Program Evaluation with Remotely Sensed Outcomes*

Ashesh Rambachan Rahul Singh Davide Viviano†

November 17, 2024

Abstract

While traditional program evaluations typically rely on surveys to measure outcomes, certain economic outcomes such as living standards or environmental quality may be infeasible or costly to collect. As a result, recent empirical work estimates treatment effects using remotely sensed variables (RSVs), such mobile phone activity or satellite images, instead of ground-truth outcome measurements. Common practice predicts the economic outcome from the RSV, using an auxiliary sample of labeled RSVs, and then uses such predictions as the outcome in the experiment. We prove that this approach leads to biased estimates of treatment effects when the RSV is a post-outcome variable. We nonparametrically identify the treatment effect, using an assumption that reflects the logic of recent empirical research: the conditional distribution of the RSV remains stable across both samples, given the outcome and treatment. Our results do not require researchers to know or consistently estimate the relationship between the RSV, outcome, and treatment, which is typically mis-specified with unstructured data. We form a representation of the RSV for downstream causal inference by predicting the outcome and predicting the treatment, with better predictions leading to more precise causal estimates. We re-evaluate the efficacy of a large-scale public program in India, showing that the program's measured effects on local consumption and poverty can be replicated using satellite imagery.

*We thank Isaiah Andrews, Ben Olken, and Jesse Shapiro for helpful comments and discussion. Leonard Mushunje and Sammi Zhu provided excellent research assistance.

†Rambachan: Massachusetts Institute of Technology, asheshr@mit.edu. Singh: Harvard University, rahul_singh@fas.harvard.edu. Viviano: Harvard University, dviviano@fas.harvard.edu

1 Introduction

While traditional program evaluations typically rely on surveys to measure impact, important economic outcomes such as living standards or environmental quality may be infeasible or costly to collect. As a result, recent empirical work estimates treatment effects using remotely sensed variables (RSVs). Examples include night lights as a measure of economic activity (e.g., Henderson, Storeygard and Weil, 2011; Asher et al., 2021), satellite images as a measure of deforestation and air pollution (e.g., Jayachandran et al., 2017; Zheng et al., 2020), color saturation as a measure of agricultural land use (e.g., Walker et al., 2022; Jack et al., 2022), and roof material as a measure of housing quality (e.g., Huang, Hsiang and Gonzalez-Navarro, 2021). Other forms of digital traces have been used as indirect outcome measurements (Aiken et al., 2023). This raises the question of how researchers should rigorously estimate treatment effects in experiments using such RSVs.

A common practice – in environmental, urban, and development economics – is to predict the outcome from the RSV, then to use the predicted outcome in lieu of the economic outcome in the experiment. The researcher forms such predictions using an auxiliary (observational) data set containing the RSV and its true outcome label. The predictor is typically complex, with unknown statistical guarantees.

We prove that this empirical strategy can lead to arbitrary bias when the RSV is a post-outcome variable. Using the predicted outcome in lieu of the true outcome implicitly treats the RSVs as a short term mediator between the treatment and the outcome (i.e., as a surrogate). However, this practice does not fit the logic of empirical applications where the RSV is a post-outcome variable.

For intuition, consider a binary outcome indicating agricultural land use, and an RSV capturing color saturation in satellite images. What we observe in satellite images depends on land use, and not vice versa. Suppose we form predictions of land use in the experimental sample using a predictor trained on labeled satellite images from an observational study. We then compute the difference of predicted outcomes between treated and control units in the experimental sample. This estimator will depend on the composition of the effect of the treatment on the outcome and the correlation between the outcome on the satellite image. In the extreme case where there is no effect from the outcome to satellite images – a case for which ideally we would like the procedure to report infinite standard errors – such estimation method will instead report a *precise* estimate at zero, regardless of the value of the true treatment effects.

Our research question is how to develop a principled (and efficient) way to use RSVs when imputing outcomes in program evaluation. Our primary contribution is to identify treatment effects from RSV outcomes. We consider a setup in which researchers have access to an experimental sample with treatment status and the RSV, and an observational sample with the RSV and outcome label. Our goal is to conduct inference on the effect of the treatment

on the outcome, where the unit of analysis may be an individual, village or subdistrict. Our main assumption reflects the logic of the examples above: the conditional distribution of the RSV given the true outcome variable and the treatment variable is the same across both samples. It does not require that treatment is necessarily observed in both samples nor that treatment effects are the same in both samples. In addition, we require either that (i) some units receive treatment in the experimental and (possibly confounded) observational sample; or (ii) when no unit is exposed to treatment in the observational sample, the treatment only affects the RSV through the outcome. As we show, such assumptions have testable implications.

Based on these restrictions, we provide an intuitive nonparametric identification result for the average treatment effect. Consider the land use example, assuming for simplicity no direct effect of the treatment on the RSV. Changes in the treatment, conditional upon the RSV and outcome, can only be predicted by the outcome. Therefore, we can use Bayes rule to leverage information in the conditional probability of treatment given the RSV alone to learn effects on the outcome conditional on the RSV. In particular, in the case of a binary outcome, we show that the treatment effect equals a Wald estimand whose *numerator* is a function of *treatment* change probability, and the denominator is a function of predicted outcome change from the observational study. To our knowledge, this intuition is novel to the literature and allows us to achieve nonparametric identification by directly leveraging variation both in the treatment and in the outcome.

We show how this intuition extends more broadly, including contexts where treatment may directly affect RSVs. Identification corresponds to a set of moment conditions conditional on the RSV that pin down the average treatment effect. Such moment conditions take expectations of observed treatments and outcomes without needing to solve a complex estimation problem. Importantly, our approach does not require correctly specifying or consistently estimating the relationship between the treatment, outcome, and RSV – an infeasible task with unstructured data. Returning to our example above, researchers do not need to specify how land use may change the color of each image pixel. Instead, they can use arbitrary summary statistics of the image, such as average color intensity across multiple pixels, as long as this summary contains information about land use. Our identification strategy continues to hold, conditional on such summary statistics, through the law of iterated expectations.

Next, we study estimation and inference, turning to the question of which representation of the RSV to use. By applying results from existing literature on the conditional moment, and in particular Chamberlain (1987), the efficient representation of the RSV is a function of the predicted outcome and predicted treatment conditional upon the RSV, to be learned from the observational and experimental data, respectively. We allow for predictions using arbitrary machine learning techniques, common in practice. As long as such predictors converge to *some* pseudo-true value and satisfy mild regularity conditions, estimation of the

RSV's representation does not affect consistency and inference on treatment effects. However, a more precise machine learning predictor will improve efficiency for the downstream causal inference task. This intuition follows similarly to recommendations in Newey (1993) and more broadly to the argument for inference using an estimated weighting matrix in two-step generalized methods of moments.

As an empirical application, we apply our framework to illustrate how researchers may use satellite images to evaluate the efficacy of a large-scale anti-poverty program. We revisit a randomized experiment studied in Muralidharan, Niehaus and Sukhtankar (2016, 2023) that studied the impact of biometrically authenticated payments infrastructure ("Smartcards") in India. By collecting village-level coordinates, we construct a dataset with satellite and nightlight village-level data, as well as measures of village-level consumption and poverty levels. We then run our method, omitting outcome level data for roughly half of the experimental sample, corresponding to 3,000 villages. This exercise mimics empirical applications in which researchers have limited information about the target outcome in the experimental sample. Using satellite images as a remotely sensed variable, we reproduce the experiment's point estimate and confidence interval using ground-truth outcomes, for the program's effect on poverty. This amounts to reducing research costs by about three million dollars, using conservative estimates of survey costs. Finally, through a set of realistic numerical studies calibrated to this application, we demonstrate that using the surrogate approach instead of our procedure may lead to a 50% increase in mean squared error and bias of the estimated treatment effect.

Related Work: Our RSV model and nonparametric identification result complement several other auxiliary variable models with nonparametric identification, and more broadly the literature on data fusion. Whereas a surrogate is an intermediate variable between the treatment and outcome (e.g., Athey et al., 2024; Kallus and Mao, 2022), the RSV is a post-outcome variable. We prove that misusing an RSV as a surrogate leads to arbitrary bias. The proxy literature (Imbens et al., 2024; Ghassami et al., 2022) extends surrogate models to deal with unobserved confounding, but has the same limitation; it may have arbitrary bias when the RSV is a post-outcome variable. It appears that only one strategy within the proxy literature uses only one auxiliary variable (Park, Richardson and Tchetgen Tchetgen, 2024). However, that strategy provides no guidance on how to deal with incomplete observations in program evaluation. Relative to a vast literature on data fusion (e.g. Cross and Manski, 2002; Molinari and Peski, 2006; D'Haultfœuille, Gaillac and Maurel, 2024; Bareinboim and Pearl, 2016), here we focus on, to our knowledge, a novel problem studying causal inference with no outcome information from the experimental dataset, and *post*-outcome information from an observational dataset.

This work share some themes with a growing literature on unstructured data use in economics, which uses generative models for inference on outcomes of interest (Gentzkow, Shapiro and Taddy, 2019; Battaglia et al., 2024). The main difference from this literature

is that we do not require a correctly specified generative model for how the treatment and outcome may affect the RSV, which may be prone to misspecification with high-dimensional data. Additional references include Fong and Tyler (2021); Singh and Vijaykumar (2023); Egami et al. (2024); Zhang et al. (2023) who study prediction problems where the covariates (instead of outcomes) are mismeasured or indirectly represented. Allon et al. (2023) assume instead the outcome is measured for some individuals in the main study (experiment) and present a doubly-robust estimators under a missing at random assumption. Different from these latter references, here we do not observe outcomes from the main experiment, and the outcomes from the observational study may present confounding bias and/or no exposure to treatment.

More broadly, this work connects to the rich literature on nonclassical measurement error (Hu, 2008; Molinari, 2008; Schennach, 2020; Hu, 2017). Different from this literature, we do not require a repeated measurement yet preserve point identification. This is possible by leveraging the combination of two complementary datasets for causal inference.

Finally, we complement a large empirical literature on the use of RSV for program evaluation (e.g. Donaldson and Storeygard, 2016; Proctor, Carleton and Sum, 2023), focusing on inference on direct effects of the treatment on independent units – such as villages or subdistricts.

2 Setup and Identifying Assumptions

Suppose we observe units in two samples: an experimental sample (e) consisting of n_e units and an observational sample (o) consisting of n_o units. Let $S \in \{e, o\}$ denote an indicator for whether a unit belongs to the experimental or observational sample, and $n = n_e + n_o$ is the total sample size across the experimental and observational samples.

Each unit is associated with a binary treatment $D \in \{0, 1\}$ and a target outcome $Y \in \mathcal{Y}$. The target outcome may either be scalar or vector-valued. To simplify exposition, we assume the support of the target outcome $\mathcal{Y}$ is discrete in the main text, and Appendix A.2 generalizes our results to the continuous case. Following the potential outcome framework (Imbens and Rubin, 2015), each unit is associated with a pair of potential outcomes $(Y(0), Y(1))$, and the realized outcome satisfies $Y = DY(1) + (1 - D)Y(0)$. The target outcome is not observed for units in the experimental sample, and the treatment may not be observed in the observational sample. Each unit is also associated with additional (low-dimensional) pre-treatment covariates $X \in \mathcal{X}$ that may be used to randomize units into treatment in the experimental sample. We assume that the support $\mathcal{X}$ is discrete, which is typical in stratified experiments; our results may be extended to allow for continuous covariates though we omit that extension for brevity.

Importantly, for all units, we additionally observe a *remotely sensed variable* (RSV) for the target outcome, denoted by $R \in \mathcal{R}$. The remotely sensed variable R is typically high-

dimensional, such as a satellite image of a small village or high-frequency mobile phone trace data on messages and phone calls sent and received. We would like to use the remotely sensed variable as a noisy measurement for the target outcome in the experimental sample. As a technical condition, we assume $\mathbb{P}(R = r \mid X)$ is a positive density bounded away from zero almost surely for all $r \in \mathcal{R}$, where $\mathbb{P}(\cdot)$ defines the density function for a continuous variable and its probability mass function for a discrete one.

Altogether, in this setting, units are characterized by values $(Y(0), Y(1), D, X, R, S)$. For units in the experimental sample $(S = e)$, we observe (X, D, R), whereas we observe (X, D, R, Y) for units in the observational sample $(S = o)$. The treatment D could be potentially constant (equal to zero) for all units in the observational sample. For each unit, we therefore observe the random vector

$$(Y1\{S = o\}, D, R, X, S)\,. \tag{1}$$

Table 1 summarizes this data environment. If an observation includes (D, Y) where D has been randomly assignment, we will relabel this observation as $S = o$ for expositional convenience.[1]

Sample S	Covariates X	Treatment D	Outcomes Y	Remotely sensed variable R
Experimental	✓	✓	×	✓
Observational	✓	✓	✓	✓

Table 1: Summary of the data environment, where ✓ denotes the variable is observed and × denotes the variable is missing. The observational sample may include no unit exposed to treatment.

2.1 Causal Estimands

In this setting, we are interested in studying the causal effect of the treatment D on the target outcome Y in the experimental sample, and so we will study a large class of possible causal estimands. For some known transformation $w(\cdot)\colon \mathcal{Y} \to \mathbb{R}$, the *target estimand* is

$$\theta_w := \theta_w(1) - \theta_w(0), \text{ where } \theta_w(d) := \mathbb{E}[w\left(Y(d)\right) \mid S = e]. \tag{2}$$

For alternative choices of $w(\cdot)$, the target estimand captures a large class of average causal effects in the experimental sample that may be of interest empirically. For example, consider a scalar target outcome with $w(y) = y$; in this case, the target estimand simplifies to the average treatment effect in the experimental sample, $\mathbb{E}[Y(1) - Y(0) \mid S = e]$. When researchers collect multiple outcomes, it is common to either focus on one particular element

[1]Our framework also allows us to consider such an observation as part of both the experimental and observational sample to improve precision, where S takes a set valued $\{e, o\}$. See Remark 4 for a discussion.

as the primary outcome or to construct an outcome index; both of these can be captured through particular choices of the transformation $w(\cdot)$.[2]

Even though we do not directly observe the target outcome in the experimental sample, we would like to identify and conduct inference on the target estimand. How can we combine the experimental sample, in which we only observe R, and the observational sample, in which observe both (R, Y), to identify and conduct inference on the target estimand?

To make this more concrete, we map this setting into three illustrative examples that we will return to throughout the paper.

