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

David Rhys Bernard (Rethink Priorities), Jojo Lee and Victor Yaneng Wang (Global Priorities Institute, University of Oxford)

Global Priorities Institute | September 2023

GPI Working Paper No . 13-2023

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

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

David Rhys Bernard*[1], Jojo Lee†[2], and Victor Yaneng Wang‡[2]

[1]Rethink Priorities
[2]Global Priorities Institute, University of Oxford

October 2, 2023

Summary

The *surrogate index* method allows policymakers to estimate long-run treatment effects before long-run outcomes are observable. We meta-analyse this approach over nine long-run RCTs in development economics, comparing surrogate estimates to estimates from actual long-run RCT outcomes. We introduce the *M-lasso* algorithm for constructing the surrogate approach's first-stage predictive model and compare its performance with other surrogate estimation methods. Across methods, we find a negative bias in surrogate estimates. For the M-lasso method, in particular, we investigate reasons for this bias and quantify significant precision gains. This provides evidence that the surrogate index method incurs a bias-variance trade-off.

JEL Codes: C01, C53, I15, I25

Keywords: long-term treatment effects, surrogate index, meta-analysis, RCT, M-lasso.

Conflict of interest statement: The authors have no conflicts of interest to declare.

*Email: dbernard@rethinkpriorities.org
†Corresponding author. Email: jojoskilee@gmail.com
‡Email: vwang57@gmail.com

Acknowledgments: We thank Nina Ruer for excellent research assistance on this project, the researchers who made their long-term RCT microdata freely available online which enabled this project, and the Global Priorities Institute at the University of Oxford where we started this project. We also thank John Firth and Susana Campos Martins for their helpful suggestions and comments. Finally, we acknowledge support from funding by Forethought Foundation, the Effective Altruism Long-Term Future Fund, and a French government subsidy managed by the Agence Nationale de la Recherche under the framework of the Investissements d'avenir programme reference ANR-17-EURE-001. All mistakes remain our own.

1 Introduction

The long-term effects of treatments and policies are important in many different fields. In medicine, one may want to estimate the effect of a surgery on life expectancy; 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, the results would be observed too late to inform policy decisions made today.

A prominent solution to this issue is the *surrogate index*, a method for estimating long-run effects without long-run outcome data, which was originally proposed by Athey et al. (2019). Our paper contributes to the evolving literature on this method by examining its empirical performance in a wide range of RCT contexts. We also extend the discourse initiated by LaLonde (1986) on the bias of non-experimental methods, extending the set of estimators studied to those focused on long-term effects. Our findings and recommendations aim to guide practitioners intending to use the surrogate index method, thereby aiding in the development of effective long-term treatment strategies.

We test the surrogate approach on data from nine RCTs in development economics. These RCTs are selected on the basis of being long-running and having a sufficiently large sample size.

In each RCT, we first produce an unbiased estimate of the standard experimental average treatment effect by regressing long-term outcomes on treatment status. Next, we reanalyse the data using the surrogate index approach. If the surrogate estimate is close to the unbiased estimate from the experimental approach, then the surrogate index method is working well. We run meta-analyses on the difference between these estimates to understand how well the surrogate index method performs under different conditions.

We test many different implementations of the surrogate index estimator, varying (1) the set of surrogates used, (2) the first-stage prediction method used, and (3) the observational dataset used to construct the surrogate index. Notably, we introduce a new estimator called the *M-lasso*, which is specifically designed for use with the surrogate method.

When meta-analysing our results, we find that the surrogate index method is consistently negatively biased and underestimates positive long-term treatment effects by 0.05 standard deviations on average. This is the case regardless of which estimation method we use. We suggest that this is due to missing surrogates, as well as bias in the first-stage predictive model of the surrogate procedure.

