Estimating long-term treatment effects without long-term outcome data

David Rhys Bernard

Global Priorities Institute | September 2020

GPI Working Paper No. 11-2020

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

Estimating long-term treatment effects without long-term outcome data

David Rhys Bernard
Paris School of Economics
david.rhys.bernard@gmail.com

September 30, 2020

Abstract

Estimating long-term impacts of actions is important in many areas but the key difficulty is that long-term outcomes are only observed with a long delay. One alternative approach is to measure the effect on an intermediate outcome or a *statistical surrogate* and then use this to estimate the long-term effect. Athey et al. (2019) generalise the surrogacy method to work with multiple surrogates, rather than just one, increasing its credibility in social science contexts. I empirically test the multiple surrogates approach for long-term effect estimation in real-world conditions using long-run RCTs from development economics. In the context of conditional cash transfers for education in Colombia, I find that the method works well for predicting treatment effects over a 5-year time span but poorly over 10 years due to a reduced set of variables available when attempting to predict effects further into the future. The method is sensitive to observing appropriate surrogates.

1 Introduction

Many fields seek to estimate the long-term effects of a treatment or policy. In medicine one may want to estimate the effect of a surgery on life expectancy, or in economics the effect of a conditional cash transfer during childhood on adult income. One way to measure these effects would be to run a randomised controlled trial (RCT) and then wait to observe the long-run outcomes. However, typically the results would be observed too late to be relevant for the policy decision today.

One approach developed in medicine to deal with this problem is to study the effects on an intermediate outcome or a *surrogate* outcome. One can then combine results on the effect of the treatment on the surrogate and the relationship between the surrogate and the long-term outcome to estimate the effect of the treatment on the long-term outcome. For example, one could measure the effect of a surgery on the size of a tumour and the relationship between tumour size and mortality rates and use this to calculate the effect of surgery on life expectancy. To combine results in this way we must make an assumption often known as the Prentice criterion, namely that the treatment and the long-term outcome are independent, conditional on the surrogate (Prentice, 1989). In the

previous example, the size of the tumour could be a surrogate for life expectancy if life expectancy is independent of the surgery conditional on the size of the tumour.

Surrogates for long-run effect estimation are often used both formally and informally in medicine, however their use in economics is minimal, despite significant interest in the long-run effects of a variety of programmes and policies (Bouguen et al., 2018). This is likely because the surrogacy assumption is hard to justify in a social science context and there are multiple ways it could be violated. Freedman et al. (1992) show that conditional independence requires that the surrogate mediates the full effect of the treatment on the long-term outcome and if it does not, the surrogate is not valid. Others have shown that even under full mediation, if there is unobserved confounding between the surrogate and the long-term outcome then the surrogacy assumption is also invalid (VanderWeele, 2015).

Due to these issues, Athey et al. (2019) develop surrogacy methods which rely on many surrogate variables instead of just one. The idea behind this is that even though any individual variable may not be a valid surrogate, collectively they are more likely to satisfy the surrogacy assumption. They combine many short-term outcomes into a "surrogate index" which is the expected value of the long-term outcome conditional on the short-term outcomes. They show that under the assumption that the long-term outcome is independent of treatment conditional on the surrogate index, the average treatment effect on the surrogate index is the same as the average treatment effect on the long-term outcome. Based on this they develop different estimators for long-term effects when you do not observe the long-term outcome. I test these surrogacy estimators with real world data from long-run RCTs.

RCTs started to increase in popularity in development economics in the late 1990s (Banerjee et al., 2016). Recently, researchers have started to use the exogenous variation generated by these early experiments to study the effects of programs such as conditional cash transfers on long-term outcomes, such as high school graduation rates and adult income twenty years later. Bouguen et al. (2018) summarise the results of 14 different long-run development RCTs. This provides a laboratory to assess the performance of these surrogacy estimators for long-term outcomes against the unbiased benchmark of the experimental estimate.

The main strategy in this paper is to analyse an experimental dataset in two ways. First, get an unbiased estimate of the standard experimental average treatment effect by regressing long-term outcomes on treatment status. Then, manipulate the data (for example pretending we do not observe the long-term outcomes in one treatment arm) and reanalyse the data using the surrogacy approach. If the estimate from the surrogacy approach is close to the unbiased estimate from the experimental approach then this means the surrogacy approach works well. The further away the two estimates are, the poorer the performance of the surrogacy approach.

I use data from two RCTs, both from Barrera-Osorio et al. (2019) who study the effect of conditional cash transfers on medium- and long-term educational outcomes. I test many different

implementations of the surrogacy approach, varying which sample is used as the observational dataset. In these RCTs, I find that both surrogacy approaches work well when the full set of surrogates is used and the long-term effect is 4-5 years in the future. However, the performance of the method is very sensitive to the set of surrogates used and if key surrogates are missing, for instance because we are trying to predict effects further into the future the method performs poorly.

Athey et al. (2019) show that the surrogacy method works well for estimating long-term (9 years) effects on employment of a job-training program. My results show that it is more difficult to predict long-term impacts of human capital interventions from their short-term impacts. Generating more evidence like this is key to understanding when we can reliably estimate long-term effects which are critically important in many domains.

The paper proceeds as follows. Section 2 summarises the econometric theory from Athey et al. (2016, 2019). Section 3 describes in more detail the data I use from Barrera-Osorio et al. (2019). Section 4 describes the results from my different implementations of the surrogacy approach and robustness checks, while section 5 concludes.

2 Theory

Athey et al. (2016, 2019) introduce the surrogate index method for estimating long-term effects when we do not observe the long-term outcomes of an experiment.

This requires two samples: an experimental sample, with N_E units and an observational sample with N_O units. In the experimental sample, we are interested in the impact of a binary treatment $T_{E,i} \in \{0, 1\}$ on the long-term outcome $Y_{E,i}$. The problem is that we do not observe $Y_{E,i}$ in the experimental sample.[1] However, we do observe many intermediate outcomes or surrogates, $S_{E,i}$. Additionally we observe pre-treatment covariates that are unaffected by treatment, $X_{E,i}$.

In the observational sample we observe the same pre-treatment covariates and surrogates, as well as the long-term outcome, $(X_{O,i}, S_{O,i}, Y_{O,i})$. The observational units do not have to be exposed to any treatment and we do not need to know their treatment status. The data requirements are shown in figure 1.

We follow the potential outcomes framework and are interested in the effect of the treatment on the outcome $Y^1_{E,i} - Y^0_{E,i}$. As we cannot observe both potential outcomes for any given individual, we focus on the average treatment effect across the sample, $\tau_Y = E(Y^1_{E,i} - Y^0_{E,i})$. Note that the surrogates also have two potential outcomes, $S^1_{E,i}$ and $S^0_{E,i}$, and we can similarly define $\tau_S = E(S^1_{E,i} - S^0_{E,i})$.

