Pr(Y=y, S=o|X=x)} \,|\, R=r, X=x\right] =$$

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

Consequently, it immediately follows that

$$\mathbb{E}\left[\frac{1\{D=1,S=e\}}{\Pr(D=1,S=e)}-\frac{1\{D=0,S=e\}}{\Pr(D=0,S=e)}\,|\,R=r,X=x\right]=$$

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

proving the result as desired. □

C.4 Proof of Theorem 2

The proof follows the same steps as the proofs of Lemma 1 and Theorem 1. Applying the law of total probability, for $d\in\{0,1\}$ and $r\in\mathcal{R}$, we have

$$\delta_d^e(r,x)=\delta_{0,d}^e(r,x)+\left\{\delta_{1,d}^e(r,x)-\delta_{0,d}^e(r,x)\right\}\Pr\{Y(d)=1\,|\,S=e,X=x\}$$

for $\delta_{y,d}^e(r,x):=\Pr(R=r\,|\,Y=y,D=d,S=e,X=x)$. By Assumptions 2 and 3(i),

$$\Pr(R=r\,|\,Y=y,D=d,S=e,X=x)=\Pr(R=r\,|\,Y=y,D=d,S=o,X=x).$$

Combining the previous two displays, we have

$$\delta_d^e(r,x)=\delta_{0,d}^o(r,x)+\left(\delta_{1,d}^o(r,x)-\delta_{0,d}^o(r,x)\right)\Pr\{Y(d)=1\,|\,S=e,X=x\}.$$

We next apply Bayes' rule to rewrite

$$\delta_d^e(r,x)=\frac{\Pr(D=d,S=e\,|\,R=r,X=x)\Pr(R=r|X=x)}{\Pr(D=d,S=e|X=x)}$$
$$\delta_{y,d}^o(r,x)=\frac{\Pr(Y=y,D=d,S=o\,|\,R=r,X=x)\Pr(R=r|X=x)}{\Pr(Y=y,D=d,S=o|X=x)}.$$

It therefore follows that, for $d\in\{0,1\}$ and $r\in\mathcal{R}$,

$$\frac{\Pr(D=d,S=e\,|\,R=r,X=x)}{\Pr(D=d,S=e|X=x)}-\frac{\Pr(Y=0,D=d,S=o\,|\,R=r,X=x)}{\Pr(Y=0,D=d,S=o|X=x)}=$$

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

By iterated expectations, this can be further rewritten as

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

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

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 optimal 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):

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

(d) Learn the optimal 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 an efficient 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 an efficient 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):

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

(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,

$$(\Delta^e - \Delta^o\theta)^2 = (\Delta^e)^2 - (2\theta)\Delta^e\Delta^0 + \theta^2(\Delta^o)^2 = (\Delta^e)^2 + \theta^2(\Delta^o)^2.$$

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 $\hat{Z} = \widehat{H}(R)$. 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

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

In the inequality, we use $E(U_i^2|Z_i,I_{-\ell})=\mathbb{E}(U_i^2|Z_i)\leq\bar{\sigma}_U^2$, where $\bar{\sigma}_U^2$ exists by hypothesis. 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. The latter is finite by hypothesis, 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

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

We use $|X_i|\leq\bar{X}$ almost surely, which follows from our assumptions. □

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 and $\mathbb{E}(Z_i^2)<\infty$. □

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 $\mathbb{E}_n(\hat{X}\hat{Z}) - \mathbb{E}_n(\hat{X}Z) = o_p(1)$, $\mathbb{E}_n(\hat{X}Z) \xrightarrow{p} \mathbb{E}_n(XZ) = o_p(1)$, 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 $n^{1/2}(\hat{\theta} - \theta) \rightsquigarrow \mathcal{N}\left[0, \frac{V}{\{\mathbb{E}(XZ)\}^2}\right]$ by Slutsky's theorem. In the following lemmas, we prove the remaining high probability statements. □

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, if $\mathbb{E}(Z^2)<\infty$, then 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,