While it is important to understand this negative bias as a potential shortcoming of the surrogate approach, we would not necessarily take it to dissuade researchers from this method altogether. Instead, one could interpret surrogate estimates as a reasonable lower bound on the true long-term treatment effect. Furthermore, there is often no better alternative for estimating the true effect.

We also study potential determinants of the surrogate bias for the M-lasso estimator. In particular, we find suggestive evidence that M-lasso bias is smaller for simpler interventions. However, we do not find that this bias depends on the predictive accuracy of the first-stage model in the observational dataset. Our evidence is also inconclusive about how bias is affected by longer time horizons between the surrogates and the outcomes.

We further show that despite the potential bias from using the surrogate index method, it results in significant precision gains, with standard errors on average 52% the size of those from the long-term RCT estimates. Hence, even if researchers had access to long-term outcomes, they might still choose to use the surrogate index, depending on their willingness to trade off bias and variance.

The rest of this paper proceeds as follows. Section 2 discusses related literature. Section

3 summarises the econometric theory behind the surrogate index approach, and section 4 describes in more detail the data we use. Section 5 explains the methods we use to estimate comparable long-term RCT and surrogate index estimates. Section 6 presents results of the meta-analysis over 9 RCTs for different implementations of the surrogate index. In it, we empirically characterise the bias and standard errors for the surrogate method, as well as examine which surrogates are selected by the M-lasso. Finally, section 7 concludes.

2 Related literature

Using surrogates or intermediate outcomes for long-term effects is an approach pioneered in medicine to deal with the difficulties of long-term effects. One can 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 whether a patient received the surgery, conditional on the size of the tumour.

Such methodological questions are also of interest to economists, many of whom may be interested in the long-run impacts of different policies (Bouguen et al., 2019). Indeed, the use of surrogates in economics is now rapidly growing: Guzman et al. (2020), Dynarski et al. (2021) and Otero et al. (2021) have recently applied the surrogate index method from Athey et al. (2019) in the contexts of pro-social motivations, college admissions and affirmative action respectively. However, one difficulty is that the Prentice criterion is hard to justify in a social science context and there are multiple ways it can 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 if there is unobserved confounding between the surrogate and the long-term outcome, even under full mediation, the surrogacy assumption is also invalid (VanderWeele, 2015).

Due to these issues, Athey et al. (2019) develop surrogacy methods that utilise many surrogate variables instead of just one, as well as controlling for many potential confounders. The idea behind their approach is that even though any individual variable may not be a valid surrogate by itself, collectively they are more likely to fully mediate the treatment effect and 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 index, they develop different estimators for long-term effects when the long-term outcome is not observable. Furthermore, they also show that treatment effect estimates on the surrogate index, by discarding random noise variation in the long-term outcome that is orthogonal to treatment, can be more precise than treatment effects on the true long-term outcome. We test these surrogacy estimators with real-world data from multiple long-run RCTs in economics.

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 this early wave of experiments to study the effects of programs such as conditional cash transfers on various long-term outcomes (e.g. adult income twenty years later). Some further examples of long-run RCTs can be found, for instance, in Bouguen et al. (2019).

3 Theory

Athey et al. (2019) introduce the surrogate index method for estimating long-term effects when we do not observe the long-term outcomes of an experiment. In section 3.1, we first provide an intuitive explanation of how the surrogate index method works. Section 3.2 then discusses how to construct the surrogate estimator for long-term treatment effects.