We define the propensity score as the conditional probability of receiving treatment and make the following standard ignorability assumption.

[1] Note that $Y_{E,i}$ could also be a contemporaneous outcome that is unobserved (possibly because it is costly to measure) in the experimental dataset but is observed in the observational dataset. I focus on the case where $Y_{E,i}$ is unobserved because it occurs in the future, but the analysis could also be done for contemporaneous outcomes.

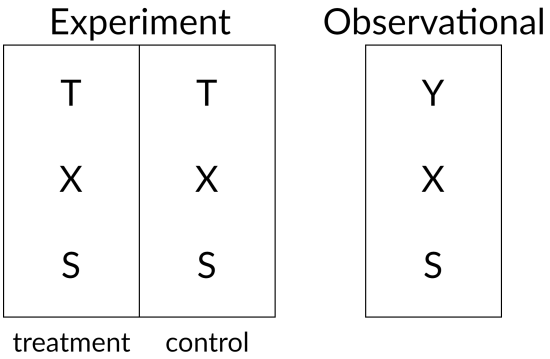


Figure 1: Data required for surrogacy approach

Definition 1. Propensity score
$e(x) = Pr(T_{E,i} = 1 \mid X_{E,i} = x)$

Assumption 1. Ignorable treatment assignment
(i) $T_{E,i} \perp\!\!\!\perp (Y^0_{E,i}, Y^1_{E,i}, S^0_{E,i}, S^1_{E,i}) \mid X_{E,i}$
(ii) $0 \leq Pr(T_{E,i} = 1 \mid X_{E,i} = x) \leq 1$ for all $x \in \mathbb{X}_E$

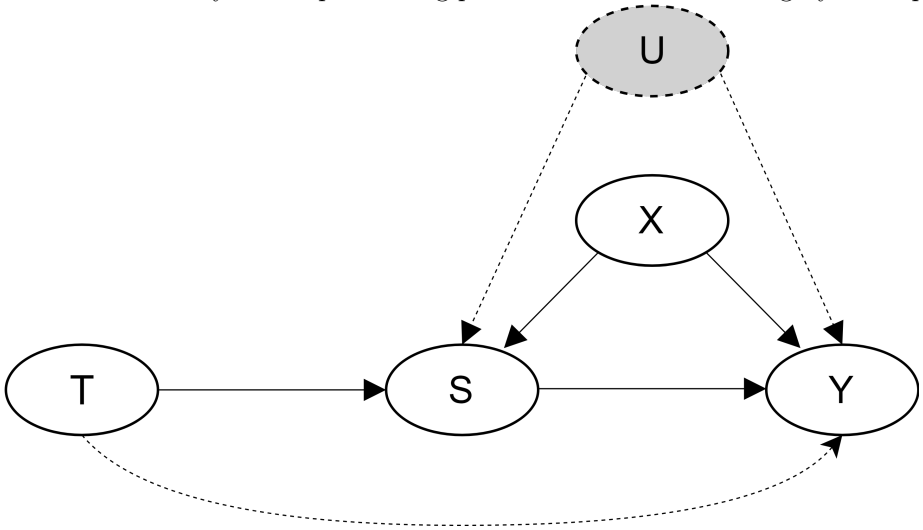
As we use RCTs for the analysis, this assumption is true by design as randomisation ensures the independence of treatment status and potential outcomes. If we did observe the long-term outcome $Y_{E,i}$ in the experimental dataset, this assumption would be sufficient for identifying the ATE on the long-term outcome. However, as we do not observe $Y_{E,i}$ in the experimental sample we must rely on the surrogates and make further assumptions.

Assumption 2. Surrogacy
$T_{E,i} \perp\!\!\!\perp Y_{E,i} \mid S_{E,i}, X_{E,i}$

This assumption states that once we condition on the surrogates and the baseline covariates, the treatment and the long-term outcome are independent. This assumption is represented with a directed acyclic graph (DAG) in figure 2. We can see that this assumption requires two things. Firstly, that there is no direct effect of the treatment on the long-term outcome that is not mediated through the surrogates. In DAG terms, there must be no arrow $T \to Y$. Secondly, it requires that there are no unobserved confounders of the surrogates and long-term outcome. In the DAG this is represented as there being no unobserved mediator-outcome confounders U, outside the set of observed baseline characteristics X.

This assumption is recognised as critical but previous work has only allowed for one surrogate. If either (1) the one surrogate does not mediate the full effect of the treatment on the outcome or (2) there are confounders between the surrogate and the outcome then the surrogacy assumption is

Figure 2: Directed Acyclic Graph showing potential violations of surrogacy assumption



Notes: Observed confounders represented by X, unobserved confounders represented by U. Treatment is T, surrogates are S and the long-term outcome is Y.

invalid (Prentice, 1989; Frangakis and Rubin, 2002; Joffe and Greene, 2009; VanderWeele, 2015). In social science contexts, it is highly unlikely that one variable will ever fully mediate the effect of a treatment on an outcome. By moving to a method which allows for multiple surrogates we make it more likely that the surrogates jointly mediate the full effect of treatment. Furthermore, we can control for observed confounders reducing the risk of the surrogacy assumption being violated by surrogate-outcome confounding. To motivate the multiple surrogates approach and the necessary surrogacy assumption, Athey et al. (2016, p 9.) write:

> We view it as similar in spirit to the unconfoundedness assumption. It is unlikely to be satisfied exactly in any particular application, but, especially in cases with a large number of intermediate variables as well as pretreatment variables, it may be a reasonable approximation... Moreover, there is often no reasonable alternative. From our perspective it is useful to view the problem of identifying and estimating $\tau = \mathbb{E}_E[Y^1_{E,i} - Y^0_{E,i}]$ as a missing data one. The outcome $Y_{E,i}$ is missing for all units in the experimental sample, and any estimator of the treatment effect τ ultimately relies on imputing these missing outcomes.

We can view this paper as testing in which applications and contexts, and with what variables we might expect the surrogacy assumption to be (approximately) satisfied.

To move from the one surrogate to the multiple surrogate case, Athey et al. (2016) introduce two new concepts. The first is the surrogate score, the conditional probability of having received

treatment given the baseline covariates and the surrogates. This differs from the propensity score as propensity scores are conditional only on pre-treatment variables whereas the surrogacy score is also conditional on post-treatment variables.