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

Third, the derivative is

$$\{\nabla h(B)\}^\top=\Big(B_5^{-1},\ \ -B_6^{-1},\ \ -\theta B_7^{-1},\ \ \theta B_8^{-1},\ \ -B_1B_5^{-2},\ \ B_2B_6^{-2},\ \ \theta B_3B_7^{-2},\ \ -\theta B_4B_8^{-2}\Big).$$

Therefore, by the delta method,

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

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. $\square$

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

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 $\mathbb{E}_n(Z^2)\xrightarrow{p}\mathbb{E}(Z^2)$ 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 $\mathbb{E}_\ell\{(\hat{X}_i-X_i)^2\}$,

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

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 $\mathbb{E}_\ell\{(\hat{X}-X)^2\}\xrightarrow{p}0$. 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

Our identification results directly extend from $\mathcal{Y}=\{0,1\}$ to $\mathcal{Y}=\{y_1,...,y_{|\mathcal{Y}|}\}$.

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

1. If Assumption 3(i) holds, then

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

2. If Assumption 3(ii) holds, then

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

Proof. By iterated expectations, we write

$$\begin{aligned}\Pr(R\,|\,D=d,S=e,X)&=\sum_{y\in\mathcal{Y}}\Pr(R\,|\,Y=y,D=d,S=e,X)\Pr(Y=y\,|\,D=d,S=e,X)\\&=\sum_{y\in\mathcal{Y}}\Pr(R\,|\,Y=y,D=d,S=e,X)\Pr\{Y(d)=y\,|\,S=e,X\},\end{aligned}$$

where the second equality uses Assumption 1. Under Assumptions 2 and 3(i), $\Pr(R\,|\,Y=y,D=d,S=e,X)=\Pr(R\,|\,Y=y,D=d,S=o,X)$, delivering the first result. Under Assumptions 2 and 3(ii), it follows that $(S,D)\perp\!\!\!\perp R\,|\,X,Y$, and so $\Pr(R\,|\,Y=y,S=e,D=d,X)=\Pr(R\,|\,Y=y,S=o,X)$, delivering the second result. □

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

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

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

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

Proof. Consider the first statement. By Bayes' rule,

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

Rewriting Lemma E.1.1 and canceling $\Pr(R=r\,|\,X)$ from both sides yields, as desired,

$$\begin{aligned}&\frac{\Pr(D=d,S=e\,|\,R,X)}{\Pr(D=d,S=e\,|\,X)}\Pr(R=r\,|\,X)\\ &= \sum_{y\in\mathcal{Y}} \frac{\Pr(Y=y,D=d,S=o\,|\,R,X)}{\Pr(Y=y,D=d,S=o\,|\,X)}\Pr(R=r\,|\,X)\Pr\{Y(d)=y\,|\,S=e,X\}.\end{aligned}$$

Consider the second statement. Again, by Bayes' rule,

$$\Pr(R=r\,|\,Y=y,S=o,X) = \frac{\Pr(Y=y,S=o\,|\,R,X)}{\Pr(Y=y,S=o\,|\,X)}\Pr(R=r\,|\,X).$$

The desired result follows by rewriting Lemma E.1.2 and canceling $\Pr(R=r\,|\,X)$. □

The previous two lemmas give the generalization of our main result, expressing the causal parameter as a conditional moment and unconditional moment.

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

1. If Assumption 3(i) holds, then, almost surely,

$$\mathbb{E}\left[\frac{1\{D=d,S=e\}}{\Pr(D=d,S=e\,|\,X)} - \sum_{y\in\mathcal{Y}} \frac{1\{Y=y,D=d,S=o\}}{\Pr(Y=y,D=d,S=o\,|\,X)}\Pr\{Y(d)=y\,|\,X,S=e\}\,|\,X,R\right] = 0.$$

2. If Assumption 3(ii) holds, then, almost surely,

$$\mathbb{E}\left[\frac{1\{D=d,S=e\}}{\Pr(D=d,S=e\,|\,X)} - \sum_{y\in\mathcal{Y}} \frac{1\{Y=y,S=o\}}{\Pr(Y=y,S=o\,|\,X)}\Pr\{Y(d)=y\,|\,X,S=e\}\,|\,X,R\right] = 0.$$

Proof. The result follows from Lemma E.2 and iterated expectations. □

Let $\theta_d(X) := \big(\Pr\{Y(d)=y_1 \,|\, S=e,X\},...,\Pr\{Y(d)=y_{|\mathcal{Y}|} \,|\, S=e,X\}\big)^\top$.