Example 1 (Satellite images of household wealth)**.** Consider a large-scale randomized experiment evaluating an unconditional cash transfer, which was conducted across 653 villages in Kenya between 2014-2017 by GiveDirectly (Egger et al., 2022). Treatment was randomized at the village level, with 328 villages selected into treatment. Researchers would like to study the causal effect of the cash transfer D on household wealth Y. In revisiting this experiment, Huang, Hsiang and Gonzalez-Navarro (2021) use satellite images of buildings as a remotely sensed variable R for household wealth. Rather than collecting household wealth Y in the experimental sample through costly end-line surveys, suppose we only observed the satellite images R. In an observational sample of villages, we observe both (Y, R), which may be assembled by linking measures of household wealth from an existing census to satellite images. We would like to combine these two data sources in order to identify and conduct inference on the causal effect of the cash transfer on household wealth in the experimental sample. ▲

Example 2 (Satellite images to measure deforestation)**.** Jayachandran et al. (2017) study the causal effect of a payment program in Uganda that offered forest-owning households cash payments in exchange for refraining from deforestation. Access to the payment program was randomized at the village level, with 60 villages select into treatment. Researchers would like to study the causal effect of the payment program D on village-level deforestation Y. Rather than having surveyors return to treated villages to manually measure changes in forestland, we may use satellite images of forest cover R as a measurement of village-level deforestation in the experimental sample. Additionally, we have access to an observational sample where the pair (Y, R) are observed — for example, perhaps there exists surveys of forestland in other regions of Uganda that can be linked to satellite images. How can we leverage both data sources to infer the causal effect of the payment program in the experiment? ▲

Example 3 (Mobile phone traces of consumption)**.** Consider a large-scale randomized experiment that evaluated an unconditional cash transfer across individuals in Togo between 2020-2021 by GiveDirectly-Novisi (Aiken et al., 2023). Roughly 49,000 eligible individuals

[2]Our analysis can be naturally extended to other causal estimands that may be of interest, such as average causal effects on the treated units in the experimental sample, here omitted for brevity.

were randomized into receiving the cash transfer. We would like to study the causal effect of the cash transfer D on individual consumption and food security Y. Previous work has found that mobile phone trace data (e.g., high-frequency records of text messages and phone calls sent and received) can predict survey measurements of consumption (e.g., Aiken et al., 2022). Rather than collecting consumption data directly, Aiken et al. (2023) use mobile-phone trace data as a remotely sensed variable R for consumption and food security in the experiment. The authors additionally collect an observational dataset that links mobile phone trace data to existing measurements of consumption and food security. How can we leverage both data sources to infer the causal effect of the cash transfer in the experiment? ▲

2.2 Identifying Assumptions

We introduce our main assumptions to identify the target estimand in the experimental sample by leveraging the remotely sensed variable and the observational sample.

Our first assumption is about sampling, randomization in the experimental sample, and a common support assumption in the target outcome between the observational and experimental samples.

Assumption 1 (Experimental and observational overlap)**.** Let the following hold:

1. (Sampling) Let $\left\{Y(1), Y(0), D, X, R\right\} \Big| S = s \sim \mathcal{P}_s$, $1\{S_i = e\} \sim \text{Bern}(p_i)$ for some unknown distributions $\mathcal{P}_e, \mathcal{P}_o$, and $p_i \in (\bar{p}, 1-\bar{p}), \bar{p} \in (0,1), i \in \{1, \cdots, n\}$;

2. (Common support) Assume that $\mathbb{P}(Y = y|X, S = o) \geq \underline{y} > 0$ almost surely for all $y \in \mathcal{Y}$, and $\mathbb{P}(Y \in \mathcal{Y}|X, S = e) = 1$ almost surely;

3. (Randomized treatment) $D \perp\!\!\!\perp \{Y(0), Y(1)\} \mid X, S = e$;

4. (Overlap) $\mathbb{P}(D = 1 \mid X, S = e) \in (\delta, 1-\delta)$ almost surely for some $\delta \in (0,1)$.

Assumption 1.1 assumes that units are sampled independently with a distribution that depends on the sampling indicator. Note that the sampling indicator may not be identically distributed; therefore, we allow for settings with selective sampling of the experiment and observational study (e.g., site selection).[3] Assumption 1.2 states that the outcome in the observational study has a common or larger support than the outcome in the experiment.

Assumption 1.3 and Assumption 1.4 are the standard unconfoundedness condition and overlap in the experiment. Throughout the text, we will not assume that the treatment is independent of potential outcomes in the observational study. Under these conditions, if we were to observe the target outcome in the experimental sample, we could identify the target estimand using standard arguments in causal inference.

[3]Our results naturally extend to independent but not identically distributed data, here omitted for notational convenience.

However, since we do not observe the target estimand, we would instead like to use the remotely sensed variable R as a noisy measurement for the target variable. We next introduce our main identifying assumption that formalizes this intuition.

Assumption 2 (RSV Stability)**.** Let $S \perp\!\!\!\perp R \mid Y, X, D$.

Assumption 2 imposes that the remotely sensed variable R is independent of the experimental condition (experiment or observational study) S conditional on the target outcome Y, pretreatment covariates X, and treatment D. This assumption allows us to use the remotely sensed variable and information learned from the observational sample in order to derive causal inferences about the experimental sample, even though we do not directly observe the target outcome. In words, Assumption 2 states that the relationship between the target outcome Y and the remotely sensed outcome R is *stable* between the experiment and observational study. It does not require, however, that such a stability condition holds for treatment effects (effect of D on Y), which may differ between the two experimental conditions.

Returning to Example 1, Assumption 2 states that $R|Y, X, D, S = e$ follows the same distribution as $R|Y, X, D, S = o$; the conditional distribution of satellite images given wealth, baseline covariates, and cash transfer status is stable across the two samples. Assumption 2 holds when wealth leads to similar patterns observed from satellite images and it fails if, for example, villages with the same level of wealth may exhibit different building structures because of unobserved village-level norms.

While restrictive, we will show that Assumption 2 is in fact testable in our setting (as we discuss in Remark 2). Consequently, before attempting to use a remotely sensed variable to draw causal inferences from the experimental sample, researchers can check whether our main identifying assumption is falsified.

Our final assumption imposes restrictions on the treatment in the observational study.

Assumption 3 (Treatment in the observational sample)**.** Suppose that *either* condition holds:

1. Either $P(D = 1|X, S = o) \in (\delta, 1 - \delta)$ for a constant $\delta > 0$;
2. or $D \perp\!\!\!\perp R|Y, X$.

Assumption 3 imposes *either* 3.1 or 3.2. Assumption 3.1 that *some* units in the observational sample received the treatment, where, importantly, the treatment can be confounded for such units. That is, Assumption 1.3 can be violated in the observational sample. Returning to Example 1, this amounts to assuming that we observe consumption (Y) for *some* individuals who received cash-transfers ($D = 1$). For such individuals, we do not require that the treatment be randomly assigned. However, because we require that the stability conditions

holds (Assumption 2), the direct effect between the treatment and RSV (i.e., dotted line in Figure 1) must be "stable" between the two samples.

Assumption 3.1 holds if researchers obtain outcome data for treated units either through observational data where some cash-transfers were conducted, or augment the observational sample by including *some* treated units in the experiment by collecting Y for such units.

A failure of Assumption 3.1 means that we have no "complete" observations including Y and nonzero D for the same unit. Therefore a further assumption is necessary about how (Y, D) relate. For this case, Assumption 3.2 states that the treatment has no effect on the remote sensed variable R other than through the outcome Y.

Assumptions 2 and 3.2 together imply $(S, D) \perp\!\!\!\perp R|X, Y$, whereas Assumptions 2 and 3 together allows for $D \not\perp\!\!\!\perp R|X, Y$. Table 2 summarizes the implications of our identifying assumptions for how potential outcomes, propensity scores, and the remotely sensed variable may vary across the experimental and observational samples.

	Experiment	Observational study	Description
POs $\mathbb{E}[Y(d)\|D=d,S]$	$\mathbb{E}[Y(d)\|S=e]$	$\mathbb{E}[Y(d)\|D=d,S=o]$	Can change with S
(Corresponding assumption)	(Unconfounded $Y(d) \perp D\|X$)	(Possibly endogenous $Y(d) \not\perp D\|X$)	
Propensity score $\mathbb{P}(D=1\|X,S)$	$\mathbb{P}(D=1\|X,S=e)$	$\mathbb{P}(D=1\|X,S=o)$	Can change with S
(Corresponding assumption)	(D Randomized)	(D not necessarily randomized)	
RSV $\mathbb{P}(R\|Y,D,X)$	$\mathbb{P}(R\|Y,X)$	$\mathbb{P}(R\|Y,X)$	Stable with D
(Corresponding assumption)	(no D direct effect)	(no D direct effect)	if D is unobserved in obs study
			and/or has no variation in obs study
RSV $\mathbb{P}(R\|Y,D,X)$	$\mathbb{P}(R\|Y,X,D)$	$\mathbb{P}(R\|Y,X,D)$	Can change with D
(Corresponding assumption)	(possible D direct effect)	(possible D direct effect)	if D is observed in obs study
			and overlap holds in obs study
RSV $\mathbb{P}(R\|Y,D,S,X)$	$\mathbb{P}(R\|Y,D,X)$	$\mathbb{P}(R\|Y,D,X)$	Stable with S
(Corresponding assumption)	(RSV representability)	(RSV representability)	

Table 2: Stability conditions between experiment and observational study.

Figure 1 (a) illustrates the underlying mechanism: the treatment affects the outcome, which *then* affects the RSV. It is possible that the treatment also affects R whenever we have some variation in D in the observational sample.

2.3 Common Empirical Practice Can Lead to Arbitrary Biases

It is common in existing work for researchers to use remotely sensed variables to draw inferences about the target estimand in a simple two-step procedure: first, in the observational sample, researchers predict the target outcome using the remotely sensed variables (possibly with supervised machine learning techniques), obtaining $\hat{f}(R)$; second, the constructed prediction function is applied on the experimental sample, using the predictions as the con-

structed outcome of interest. Indeed, this empirical strategy is used in some of the references given in Examples 1, 2, 3.

(a) DAG for RSVs (b) DAG for surrogates

Figure 1: Comparison of the causal mechanism for RSVs and surrogates (e.g. Athey et al., 2024) Our framework inverts the role of the outcome and RSV compared to the surrogate methods. It also allows for direct effects of D on R when D varies in the observational study. As we show in Theorem 2.1, wrongly assuming a surrogate model in this context can lead to arbitrary biases.

While intuitive, we now show that this empirical strategy can lead to arbitrary biases in the resulting estimates of the target estimand under the identifying assumptions in Section 2.2. To see this, let us consider the simple setting in which the target outcome is a scalar, there are no covariates, and that $\mathbb{P}(D = 0 \mid S = o) = 1$ (i.e., no treatment is implemented in the observational dataset). We would like to estimate the average potential outcome $\mathbb{E}[Y(1) \mid S = e]$. In this case, common empirical practice using remotely sensed variables can be interpreted as implicitly targeting the reduced-form estimand

$$\widetilde{\theta}(1) = \mathbb{E}\left[\mathbb{E}[Y \mid R, S = o] \mid D = 1, S = e\right]. \tag{3}$$

By predicting the target outcome using the remotely sensed variable in the observational sample, our best hope would be to recover the true conditional expectation $\mathbb{E}[Y \mid R, S = o]$. Unfortunately, under the identifying assumptions in Section 2.2, the implicit target in (3) can incur biases with arbitrary sign.

Theorem 2.1. *Let $X = 1$ almost surely (no covariates for expositional convenience), and suppose that Assumptions 1, 2, 3.2 hold. Then there exists a data-generating process under which $S \perp\!\!\!\perp (R, Y)|D$ (i.e., individuals are randomly allocated to the experiment or observational study) such that*

$$\widetilde{\theta}(1) - \mathbb{E}[Y(1) \mid S = e] > 0$$

and a (different) data-generating process under which $S \perp\!\!\!\perp (R, Y)|D$ such that

$$\widetilde{\theta}(1) - \mathbb{E}[Y(1) \mid S = e] < 0.$$

Proof. See Appendix 2.1. □

Theorem 2.1 shows that under the RSV framework, common empirical practice may incur bias with arbitrary signs for estimating the expected value of the potential outcome $Y(1)$ in the experimental sample. Because $\mathbb{E}[Y(0)|S = e]$ is directly identified from the observational study under unconfoundedness of S, this result translates into similar bias for the average effect. Provided researchers believe that the remotely sensed variable is a noisy measurement

of the target outcome in the sense of Figure 1(a), then conclusions about the target estimand based on common practice can be misleading. Indeed, Theorem 2.1 could explain recent failures in using remotely sensed variables to replicate experimental findings (e.g., Aiken et al., 2023).

In light of this result, it is natural to ask: are there alternative identifying assumptions under which current empirical practice using remotely sensed variables would be valid? Yes – through the lens of recent causal inference work, common empirical practice implicitly views the remotely sensed variable as a *surrogate* for the target outcome (e.g., Athey et al., 2024; Kallus and Mao, 2022). Consequently, existing results on surrogates implies that the implicit target in (3) recovers the average potential outcome $\mathbb{E}[Y(1) \mid S = e]$ if

$$(D, S) \perp\!\!\!\perp Y \mid R, X \quad \text{(Surrogacy and surrogate comparability)} \tag{4}$$

is satisfied. It is immediate to see that Equation (4) and the identifying assumptions in Section 2.2 are non-nested.

Importantly, the identifying assumptions in Section 2.2 and Equation (4) place different causal assumptions on the remotely sensed variable. Assumption 3 states that the primary outcome fully mediates the effect of the treatment on the RSV – in Example 1, randomized cash transfers only affect satellite images of houses through their causal effect on household wealth. By contrast, Equation (4) states the exact opposite: the surrogate (i.e., the RSV) fully mediates the effect of the treatment (i.e., cash transfer) on the outcome (i.e., wealth). Assuming that the treatment affects the RSV, which then affect the outcome, clearly fails in our relevant applications.

Remark 1 (Negative controls in causal inference)**.** It is interesting to compare our framework to recent literature on negative controls, whose (different) goal is to control for confounding bias in observational studies. The only context where these models have been implemented in contexts of two (incomplete) datasets as those studied here are the negative control surrogate models (e.g. Imbens et al., 2024). Negative control surrogate models are an extension of surrogate models (Imbens et al., 2024); they nest the surrogate model as a special case. They suffer from the same drawback of surrogate models formalized in Theorem 2.1.

Once we contrast instead to standard negative control models (with one auxiliary variable), as in Park, Richardson and Tchetgen Tchetgen (2024), there are two main differences motivated by the different goals that these methods have. Unlike negative control models, here we allow the treatment variable to affect both the outcome and the auxiliary variable when Assumption 3.1 holds. When no complete observation exists, i.e. the setting covered by Assumption 3.2, the single proxy model provides no guidance of how to proceed.

Finally, none of these negative control models have testable implications for the main identifying assumption, whereas our RSV model does. □

3 Main Result

In this section, we provide our non-parametric identification result establishing that we can identify the target estimand by combining the remotely-sensed variable in the experimental sample with the observational sample.

3.1 Special Case: Identification for a Binary Outcome

Let us first consider the special case in which the target outcome is binary $Y \in \{0, 1\}$, and there are no pre-treatment covariates X. Suppose Assumptions 1, 2 hold and we want to recover the average treatment effect $\theta = \mathbb{E}[Y(1) - Y(0) \mid S = e]$.

Step 1: Mixtures of causal estimands As a first step, the identified distribution of the remotely-sensed variable in the experimental sample can be written as a *mixture* of the average potential outcomes we would like to identify:

$$\begin{aligned}\mathbb{P}(R = r \mid D = d, S = e) =& \mathbb{P}(R = r \mid Y = 1, D = d, S = e)\mathbb{P}(Y(d) = 1 \mid S = e) \\ &+ \mathbb{P}(R = r \mid Y = 0, D = d, S = e)\mathbb{P}(Y(d) = 0 \mid S = e).\end{aligned} \tag{5}$$

Though this implies the distribution of the remotely-sensed variable in the experimental sample is informative about the target estimand we would like to identify, there is a catch: the weights in this mixture depend on the relationship between the remotely-sensed variable and the target outcome in the experimental sample. Since we do not observe the target outcome, these weights are not identified using only the experimental sample.