Definition 2. Surrogate score
$r(s,x) = Pr(T_{E,i} = 1 \mid S_{E,i} = s, X_{E,i} = x)$

The key property that the surrogate score shares with the propensity score is that it allows us to collapse high-dimensional surrogates and covariates into a one-dimensional scalar and use this in our estimators. This property relies on the proof from Rosenbaum and Rubin (1983) showing that if treatment assignment is ignorable given x, then it is ignorable given a balancing score $b(x)$. This implies:

Proposition 1. Surrogacy score. Under surrogacy (Assumption 2) we have
$T_{E,i} \perp\!\!\!\perp Y_{E,i} \mid r(S_{E,i}, X_{E,i})$

The question then arises, how can we use the observational sample to help estimate τ_Y in the absence of $Y_{E,i}$ in the experimental sample. The surrogacy assumption is necessary but not sufficient for this. We also need to assume that the samples are comparable. By this I mean that the conditional distribution of $Y_{E,i}$ given $(S_{E,i}, X_{E,i})$ is the same as the conditional distribution of $Y_{O,i}$ given $(S_{O,i}, X_{O,i})$. It must also be the case that the support of X and S ($\mathbb{X}$ and $\mathbb{S}$ respectively) in the experiment is contained within the support of X and S in the observational dataset.

Assumption 3. Comparability of samples
$Y_{E,i} \mid S_{E,i}, X_{E,i} \sim Y_{O,i} \mid S_{O,i}, X_{O,i}$
and $\mathbb{X}_E \in \mathbb{X}_O$, and $\mathbb{S}_E \in \mathbb{S}_O$.

The second key concept is the surrogate index, the conditional expectation of the primary outcome given the covariates and surrogates in the observational sample:

Definition 3. The surrogate index
$h_O(s,x) = E_O(Y_{O,i} \mid S_{O,i} = s, X_{O,i} = x)$
and $h_E(s,x) = E_E(Y_{E,i} \mid S_{E,i} = s, X_{E,i} = x)$

$h_E(.)$ is not estimable as we do not observe $Y_{E,i}$ in the experimental sample. However, we do observe $Y_{O,i}$ in the observational sample so $h_O(.)$ is estimable. We can define the same conditional expectation within each treatment arm of the experiment, so we have the conditional expectation of the primary outcome given pre-treatment covariates, surrogates and the treatment:

$\mu_E(s,x,t) = E_E(Y_{E,i} \mid S_{E,i} = s, X_{E,i} = x, T_{E,i} = t)$

It therefore follows that:

Proposition 2. Surrogate index

(i) Under surrogacy (assumption 2) we have

$\mu_E(s,x,0) = \mu_E(s,x,1) = h_E(s,x)$, for all $s \in \mathbb{S}_E$, $x \in \mathbb{X}_E$

(ii) Under comparability (assumption 3) we have

$h_E(s,x) = h_O(s,x)$ for $s \in \mathbb{S}_E$, and $x \in \mathbb{X}_E$

(i) says that if the surrogacy assumption is true, the conditional expectation of treated and untreated people within the experiment, conditional on surrogates and baseline covariates is the same. (ii) adds that if the comparability of samples assumption is also true, then the conditional expectation is the same in the experimental dataset and the observational dataset.

The results so far are sufficient for the first surrogacy estimator, the surrogate index, which is the one I focus on in the empirical section of the paper. A future extension for this work will be implementing and testing the other surrogacy estimators as well.

Athey et al. (2016) also further develop the theory and introduce one additional concept to propose an alternative estimator. We can define the sampling score, $t(s,x)$ as the probability of being in the experimental sample $P_i = E$ conditional on surrogates and baseline covariates. We make the assumption that the sampling score is less than one. This assumption is a purely technical one, necessary to make the estimators estimable as $(1 - t(s,x))$ appears in the denominator of one of the estimators.

Definition 4. Sampling score

$t(s,x) = Pr(P_i = E \mid S_i = s, X_i = x)$

Assumption 4. Sampling score

$t(s,x) < 1$ for all s, x

We have defined four conditional expectations, (1) the propensity score, (2) the surrogate index, (3) the surrogate score, and (4) the sampling score, which we can use for the two surrogacy estimators.

$$\text{Propensity score: } e(x) = Pr_E(T_{E,i} = 1 \mid X_{E,i} = x) \tag{1}$$

$$\text{Surrogate index: } h_O(s,x) = \mathbb{E}_O(Y_{O,i} \mid S_{O,i} = s, X_{O,i} = x) \tag{2}$$

$$\text{Surrogate score: } r(s,x) = Pr_E(T_{E,i} = 1 \mid S_{E,i} = s, X_{E,i} = x) \tag{3}$$

$$\text{Sampling score: } t(s,x) = Pr(P_i = E \mid S_i = s, X_i = x) \tag{4}$$

2.1 Estimator based on surrogate index

We can estimate the surrogate index as $\hat{h}_O(s,x)$. We can then naturally construct an estimator as:

$$\hat{\tau}^E = \frac{1}{\sum_{i=1}^{N_E} T_{E,i}/\hat{e}(X_{E,i})} \sum_{i=1}^{N_E} \hat{h}_O(S_{E,i}, X_{E,i}) \cdot \frac{T_{E,i}}{\hat{e}(X_{E,i})} - \frac{1}{\sum_{i=1}^{N_E}(1-T_{E,i})/(1-\hat{e}(X_{E,i}))} \sum_{i=1}^{N_E} \hat{h}_O(S_{E,i}, X_{E,i}) \cdot \frac{1-T_{E,i}}{1-\hat{e}(X_{E,i})} \quad (5)$$

This is the surrogate index estimator. It uses the observational data to estimate a model for the surrogate index $Y_O = h_O(S_O, X_O)$. It then takes this model estimated on the observational data and fits it to the experimental data to predict $\hat{Y_E} = \hat{h}_O(S_{E,i}, X_{E,i})$ for each individual in the experiment. We then simply take the mean of predicted $\hat{Y_E}$ in the treatment, the mean in the control group and then take the difference between them (appropriately weighted by the probability of treatment if this differs according to X). In equation (5) the first term then corresponds to taking the mean of treated individuals' predicted long-term outcome, while the subtracted second term is the mean for control individuals. We can estimate the surrogate index h_o using linear regression, but if the number of surrogates and baseline covariates is large we may wish to use regularisation or machine learning methods to estimate it.

2.2 Estimator based on surrogate score

We can alternatively construct an estimator which does not rely on the surrogate index, but instead on the surrogate and sampling score. For this surrogate score estimator, we need to estimate $\hat{e}(x)$, $\hat{r}(s,x)$ and $\hat{t}(s,x)$. As these are all binary outcomes, we can use logistic regression, but again we may choose to use machine learning methods to select covariates, allow for more flexible functional forms and avoid overfitting.