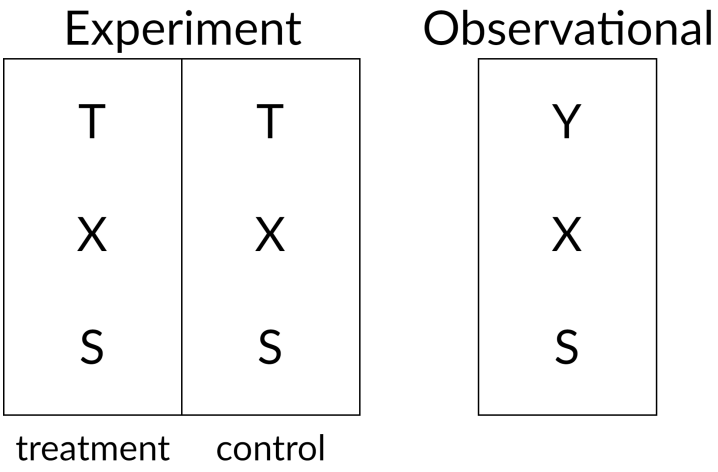
A more formal summary of the theory behind the estimators of Athey et al. (2019) is provided in Appendix G. McKenzie (2020) also provides an accessible introduction to the surrogate index method.

3.1 Intuitive overview of surrogate index

This section gives an intuitive overview of the surrogate index. As a working example, consider a randomised unconditional cash transfer to parents of children aged 8. Suppose we are interested in the effect of the transfer on children's high school graduation 10 years later when the children are 18. We only observe children's outcomes two years after the transfer; such outcomes might include school enrollment, test scores, and height.

Two samples are required for the surrogate index approach as shown in figure 1. First is the short-run experimental sample. In this sample, we must observe individual treatment status (T, e.g. a cash transfer), a set of short-run outcomes (S, e.g. children's enrollment, test scores, and height at age 10), and optionally pre-treatment covariates (X e.g. parent's education, gender, test scores at baseline). However, we do not observe the long-run outcome that we care about (Y, e.g. children's high school graduation).

Figure 1: Data required for surrogacy approach



Note: The surrogacy approach requires an experimental dataset and an observational dataset. Treatment status, baseline covariate, and surrogate variable information should be known for each individual in the experimental dataset. Outcome, baseline covariate, and surrogate variables should be known for each individual in the observational dataset.

Secondly, we need the observational sample. This is a dataset for a separate sample of individuals, where we observe the long-run outcome that we care about plus the same set of

short-run outcomes as in the experimental sample (and optionally the same pre-treatment covariates). In our example, this would be a dataset where we observed whether children had graduated high school at age 18, along with their enrollment status, test scores, and height at age 10.

There are three stages to the surrogate index approach.

Stage 1: Use the observational sample to predict the long-run outcome as a function of the short-run outcomes, creating a surrogate index model. The surrogate index is the conditional expectation of the long-term outcome conditional on the short-run outcomes (and potential pre-treatment covariates). In this example, we could use linear regression to estimate the surrogate index in the observational sample as follows:

$$Graduation = \alpha + \beta_1 \cdot Enrollment + \beta_2 \cdot TestScore + \beta_3 \cdot Height + \varepsilon \tag{1}$$

One could alternatively use supervised machine learning methods such as lasso or random forest in this first stage to estimate the predictive model.

Stage 2: Predict the surrogate index in the experimental sample using the predictive model from stage 1. As we have the same surrogates in the experimental and observational samples, we can deploy the model trained on the observational sample to produce predictions of the long-run outcome in the experimental sample. This is essentially combining the short-run outcomes into an index that is also a prediction of the long-run outcome, which we call the surrogate index in the experimental sample. In our example, we would create the following variable in the experimental sample:

$$SurrogateIndex = \hat{\alpha} + \hat{\beta}_1 \cdot Enrollment + \hat{\beta}_2 \cdot TestScore + \hat{\beta}_3 \cdot Height \tag{2}$$

where $\hat{\beta}_1$ is the estimate of β_1 from the observational sample in equation (1) and so on.