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

1. If Assumption 3(i) holds, then

$$\mathbb{E}\left[H_d(X,R)\frac{1\{D=d,S=e\}}{\Pr(D=d,S=e \,|\, X)} \,|\, X\right] = \mathbb{E}\{H_d(X,R)\Delta^o(d) \,|\, X\}\theta_d(X)$$

for $\Delta^o(d) := \left(\frac{1\{Y=y_1,D=d,S=o\}}{\Pr(Y=y_1,D=d,S=o)},...,\frac{1\{Y=y_{|\mathcal{Y}|},D=d,S=o\}}{\Pr(Y=y_{|\mathcal{Y}|},D=d,S=o)}\right)$.

2. If Assumption 3(ii) holds, then

$$\mathbb{E}\left[H_d(X,R)\frac{1\{D=d,S=e\}}{\Pr(D=d,S=e \,|\, X)} \,|\, X\right] = \mathbb{E}\{H_d(X,R)\Delta^o \,|\, X\}\theta_d(X)$$

for $\Delta^o := \left(\frac{1\{Y=y_1,S=o\}}{\Pr(Y=y_1,S=o)},...,\frac{1\{Y=y_{|\mathcal{Y}|},S=o\}}{\Pr(Y=y_{|\mathcal{Y}|},S=o)}\right)$.

As in the main text, Theorem E.1 and Corollary E.1 imply that we can use existing results on conditional moment restrictions e.g. Chamberlain (1987); Newey (1993) for identification and inference on the average potential outcomes in the experimental sample. By further averaging over the conditional distribution $R \,|\, X$, we can construct unconditional moment restrictions using representations $H_d(X,R)$. Importantly, any choice of representation such that $\mathbb{E}\{H_d(X,R)\Delta^o(d) \,|\, X\}$ or $\mathbb{E}\{H_d(X,R)\Delta^o \,|\, X\}$ has full rank suffices for identification and valid asymptotic inference. The optimal representation is an obvious extension of our characterization in the main text.

As in the main text, Theorem E.1 and Corollary E.1 provide testable implications of the identifying assumptions through over-identifying restrictions: we can compare whether alternative choices of the representation H_d yield significantly different estimates of the average potential outcomes in the experimental sample.

F Continuous outcomes

We extend our results outcomes with a continuous and bounded support.

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 $s \in \{e,o\}$, $Y(d) \,|\, X,S=s,R$ has a positive density over $[-U,U]^p$ almost surely.

Let $f_{Y(d)|X,S,R}(y)$ denote the conditional density of $Y(d)\,|\,X,S=s,R$.

To extend our analysis to continuous outcomes, we construct bins over the support such that $y\in\mathcal{Y}_\varepsilon$ with $B_y(\varepsilon)$ defining an l_∞-ball of radius $\varepsilon>0$ around the value y, meaning $||\widetilde{y}-\widetilde{y}'||_\infty\leq\varepsilon$ for any two values $\widetilde{y},\widetilde{y}'\in B_y(\varepsilon)$. Consequently, $\mathcal{Y}_\varepsilon$ defines an ε-cover of the support $[-U,U]^p$.

Let $\mathcal{Y}_\varepsilon=\{y_1,\cdots,y_{|\mathcal{Y}_\varepsilon|}\}$ denote the support of the ε-cover. Furthermore, let

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

Proposition F.1 (Binning)**.** Suppose Assumptions 1, 2, and F.1 hold.