The surrogate score estimator is based on averaging over the observational sample:

$$\hat{\tau}^O = \frac{1}{\sum_{i=1}^{N_O} \omega_{1,\hat{r},\hat{e},\hat{t}}} \sum_{i=1}^{N_O} Y_{O,i} \cdot \omega_{1,\hat{r},\hat{e},\hat{t}} - \frac{1}{\sum_{i=1}^{N_O} \omega_{0,\hat{r},\hat{e},\hat{t}}} \sum_{i=1}^{N_O} Y_{O,i} \cdot \omega_{0,\hat{r},\hat{e},\hat{t}} \quad (6)$$

The weights ω_i are defined such that they are higher for individuals who are more similar to those in the experiment, in other words, they have a higher sampling score. Furthermore, ω_0 is higher for those individuals who are more similar to the control group in the experiment (a lower surrogacy score), while ω_1 is higher for those similar to the treatment group (a higher surrogacy score).

$$\omega_{\omega,\hat{r},\hat{e},\hat{t}} = \frac{\hat{r}(S_{O,i}, X_{O,i})^{\omega} \cdot (1 - \hat{r}(S_{O,i}, X_{O,i}))^{1-\omega} \cdot \hat{t}(S_{O,i}, X_{O,i}) \cdot (1-q)}{\hat{e}(X_{O,i})^{\omega} \cdot (1 - \hat{e}(X_{O,i}))^{1-\omega} \cdot (1 - \hat{t}(S_{O,i}, X_{O,i})) \cdot q} \tag{7}$$

where $q = N_E/(N_E + N_O)$

The first term of equation (6) takes the weighted average of the long-term outcome for all individuals in the observational dataset, giving more weight to those who are similar to the treatment group. The second subtracted term does the same but giving more weight to those similar to the control group. I do not empirically test this estimator in the current version of this paper, but will in future work.

3 Data

To test the two estimators proposed above, I use data from Barrera-Osorio et al. (2019). The paper uses two experiments to study the impact of three different forms of conditional cash transfer (CCT) on educational outcomes in Bogota, Colombia, with the experimental designs shown in figure 3. The first treatment ('basic') is a standard CCT that pays a $30 transfer every two months conditional on attending school and enrolling in secondary school, for 10 months of the year for a total of $150. The second treatment ('savings') is the same but it gives only $20 every two months, giving the remaining $50 at the time of enrollment next year. The aim of this is to reduce liquidity constraints at the time of enrollment which comes with various expenses such as uniforms and school supplies and thus increase the likelihood of enrolling in the following school year. These are both studied in San Cristobal relative to one control group. The third treatment ('incentive') is studied in a separate experiment in Suba with a different control group. It gives students $20 every two months, but additionally gives students a $300 monetary incentive to graduate from secondary school and enroll in tertiary education. If they did not enter tertiary education, the $300 was delayed one year. San Cristobal and Suba are two of the poorest localities in Bogota.

In San Cristobal (SC), students entering grades 6-11 were randomly assigned to either the basic treatment, savings treatment, or control group. In Suba, only students entering grades 9-11 (upper secondary school) were included in the experiment. There were 10,947 students in San Cristobal (basic = 3,437, savings = 3,438, control = 4,072) and 2,544 in Suba (incentive treatment = 1,140, control = 1,404). For simplicity, I drop all individuals with any missing values losing roughly 10% of the sample, ending up with sample sizes: SC basic = 3,108, SC savings = 3,134, SC control = 3,695, Suba incentive treatment = 1,046, Suba control = 1,299.

The authors study both medium-term (2-8 years) and long-term (12 years) results. I am interested in the effects on long-term outcomes as I use the medium-term outcomes as surrogates. The long-term outcomes are tertiary enrollment, tertiary enrollment on time, and tertiary graduation. They find fairly precise null effects for the basic and savings treatment and larger, positive effects

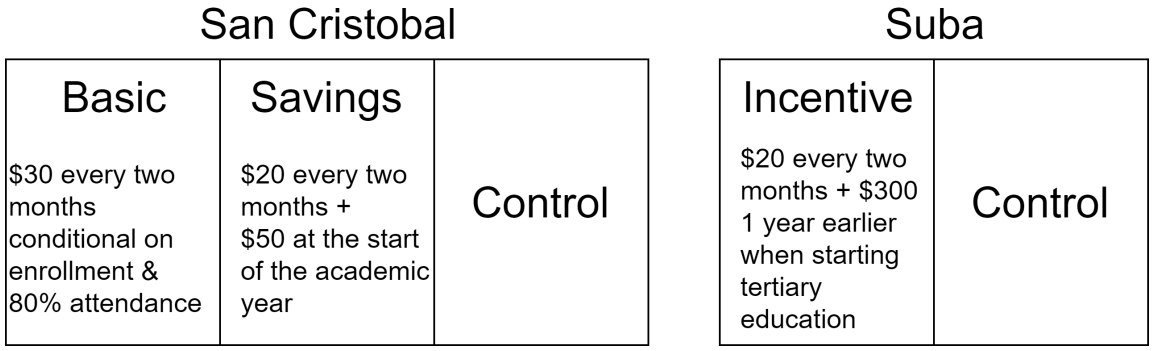


Figure 3: Experimental design in Barrera-Osorio et al. (2019)

for the incentive treatment (but only statistically significant for on time enrollment).

The medium-term outcomes come from three different datasets. First, we have annual secondary school enrollment data, containing information on enrollment in 2006-2008, so 2-4 years after the treatments were started. Second, we have administrative data from the institute that organises secondary school exit examinations so we know which children took the exam. This is a proxy for secondary school graduation. Third, we have panel data which tracks students who enroll in college and we see all students from 1998 to 2016. This is where we get the long-term outcomes observed in 2016, but information whether students had enrolled in tertiary education by 2012. I also utilise the 17 baseline variables including measures of household wealth, income, education and sizes as well as the students' age and gender.

4 Results

4.1 Direct effect

The total effect of a variable X on Y can be split up into the indirect effect, which is mediated through other variables, and the direct effect, that which remains after mediating variables have been adjusted for. The surrogacy assumption requires that the surrogates mediate the full effect of the treatment on the outcome variables. This means that the direct effect of the treatment on the outcome should be 0, i.e. the indirect effect should equal the total effect. I test this by testing whether the direct effect of the treatment is 0 after controlling for the surrogates. To do this, I run the regression:

$$Y = \alpha + \beta T + \gamma S + \delta X + \varepsilon \tag{8}$$

I test whether the direct effect, $\beta = 0$ after controlling for the surrogates and baseline covariates.

I do not include all surrogates and covariates, but instead follow the post-double-selection procedure of Belloni et al. (2014) to select which surrogates and basline covariates to control for. This amounts to running a lasso on outcome (Y) and a lasso on treatment (T) and allowing all surrogates (S) and baseline covariates (X) to be the potential controls. We then take the union of surrogates and baseline covariates with non-zero coefficients from both lassos and run the regression above and test whether the direct effect i.e., the coefficient on T, $\beta = 0$.