Step 2: (Infeasible) identification with an observational sample Our first insight is that we can leverage the observational sample to identify these weights in the experimental sample. Under Assumption 2, the relationship between the remotely-sensed variable and the target outcome is *stable*, meaning that

$$\mathbb{P}(R = r \mid Y = y, D = d, S = e) = P(R = r \mid Y = y, D = d, S = o). \tag{6}$$

Under invertibility conditions, it then follows that

$$\mathbb{P}(Y(d) = 1 \mid S = e) = \frac{\mathbb{P}(R = r \mid D = d, S = e) - \mathbb{P}(R = r \mid Y = 0, D = d, S = o)}{\mathbb{P}(R = r \mid Y = 1, D = d, S = o) - \mathbb{P}(R = r \mid Y = 0, D = d, S = o)}. \tag{7}$$

This identification result leaves us with a potentially challenging estimation problem as it depends on the conditional distribution of a high-dimensional object R. On its own, this would be a challenging machine learning problem that relies on a particular generative model for R (e.g., satellite images). In addition, it is unclear whether the resulting generative models would satisfy known regularity conditions to consistently estimate and conduct conventional inference on the average treatment effect.

Step 3: Translation to simple conditional moments Our next insight is that we can further rewrite this identification result in terms of simple conditional moment conditions that sidestep having to estimate the distribution of $R|Y, D, S$. Formally, using Bayes theorem, we can write for any R

$$\mathbb{P}(Y(d) = 1|S = e)\mathbb{E}\left[V_i^1(d) - V_i^0(d)\middle|R\right] = \mathbb{E}\left[U_i(d) - V_i^0(d)\middle|R\right] \tag{8}$$

where

$$V_i^y(d) = \frac{1\{Y = y, D = d, S = o\}}{\mathbb{P}(Y = y, D = d, S = o)}, \quad U_i(d) \quad = \frac{1\{D = d, S = e\}}{\mathbb{P}(D = d, S = e)}. \tag{9}$$

That is, we reduced a complex machine learning problem for R to a set of *simple* conditional moment equality that depend on indicators and *unconditional* probabilities that can be easily estimated.

Step 4: Choosing a representation For identification, we introduce a univariate function $H_d(R)$ that we will refer to as a *representation* of R. For identification, such representation can be arbitrary as long as $\mathbb{E}[H_d(R)(V^1(d) - V^0(d))] \neq 0$, i.e., as long as it is predictive of the outcome of interest. We can write

$$\mathbb{P}(Y(d) = 1|S = e) = \frac{\mathbb{E}[H_d(R)(U(d) - V^0(d))]}{\mathbb{E}[H_d(R)(V^1(d) - V^0(d))]}$$

without imposing any model on how R varies with Y and D.

While any choice of representation $H_d(R)$ identifies the average treatment effect, simple choices may be inefficient, leading to needlessly large standard errors. In Section 4, following the long-standing literature on conditional moment equalities (Chamberlain, 1987; Newey, 1993) we discuss how researchers can exploit Equation (10) and learn the efficient representation $H_d(R)$. We therefore connect the task of representation learning (e.g. Johannemann et al., 2019; Vafa, Athey and Blei, 2024) to the downstream of causal inference by motivating the representation based on the efficiency properties of our final estimator.

Finally, because such equality must hold for all choices $H_d(\cdot)$, it also provides us with testable implications of Assumption 2.

Example: "Inverse" instrumental variable problem To gain further insight, suppose that all units in the observational study are not exposed to treatment and in addition to the conditions discussed so far, also Assumption 3.2 holds. Then we can show that the conditional moment equalities further simplify to (for any representation $H_d(R)$)

$$\theta = \frac{\mathbb{E}[H_d(R)\Delta(e)]}{\mathbb{E}[H_d(R)\Delta(o)]} \tag{10}$$

for

$$\begin{aligned} \Delta(o) &= \left(\frac{1\{Y = 1, S = o\}}{\mathbb{P}(Y = 1, S = o)} - \frac{1\{Y = 0, S = o\}}{\mathbb{P}(Y = 0, S = o)}\right) \\ \Delta(e) &= \left(\frac{1\{D = 1, S = e\}}{\mathbb{P}(D = 1, S = e)} - \frac{1\{D = 0, S = e\}}{\mathbb{P}(D = 0, S = e)}\right). \end{aligned} \tag{11}$$

Returning to Example 1, suppose that researchers only have access to satellite images of roofs in a village rather than direct measurements of health. Researchers could choose to construct a lower-dimensional representation of the images, for example, measuring the material used to build roofs $H(R)$ to identify the average treatment effect.

If we interpret $\Delta(e)$ as the "treatment", $\Delta(o)$ as the "outcome" and $H_d(R)$ as an "instrument", Equation (11) can be interpreted as an *inverse* instrumental variable regression problem, to, our knowledge, novel in the literature, where $\theta = \mathbb{E}[\tilde{D}\tilde{Z}]/\mathbb{E}[\tilde{Y}\tilde{Z}]$. ▲

3.2 General Identification Result

Having established that we can identify the average treatment effect using the remotely-sensed variable in a special case, we now generalize our identification result allowing for pre-treatment covariates, a discrete-valued outcome and any choice of target estimand.

As a first step, we write the distribution of the remotely-sensed variable as a mixture of the average potential outcome in the experimental sample with weights depending on the relationship between the remotely-sensed variable and the target outcome in the observational sample.

Lemma 3.1. *Suppose Assumptions 1 and 2 hold.*

1. *If Assumption 3.1 holds then, for $d \in \{0,1\}$ and all $r \in \mathcal{R}$,*

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

2. *If instead Assumption 3.2 holds, then, for $d \in \{0,1\}$ and all $r \in \mathcal{R}$,*

$$\mathbb{P}(R = r \mid D = d, S = e, X) = \sum_{y \in \mathcal{Y}} \mathbb{P}(R = r \mid Y = y, S = o, D = 0, X)\mathbb{P}\{Y(d) = y \mid S = e, X\}.$$

Proof. See Appendix C.1. □

The main challenge is that the probability distribution of R can be difficult to estimate because R can be high-dimensional. However, we can invert the expressions above using Bayes theorem and express the conditional distribution of $Y(d)$ as a function of "predictive" probabilities of treatment and experiment status. Define the treatment prediction weight π_d and the outcome prediction weight $\gamma_{y,d}$ as

$$\pi_d(X,R) = \frac{\mathbb{P}(D = d, S = e \mid R, X)}{\mathbb{P}(D = d, S = e \mid X)}, \quad \gamma_{y,d}(X,R) = \frac{\mathbb{P}(Y = y, S = o, D = d \mid R, X)}{\mathbb{P}(Y = y, S = o, D = d \mid X)}.$$

Theorem 3.1. *Suppose Assumptions 1, 2 hold. For $d \in \{0,1\}$,*

1. *If Assumption 3.1 holds*

$$\pi_d(X,R) = \sum_{y \in \mathcal{Y}} \gamma_{y,d}(X,R)\mathbb{P}\{Y(d) = y \mid S = e, X\}. \tag{12}$$

2. If instead Assumption 3.2 holds

$$\pi_d(X,R) = \sum_{y\in\mathcal{Y}} \gamma_{y,d=0}(X,R)\mathbb{P}\{Y(d)=y \mid S=e, X\}. \tag{13}$$

Proof. See Appendix C.3. □

Here, the left-hand side of Equation (12) depends on the treatment prediction weight. Intuitively, because R is a function of the treatment through the outcome, whenever the treatment affects the outcome, we may expect variation in these (predictive) probabilities, even if the experiment was completely randomized. The right-hand side depends on the outcome prediction weight, applied to the causal distribution of interest.

This identification result enables us to express the causal parameter in terms of *simple* conditional moment restrictions.

Theorem 3.2 (Identification via conditional moments)**.** *Let Assumptions 1, 2, 3 hold. For* $d \in \{0,1\}$

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

almost surely, where $\bar{d}=d$ *under Assumption 3.1 and* $\bar{d}=0$ *under Assumption 3.2.*

Proof. The result is immediate from the law of iterated expectations and Theorem 3.1. □

Theorem 3.2 implies that we can construct estimators that can leverage arbitrary information from the RSV R for identification. In particular, we can use existing results for estimation with conditional moments, e.g., Newey (1993); Chamberlain (1987) to leverage most of information about the RSV for efficiency considerations, while allowing us to conduct valid inference without having to correctly predict R or Y conditional on available information.

That is, rather than working directly with the conditional moment restrictions, we can construct moment restrictions by averaging over $R \mid X$ and constructing instrument functions or representations $H_d(X,R)$ in the spirit of Newey (1993); Ai and Chen (2003); Donald, Imbens and Newey (2003); Domínguez and Lobato (2004); Kitamura, Tripathi and Ahn (2004) among others. Such functions do not require that we correctly predict the outcome or the treatment conditional on R.

The following corollary discusses the more general case with multi-valued outcomes.

Corollary 3.1 (Corresponding unconditional moments)**.** *Let Assumptions 1, 2, 3 hold. Then, for any measurable matrix* $H_d(\cdot)\colon \mathcal{X}\times\mathcal{R}\to\mathbb{R}^{K\times 1}, K\geq|\mathcal{Y}|$,

$$\mathbb{E}\left[H_d(X,R)\frac{1\{D=d,S=e\}}{\mathbb{P}(D=d,S=e\mid X)} \mid X\right] = \mathbb{E}\left[H_d(X,R)W_{\bar{d}}^{\top} \mid X\right]\tau_d^{\star}(X).$$

where

$$W_{\bar{d}} = \Big(\frac{1\{Y = y_1, S = o, D = \bar{d}\}}{\mathbb{P}(Y = y_1, S = o, D = \bar{d} \mid X)}, \cdots, \frac{1\{Y = y_{|\mathcal{Y}|}, S = o, D = \bar{d}\}}{\mathbb{P}(Y = y_{|\mathcal{Y}|}, S = o, D = \bar{d} \mid X)}\Big)^\top.$$
$$\tau_d^\star(X) = \Big(\mathbb{P}\{Y(d) = y_1 | X, S = e\}, \cdots, \mathbb{P}\{Y(d) = y_{|\mathcal{Y}|} | X, S = e\}\Big),$$

and $\bar{d} = d$ *under Assumption 3.1 and* $\bar{d} = 0$ *under Assumption 3.2.*

Proof. See Appendix C.5. □

Corollary 3.1 suggests that we can estimate the distribution of potential outcomes either through a conditional moment restriction or by constructing moment restrictions (unconditional on R) based on instrument functions $H_d(X, R)$. Remarkably, such instrument functions can be *arbitrary* to guarantee consistency and valid asymptotic inference as long as $\mathbb{E}[H_d(X, R)W^\top | X]$ is full rank. Efficiency considerations follow similar arguments as those in Chamberlain (1987) and discussed for the binary case in the following section (see Appendix A for the general case).

Remark 2 (Testable implications)**.** The conditional moment restriction in Theorem 3.2 provides us with a testable implication for Assumption 2. For any measurable $H_d(X, R) \in \mathbb{R}^{|\mathcal{Y}|\times 1}$ such that $\mathbb{E}\left[H_d(X, R)\mathbb{E}(W_{\bar{d}}|X, R)^\top\right]$ is invertible,

$$\mathbb{E}\Big[H_d(X, R)\mathbb{E}(W_d|X, R)^\top\Big]^{-1}\mathbb{E}\Big[H_d(X, R)\gamma_d(X, R)\Big]$$

is a constant function as we vary $H_d(\cdot)$, which is testable. For instance, returning to the binary outcome example in Section 3.1, we know that we must have

$$\Big(\mathbb{E}[\Delta(o)H_d(R)]\Big)^{-1}\mathbb{E}[\Delta(e)H_d(R)]$$

constant as a function of $H_d(\cdot)$. This provides over-identifying restrictions that can be tested. In particular, we can compare alternative choices $H_d(\cdot)$ (i.e., the transformation of the RSV variable R we consider) and test whether this would provide significantly different results. If they do, then it would indicate violation of the identifying conditions.

4 Estimation and Inference

In this section, we turn to estimation and inference with a binary outcome and no covariates, our leading case in our application. For expositional convenience, we focus on the case in which Assumption 3.2 holds, although the discussion directly extends to settings where Assumption 3.1 holds (see Remark 5). The more general case with covariates and non-binary outcomes is studied in Appendix A and requires additional notation.

Let the average treatment effect in the experimental sample, $\theta = \mathbb{E}[Y(1) - Y(0)|S = e]$, be the estimand of interest. We provide a description of our procedure below and outline the exact steps in Algorithm 1.

Estimation through an "inverse" representational variable problem: Following Theorem 3.2, and Equation (10), our target parameter is defined by the following (linear) conditional moment condition

$$g(\theta, R) := \mathbb{E}\Big[\Delta(e) - \theta\Delta(o)\Big|R\Big] = 0,$$

where $\Delta(e)$ and $\Delta(o)$ are as defined in Equation (11). The conditional moment condition implies that, for any function $H(r)$,

$$\mathbb{E}\Big[H(R)g(\theta, R)\Big] = 0,$$

providing us with a moment condition for θ.

Efficient (but infeasible) choice of the representation: Which function $H(\cdot)$ should we choose? In principle, we could choose any function $H(\cdot)$, but of course we want $H(\cdot)$ that maximizes efficiency. To achieve this goal, we use efficiency results in Chamberlain (1987) and Newey (1993), keeping in mind, however, that here R is possibly high-dimensional. Chamberlain (1987) shows that it is sufficient (in our case) to find a univariate function $H(\cdot)$ to achieve the semi-parametric efficiency bound. Mapping results in Chamberlain (1987) to our case, the choice $H^*(\cdot)$ that achieves the semi-parametric efficiency bound takes the following form:

$$H^*(R) = \frac{\mathbb{E}[\Delta(o)|R]}{\sigma^2(\theta, R)}, \quad \sigma^2(\theta, R) = \mathbb{E}\left[\Big(\Delta(e) - \theta\Delta(o)\Big)^2\Big|R\right], \tag{14}$$

where it is immediate that

$$\sigma^2(\theta, R) = \mathbb{E}\left[\Delta(e)^2 + \theta^2\Delta(o)^2\Big|R\right].$$

To gain insight in the structure of the optimal representation $H^*(\cdot)$, note that we can write

$$\mathbb{E}[\Delta(o)|R] = \frac{\mathbb{P}(S = o|R)}{\mathbb{P}(S = o)\kappa(o)}\Big(\mathbb{P}(Y = 1|S = o, R) - \mathbb{P}(Y = 1|S = o)\Big),$$

where $\kappa(o) = \mathbb{P}(Y = 1|S = o)(1 - \mathbb{P}(Y = 1|S = o))$. That is, we can think of the efficient representation as a function of the conditional probability of $Y = 1$ given the remotely sensed variable, with appropriate weights that also depend on the remote sensed variable. We can further interpret $\sigma^2(\theta, R)$ as the inverse variance of our regression, which is also a function of three (unknown) nuisance functions:

$$\mathbb{P}(D = 1|R, S = e), \quad \mathbb{P}(Y = 1|R, S = o), \quad \mathbb{P}(S = o|R).$$

The structure of the optimal representation $H^*(\cdot)$ is intuitive: although we can choose many representations, the efficient choice depends on functions of the conditional probability of treatment and outcome given the RSV R.

Finally, note that we can also control for baseline characteristics X to possibly improve prediction accuracy without affecting consistency or inference.