1. If Assumption 3(i) holds, $\pi_d(X,R)=\sum_{y\in\mathcal{Y}_e}\gamma_{y,d}(X,R)\Pr\{Y(d)\in B_y(\epsilon)\,|\,S=e,X\}\mathrm{d}y$.
2. If Assumption 3(ii) holds, $\pi_d(X,R)=\sum_{y\in\mathcal{Y}_e}\gamma_y(X,R)\Pr\{Y(d)\in B_y(\epsilon)\,|\,S=e,X\}\mathrm{d}y$.

Proof. The proof proceeds analogously to the proofs of Lemmas E.1 and E.2, after replacing $Y=y$ events with $Y\in B_y(\varepsilon)$ events. Assumption F.1 and $\varepsilon>0$ guarantee that $Y\,|\,X,S,R\in B_y(\varepsilon)$ has a positive probability. □

Consequently, for any fixed $\epsilon>0$, we can estimate and conduct inference on the estimand

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

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

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

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 $\varepsilon>0$.

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

Proof. To prove this result, let $f_{Y(d)|S}(y_d)$ denote the conditional density of $Y(d)$ given S. As a first step, let us write

$$\begin{aligned}\theta(\epsilon)&:=\sum_{\tilde{y}_1\in\mathcal{Y}_e}\tilde{y}_1\Pr\{Y(1)\in B_{\tilde{y}_1}(\epsilon)\,|\,S=e\}-\sum_{\tilde{y}_0\in\mathcal{Y}_e}\tilde{y}_0\Pr\{Y(0)\in B_{\tilde{y}_0}(\epsilon)\,|\,S=e\}\\&=\sum_{\tilde{y}_1\in\mathcal{Y}_e}\int_{B_{\tilde{y}_1}(\epsilon)}\tilde{y}_1f_{Y(1)|S=e}(y_1)\mathrm{d}y_1-\sum_{\tilde{y}_0\in\mathcal{Y}_e}\int_{B_{\tilde{y}_0}(\epsilon)}\tilde{y}_0f_{Y(0)|S=e}(y_0)\mathrm{d}y_0,\end{aligned}$$

and analogously

$$\begin{aligned}\theta &= \int y_1 f_{Y(1)|S=e}(y_1)\mathrm{d}y - \int y_0 f_{Y(0)|S=e}(y_0) \\ &= \sum_{\tilde{y}_1 \in \mathcal{Y}_e} \int_{B_{\tilde{y}_1}(\epsilon)} y_1 f_{Y(1)|S=e}(y_1)\mathrm{d}y_1 - \sum_{\tilde{y}_0 \in \mathcal{Y}_e} \int_{B_{\tilde{y}_0}(\epsilon)} y_0 f_{Y(0)|S=e}(y_0)\mathrm{d}y_0.\end{aligned}$$

Consequently, it follows that

$$\begin{aligned}|\theta(\epsilon)-\theta| &\overset{(1)}{\leq} \left| \sum_{\tilde{y}_1 \in \mathcal{Y}_e} \int_{B_{\tilde{y}_1}(\epsilon)} (\tilde{y}_1 - y_1) f_{Y(1)|S=e}(y_1)\mathrm{d}y_1 \right| + \left| \sum_{\tilde{y}_0 \in \mathcal{Y}_e} \int_{B_{\tilde{y}_0}(\epsilon)} (\tilde{y}_0 - y_0) f_{Y(0)|S=e}(y_0)\mathrm{d}y_0 \right| \\ &\leq \max_{d\in\{0,1\}} 2 \left| \sum_{\tilde{y}_d \in \mathcal{Y}_e} \int_{B_{\tilde{y}_d}(\epsilon)} (\tilde{y}_d - y_d) f_{Y(d)|S=e}(y_d)\mathrm{d}y_d \right| \overset{(2)}{\leq} \max_{d\in\{0,1\}} 2 \sum_{\tilde{y}_d \in \mathcal{Y}_e} \left| \int_{B_{\tilde{y}_d}(\epsilon)} (\tilde{y}_d - y_d) f_{Y(d)|S=e}(y_d)\mathrm{d}y_d \right| \\ &\overset{(3)}{\leq} \max_{d\in\{0,1\}} 2 \sum_{\tilde{y}_d \in \mathcal{Y}_e} \int_{B_{\tilde{y}_d}(\epsilon)} |\tilde{y}_d - y_d| f_{Y(d)|S=e}(y_d) dy_d \overset{(4)}{\leq} \epsilon,\end{aligned}$$