The limitation of this test is that the direct effect is only identified if there is no unobserved mediator-outcome confounding. This is shown graphically in figure 2 (above). To identify the direct effect $T \to Y$, the indirect effect $T \to S \to Y$ needs to be identified. The first stage of the indirect effect $T \to S$ is identified as T is randomised. However, the second stage $S \to Y$ has the potential to be confounded as S is not randomised. If all $S \to Y$ confounders are observed (i.e. in X) then we can identify the effect $S \to Y$. However, if there are unobserved confounders, then we cannot identify the second stage, and therefore cannot identify either the indirect effect or the direct effect. Note that this is similar to trying to test the exclusion restriction in an instrumental variables case.

As I allow the post-double selection lasso to control for observed confounders that are predictive of the outcome (and the treatment although none should be as treatment is randomised), I eliminate at least some of mediator-outcome confounding.[2] However, it is possible that there are still important unobserved confounders, as is generally the case with mediation analysis, so the estimate of direct effect may still be biased.

With this limitation in mind, I present the estimates of the total effect and the direct effect in figure 4. We can see that the estimate of the direct effect is not statistically different from 0 in any of the treatment-outcome combinations. Furthermore, it is almost always closer to 0 than the total effect, meaning that the surrogates are mediating the treatment effect.

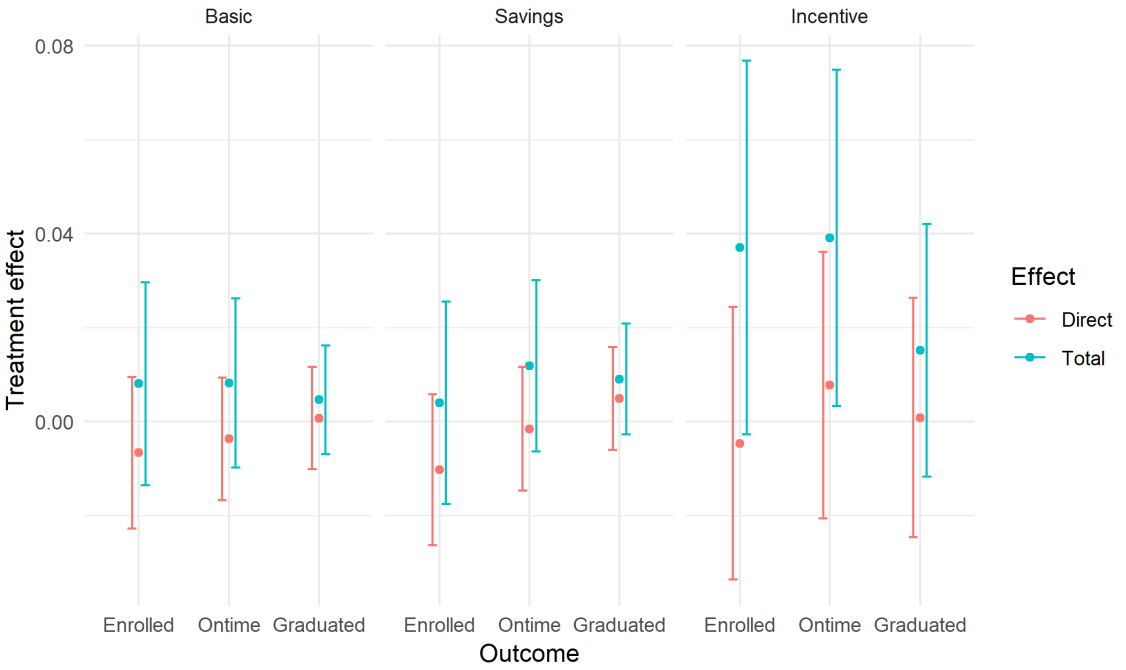
However, the direct effect is also never statistically significantly different from the total effect. This mostly occurs because, in line with the original results of the paper, the total effect of the basic and savings treatments on the long-term outcomes are close to 0 and not significant. On the other hand, in the incentive treatment, the effects are larger and more significant so the fact that the point estimates of the direct effect are very close to 0 suggests that the treatment effect is fully mediated by the surrogates and that therefore the surrogacy assumption is satisfied.

4.2 Random sampling

In this section and the following I implement a more practical test of the surrogacy approach by actually implementing the surrogate index estimator and checking whether the estimated treatment

[2] An improvement to the post double selection lasso used in this case could be made. Instead of selecting surrogates and baseline variables that are predictive of treatment and outcome, we could instead separately select surrogates that are predictive of T and Y, and select baseline controls that are predictive of S and Y. This would be complicated if S is high dimensional though.

Figure 4: Total effects and direct effects of treatments in Barrera-Osorio et al. (2019)



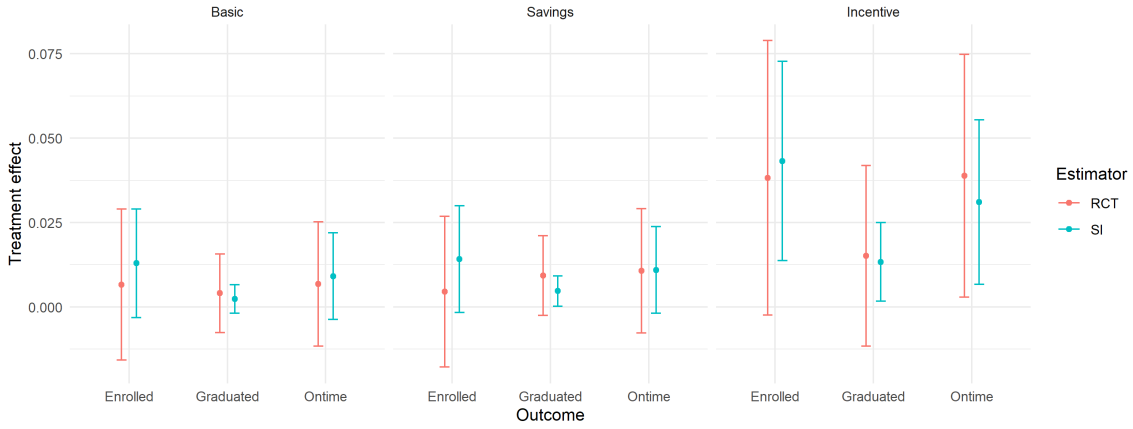
Notes: 95% confidence intervals around estimate shown. Total effect is from RCT with baseline covariates as controls selected by post double selection lasso. Direct effect is from RCT with baseline covariates *and* surrogates as controls selected by post double selection lasso.

effects are similar to those produced by the long-run RCT. The tests differ in how I construct the observational group. In this section, for each treatment-experiment (basic + San Cristobal control, savings + San Cristobal control, and incentive + Suba control), I construct the observational group by randomly sampling half of the observations and deleting the information on treatment status. For the other half, I delete the information on the long-term outcomes and this therefore simulates the short-run experiment.