Constructing a representation using machine learning techniques: If we knew the function $H(R)$, we could construct an efficient estimator by taking as our efficient moment

condition (Newey, 1993)

$$g_n(\theta, H) := \frac{1}{n}\sum_{i=1}^{n}\Big(\Delta_i(e) - \theta\Delta_i(o)\Big)H(R_i).$$

Unfortunately, we do not know $H(\cdot)$ and we certainly do not want that the high-dimensionality or complexity of R may affect $\sqrt{n}$-converge rates and valid asymptotic inference. To construct a representation $H(\cdot)$ we can follow two alternative approaches:

(i) Imposing that $H \in \mathcal{H}$ for a pre-specified function space $\mathcal{H}$ with bounded complexity to guarantee that the corresponding estimate representation $\hat{H}(R)$ satisfy

$$\sqrt{n}\Big(g_n(\theta, H) - g_n(\theta, \hat{H})\Big) = o_p(1). \tag{15}$$

That is, we want a representation $H(\cdot)$ that guarantees that for *some* $H(\cdot)$ the second step estimation error from $\hat{H}$ is asymptotically negligible (which holds for a large class of machine learning estimators, see Chernozhukov et al., 2020);

(ii) using cross fitting to estimate $H(\cdot)$ without any restriction on its function class, therefore allowing for very complex estimators such as deep neural networks.

Here, we focus on (i) and provide a formal discussion about (ii) in Appendix B.

Example of representations for (i): We note that Equation (15) does not pose a condition on the data-generating process, but rather it only amounts to restricting the set of representation researchers are willing to consider. See for example, Foster and Syrgkanis (2023) Hastie, Tibshirani and Wainwright (2015) Chernozhukov et al. (2020) for discussions on relevant high-level conditions on machine learning estimators (e.g., lasso, Random Forest or some neural networks) required so that Equation (15) is attained. To give an example for (i), we may use lasso by projecting $\Delta(o)$ onto a linear combination of R, imposing sparsity on the coefficient we estimate. Even when the true regression is non-sparse, the estimated coefficient converges to its pseudo-true value, and, H would denote the probability limit value of $\hat{H}$. Under Equation (15), consistency of the estimated causal effect is guaranteed by any function H (under mild invertibility conditions), despite sparsity not holding. One can generalize this argument to other machine learning estimators.[4]

[4]For a simple illustration for (i), suppose that we chose $H(R) = R^\top\beta$ linear in R, where we impose a sparse representation of β, and estimate $\hat{\beta}$ by projecting $\Delta(o)$ onto R. Here, sparsity of β does not entail restrictions on the data generating process (and failure of sparsity would not affect consistency of our estimator), but rather it is an intuitive condition we impose on the class of $H(\cdot)$ to trade-off bias with the variance of the estimated representation $\hat{H}$. Then we can write

$$\sqrt{n}\Big|g_n(\theta, H) - g_n(\theta, \hat{H})\Big| \leq \Big|\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\Delta_i(e) - \Delta_i(o)\Big)R_i\Big|\Big|_\infty ||\beta - \hat{\beta}||_1$$

where β is the probability limit of $\hat{\beta}$ within the constraint set defined by the sparsity condition imposed on estimation. We typically expect $||\beta - \hat{\beta}||_1 = \sqrt{\log(p)/n}$ where p is the dimension of R (Hastie, Tibshirani

When instead we think that the estimated representation would not admit a probability limit so that Equation (15) fails, as for example it may occur with deep neural networks, one should use cross-fitting derived in Appendix B.

Conducting inference: For estimated representations $\hat{H} \in \mathcal{H}$ that satisfy a bounded complexity restriction so that Equation (15) holds, it directly follows that from standard arguments (Newey and McFadden, 1994)

$$\sqrt{n} g_n(\theta, \hat{H}) = \sqrt{n} g_n(\theta, H) + o_p(1) \to \mathcal{N}(0, \mathbb{V}_\theta(H))$$

for a variance matrix $\mathbb{V}_\theta(H)$ as a function of H. We can therefore directly conduct standard inference for methods of moments estimators as described in Newey and McFadden (1994).

In summary, the procedure uses many of the key steps of empirical practice—predicting treatments and outcomes from RSVs—yet combines them in theoretically principled ways. In particular, these predictions appear in a specific form in the construction of the optimal representations for causal estimation.

Remark 3 (Estimating the unconditional probability)**.** So far, we assumed that we know (or could estimate at a fast rate) $\mathbb{P}(D = 1, S = e), \mathbb{P}(Y = 1, S = o)$. When this is not the case, we can use our same reasoning and replace $\Delta(o), \Delta(e)$ with

$$\hat{\Delta}_i(o) = \left(\frac{1\{Y_i = 1, S_i = o\}}{\mathbb{P}_n(Y = 1, S = o)} - \frac{1\{Y_i = 0, S_i = o\}}{\mathbb{P}_n(Y = 0, S = o)} \right)$$
$$\hat{\Delta}_i(e) = \left(\frac{1\{D_i = 1, S_i = e\}}{\mathbb{P}_n(D = 1, S = e)} - \frac{1\{D_i = 0, S_i = e\}}{\mathbb{P}_n(D = 0, S = e)} \right)$$

where $\mathbb{P}_n(x)$ denotes the empirical probability of event x. The only difference here is that the variance calculation should account for randomness when estimating each denominator. To do so, we can apply directly the Delta method, or use the non-parametric bootstrap. □

Remark 4 (Observing the outcome for some units in the experiment)**.** In some applications, researchers may collect outcome information for a small *subsample* of the experimental participants. We can integrate this information directly into our estimation procedure by using such observations to both construct $\Delta_i(e)$ as well as to construct $\Delta_i(o)$ without affecting valid inference. Formally, we can define $S \in \{e, o, \{e, o\}\}$, where $S = \{e, o\}$ indicates that the unit is in the experiment, but for this unit we also observe the outcome Y (and D). Our framework extends verbatim to these scenarios by replacing the conditions $S = e, S = o$ with $e \in S, o \in S$ in the indicator functions $1\{D_i = d, S_i = e\}$ and $1\{Y_i = y, S_i = o\}, d, y \in \{0, 1\}$, and in their corresponding conditional and unconditional expectation.[5] □

and Wainwright, 2015) and similarly we expect $\left|\left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big(\Delta_i(e) - \Delta_i(o) \Big) R_i \right|\right|_\infty$ under sub-gaussianity to be of order $\log(p)$, so that the expression is of order $o_p(1)$ if $\log(p)/n = o(1)$ (see e.g. Athey, Imbens and Wager, 2018).

[5] In this case, the efficient $H(\cdot)$ depends on the correlation between $\Delta_i(e)$ and $\Delta_i(o)$. In practice, because

Remark 5 (Estimation in the presence of direct effects (Assumption 3.1))**.** Whenever Assumption 3.1 holds, i.e., D may generate direct effects on R, we can follow this same argument verbatim. In particular, we replace $\Delta(o)$ and $\Delta(e)$ with

$$\widetilde{\Delta}(o) = \Big[V_i^1(1) - V_i^0(1)\Big] - \Big[V_i^1(0) - V_i^0(0)\Big], \quad \widetilde{\Delta}(e) = \Big[U_i(1) - U_i(0)\Big] - \Big[V_i^0(1) - V_i^0(0)\Big]$$

where V_i, U_i are defined in Equation (9). The efficient representation follows as in Equation (14) with $\widetilde{\Delta}(e), \widetilde{\Delta}(o)$ in lieu of $\Delta(e), \Delta(o)$ and σ^2 that also depends on the correlation between $\widetilde{\Delta}(e), \widetilde{\Delta}(o)$. □

5 Estimating the Effect of a Large Public Employment Program using Satellite Images

In this section, we apply our econometric framework to illustrate how researchers may use satellite images to evaluate the efficacy of a large-scale anti-poverty program.

We revisit a randomized experiment conducted in Muralidharan, Niehaus and Sukhtankar (2016) and revisited in Muralidharan, Niehaus and Sukhtankar (2023) that studied the impact of biometrically authenticated payments infrastructure ("Smartcards") on the beneficiaries of employment and pension programs in the Indian state of Andhra Pradesh. We find that the documented effects of the Smartcard program on poverty can be reproduced instead using satellite images as remotely-sensed variables for measures of local income and consumption.

5.1 Experiment and Data Description

In order to evaluate its effects, Muralidharan, Niehaus and Sukhtankar (2016) randomized the rollout of the Smartcard system across subdistricts ("mandals") Andhra Pradesh across three waves between 2010 and 2012. The authors divided approximately 450 mandals in Andhra Pradesh into four groups. First, a group of mandals were selected to remain outside of the study, and the authors refer to these as "non-study" mandals. The remaining mandals were then randomized into one of three groups: a control group in which the Smartcard program was not deployed until the end of 2012 (44 mandals), a "buffer control" control group in which Smartcards were deployed only in the second wave (136 mandals), and a treatment group in which Smartcards were deployed in the first wave in 2010 (111 mandals). Figure 2 summarizes the geographic location of mandals and their group assignments.

we have freedom in choosing $H(\cdot)$, we can omit this term without affecting consistency, and without affecting efficiency if the number of experimental units for which we observe the outcome of interest is small relative to the overall size of the dataset.

Algorithm 1 Two-step estimation procedure with binary outcome under Assumption 3.2

Require: Observations $\Big(1\{S=o\}Y, S, D, X, R\Big)_{i=1}^{n}$.

1: Estimate $\mathbb{P}(Y=1|R, S=o), \mathbb{P}(S=o|R)$ and $\mathbb{P}(D=1|R, S=e)$ using arbitrary machine learning estimators. Denote such estimators as $\hat{P}_Y(R), \hat{P}_{S=o}(R), \hat{P}_D(R)$.

2: Denote $\hat{p}_D(1) = \mathbb{P}_n(D=1, S=e), \hat{p}_D(0) = \mathbb{P}_n(D=0, S=e)$, $\hat{p}_Y(1) = \mathbb{P}_n(Y=1, S=o), \hat{p}_Y(0) = \mathbb{P}_n(Y=0, S=o)$ where $\mathbb{P}_n(x)$ denotes the empirical probability of event x.

3: Define

$$\hat{Q}_D(R) = \Big(1-\hat{P}_{S=o}(R)\Big)\Big(\frac{\hat{P}_D(R)}{\hat{p}_D(1)} - \frac{1-\hat{P}_D(R)}{\hat{p}_D(0)}\Big), \quad \hat{Q}_Y(R) = \hat{P}_{S=o}(R)\Big(\frac{\hat{P}_Y(R)}{\hat{p}_Y(1)} - \frac{1-\hat{P}_Y(R)}{\hat{p}_Y(0)}\Big).$$

4: For all units i

$$\hat{\theta}^{\text{first step}} = \arg\min_{\theta} \sum_i \Big(\hat{Q}_D(R_i) - \hat{Q}_Y(R_i)\theta\Big)^2.$$

5: Construct the function

$$r \mapsto \hat{H}(r) = \frac{\hat{Q}_Y(r)}{\hat{\sigma}^2(r)},$$

$$\hat{\sigma}^2(r) = \Big(1-\hat{P}_{S=o}(R)\Big)\Big(\frac{\hat{P}_D(r)}{\hat{p}_D(1)^2} + \frac{1-\hat{P}_D(r)}{\hat{p}_D(0)^2}\Big) + \hat{P}_{S=o}(R)(\hat{\theta}^{\text{first step}})^2\Big(\frac{\hat{P}_Y(r)}{\hat{p}_Y(1)^2} + \frac{1-\hat{P}_Y(r)}{\hat{p}_Y(0)^2}\Big).$$

6: Estimate $\hat{\Delta}_i(e), \hat{\Delta}_i(o)$ for all units i in fold k

$$\hat{\Delta}_i(o) = \Big(\frac{1\{Y_i=1, S_i=o\}}{\mathbb{P}_n(Y=1, S=o)} - \frac{1\{Y_i=0, S_i=o\}}{\mathbb{P}_n(Y=0, S=o)}\Big),$$
$$\hat{\Delta}_i(e) = \Big(\frac{1\{D_i=1, S_i=e\}}{\mathbb{P}_n(D=1, S=e)} - \frac{1\{D_i=0, S_i=e\}}{\mathbb{P}_n(D=0, S=e)}\Big).$$

7: Estimate

$$\hat{\theta} = \frac{\sum_i \hat{\Delta}_i(e)\hat{H}(R_i)}{\sum_i \hat{\Delta}_i(o)\hat{H}(R_i)}.$$

8: Estimate the standard error of $\hat{\theta}$ either through the Delta method or by Bootstrapping units i (while fixing the estimated function $\hat{H}$ whose estimation error is of second order under Equation (15)). Denote $\hat{v}$ the estimated standard error of $\hat{\theta}$.

9: Compute a α-level confidence interval as $\hat{C}(\alpha)$ as

$$\hat{C}(\alpha) = \Big[\hat{\theta} - \Phi^{-1}(1-\alpha/2)\hat{v}, \hat{\theta} + \Phi^{-1}(1-\alpha/2)\hat{v}\Big].$$

Treatment: In our empirical analysis, we define the treatment as whether the Smartcard program was rolled out in a mandal as part of the first wave in 2010. Mapping into the notation of our framework, we define $D=1$ for a mandal assigned to the treatment group, and we define $D=0$ for a mandal assigned to either the control or the buffer control group. Consequently, we interpret the target estimand as measuring the causal effect of the

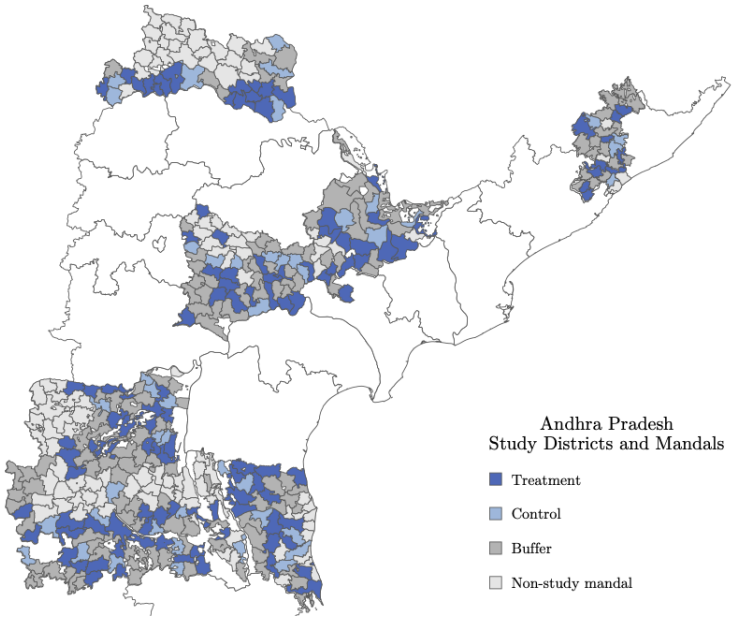

Figure 2: Map of subdistricts ("mandals") in Andhra Pradesh, reproduced from Figure C.1 in Muralidharan, Niehaus and Sukhtankar (2016), and their assignments to the non-study, control, buffer control, and treatment groups.

a long-term (i.e., two-year) adopt of the Smartcard program against either no adoption or short-term adoption. We further define the experimental sample $S = e$ as all mandals in either the treatment, control, or buffer control group; and we define the observational sample $S = 0$ as all non-study mandals.