Stage 3: Estimate the treatment effect on the surrogate index in the experiment. Now, we can estimate the treatment effect on the surrogate index in the experimental sample.

$$SurrogateIndex = \gamma + \delta \cdot Treatment + \upsilon \tag{3}$$

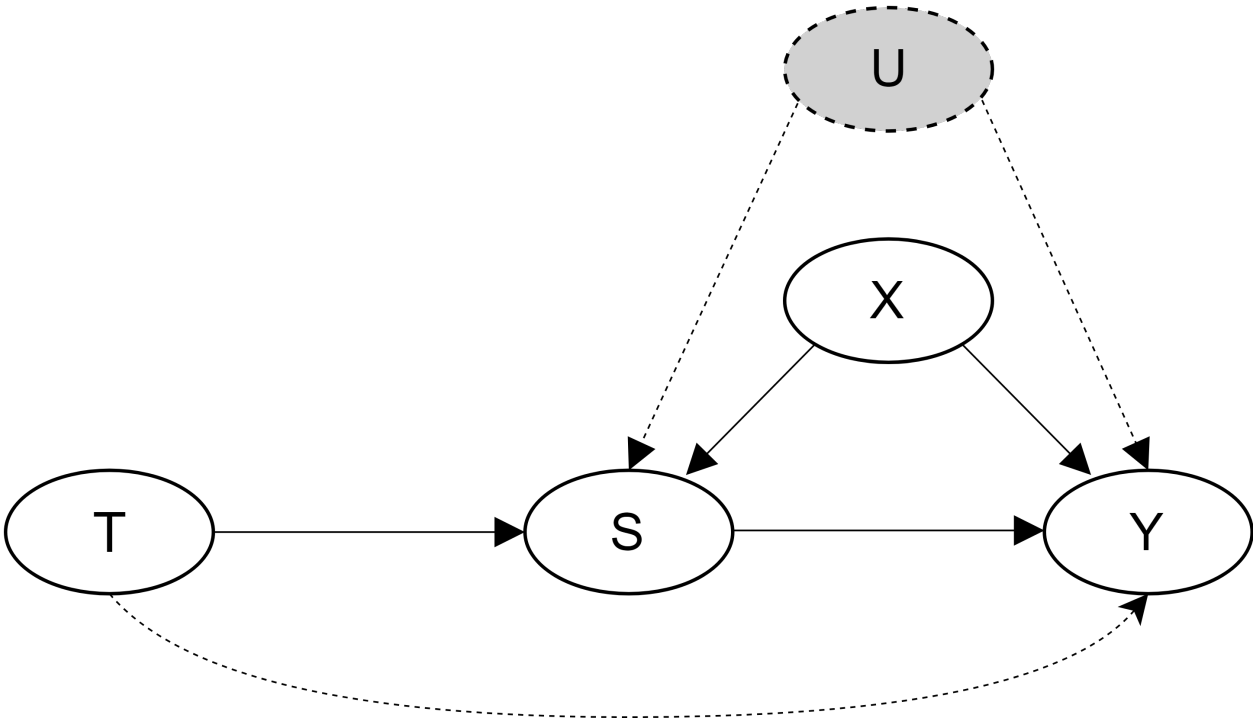
δ is the estimated effect of the treatment on the long-term outcome.

This approach requires three assumptions to work (note these are also presented more formally and discussed further in Appendix G):

Assumption 1: Unconfoundedness. Unconfoundedness is the standard assumption that the potential outcomes of the long-term outcome are independent of treatment, conditional on covariates. In our example of a randomised experiment, this assumption will be true by design. However, in cases where the experimental sample is a non-randomised study, this assumption will have to be justified in the standard way. We also assume common support, i.e. that each observation has some probability of being treated and of not being treated, conditional on covariates.

Assumption 2: Comparability of samples. This assumption requires that the conditional distributions of the long-term outcome, conditional on the short-term outcomes and covariates, are the same in the observational and experimental samples. This assumption is necessary as we train a predictive model on the observational sample, but use it to make predictions on the experimental sample. This assumption is what allows us to transfer the model from one sample to the other.

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



Notes: Treatment is represented by T, surrogates are S, the long-term outcome is Y, observed surrogate-outcome confounders are X, and unobserved surrogate-outcome confounders are U. Dotted lines are causal paths ruled out by the surrogacy assumption.