The benefit of this approach is that it guarantees that the comparability of samples assumption is true in expectation. Recall that the comparability of samples assumption requires that the distribution of the long-term outcome conditional on the surrogates and baseline covariates is the same in the experimental and the observational group ($Y_{E,i} \mid S_{E,i}, X_{E,i} \sim Y_{O,i} \mid S_{O,i}, X_{O,i}$). In this case it is true because the experimental and observational datasets are drawn from the same population and have the same treatment distribution. This means we are only testing the surrogacy assumption.

As the method could by chance perform poorly in any one random split, I take 1000 random splits and take the average treatment effect across these to get a less noisy estimate and provide a variance, similar in spirit to bootstrapping. The results are shown in figure 5.

Figure 5: Comparison of RCT surrogate index (SI) estimates from random sampling approach using Barrera-Osorio et al. (2019) data



Notes: RCT estimates are from regression with controls selected by post-double selection lasso with 95% confidence interval shown. As outcomes are binary, I use logistic regression for estimating the surrogate index, surrogate score, sampling and propensity scores. Point estimates of surrogate index (SI) average of point estimates from 1000 random splits and the error bars are 1.96 × the standard deviation of the 1000 estimates.

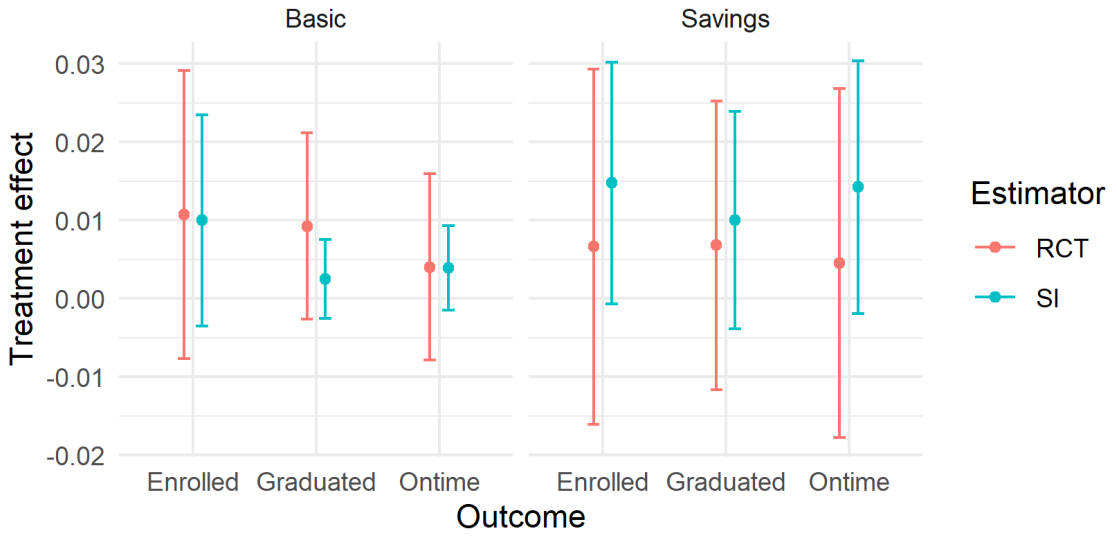
Here we can see that the estimates from the surrogate index estimator are generally close to the unbiased RCT estimate and never statistically significantly different from it. This is true for all treatments and all outcomes. There is no particular pattern of the surrogate index estimator over or underestimating the RCT estimate.

4.3 Multi-arm design

In this section I exploit the fact that in San Cristobal there are two treatment arms and a control group. First, I construct the experiment as the basic treatment and the control group and treat the savings arm as the observational dataset. In other words, I pretend I do not have the long-term outcome in the basic treatment and control group, and pretend I do not have the treatment indicator in the savings arm. I run the surrogacy estimators on these artificial datasets and compare them to the true experimental results. Then I do the reverse and construct the experiment as the savings and control group and let the basic arm be the observational dataset. Results are shown in figure 6.

This set-up is more demanding but also more realistic than the random sampling case as the comparability of samples assumption is no longer true by design, although it is true if surrogacy holds. This is because in reality there may be interactions between the treatments and the surrogates in the conditional expectations of the long-term outcome. This means that the relationship between the surrogates and the long-term outcome may be different in the observational and experimental dataset.

Figure 6: Comparison of RCT and surrogacy estimates from Barrera-Osorio et al. (2019) San Cristobal data



Notes: Data only from San Cristobal used. For basic, savings is used as observational data and vic-versa. I use linear regression for estimating the surrogate index. 95% confidence intervals shown generated from 1000 bootstrap replications.

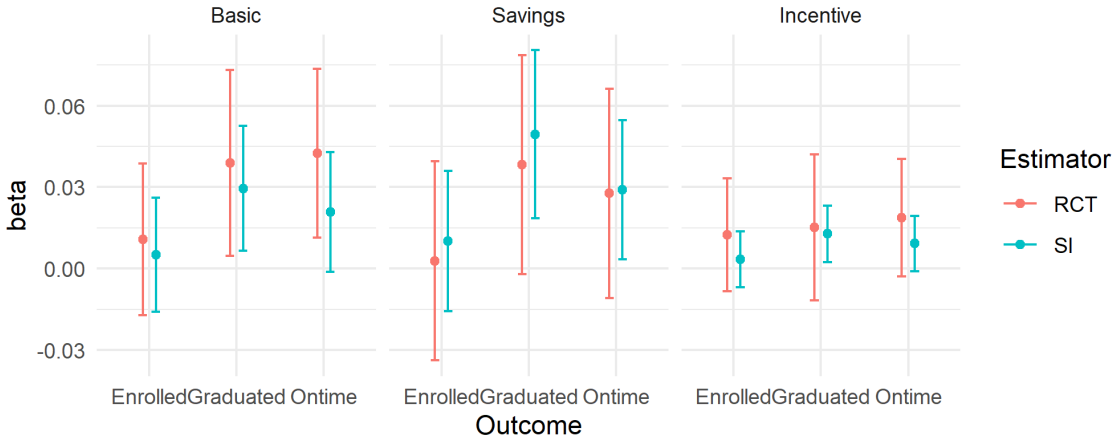
In figure 6 a couple of results are clear. In general the surrogate estimator performs well. All estimates fall within the 95% confidence interval of the RCT estimate. Furthermore, they fall close to the centre meaning that the point estimates are very similar to the RCT point estimates. This

suggests that the worry about there being treatment-surrogate interactions leading to differential relationships between the surrogates and long-term outcome is not practically significant in this case. We can also note that the confidence intervals on the surrogate index estimator are tighter. This is in line with theoretical results from Athey et al. (2019).