Extracting village-level coordinates: For each mandal in the study, we collect all village-level (i.e., "Shrid") identifiers using data from the Socioeconomic High-Resolution Rural-Urban Geographic Platform for India (SHRUG) (Asher et al., 2021).[6] Each shrid corresponds to a village: in our analyses, we focus on shrids with a population of at least 100 individuals, removing 2% of shrids with a population below 100 individuals.

In Table 3, we summarize the number and composition of shrids across the treatment groups (non-study, control, buffer control, and treatment). Altogether, our sample contains over 8,300 shrids, of which approximately 6,000 are in the experimental sample (i.e., either in a treatment, control or buffer control mandal). On average, shrids have a population of approximately 2,000 individuals. Although villages tend to be located in rural areas, the populated area within a typical shrid tends to be geographically concentrated – see, for example, Figure XX which provides a satellite image of a particular shrid in our sample.

Target outcome: For each shrid, we use the 2012-2013 Social Economic and Caste Census SECC) and the 2013 Indian Economic Census to obtain shrid-level variables for the

[6]The data is available at https://www.devdatalab.org/shrug.

Table 3: Summary statistics about villages (i.e., shrids) based on their treatment status.

	Outside Study	Control	Buffer Control	Treatment
Number of Shrids	2,260	853	2,931	2,276
Number mandals	105	44	136	111
Average pop	2,143	2,296	2,285	2,604
Urban area	0.002	0.001	0.003	0.004
Average male pop	0.512	0.508	0.506	0.508
Average female pop	0.489	0.492	0.495	0.493

Figure 3: Satellite image of an illustrative shrid. The blue line denotes the polygon used to collect satellite information. Although shrids are typically located in rural areas, the population tend to be concentrated in a more dense region within the shrid.

target outcome Y. As our main target outcome, we study whether, in a given shrid the average per-capita consumption is below its first quartile in the data. Average per-capita consumption is estimated the Shrid level by the SHRUG project following Asher and Novosad (2020) using the SECC.

We also study the effect of the Smartcard program on two additional target outcomes that we obtain directly from the SECC. To measure poverty, the SECC classifiers households into low, middle, and higher-income categories based on whether the highest-earning member of the household reported monthly earnings below Rs. 5,000, between Rs. 5,000 and 10,,000, and greater than Rs. 10,000 respectively. Our two additional target outcomes analyze whether the shrid has (i) no household above Rs. 5,000; and (ii) no household above Rs. 10,000. Of the Shrids in the pure control group, 6% have no household with an income level above Rs. 5,000, and 21% with no household above Rs. 10,000. Intuitively, the share of households above each income bracket provides us with a measure of the economic

development of the specific village and emulates the exercise using individual-level data in Muralidharan, Niehaus and Sukhtankar (2023).

Remotely-sensed variable: For each shrid, we extract its longitude and latitude coordinates obtained from the SHRUG platform in order to define its geographical boundaries as a polygon. In Figure 3, we illustrate the boundaries of a particular shrid in our sample. We then construct remotely-sensed variable R that consists of shrid-specific summary statistics of nightlight images from 2012 to 2020 as well as embeddings of daytime satellite images collected in 2019.

More specifically, it is well-known that nightlights can be used to predict local economic conditions (e.g., Henderson, Storeygard and Weil, 2011), and so it is natural to suspect that these could be used as a noisy measurement for local poverty in the experimental sample. For each shrid, we therefore calculate the average pixel-level estimate of different luminosity measures following Asher et al. (2021), which are available through the SHRUG platform.

We additionally generate remotely-sensed variables based on embeddings of daytime satellite images of each Shrid using MOSAIKS (short for "Multi-Task Observation Using Satellites and Kitchen Sinks"), developed by Rolf et al. (2021). Daytime satellite imagery has been found contain ample signal for predicting local poverty (Jean et al., 2016), and more recently the MOSAIKS embeddings have been shown to accurately predict outcomes that enter into the United Nations Human Development Index at a granular level (Sherman et al., 2023). By passing the polygon coordinates of each shrid to the MOSAIKS API, we obtain a high-dimensional (4,000 dimensions) embedding of daytime satellite imagery for each Shrid that we use for our prediction tasks together with the nightlight features.

5.2 Main Empirical Results

To illustrate our framework, we consider the following thought experiment: suppose the researchers only collected the target outcome in roughly half of the shrids in the experimental sample. Could we use the remotely-sensed variable and the observational sample to still recover the average treatment effect of the Smartcard system on average per-capita consumption and two measures of the income distribution? This exercise mimics empirical applications in which researchers have limited information about the target outcome in the experimental sample.

We specifically consider two scenarios to illustrate our framework:

(A) The researcher only observes the target outcomes for half of the shrids in the experimental sample ("random subset");

(B) The researcher observes the target outcomes only for individuals in the observational and in the buffer control group ("Buffer and Holdout").

	RSV Experiment
Treatment D	Smartcard system (wave 1)
Outcome Y	Two measures of earnings, one measure of consumption
Remote sensed variable R	Village-level satellite images (day and nightlight)
Data Sources	SECC 2012 + online satellite images
Validation exercise	Estimate effects without Y for half of villages in the experiment
Number of villages with missing Y	$\sim 3,000$
Cost saving (0.5$/person)	3 million dollars

Table 4: Illustration of how our theoretical framework maps to this empirical exercise. We simulate an environment where we have no access to the outcome of the pure control and treatment group (about 3,000 shrids). If we were to collect outcome data for all of such shrids, the data collection cost would be about 3 million dollars.

In both scenarios, we are censoring the target outcome for about 3,000 shrids. Table 4 summarizes our exercise. Assuming a conservative lower bound that it costs only $0.50 per person to collect the target outcomes using traditional surveying strategies, not requiring outcome information for about 3,000 villages has a total economic value in survey costs of about 3 million dollars (since there are on average 2,000 individuals per village).[7]

Representation learning: For both scenarios, we implement our procedure as described in Section 4, following Remark 4 to account for observing the target outcome for some units in the experimental sample. Even though the remotely-sensed variable in this application is extremely high-dimensional, it suffices to compress R into a scalar representation $H(R)$ to achieve the semiparametric efficiency bound. To construct this efficient representation, we follow Algorithm 1 and use Random Forests to build our prediction functions for the target outcome and the treatment using the remotely-sensed variable.

Focusing first on Scenario (A), Figure 4 provides a bin-scatter plot of the predicted probability the consumption outcome equals one (i.e., average per-capita consumption falls below the first quartile) against the continuous, log average per-capita consumption observe in the data. The left panel Figure 4 shows a strong correlation between the predicted probabilities and log consumption; suggesting that there is substantial signal for predicting

[7]Survey costs may typically be larger, for example *phone* survey costs in Pakistan collected by Viviano and Rudder (2024) amount to total costs of about 7$ per respondent.

local consumption using satellite images in this sample.

How does the efficient representation $H(R)$ differ from directly predicting the target outcome given the remotely-sensed variable? The right panel of Figure 4 reports the estimate of the optimal representation $H(R)$ outputted by our algorithm plotted against the log-consumption. We would like the optimal representation $H(R)$ to be the one that is most correlated with our target binary outcome of whether consumption is below its first quartile. In Figure 4 consumption is below its first quartile if below the horizontal dotted red line. Notice that the estimated optimal representation differs from directly predicting the target outcome in a particular way: the optimal representation takes negative value as the target outcome equals zero and positive as the outcome equals one, and there is a sharp change its values around the first quartile. Intuitively, the optimal representation transforms the predictions we obtain for the outcome and treatment to maximize the amount of information we can learn to recover treatment effects on our binary outcome.

Defining the true regression: Since we actually observe the target outcome for all units in the experimental sample, we can compare our procedure to the estimated average treatment effect in the experimental sample. In particular, define the difference-in-means between the treatment group and the full control group as

$$\hat{\theta}^{\text{true}} = \frac{1}{\sum_i D_i 1\{e \in S_i\}} \sum_i D_i 1\{e \in S_i\} Y_i - \frac{1}{\sum_i (1-D_i) 1\{e \in S_i\}} \sum_i 1\{e \in S_i\}(1-D_i) Y_i,$$

where $e \in S_i$ if i is either in the buffer control, control, or treatment group. Here, $\hat{\theta}^{\text{true}}$ is the estimated effect from regression the outcomes in the experimental sample in the (randomized) treatments. Clearly, $\hat{\theta}^{\text{true}}$ is unbiased for $\theta = \mathbb{E}[Y(1) - Y(0) | e \in S_i]$.

Comparing RSV and the true regression: Figure 5 reports the estimated effect using our proposed method (RSV) against $\hat{\theta}^{\text{true}}$, and the corresponding confidence interval obtained as in Algorithm 1 with standard errors clustered at the mandal level for both our method and $\hat{\theta}^{\text{true}}$.[8] For either (A) (red line) or (B) (green line), our method uses the outcomes from the experiment of only about 50% of shrids. For (B), our method does not have access to the outcome of any treated shrid.

Results are consistent with findings in Muralidharan, Niehaus and Sukhtankar (2023), with the program decreasing overall poverty level and increasing consumption. We find no statistical difference between the true regression and the proposed method across the three outcomes of interest. Point estimates preserve the same signs and magnitudes across the three outcomes. The length of confidence intervals is comparable and smaller as we contrast our procedure to the regression that observes the outcomes of all individuals in the experiment, despite our method having access to the outcomes of only 50% of individuals in

[8] For our method, we construct confidence intervals using the non-parametric bootstrap at the sub-district level under the assumption that Random Forest satisfies Equation (15), see Foster and Syrgkanis (2023) for a discussion.

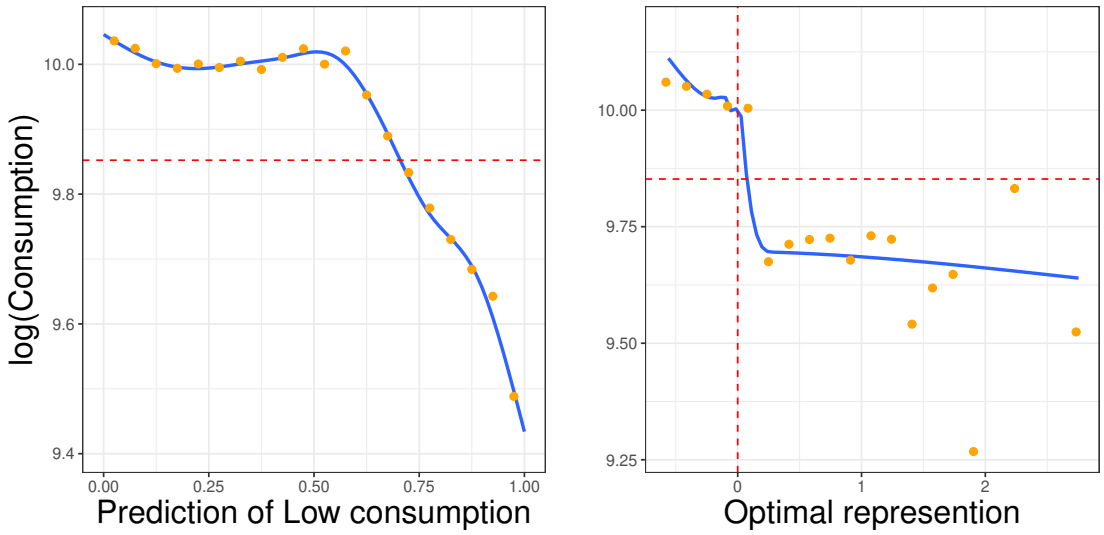

Figure 4: The left-hand side panel reports the bin-scatter plot of predicting whether consumption is below its first quartile (x-axis) against the continuous value of log-consumption. We form predictions with Random Forest using half of the experimental data, with class weights ten-to-one to balance unbalanced labels of the outcome. The red-dotted line reports the value of log-consumption corresponding to its first quartile. The right-hand side reports the estimated optimal representation $H(R)$ (x-axis) against log-consumption. The optimal representation transforms the predictions we obtain through Random Forest to maximize correlation with the target binary outcome. Here, $H(R)$ sharply changes from negative to positive values as the binary outcome switches its sign (it is below the horizontal dotted red line) to maximize information we learn from the data.

the experiment. By using information from easily accessed satellite images, we can recover the same treatment effect with about the same (or better) level of precision, even in the absence of outcomes of interest for half of the experimental sample.

5.3 Simulations to Compare with Common Empirical Practice

As a final exercise, we contrast our method with current empirical practice involving remotely-sensed variable — i.e., plugging in the predicted outcome and interpreting it as a surrogate problem as discussed in Section 2.3. To do so, we focus on the consumption target outcome, and design a Monte Carlo simulation calibrated to our empirical application. In particular, we randomly generate shrid-level datasets in the following manner:

- We draw treatments independently across shrids with probability 1/2;
- We draw potential outcomes with replacement $Y_i(1), Y_i(0)$ independently from the set of outcomes in the treatment and control group in the experiment;
- We draw with replacement remotely-sensed variables $R_i(1), R_i(0)$ independently from

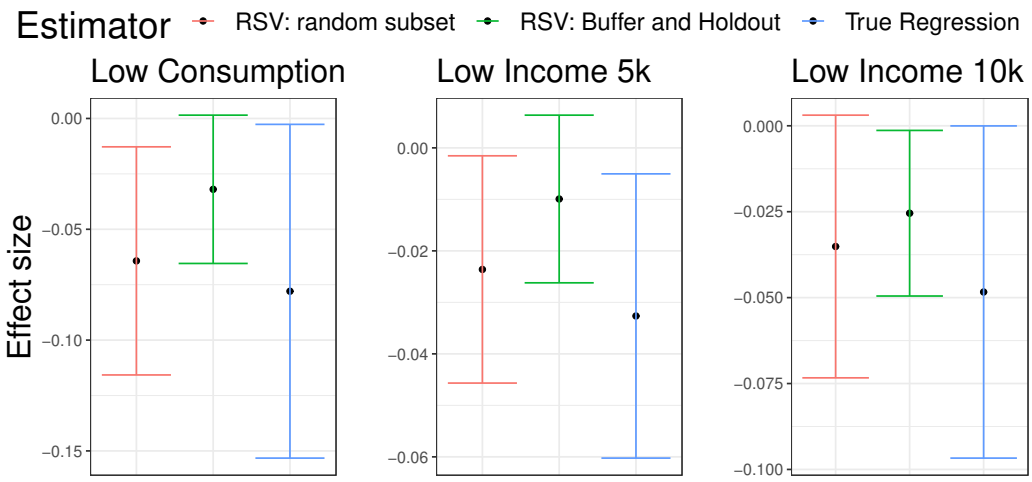

Figure 5: Main result from the empirical application. Each panel corresponds to a different outcome of interest, with low consumption indicating whether consumption is below its first quartile, Low Income 5k indicating that no individual has income above Rs. 5,000 in the Shrid, and Low Income Rs. 10k indicating no individual has income of above Rs. 10,000. The red lines report point estimates and 90% confidence intervals of our proposed method (RSV) that uses outcome information only from the first half of subdistricts. The green lines report point estimates and 90% confidence intervals obtained through our method (RSV) that uses outcome information only from the control buffer and holdout sample group. Therefore, neither two cases observe the outcome of about $\sim 3,000$ Shrids in the main experiment. The blue lines report the estimated effect and 90% confidence interval of the regression that uses outcome information for all units in the experiment. Standard errors are clustered at the mandal (sub-district) level.

the set of units with outcome equal to one and zero respectively;

- We denote $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$ and $R_i = Y_i R_i(1) + (1 - Y_i) R_i(0)$;

On each simulated dataset, we consider scenario (B) in which the target outcome is not observed for any treated unit by the researcher. For each dataset, we compare our RSV method against empirical practice that treats the remotely-sensed variable as a surrogate.