where (1) and (2) use the triangle inequality, (3) follows by Hölder's inequality, and (4) by the construction of the ε-cover. □

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.

G Empirical application details

G.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 unit 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 less 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 5 summarizes village characteristics. We study 8,320 villages, with an average population exceeding 2,000 individuals per village. Villages are typically located in rural areas. The populated area within a typical village is geographically concentrated.

Table 5: Village summary statistics.

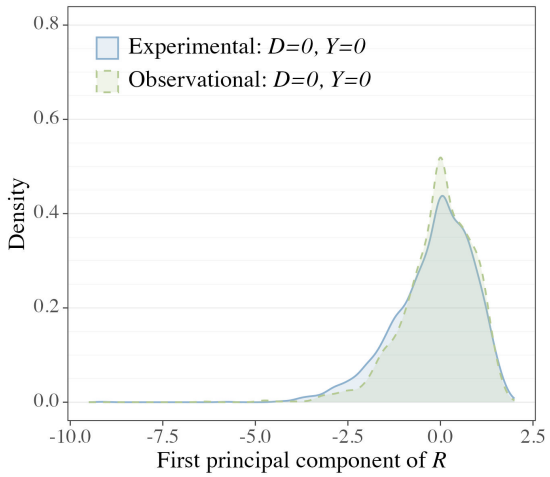
Sample	Observational	Experimental: Untreated	Experimental: Untreated	Experimental: Treated
Smartcards	N/A	2012	2011	2010
Number of villages	2,260	853	2,931	2,276
Average population	2,143	2,296	2,285	2,604
Average fraction female	0.489	0.492	0.495	0.493
Average fraction urban	0.002	0.001	0.003	0.004

For each village, we collect poverty measurements to serve as the outcome. Following Asher and Novosad (2020), the data sources are the 2012-2013 Social 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.

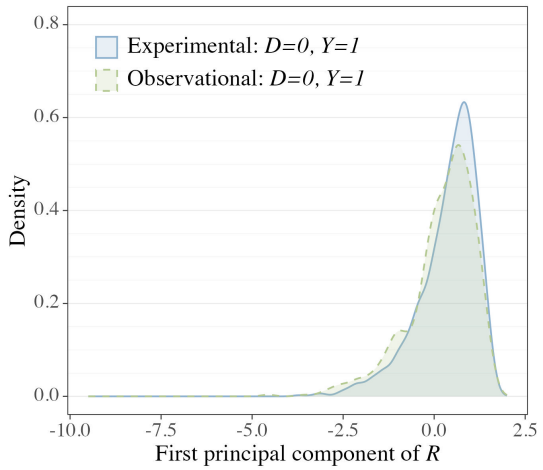
G.2 Real satellite images

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 from 2012 to 2020, summarized as a scalar in $\mathbb{R}$ (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.

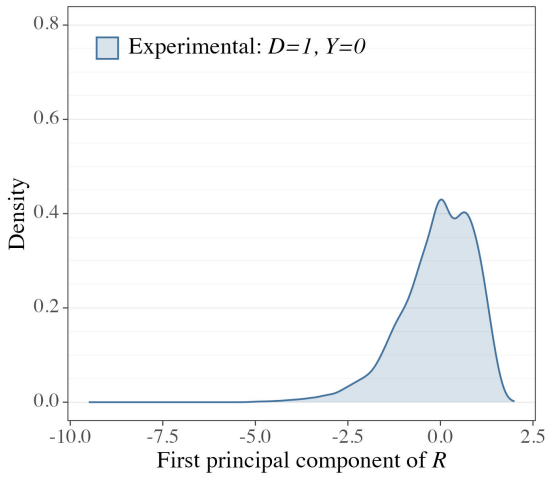
In the first exercise, we truncate the satellite image vector to $\mathbb{R}^{1000}$ for computational tractability. In the second and third exercise, we use the full RSV.