4.4 Cross-experiment design

Next, I exploit the fact that I have two experiments with exactly the same variables. I use San Cristobal as the experiment and Suba as the observational data and vice-versa. However, recall that in Suba only students in grades 9-11 at the start of the experiment were eligible for treatment. Thus I also limit the San Cristobal data to students in those grades to ensure common support on age. This case is probably the most realistic of those done so far as the observational and experimental datasets are drawn from different populations, even if these different populations are just different neighbourhoods in the same city. This also means that it is the most demanding as these populations may not be comparable. Results are shown in figure 7.

Figure 7: Comparison of RCT and surrogacy estimates from full Barrera-Osorio et al. (2019) data



Notes: Only individuals from grade 9-11 are used from San Cristobal to match Suba. Suba forms the observational data for basic and savings. San Cristobal forms the observational group for incentive. I use linear regression for estimating the surrogate index. 95% confidence intervals shown generated from 1000 bootstrap replications.

Note that the RCT estimates for the basic and savings treatments are higher than before as they are estimated on the subsample of students in San Cristobal who started receiving in grades 9-11. Again we see that the surrogate index performs relatively well in this context. We can see that for the basic and incentive treatments, the surrogate point estimate is lower than the experimental point estimate for all outcomes, while it is higher for all outcomes in the savings arm, but the difference is never statistically significant and the point estimates are close in magnitude. We again see that the

surrogate estimate is more precise than the RCT estimate.

4.5 Reducing number of surrogates

In this section, I test how well the surrogacy methods perform when the set of surrogates to choose from is reduced. To do this, I rerun the surrogacy estimators with different subsets of the surrogates, but always include all the baseline covariates.

As a reminder, in Barrera-Osorio et al. (2019), the medium-term outcomes come from three different datasets. First, there is annual secondary school enrollment data, containing information on enrollment in 2006-2008, so 2-4 years after the treatments were started. Second, there is administrative data from the institute that organises secondary school exit examinations so they know which children took the test which is a proxy for secondary school graduation. Third, there is panel data which tracks students who enroll in college, in which we can see all students from 1998 to 2016. From this we get the long-term outcomes observed in 2016, but also whether students had enrolled in tertiary education by 2012.

I divide the surrogates into four categories. First is 'Enrolled' which contains variables measuring enrollment in 2006, 2007 and 2008. Second is 'Repetition' which contains variables concerning whether students were enrolled ontime or heldback or dropped out of secondary school from 2006-2008. Third is 'Graduation' which captures whether students sat the secondary school exit exam. Finally, we have 'Tertiary', which contains the information on whether students had enrolled in tertiary education by 2012, and whether it was vocational, university or other education.

I test the surrogacy approach with the different categories of surrogates. Figure 8 shows the results for when the basic treatment arm and control group in San Cristobal is the experiment and Suba is the observational data, 9 for savings and control in San Cristobal being the experiment and Suba being the observational data, while figure 10 shows when the incentive arm and control group in Suba is the experiment and all three arms in San Cristobal are used as observational data.

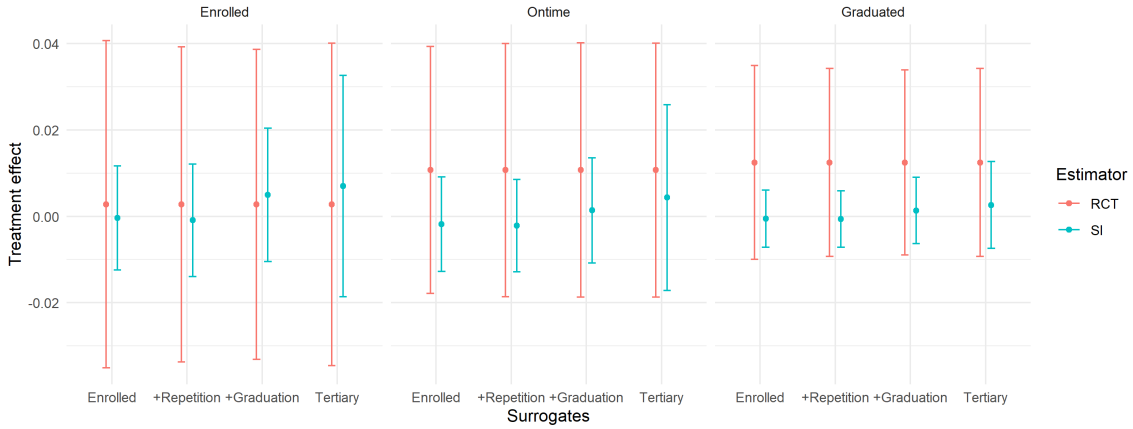
I start with 'Enrolled', the 3 variables on enrollment from 2006-2008. We can see in all three figures, for all outcomes this produces a treatment effect estimate precisely on 0, while the RCT estimates range from 0.01 to 0.04, meaning that the surrogate estimates are poor. This is because the secondary school enrollment variables are poor predictors of tertiary enrollment and graduation almost 10 years later. This means, they do not get close to satisfying the surrogacy assumption as most of the effect of the conditional cash transfers on tertiary education outcomes is not mediated through short-term secondary school enrollment. In fact, the zero surrogate index estimate suggests that none of the effect is mediated through secondary school enrollment.

I then add to the set of surrogates used for estimating the surrogate index 'Repetition', the 3 variables containing information about whether people enrolled on-time, were held-back or dropped out. We again see similar poor results, suggesting that these variables also do not predict the

long-term outcomes. When I further add 'Graduation', the variable about taking the secondary school exit examination, performance is slightly improved for some outcomes with the surrogate estimates moving slightly closer to the estimates from the long-term RCT. This suggests that this variable is a surrogate and does mediate some of the effect of treatment. This is intuitive as those who do not graduate secondary school are unlikely to attend tertiary education. However, there are clearly still other major channels through which the treatment affects the long-term outcomes as surrogacy estimates are still far from the RCT estimates.

Finally, I consider only 'Tertiary', the 4 variables from 2012 containing information on tertiary enrollment up to that point, and ignore all the previous possible surrogates. Here, we see the performance of the surrogacy estimators is much better, being much closer to the RCT estimates. This means that almost all of the effect of the treatment on the long-term outcomes is mediated by these surrogates and that they are predictive of the long-term outcomes.