Assumption 3: Surrogacy. This assumption requires that the long-term outcome is independent of the treatment, conditional on the short-run outcomes and covariates. In our example, this would mean that the treatment had no additional explanatory power for high school graduation once enrollment, test scores, and height at age 10 were controlled for. Essentially, this assumption requires us to observe all of the mediating causal pathways between the treatment and the long-term outcomes. Note that the surrogate index incorporates multiple short-term outcomes, making it more likely that all of the causal paths between the treatment and long-term outcome are observed. However, in our example, we may be concerned that socio-emotional skills are not observed and these could be an important alternative path through which later outcomes are affected.

To illustrate the surrogacy condition further, we use the following notation: treatment status is represented by T, surrogates are S, long-term outcomes are Y, observed surrogate-outcome confounders are X, and unobserved surrogate-outcome confounders are U. We can then represent the surrogacy assumption with a directed acyclic graph (DAG) as in figure 2.[1]

The surrogacy condition can be viewed as entailing two sub-requirements:

1. The effect of S on Y is causally identified. This means that we have to control for any observable confounders X that are correlated with both S and Y. If there are any *unobserved* confounders U, then the surrogacy condition will fail.

2. S fully mediates the effect of T on Y, i.e. there is no causal pathway from T to Y that does not go through S. Another way to phrase this is that the surrogate variables must 'span the entire causal pathway' from treatment to long-run outcome.

[1] An explanation of directed acyclic graph approaches to causal inference is contained in Imbens (2020).

S is more likely to span the whole causal pathway from T to Y when we use multiple surrogate variables, rather than just one. Even then, though, this condition is unlikely to hold perfectly in practice. However, Athey et al. (2019) demonstrate that small violations of the surrogacy condition only lead to small biases in the surrogate estimator of the treatment effect. This is analogous to the exclusion condition with instrumental variables – while the condition is unlikely to hold exactly, the overall method performs reasonably well as long as we come close to satisfying it.

3.2 Estimator based on surrogate index

Recall that the surrogate index is the conditional expectation of the long-run outcome, given the covariates and surrogates in the observational sample. We formally define this as follows:

Definition. The surrogate index.
$h_O(s, x) = \mathbb{E}[Y_i \mid S_i = s, X_i = x, P_i = O]$

where Y_i, S_i, and X_i are respectively the long-term outcome, surrogates, and covariates for individual i. Finally, P_i is a binary variable that equals O if we are using data from the observational dataset and E if we are using data from the experimental dataset.

Recall next that in the first stage of the surrogate approach, we estimate the surrogate index in the observational dataset as $\hat{h}_O(s, x)$. We subsequently use this to construct an *surrogate index estimator* for the long-term treatment effects in the experimental dataset. This is defined as:

$$\hat{\tau}^E = \frac{1}{\sum_{i=1}^{N_E} T_i/\hat{e}(X_i)} \sum_{i=1}^{N_E} \hat{h}_O(S_i, X_i) \cdot \frac{T_i}{\hat{e}(X_i)} - \frac{1}{\sum_{i=1}^{N_E}(1-T_i)/(1-\hat{e}(X_i))} \sum_{i=1}^{N_E} \hat{h}_O(S_i, X_i) \cdot \frac{1-T_i}{1-\hat{e}(X_i)} \quad (4)$$

where $\hat{e}(x)$ is an estimate of the propensity score. The propensity score is the conditional probability of an individual in the experimental dataset being treated, conditional on their covariates; it is formally defined as $e(x) = Pr(T_i = 1 \mid X_i = x, P_i = E)$.

Intuitively, the surrogate index estimator proceeds as follows. Firstly, it takes the surrogate index model and fits it to the experimental dataset to predict experimental long-run outcomes $\hat{Y}_E = \hat{h}_O(S_E, X_E)$. In equation (4) the first term then corresponds to taking the mean of treated individuals' predicted long-term outcomes, while the second term is the mean for control individuals. Note that these terms are also appropriately weighted by the probability of treatment if this differs according to X. Finally, we take the difference between these two means.