This design allows us to conduct simulations that are most closely tied to our empirical exercise. In particular, we do not have to posit any data-generating process about how the outcomes affect the remote sensed variable, as we non-parametrically draw such variables directly from our sample. In addition, because we draw the outcomes from the set of treated and control units, we know the true treatment effect in this simulation exercise, corresponding to the treatment effect in the experiment.

Both our procedure and surrogate use the same machine learning estimator to predict the outcomes, which is the same as what was used in our empirical exercise. We report the results in Figure 6 over two hundred replications. Surrogate has a larger bias and mean-squared

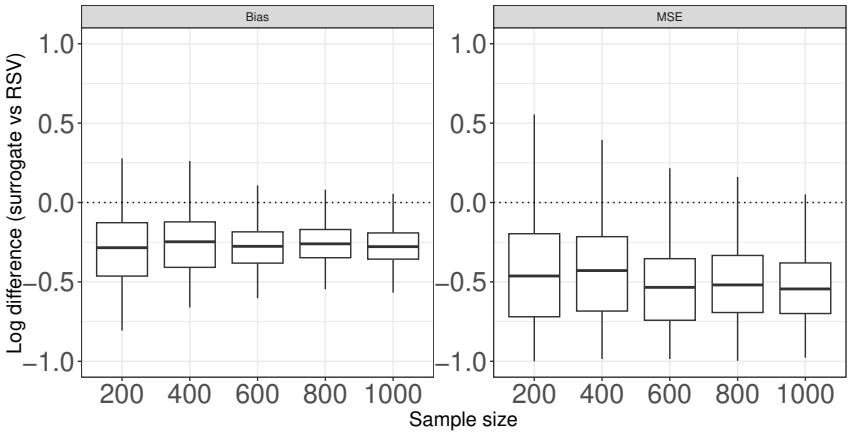

Figure 6: Simulations calibrated to the experiment over two-hundred replications to compare our method to surrogate methods. The x-axis reports the sample size for each simulation and the y-axis reports the different in log scale between the surrogate and the RSV method. The left-hand side panel reports the difference in bias and the right-hand side reports the difference in mean-squared error.

error on average with respect to our method. As the sample size increases, we see that the increase in the mean-squared error of the surrogate converges to about 50% with respect to RSV. This result corroborates our theoretical findings in Section 2.3 in our application.

References

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

Aiken, Emily, Suzanne Bellue, Dean Karlan, Chris Udry, and Joshua E. Blumenstock. 2022. "Machine learning and phone data can improve targeting of humanitarian aid." *Nature*, 603(7903): 864–870.

Aiken, Emily, Suzanne Bellue, Joshua Blumenstock, Dean Karlan, and Christopher R Udry. 2023. "Estimating Impact with Surveys versus Digital Traces: Evidence from Randomized Cash Transfers in Togo." National Bureau of Economic Research Working Paper 31751.

Allon, Gad, Daniel Chen, Zhenling Jiang, and Dennis Zhang. 2023. "Machine learning and prediction errors in causal inference." *The Wharton School Research Paper Forthcoming*.

Asher, Sam, and Paul Novosad. 2020. "Rural roads and local economic development." *American economic review*, 110(3): 797–823.

Asher, Sam, Tobias Lunt, Ryu Matsuura, and Paul Novosad. 2021. "Development research at high geographic resolution: an analysis of night-lights, firms, and poverty in India using the shrug open data platform." *The World Bank Economic Review*, 35(4): 845–871.

Athey, Susan, Guido W Imbens, and Stefan Wager. 2018. "Approximate residual balancing: debiased inference of average treatment effects in high dimensions." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4): 597–623.

Athey, Susan, Raj Chetty, Guido W Imbens, and Hyunseung Kang. 2024. "The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely." National Bureau of Economic Research Working Paper 26463.

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

Battaglia, Laura, Timothy Christensen, Stephen Hansen, and Szymon Sacher. 2024. "Inference for regression with variables generated from unstructured data." *arXiv preprint arXiv:2402.15585*.

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

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/debiased machine learning for treatment and structural parameters."

Chernozhukov, Victor, Whitney Newey, Rahul Singh, and Vasilis Syrgkanis. 2020. "Adversarial estimation of riesz representers." *arXiv preprint arXiv:2101.00009*.

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

Domínguez, Manuel A, and Ignacio N Lobato. 2004. "Consistent estimation of models defined by conditional moment restrictions." *Econometrica*, 72(5): 1601–1615.

Donaldson, Dave, and Adam Storeygard. 2016. "The view from above: Applications of satellite data in economics." *Journal of Economic Perspectives*, 30(4): 171–198.

Donald, Stephen G, Guido W Imbens, and Whitney K Newey. 2003. "Empirical likelihood estimation and consistent tests with conditional moment restrictions." *Journal of Econometrics*, 117(1): 55–93.

D'Haultfœuille, Xavier, Christophe Gaillac, and Arnaud Maurel. 2024. "Partially linear models under data combination." *Review of Economic Studies*, rdae022.

Egami, Naoki, Musashi Hinck, Brandon Stewart, and Hanying Wei. 2024. "Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models." *Advances in Neural Information Processing Systems*, 36.

Egger, Dennis, Johannes Haushofer, Edward Miguel, Paul Niehaus, and Michael Walker. 2022. "General equilibrium effects of cash transfers: experimental evidence from Kenya." *Econometrica*, 90(6): 2603–2643.

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

Foster, Dylan J, and Vasilis Syrgkanis. 2023. "Orthogonal statistical learning." *The Annals of Statistics*, 51(3): 879–908.

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. 2019. "Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech." *Econometrica*, 87(4): 1307–1340.

Ghassami, AmirEmad, Alan Yang, David Richardson, Ilya Shpitser, and Eric Tchetgen Tchetgen. 2022. "Combining experimental and observational data for identification and estimation of long-term causal effects." *arXiv preprint arXiv:2201.10743*.

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright. 2015. "Statistical learning with sparsity." *Monographs on statistics and applied probability*, 143(143): 8.

Henderson, J. Vernon, Adam Storeygard, and David N. Weil. 2011. "A Bright Idea for Mesuring Economic Growth." *American Economic Review*.

Huang, Luna Yue, Solomon M Hsiang, and Marco Gonzalez-Navarro. 2021. "Using Satellite Imagery and Deep Learning to Evaluate the Impact of Anti-Poverty Programs." National Bureau of Economic Research Working Paper 29105.

Hu, Yingyao. 2008. "Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution." *Journal of Econometrics*, 144(1): 27–61.

Hu, Yingyao. 2017. "The econometrics of unobservables: Applications of measurement error models in empirical industrial organization and labor economics." *Journal of econometrics*, 200(2): 154–168.

Imbens, Guido, Nathan Kallus, Xiaojie Mao, and Yuhao Wang. 2024. "Long-term causal inference under persistent confounding via data combination." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, qkae095.

Imbens, Guido W, and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences.* Cambridge university press.

Jack, B. Kelsey, Seema Jayachandran, Namrata Kala, and Rohini Pande. 2022. "Money (Not) to Burn: Payments for Ecosystem Services to Reduce Crop Residue Burning." National Bureau of Economic Research Working Paper 30690.

Jayachandran, Seema, Joost de Laat, Eric F. Lambin, Charlotte Y. Stanton, Robin Audy, and Nancy E. Thomas. 2017. "Cash for carbon: A randomized trial of payments for ecosystem services to reduce deforestation." *Science*, 357(6348): 267–273.

Jean, Neal, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. 2016. "Combining satellite imagery and machine learning to predict poverty." *Science*, 353(6301): 790–794.

Johannemann, Jonathan, Vitor Hadad, Susan Athey, and Stefan Wager. 2019. "Sufficient representations for categorical variables." *arXiv preprint arXiv:1908.09874*.

Kallus, Nathan, and Xiaojie Mao. 2022. "On the role of surrogates in the efficient estimation of treatment effects with limited outcome data."

Kitamura, Yuichi, Gautam Tripathi, and Hyungtaik Ahn. 2004. "Empirical likelihood-based inference in conditional moment restriction models." *Econometrica*, 72(6): 1667–1714.

Molinari, Francesca. 2008. "Partial identification of probability distributions with misclassified data." *Journal of Econometrics*, 144(1): 81–117.

Molinari, Francesca, and Marcin Peski. 2006. "Generalization of a result on "regressions, short and long"." *Econometric Theory*, 22(1): 159–163.

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

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

Newey, Whitney K. 1993. "16 Efficient estimation of models with conditional moment restrictions."

Newey, Whitney K, and Daniel McFadden. 1994. "Large sample estimation and hypothesis testing." *Handbook of econometrics*, 4: 2111–2245.

Park, Chan, David B Richardson, and Eric J Tchetgen Tchetgen. 2024. "Single proxy control." *Biometrics*, 80(2): ujae027.

Proctor, Jonathan, Tamma Carleton, and Sandy Sum. 2023. "Parameter Recovery Using Remotely Sensed Variables." National Bureau of Economic Research Working Paper 30861.

Rolf, Esther, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. 2021. "A generalizable and accessible approach to machine learning with global satellite imagery." *Nature communications*, 12(1): 4392.

Schennach, Susanne M. 2020. "Mismeasured and unobserved variables." In *Handbook of econometrics*. Vol. 7, 487–565. Elsevier.

Sherman, Luke, Jonathan Proctor, Hannah Druckenmiller, Heriberto Tapia, and Solomon M Hsiang. 2023. "Global High-Resolution Estimates of the United Nations Human Development Index Using Satellite Imagery and Machine-learning." National Bureau of Economic Research Working Paper 31044.

Singh, Rahul, and Suhas Vijaykumar. 2023. "Kernel ridge regression inference." *arXiv preprint arXiv:2302.06578*.

Vafa, Keyon, Susan Athey, and David M. Blei. 2024. "Estimating Wage Disparities Using Foundation Models."

Viviano, Davide, and Jess Rudder. 2024. "Policy design in experiments with unknown interference."

Walker, Kendra, Ben Moscona, Kelsey Jack, Seema Jayachandran, Namrata Kala, Rohini Pande, Jiani Xue, and Marshall Burke. 2022. "Detecting Crop Burning in India using Satellite Data."

Zhang, Jingwen, Wendao Xue, Yifan Yu, and Yong Tan. 2023. "Debiasing Machine-Learning-or AI-Generated Regressors in Partial Linear Models." *Available at SSRN*.

Zheng, Tongshu, Michael H Bergin, Shijia Hu, Joshua Miller, and David E Carlson. 2020. "Estimating ground-level PM2. 5 using micro-satellite images by a convolutional neural network and random forest approach." *Atmospheric Environment*, 230: 117451.

A Estimation with Non-Binary Outcome

This section studies estimation and inference for non-binary outcomes. We first investigate the case where the outcome has a discrete support. We then extend this to non-discrete support in Appendix A.2.

A.1 Moment Conditions

Rather than working directly with the conditional moment restrictions, we can construct moment restrictions by averaging over $R \mid X$ and constructing instrument functions $H_d(X, R)$ of dimension K for arbitrary $K \geq |\mathcal{Y}|$ in the spirit of Newey (1993); Ai and Chen (2003); Donald, Imbens and Newey (2003); Domínguez and Lobato (2004); Kitamura, Tripathi and Ahn (2004) among others.

Let us now define

$$\tau_d^\star(X) = \left\{\tau_{d,y_1}^\star(X), \cdots, \tau_{d,y_{|\mathcal{Y}|}}^\star(X)\right\}, \quad \tau_{d,y}^\star(X) = \mathbb{P}\{Y(d) = y | X, S = e\},$$

for $y \in \mathcal{Y}$ the main estimands of interest.

Corollary 3.1 suggest that we can estimate the distribution of potential outcomes either through a conditional moment restriction or by constructing moment restrictions (unconditional on R) based on instrument functions H. Such instrument functions can be *arbitrary* to guarantee consistency and valid asymptotic inference as long as $\mathbb{E}[H_d(X, R)W_{\bar{d}}^\top | X]$ is full rank, where $\bar{d} = 0$ if Assumption 3.2 holds and $\bar{d} = d$ otherwise.

A.1.1 First step estimation for given represention $H(\cdot)$

First, we introduce the estimator fixing the representation function H.

Our main estimation strategy uses two sets of moment conditions. The first set of moment conditions estimates the low dimensional parameters $P(X = x), P(S = e|X), P(Y = y, S = o|X), P(D = d, S = e|X)$. We write these as

$$M_{0,i}^{\bar{d}}(\eta, x) = \begin{bmatrix} 1\{Y = y_1, S = o, X = x, D = \bar{d}\} - \eta_1(x) \\ \vdots \\ 1\{Y = y_{|\mathcal{Y}|}, S = o, X = x, D = \bar{d}\} - \eta_{|\mathcal{Y}|}(x) \\ 1\{D = 1, S = e, X = x\} - \eta_{|\mathcal{Y}|+1}(x) \\ 1\{S = e, X = x\} - \eta_{|\mathcal{Y}|+2}(x) \\ 1\{X = x\} - \eta_{|\mathcal{Y}|+3}(x) \end{bmatrix}$$

and write in compact form

$$M_{0,i}^{\bar{d}}(\eta) = \left[M_{0,i}^{\bar{d}}(x_1), \cdots, M_{0,i}^{\bar{d}}(x_{|\mathcal{X}|}), 1\{S = e\} - \eta_{|\mathcal{Y}|+4}\right]$$

where η denotes a low dimensional parameter that captures simple conditional expectations. Denote $\eta^\star$ the true parameter such that $\mathbb{E}[M_{0,i}^{\bar{d}}(\eta^\star)] = 0$ (unique by construction). Under

Assumption 3.2, because the observational group has no treated units, we will only focus on $\bar{d} = 0$. We return to estimation under Assumption 3.1 in Remark 6.

The second set of moment conditions are those corresponding to the conditional moments, which we write as

$$M_{1,i}^{\bar{d}}(\tau, \eta, d, x) = \left[\sum_{g=1}^{|\mathcal{Y}|} \frac{1\{Y = y_g, S = o, X = x, D = \bar{d}\}}{\eta_g(x)} \tau_{d,y_g}(x) - \frac{1\{D = d, S = e, X = x\}}{(1-d) \times (1 - \eta_{|\mathcal{Y}|+1}(x)) + d \times \eta_{|\mathcal{Y}|+1}(x)} \right]$$

and

$$M_{1,i}^{\bar{d}}(\tau, \eta) = \left[M_{1,i}^{\bar{d}}(\tau, \eta, d = 0, x_1), \cdots, M_{1,i}^{\bar{d}}(\tau, \eta, d = 0, x_{|\mathcal{X}|}), M_{1,i}^{\bar{d}}(\tau, \eta, d = 1, x_1), \cdots, M_{1,i}^{\bar{d}}(\tau, \eta, d = 1, x_{|\mathcal{X}|}) \right].$$

For now, consider a weighting matrices $\left[H_0(X, R), H_1(X, R) \right] \in \mathbb{R}^{2|\mathcal{Y}| \times 1}$, and the corresponding estimator $\left(\hat{\tau}_H, \hat{\eta}_H \right)$ such that

$$\sum_{i=1}^{n} M_{0,i}^{\bar{d}=0}(\hat{\eta}_H) = 0, \quad \sum_{i=1}^{n} H(X, R) M_{1,i}^{\bar{d}=0}(\hat{\tau}_H, \hat{\eta}_H) \quad = 0,$$

subsuming Assumption 3.2 holds. This corresponds to a simple method of moments with as many moments as the number of parameters. We write our estimated effect (for given weights w) as

$$\theta_w(\hat{\tau}, \hat{\eta}) = \sum_{x \in \mathcal{X}, g \in \{1, \cdots, |\mathcal{Y}|\}} \frac{\hat{\eta}_{|\mathcal{Y}|+2}(x)}{\hat{\eta}_{|\mathcal{Y}|+3}(x)} w(y_g) \Big(\hat{\tau}_{1,g}(x) - \hat{\tau}_{0,g}(x) \Big).$$

The main advantage of our identification strategy is that it does not require consistently estimating $\mathbb{P}(Y = y|R, X)$ or $\mathbb{P}(D = 1|R, X)$ for valid asymptotic inference. As noted in a long-standing literature on conditional moment equality (Chamberlain, 1987; Newey, 1993), it suffices that the choice of the matrix $H(X, R)$ presents sufficient variation in R to be able to construct consistent estimators.