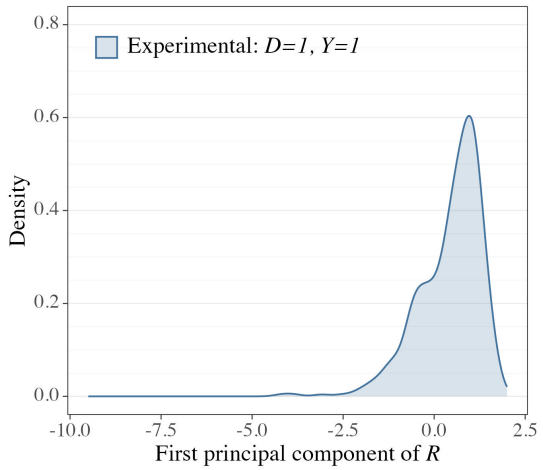
(a) Densities of $R|S$, $D=0$, $Y=0$.



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



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

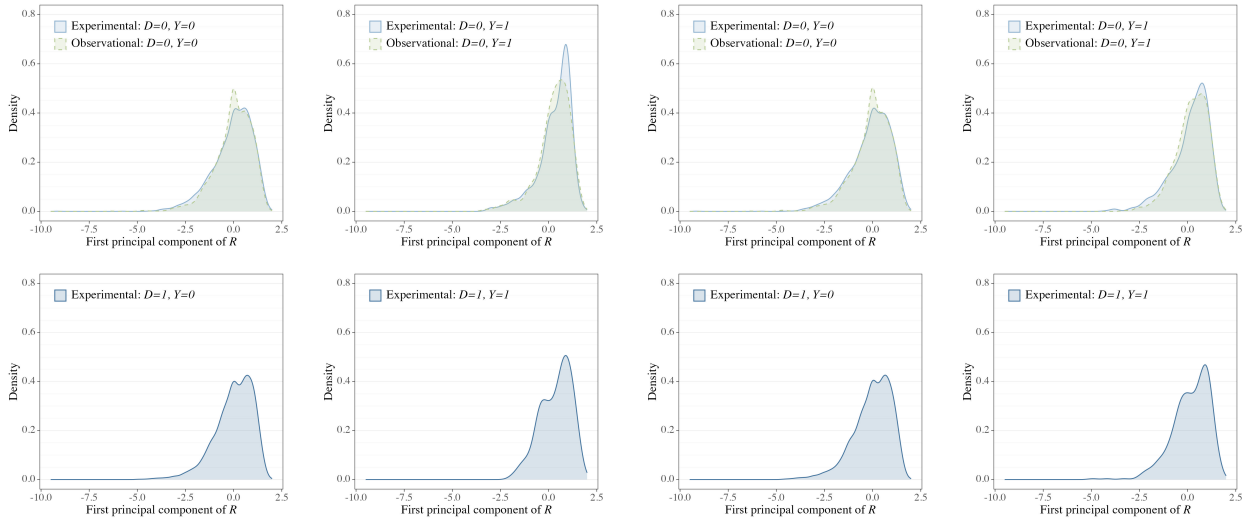


(d) Densities of $R|S=e$, $D=1$, $Y=1$.

Figure 11: No direct effects (Assumption 3(ii)) is plausible in the second and third exercises for the low consumption outcome. Within each plot, we compare $\Pr(R|S=s,D=0,Y=y)$ in the top row with $\Pr(R|S=s,D=1,Y=y)$ in the bottom row, using data from Muralidharan et al. (2023). Because the satellite image $R \in \mathbb{R}^{4000}$ is high dimensional, we visualize the density of its standardized first principal component.