Figure 8: Comparison of estimates of basic treatment arm, limiting surrogates used

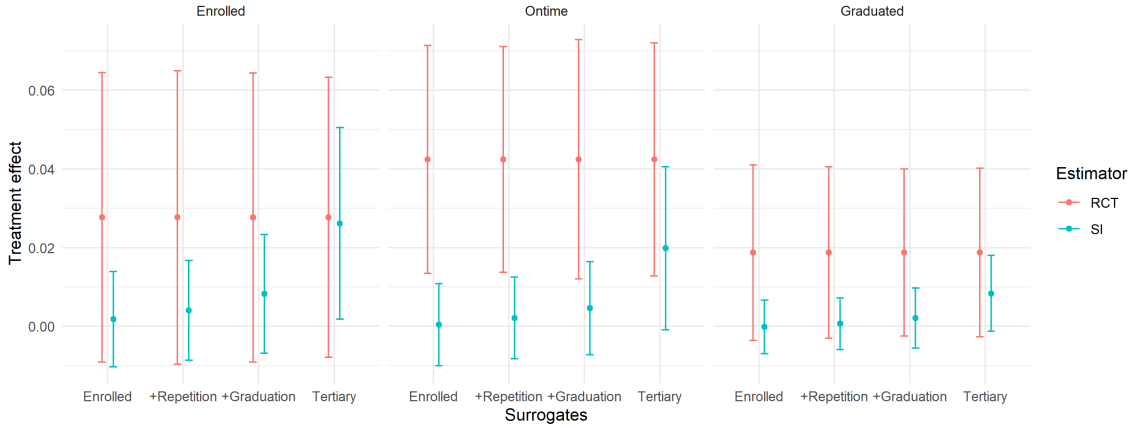


Notes: Only individuals from grade 9-11 are used from San Cristobal to match Suba. I use linear regression for estimating the surrogate index. 95% confidence intervals shown generated from 1000 bootstrap replications.

It is perhaps unsurprising that tertiary enrollment by 2012 is a good surrogate for tertiary enrollment and graduation by 2016. If you had enrolled by 2012, then it is purely deterministic that you must also have enrolled by 2016. It is also unsurprising that people who had enrolled by 2012 were more likely to graduate by 2016, than those who had not enrolled by 2012, especially considering the fact that an undergraduate degree typically takes four or five years in Colombia.

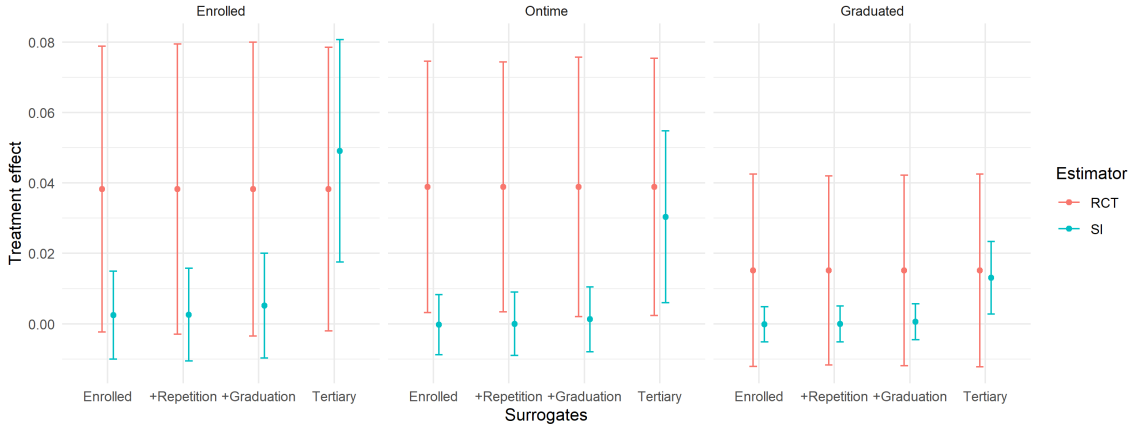
The surrogates that do most of the work in this study have a clear *ex ante* connection to the long-term outcome and they are observed only four years before the long-term outcome. This gives cause for concern about the performance of surrogacy approaches when we only observe variables that are weakly related to the long-term outcomes and temporally further away. However, the number of intermediate outcomes (11) collected by Barrera-Osorio et al. (2019) is small relative to

Figure 9: Comparison of estimates of savings treatment arm, limiting surrogates used



Notes: Only individuals from grade 9-11 are used from San Cristobal to match Suba. I use linear regression for estimating the surrogate index. 95% confidence intervals shown generated from 1000 bootstrap replications.

Figure 10: Comparison of estimates of incentive treatment arm, limiting surrogates used



Notes: Only individuals from grade 9-11 are used from San Cristobal to match Suba. I use linear regression for estimating the surrogate index. 95% confidence intervals shown generated from 1000 bootstrap replications.

the number that are typically collected in development RCTs so we might expect better performance elsewhere.

5 Conclusion

Surrogacy approaches are a potentially very valuable method for estimating long-term effects. I developed methods to test surrogacy approaches using RCTs. I tested the surrogate index approach introduced in Athey et al. (2016) using experiments from Barrera-Osorio et al. (2019). When the full set of surrogates is used the surrogacy approaches perform well, replicating the unbiased experimental estimate. However, the performance of the method is very sensitive to the set of surrogates included, with tertiary enrollment by 2012 being a key variable which determines the performance of the surrogacy methods.

There are a number of avenues for further research on this topic. Firstly, I plan to repeat the analysis in other datasets. This will improve our understanding of what topics the surrogacy approach can be used successfully in. Secondly, I also plan to explore whether surrogacy can be used to estimate precise heterogeneous long-term treatment effects on top of the average treatment effect. Thirdly, I will code and implement the two new surrogate estimators introduced in Athey et al. (2019) along with the bounding approach inspired by Oster (2019).

References

Athey, S., Chetty, R., Imbens, G., and Kang, H. (2016). Estimating treatment effects using multiple surrogates: The role of the surrogate score and the surrogate index. *arXiv preprint:1603.09326*.

Athey, S., Chetty, R., Imbens, G., and Kang, H. (2019). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Technical report, Opportunity Insights.

Banerjee, A., Duflo, E., and Kremer, M. (2016). The influence of randomized controlled trials on development economics research and on development policy. In *The State of Economics, the State of the World Conference at the World Bank*.

Barrera-Osorio, F., Linden, L. L., and Saavedra, J. E. (2019). Medium-and long-term educational consequences of alternative conditional cash transfer designs: Experimental evidence from colombia. *American Economic Journal: Applied Economics*, 11(3):54–91.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Bouguen, A., Huang, Y., Kremer, M., and Miguel, E. (2018). Using RCTs to estimate long-run impacts in development economics. Technical report, National Bureau of Economic Research.

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

Freedman, L. S., Graubard, B. I., and Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statistics in medicine*, 11(2):167–178.

Joffe, M. M. and Greene, T. (2009). Related causal frameworks for surrogate outcomes. *Biometrics*, 65(2):530–538.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204.

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

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

VanderWeele, T. (2015). *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.