In particular, what we need is that $\mathbb{E}[H_d(X, R) W_{\bar{d}}^{\top} | X]$ with $W_{\bar{d}}$ as defined in Corollary 3.2 is full rank almost surely for $d \in \{0, 1\}$.

Inference follows directly from the literature on methods of moments (e.g. Newey and McFadden, 1994).

A.1.2 Second step estimator for efficiency improvement

Next, we turn to the question of choosing the matrix H. From Chamberlain (1987) and Newey (1993) it follows that the estimator that imposes the moments

$$\sum_{i=1}^{n} \underbrace{\mathbb{E}\left[\frac{\partial M_{1,i}^{\bar{d}=0}(\tau, \eta^\star)}{\partial \tau} \Big|_{\tau=\tau^\star} | X, R \right]^{\top}}_{:=\alpha(X,R)} \underbrace{\mathbb{E}\left[M_{1,i}^{\bar{d}=0}(\tau^\star, \eta^\star) M_{1,i}^{\bar{d}=0}(\tau^\star, \eta^\star)^{\top} \Big| X, R \right]^{-1}}_{:=\beta(X,R)} M_{1,i}^{\bar{d}}(\tau, \eta^\star) = \mathbf{0}$$

achieves the semi-parametric efficiency bound for estimating τ.

The core idea, therefore, is to use an hold-out sample (e.g., as in the cross fitting algorithm in Algorithm 1) to estimate such weights α, β. Specifically, researchers can follows the procedure:

- Divide the dataset into K fold, with $\mathcal{S}_k$ denoting the set of units in fold k.
- For each fold k, estimate $\hat{\alpha}^{-k}$, $\hat{\beta}^{-k}$ using all units except those in fold k.
- Estimate
$$\hat{\theta}^k_w(\hat{\tau}^k, \hat{\eta}^k): \quad \sum_{i \in \mathcal{S}_k} M_{0,i}(\hat{\eta}^k) = 0, \quad \sum_{i \in \mathcal{S}_k} \hat{\alpha}^{-k(i)}(X,R)\hat{\beta}^{-k(i)}(X,R) M_{1,i}(\hat{\tau}^k, \hat{\eta}^k) \quad = 0.$$
- Estimate the standard error $\hat{\sigma}_k$ of $\hat{\theta}^k_w(\hat{\tau}^k, \hat{\eta}^k)$ via bootstrap conditional or Delta method conditional on $\hat{\alpha}^{-k(i)}(X,R)\hat{\beta}^{-k(i)}(X,R)$.
- Repeat across all folds and report
$$\hat{\theta} = \frac{1}{K}\sum_{k=1}^{K} \hat{\theta}^k_w(\hat{\tau}^k, \hat{\eta}^k),$$
and construct a confidence interval of level $1-\alpha$, $\hat{C}(\alpha)$ as in Equation (B).

Remark 6 (Direct effects on R)**.** Under Assumption 3.1 we should replace $M^{\bar{d}}_{1,i}$ with

$$M_{1,i}(\tau,\eta) = \left[M^{\bar{d}=0}_{1,i}(\tau,\eta,d=0,x_1), \cdots, M^{\bar{d}=0}_{1,i}(\tau,\eta,d=0,x_{|\mathcal{X}|}), M^{\bar{d}=1}_{1,i}(\tau,\eta,d=1,x_1), \cdots, M^{\bar{d}=1}_{1,i}(\tau,\eta,d=1,x_{|\mathcal{X}|})\right]$$

and in addition impose that $\sum_i M^d_{0,i}(\hat{\eta}_H) = 0$ for $d \in \{0,1\}$. □

A.2 Continuous Outcomes

Next, we show how our results directly extend to continuous outcomes with finite support.

Assumption 4 (Continuous outcome and common support)**.** Suppose that $Y(d) \in [-U,U]^p, U < \infty$ and $Y(d)|X,S,R$ have a positive density over $[-U,U]^p$ almost surely for $d \in \{0,1\}$.

Assumption 4 allows for continuous outcomes, assuming common support between the experiment and observational study.

In the presence of continuous outcomes, we construct bins over the support of Y, such that $B_y(\varepsilon), y \in \mathcal{Y}_\varepsilon$ defines an l_∞-ball of radius ε around value y. Therefore, $\mathcal{Y}_\varepsilon$ defines an ε-cover of $[-U,U]^p$. Specifically, for any two values $x, x' \in B_y(\varepsilon), ||x - x'||_\infty \leq \varepsilon$.

Theorem A.1 (Identification with continuous outcomes)**.** *Consider continuous outcomes with $B_y(\varepsilon), y \in \mathcal{Y}_\varepsilon = \{y_1, \cdots, y_{|\mathcal{Y}_\varepsilon|}\}$ as defined in this section, for $\varepsilon > 0$. Suppose Assumptions 1, 2, 3 hold with $Y = y$ replaced by $Y \in B_y(\varepsilon), y \in \mathcal{Y}_\varepsilon$. Then, for $d \in \{0,1\}$*

$$\pi_d(X,R) = \int \gamma'_{y,\bar{d}}(X,R)\mathbb{P}(Y(d) \in B_y(\varepsilon) \mid S = e, X)dy,$$

where $\gamma'_{y,d}(X,T) = \mathbb{P}(Y \in B_y(\varepsilon), S = o, D = d|T, X)/\mathbb{P}(Y \in B_y(\varepsilon), S = o, D = d|X)$. *Here* $\bar{d} = d$ *if Assumption 3.1 holds and* $\bar{d} = 0$ *if Assumption 3.2 holds.*

Proof. The proof follows from Lemma 3.1 and Bayes theorem in the same way as Theorem 3.1, after replacing $Y = y$ events with $Y \in B_y(\varepsilon)$ events. The positive density assumption and $\varepsilon > 0$ guarantees that $Y|X, S, R \in B_y(\varepsilon)$ has a positive probability. □

Following verbatim Section 4, it follows that for fixed $\varepsilon > 0$, $\hat{\theta}_w$ consistently estimate

$$\tilde{\theta}_w(\varepsilon) = \int_{\mathcal{Y}_\varepsilon} w(y) \left(\mathbb{P}\left(Y(1) \in B_y(\varepsilon) \middle| S = e \right) - \mathbb{P}\left(Y(0) \in B_y(\varepsilon) \middle| S = e \right) \right) dy.$$

Although valid inference is guaranteed on the estimand $\tilde{\theta}_w(\varepsilon)$ from Section 4 directly, this may be a biased version of a target estimand

$$\theta_w = \int_{[-U,U]^p} w(y) \left(\mathbb{P}\left(Y(1) \in B_y(\varepsilon) \middle| S = e \right) - \mathbb{P}\left(Y(0) \in B_y(\varepsilon) \middle| S = e \right) \right) dy,$$

where now we integrate over the full support of the outcome.

Typically, we expect that these two estimands are close to each other for low-dimensional p. In this case, inference can be directly corrected, taking into account worst-case biases which is governed by ε. We illustrate this for the average effect below.

Theorem A.2. *Suppose that* $p = 1$ *and* $w_y = y$. *Then* $|\theta_w - \tilde{\theta}_w(\varepsilon)| \leq \varepsilon$.

Proof. See Appendix C.4. □

Theorem A.2 extends our results to continuous outcomes showing that discrete approximation can be directly adjusted in the construction of confidence intervals. Theorem A.2 fixes $\varepsilon > 0$ (instead of letting it converge to zero), to guarantee that $|\mathcal{Y}_\varepsilon| < \infty$. Inference is valid for any choice of $\varepsilon > 0$. Extensions for $\varepsilon \to 0$ may follow similarly and are left to future research.

B Estimation with Complex Representations

In this section we turn to estimation with binary outcomes, but allow the representation $\hat{H}^{-k}$ not to satisfy Equation (15).

In particular, Equation (15) fails if researchers pose a complex function class for $H(\cdot)$. In this case, researchers should use cross-fitting (e.g. Chernozhukov et al., 2018): we divide the sample into K folds, and, for each fold k, we estimate $\hat{H}^{-k}(R)$ using all units except those in fold k. We construct for fold k

$$g_n^k(\theta, \hat{H}^{-k}) := \sum_{i \in \mathcal{S}_k} \Big(\Delta_i(e) - \theta \Delta_i(o) \Big) \hat{H}^{-k}(R), \quad \mathbb{E}[g_n^k(\theta, \hat{H})|R] = 0$$

where $\mathcal{S}_k$ denotes the set of indexes assigned to fold k and where the second equality follows immediately from Assumption 1 (observations are independent) and Theorem 3.2. Cross-fitting here allows us to estimate $H(\cdot)$ using *arbitrary* machine learning estimators, without imposing *any* assumption on their convergence rates or probability limit. We estimate

$$\hat{\theta}^k : g_n^k(\hat{\theta}^k, \hat{H}^{-k}) = 0, \quad k \in \{1, \cdots, K\}.$$

Under independence, standard moment and full rank conditions given $\hat{H}^{-k}$ for the central limit to apply (Newey and McFadden, 1994), it follows that for any fold k

$$\frac{1}{\sqrt{|\mathcal{S}_k|}} \sum_{i \in \mathcal{S}_k} \begin{bmatrix} \Delta_i(o)\hat{H}^{-k}(R_i) - \mathbb{E}[\Delta_i(o)\hat{H}^{-k}(R)|\hat{H}^k] \\ \Delta_i(e)\hat{H}^{-k}(R_i) - \mathbb{E}[\Delta_i(e)\hat{H}^{-k}(R)|\hat{H}^k] \end{bmatrix} \to_d \mathcal{N}(0, \mathbb{V}_k(\hat{H}^k)).$$

Here, the central limit theorem is a direct consequence that $\hat{H}^{-k}$ is estimated out-of-sample. We can therefore apply the delta method or non parametric bootstrap (conditional on the estimated $\hat{H}^{-k}$) to consistently estimate the standard error of $\hat{\theta}^k$, which we define as $\hat{v}^k$.

An asymptotically valid confidence interval for θ with size $1-\alpha$ takes the form

$$\hat{C}^k(\alpha) = \left[\hat{\theta}^k - \Phi^{-1}(1-\alpha/2)\hat{v}^k, \hat{\theta}^k + \Phi^{-1}(1-\alpha/2)\hat{v}^k\right].$$

Researchers can therefore combine information across the K folds and report as point estimate and confidence interval with size $1-\alpha$

$$\hat{\theta} = \frac{1}{K}\sum_{k=1}^{K} \hat{\theta}^k, \quad \hat{C}(\alpha) = \bigcap_{k=1}^{K} \hat{C}^k(\alpha/K).$$

Because each $\hat{\theta}^k$ is unbiased for θ, also $\hat{\theta}^k$ is unbiased for θ. Validity of $\hat{C}(\alpha)$ follows immediately by the union bound which only leverages the marginal (but does not require conditions on the joint) asymptotic distribution of $\hat{\theta}^1, \cdots, \hat{\theta}^K$.[9] The complete algorithm is in Algorithm 2.

C Proofs of Main Results

C.1 Proof of Lemma 3.1

We can write

$$\mathbb{P}(R \mid D=d, S=e, X) = \int \mathbb{P}(R \mid Y=y, D=d, S=e, X)\mathbb{P}(Y=y \mid D=d, S=e, X)dy$$

Using Assumption 2 and Assumption 3.1, we can write $\mathbb{P}(R \mid Y=y, S=e, D=d, X) = \mathbb{P}(R \mid Y=y, S=o, D=d, X)$ for $d \in \{0,1\}$. Using Assumption 2 and Assumption 3.2 it follows that $(S, D) \perp R|X, Y$. Therefore, we can write $\mathbb{P}(R \mid Y=y, S=e, D=d, X) = \mathbb{P}(R \mid Y=y, S=o, D=d', X)$, for any $d, d' \in \{0,1\}$. Finally, under Assumption 1.2, $\mathbb{P}(Y=y \mid D=d, S=e, X) = \mathbb{P}(Y(d)=y \mid S=e, X)$.

[9] This is because $\mathbb{P}(\theta \notin \bigcap_{k=1}^{K} \hat{C}_k(\alpha/K)\Big) \leq \sum_{k=1}^{K} \mathbb{P}\Big(\theta \notin \hat{C}_k(\alpha/K)\Big) \to \alpha$.

C.2 Proof of Theorem 2.1

Validity of the surrogate expression under surrogate assumptions is immediate from previous work. We demonstrate that the surrogate expression may give the incorrect sign *under RSV assumptions*. As before, let us set aside covariates.

By contraction, the surrogate assumptions imply $D, S \perp Y|R$ and the RSV assumptions imply $D, S \perp R|Y$.

Since this argument is about existence, we can place the additional structure. Assume $S \perp R, Y|D$, which by weak union implies $S \perp Y|R, D$. Also assume Y is binary. Finally assume $S = o$ implies $D = 0$. This does not violate the RSV assumptions $S \perp R|Y, D$ and $D \perp R|Y$. In particular, $S \perp R, Y|D$ implies $S \perp R|Y, D$ by weak union; this extra stucture is a strengthening of our main assumption.