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 11 provides such evidence for the low consumption outcome. We compare $\Pr(R|S=s,D=0,Y=y)$ in the top row with $\Pr(R|S=s,D=1,Y=y)$ 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 $\Pr(R|S=$

$e,D=0,Y=0$) in the top row with $\Pr(R\,|\,S=e,D=1,Y=0)$ in the bottom row. The densities are similar, as desired. Figure 12 provides similar evidence for the two other poverty outcomes.



(a) Low income.

(b) Low and middle income.

Figure 12: No direct effects (Assumption 3(ii)) is plausible in the second and third exercises for the additional poverty outcomes. Within each plot, we compare $\Pr(R\,|\,S=s,D=0,Y=y)$ in the top row with $\Pr(R\,|\,S=s,D=1,Y=y)$ in the bottom row, using data from Muralidharan et al. (2023). Because the satellite image $R\in\mathbb{R}^{4000}$ is high dimensional, we visualize the density of its standardized first principal component.

G.3 Implementation details

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

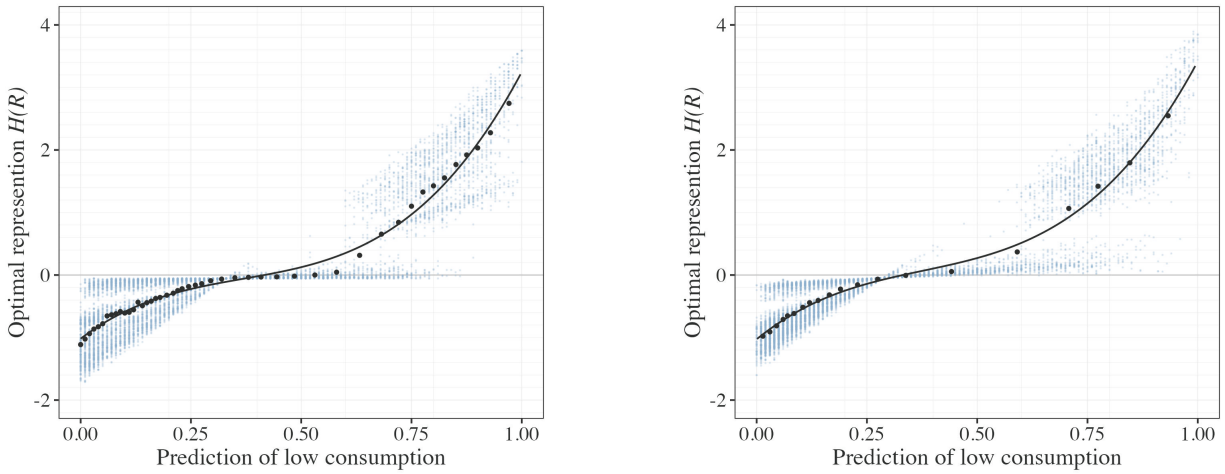
In the first exercise, we use luminosity as well as the initial 1,000 features of the satellite image embedding, for computational tractability of the 500 replications. This truncation may be viewed as regularization bias, which we alleviate with cross fitting. This procedure is justified by Proposition 3.

In the second and third exercises, we use luminosity 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 is a straightforward extension of Proposition 3, using stability in place of independence to handle the stochastic equicon-

tinuity 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 to balance the unbalanced labels of the outcome.

G.4 Comparing representations



(a) RSV: Synthetic samples

(b) RSV: Real samples

Figure 13: We contrast optimal versus simple representations of the RSV. The optimal representation combines three predictions. The simple representation uses only one prediction.

Figure 13 visualizes how our optimal representation $\widehat{H}(R)$ compares to the implicit representation of common practice $\text{PRED}_Y(R)$, using the first outcome variable. Even though the satellite image is high dimensional, its 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 13a; "synthetic samples") and for the third exercise (Figure 13b; "real samples").

There are some notable differences. The optimal 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 optimal 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.