1. The surrogate expression has bias.

   Using randomization of the experiment, write the causal expression as

$$\begin{aligned}
\mathbb{E}\{Y(1)|S = e\} &= \mathbb{P}\{Y(1) = 1|S = e\} \\
&= \mathbb{P}\{Y(1) = 1|D = 1, S = e\} \\
&= \mathbb{P}(Y = 1|D = 1, S = e) \\
&= \int \mathbb{P}(Y = 1, R = r|D = 1, S = e)\mathrm{d}r \\
&= \int \mathbb{P}(Y = 1|R = r, D = 1, S = e)\mathbb{P}\{R = r|D = 1, S = e\}\mathrm{d}r.
\end{aligned}$$

   We relate it to the surrogate expression

$$\widetilde{\theta}(1) = \int \mathbb{P}(Y = 1|R = r, S = o)\mathbb{P}(R = r|D = 1, S = e)\mathrm{d}r.$$

   By Bayes' rule,

$$\mathbb{P}(Y|R, S) = \frac{\mathbb{P}(R|Y, S)\mathbb{P}(Y|S)}{\mathbb{P}(R|S)}, \quad \mathbb{P}(R|Y, D) = \frac{\mathbb{P}(Y|R, D)\mathbb{P}(R|D)}{\mathbb{P}(Y|D)}.$$

Consider the former factor of $\widetilde{\theta}(1)$. We have

$$\begin{aligned}
\mathbb{P}(Y=1|R,S=o) &= \frac{\mathbb{P}(R|Y=1,S=o)\mathbb{P}(Y=1|S=o)}{\mathbb{P}(R|S=o)} \\
&= \frac{\mathbb{P}(R|Y=1,D=1)\mathbb{P}\{Y(0)=1|S=o\}}{\mathbb{P}(R|S=o)} \\
&= \frac{\mathbb{P}(Y=1|R,D=1)\mathbb{P}(R|D=1)\mathbb{P}\{Y(0)=1|S=o\}}{\mathbb{P}(Y=1|D=1)\mathbb{P}(R|S=o)} \\
&= \frac{\mathbb{P}(Y=1|R,D=1,S=e)\mathbb{P}(R|D=1)\mathbb{P}\{Y(0)=1|S=o\}}{\mathbb{P}(Y=1|D=1,S=e)\mathbb{P}(R|S=o)} \\
&= \frac{\mathbb{P}(Y=1|R,D=1,S=e)\mathbb{P}(R|D=1)\mathbb{P}\{Y(0)=1|S=e\}}{\mathbb{P}\{Y(1)=1|S=e\}\mathbb{P}(R|D=0)} \\
&= \frac{\mathbb{P}\{Y(0)=1|S=e\}}{\mathbb{P}\{Y(1)=1|S=e\}}\mathbb{P}(Y=1|R,D=1,S=e)\frac{\mathbb{P}(R|D=1)}{\mathbb{P}(R|D=0)}.
\end{aligned}$$

The first line uses Bayes' rule.

The second uses the RSV assumptions to write $\mathbb{P}(R|Y=1,S=o) = \mathbb{P}(R|Y=1,D=1)$. Also $S=o$ implies $D=0$, so $\mathbb{P}(Y=1|S=o) = \mathbb{P}\{Y(0)=1|S=o\}$.

The third line uses Bayes' rule again.

The fourth line uses the additional structure $S \perp R, Y|D$, as well as its implication $S \perp Y|R,D$ to write $\mathbb{P}(Y=1|R,D=1) = \mathbb{P}(Y=1|R,D=1,S=e)$ and $\mathbb{P}(Y=1|D=1) = \mathbb{P}(Y=1|D=1,S=e)$.

The fifth line makes three substitutions. In the numerator, it uses

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

because $S=o$ implies $D=0$, we have the extra structure $S \perp Y|D$ implies $Y(0) \perp S|D=0$, and finally we have randomization of the experiment. In the denominator, it uses $\mathbb{P}(Y=1|D=1,S=e) = \mathbb{P}\{Y(1)=1|S=e\}$, due to randomization of the experiment as argued above. Finally, it uses

$$\mathbb{P}(R|S=o) = \mathbb{P}(R|D=0,S=o) = \mathbb{P}(R|D=0)$$

because $S=o$ implies $D=0$, and we have the extra structure $S \perp R|D$.

The sixth line rearranges.

In summary, $\widetilde{\theta}(1)$ equals

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

whereas $\mathbb{E}\{Y(1)|S=e\}$ equals

$$\int \mathbb{P}(Y=1|R=r,D=1,S=e)\mathbb{P}\{R=r|D=1,S=e\}\mathrm{d}r.$$

We conclude that $\text{BIAS} = \widetilde{\theta}(1) - \mathbb{E}\{Y(1)|S=e\}$ equals

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

Finally, we rearrange the final factors as

$$\begin{aligned}\mathbb{P}(Y=1|R,D=1,S=e)\mathbb{P}(R|D=1,S=e) &= \mathbb{P}(Y=1,R|D=1,S=e)\\ &= \mathbb{P}(R|Y=1,D=1,S=e)\mathbb{P}(Y=1|D=1,S=e)\\ &= \mathbb{P}(R|Y=1,D=1,S=e)\mathbb{P}\{Y(1)=1|S=e\}\end{aligned}$$

using randomization of the experiment. In summary, BIAS equals

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

2. The bias can be positive or negative. We construct illustrative DGPs.

   Suppose that R is binary with

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

   This satisfies the extra structure $S \perp R, Y|D$ as well as the RSV condition $D \perp R|Y$.

   In this DGP

$$\begin{aligned}\mathbb{P}(R=1|Y=1,D=1,S=e) &= 1\\ \mathbb{P}(R=0|Y=1,D=1,S=e) &= 0.\end{aligned}$$

   Therefore the bias simplifies to

$$\text{BIAS} = \left[\frac{\mathbb{P}\{Y(0)=1|S=e\}}{\mathbb{P}\{Y(1)=1|S=e\}}\frac{\mathbb{P}(R=1|D=1)}{\mathbb{P}(R=1|D=0)} - 1\right]\mathbb{P}\{Y(1)=1|S=e\}.$$

   In this DGP,

$$\begin{aligned}\mathbb{P}(R=1|Y=1,D=1,S=e) &= 1\\ \mathbb{P}(R=1|Y=0,D=1,S=e) &= \frac{1}{2}.\end{aligned}$$

Hence by similar arguments to those above,

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

using the extra structure $S \perp R|D$, the specific DGP, and experiment randomization. Therefore

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

and hence

$$\text{BIAS} = \left[\frac{\mathbb{P}\{Y(0)=1|S=e\}}{\mathbb{P}\{Y(1)=1|S=e\}}\cdot\frac{1+\mathbb{P}\{Y(1)=1|S=e\}}{1+\mathbb{P}\{Y(0)=1|S=e\}}-1\right]\mathbb{P}\{Y(1)=1|S=e\}.$$

Lightening notation,

$$\text{BIAS} = \left(\frac{a}{b}\cdot\frac{1+b}{1+a}-1\right)b = \frac{a-b}{a+1}.$$

Thus the sign of the bias is positive when $\mathbb{P}\{Y(0)=1|S=e\} > \mathbb{P}\{Y(1)=1|S=e\}$ and negative otherwise.

C.3 Proof of Theorem 3.1

Consider proving the first result. By Bayes rule and Lemma 3.1,

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

Finally cancel the $\mathbb{P}(R|X)$. Consider proving the second. Then because $(D,S) \perp R|Y,X$

$$\begin{aligned}
&\frac{\mathbb{P}(D=d,S=e|R,X)}{\mathbb{P}(D=d,S=e|X)}\mathbb{P}(R|X) \\
&= \mathbb{P}(R|D=d,S=e,X) \\
&= \sum_{y\in\mathcal{Y}} \mathbb{P}(R=r \mid Y=y,S=o,D=0,X)\mathbb{P}\{Y(d)=y \mid S=e,X\} \\
&= \sum_{y\in\mathcal{Y}} \frac{\mathbb{P}(Y=y,S=o,D=0|X,R)\mathbb{P}(R|X)}{\mathbb{P}(Y=y,S=o,D=0|X)}\mathbb{P}\{Y(d)=y \mid S=e,X\}.
\end{aligned}$$

C.4 Proof of Theorem A.2

Define $f_{Y(d)|X,S}(y)$ denotes the conditional density of $Y(d)$ given X,S. Define $\tilde{w}_y$ a function of $y \in [-U,U]$ such that $\tilde{w}_y = \arg\min_{y'\in\mathcal{Y}_\varepsilon} |y-y'|$. Then we can write

$$\begin{aligned}
|\theta_w - \tilde{\theta}_{\tilde{w}}| &= \Big|\frac{1}{n_e}\sum_{i:S=e}\int \Big(\tilde{w}_y - w_y\Big)(f_{Y(1)|X,S=e}(y) - f_{Y(0)|X,S=e}(y))dy\Big| \\
&\leq \max_{d\in\{0,1\}} 2\Big|\frac{1}{n_e}\sum_{i:S=e}\int \Big(\tilde{w}_y - w_y\Big) f_{Y(d)|X,S=e}(y)dy\Big| \\
&\leq \max_{d\in\{0,1\}} \frac{2}{n_e}\sum_{i:S=e}\Big|\int \Big(\tilde{w}_y - w_y\Big) f_{Y(d)|X,S=e}(y)dy\Big| \\
&\leq \max_{d\in\{0,1\}} \frac{2}{n_e}\sum_{i:S=e}\int |f_{Y(d)|X,S=e}(y)|dy \times \max_{y\in\mathcal{Y}_\varepsilon}\max_{y',y''\in B_y(\varepsilon)} ||y''-y'||_\infty \leq \varepsilon,
\end{aligned}$$

where the last inequality follows from Holder's inequality and construction of the ε-cover.

C.5 Proof of Corollary 3.1

We prove the result for a scalar valued function $h_d(X,R) \in \mathbb{R}$; it therefore holds for the vector valued function $H_d(X,R)$.

Note that $\mathbb{P}\{Y(d)=y|X,S=e\}$ is a function of (X,d,y) only but not of S (since it conditions on the event $S=e$). Using Theorem 3.1, for any function $h_d(X,R)$

$$\begin{aligned}
\mathbb{E}[\pi_d(X,R)h_d(X,R)|X] &= \mathbb{E}\left[h_d(X,R)\sum_{y\in\mathcal{Y}}\gamma_{y,d}(X,R)\mathbb{P}\{Y(d)=y|X,S=e\}|X\right] \\
&= \mathbb{E}\left[h_d(X,R)\sum_{y\in\mathcal{Y}}\gamma_y(X,R)|X\right]\mathbb{P}\{Y(d)=y|X,S=e\} \\
&= \mathbb{E}\left[h_d(X,R)\sum_{y\in\mathcal{Y}}\gamma_{y,d}(X,R)|X\right]\tau^\star_{d,y}(X)
\end{aligned}$$

where the penultimate equality follows from the fact that $\mathbb{P}\{Y(d)=y|X,S=e\}$ is deterministic conditional on X.

For the final step, consider the left hand side; the right hand side is analogous. We wish to show

$$\mathbb{E}\left\{h_d(X,R)\frac{\mathbb{P}(D=d,S=e|R,X)}{\mathbb{P}(D=d,S=e|X)}|X\right\} = \mathbb{E}\left\{h_d(X,R)\frac{1(D=d,S=e)}{\mathbb{P}(D=d,S=e|X)}|X\right\}.$$

This holds by the law of iterated expectations since, for any function f and any random variable Q,

$$\mathbb{E}\{f(R,X)Q|X\} = \mathbb{E}[\mathbb{E}\{f(R,X)Q|R,X\}|X] = \mathbb{E}\{f(R,X)\mathbb{E}(Q|R,X)|X\}.$$

Algorithm 2 Two-step estimation procedure with binary outcome and complex ML repreentations

Require: Observations $\Big(1\{S=o\}Y, S, D, X, R\Big)_{i=1}^{n}$.

1: Divide the sample into K equally sized folds and define $k(i)$ the fold containing unit i, define $\mathcal{S}_k$ the set of units in fold k

2: **for** k in $\{1, \cdots, K\}$ **do**

3: Estimate $\mathbb{P}(Y=1|R, S=o), \mathbb{P}(S=o|R)$ and $\mathbb{P}(D=1|R, S=e)$ using arbitrary machine learning estimators for all units i except those in fold k Denote such estimators as $\hat{P}_Y(R), \hat{P}_{S=o}(R), \hat{P}_D(R)$.

4: Denote $\hat{p}_D(1) = \mathbb{P}_n(D=1, S=e), \hat{p}_D(0) = \mathbb{P}_n^{-k}(D=0, S=e)$, $\hat{p}_Y(1) = \mathbb{P}_n^{-k}(Y=1, S=o), \hat{p}_Y(0) = \mathbb{P}_n^{-k}(Y=0, S=o)$ where $\mathbb{P}_n^{-k}(x)$ denotes the empirical probability of event x of all units i except those in fold k

5: Define

$$\hat{Q}_D(R) = \Big(1-\hat{P}_{S=o}(R)\Big)\Big(\frac{\hat{P}_D(R)}{\hat{p}_D(1)} - \frac{1-\hat{P}_D(R)}{\hat{p}_D(0)}\Big), \quad \hat{Q}_Y(R) = \hat{P}_{S=o}(R)\Big(\frac{\hat{P}_Y(R)}{\hat{p}_Y(1)} - \frac{1-\hat{P}_Y(R)}{\hat{p}_Y(0)}\Big)$$

6: For all units i

$$\hat{\theta}^{\text{first step}} = \arg\min_{\theta} \sum_{i \notin \mathcal{S}_k} \Big(\hat{Q}_D(R_i) - \hat{Q}_Y(R_i)\theta\Big)^2.$$

7: Construct the function

$$r \mapsto \hat{H}^{-k}(r) = \frac{\hat{Q}_Y(r)}{\hat{\sigma}^2(r)},$$

$$\hat{\sigma}^2(r) = \Big(1-\hat{P}_{S=o}(R)\Big)\Big(\frac{\hat{P}_D(r)}{\hat{p}_D(1)^2} + \frac{1-\hat{P}_D(r)}{\hat{p}_D(0)^2}\Big) + \hat{P}_{S=o}(R)(\hat{\theta}^{\text{first step}})^2\Big(\frac{\hat{P}_Y(r)}{\hat{p}_Y(1)^2} + \frac{1-\hat{P}_Y(r)}{\hat{p}_Y(0)^2}\Big)$$

8: Estimate $\hat{\Delta}_i^k(e), \hat{\Delta}_i^k(o)$ for all units i in fold k

$$\hat{\Delta}_i(o) = \Big(\frac{1\{Y_i=1, S_i=o\}}{\mathbb{P}_n^{k(i)}(Y=1, S=o)} - \frac{1\{Y_i=0, S_i=o\}}{\mathbb{P}_n^{k(i)}(Y=0, S=o)}\Big)$$

$$\hat{\Delta}_i(e) = \Big(\frac{1\{D_i=1, S_i=e\}}{\mathbb{P}_n^{k(i)}(D=1, S=e)} - \frac{1\{D_i=0, S_i=e\}}{\mathbb{P}_n^{k(i)}(D=0, S=e)}\Big)$$

where $k(i)$ denotes the fold of unit i

9: Estimate

$$\hat{\theta}^k = \frac{\sum_{i \in \mathcal{S}_k} \hat{\Delta}_i(e) H^{-k}(R)}{\sum_{i \in \mathcal{S}_k} \hat{\Delta}_i(o) H^{-k}(R)}.$$

10: Estimate the standard error of $\hat{\theta}^k$ either through the Delta method or by Bootstrapping units $i \in \mathcal{S}_k$ (while fixing the function H^{-k}) and computing $\hat{\theta}^k$ across each different bootstrap iteration. Denote such estimated standard error as $\hat{v}^k$

11: Compute a α-level confidence interval as $\hat{C}^k(\alpha)$ as

$$\hat{C}^k(\alpha) = \Big[\hat{\theta}^k - \Phi^{-1}(1-\alpha/2)\hat{v}^k, \hat{\theta}^k + \Phi^{-1}(1-\alpha/2)\hat{v}^k\Big]$$

12: **end for**

13: **return** a point estimate and α level confidence interval as $\hat{\theta} = \frac{1}{K}\sum_{k=1}^{K}\hat{\theta}^k$, $\hat{C}(\alpha) = \bigcap_{k=1}^{K}\hat{C}^k(\alpha/K)$.