As our settings are all randomised controlled trials with a constant probability of treatment ($e(x) = p$), we will work with a simplified version of the estimator:

$$\hat{\tau}^E = \frac{1}{\sum_{i=1}^{N_E} T_i} \sum_{i=1}^{N_E} \hat{h}_O(S_i, X_i) \cdot T_i - \frac{1}{\sum_{i=1}^{N_E}(1-T_i)} \sum_{i=1}^{N_E} \hat{h}_O(S_i, X_i) \cdot (1-T_i) \quad (5)$$

We estimate the surrogate index using several different methods, described in more detail in section 5.2. We use two linear regression methods with either the single most correlated surrogate or all surrogates. We then use two lasso methods: one that picks surrogates that are predictive of the long-run outcome, and another that picks surrogates that are either predictive of the long-run outcome *or* predictive of the treatment. Finally, we use the XGBoost supervised learning algorithm.

Table 1: Summary of RCTs used in analysis

Paper	Intervention(s)	Country	Years	Waves	Arms
Barrera-Osorio et al. (2019)	CCT	Colombia	12	5	5
Duflo et al. (2015)	Grant & HIV course	Kenya	7	3	4
De Mel et al. (2012)	Cash grant	Sri Lanka	5	12	3
Blattman et al. (2020)	Cash grant	Uganda	9	3	2
Banerjee et al. (2021)	Graduation program	India	7	3	2
Baranov et al. (2020)	Psychotherapy	Pakistan	7	3	2
Gertler et al. (2012)	CCT	Mexico	6	7	2
Buchmann et al. (2023)	Empowerment	Bangladesh	10	3	4
Hamory et al. (2021)	Deworming	Kenya	20	4	2

Notes: Waves is the number of post-treatment survey waves; it does not include pre-treatment survey waves. Arms is the number of treatment arms and includes both treatment and control group arms. CCT stands for conditional cash transfer.

4 Data

To test the surrogate estimator proposed above, we use data from nine different long-term randomised controlled trials (RCTs). These RCTs were chosen on the basis of being long-running and having a sufficiently large sample size.

Table 1 summarises key information about each of these RCTs. Note that we achieve broad coverage: there are two studies from Latin America, four from South Asia, and three from East Africa. Furthermore, we study several different common development interventions, in areas ranging from cash transfers to health and education. The long-run outcomes stretch from 5 to 20 years after treatment.

The studies also vary in the number of arms: 4 of the 9 studies include more than one treatment arm. A detailed description of each study can be found in Appendix A.

5 Methodology

In this section, we first describe how we use the RCT datasets to test the surrogate index methodology. Secondly, we describe the different implementations of the surrogate index estimator that we use.

5.1 RCT data usage

We use the nine RCT datasets described above to test the surrogate index method. In the first step of our test, for each outcome, we estimate the ground truth treatment effect in the standard way by regressing the outcome on the randomised treatment indicator. For all outcomes in all studies, we estimate the intent-to-treat effect, even if there is non-compliance in the study.

Next, we imitate the situation of missing long-term outcome data by removing the long-run outcome from the experimental sample. We instead use the surrogate index method to impute the long-term outcome and re-estimate the treatment effect on the imputed outcome, again estimating the intent-to-treat effect. We describe the different implementations of the surrogate index for predicting the long-term outcome in section 5.2 below.

We vary the number of surrogates we use for estimating the surrogate index. Suppose we have an RCT with five post-treatment waves as depicted in table 2. β_{ij} represents an

Table 2: Possible estimates from surrogacy approach

		Surrogates			
	$Wave$	1	2	3	4
Outcomes	1				
	2	β_{21}			
	3	β_{31}	β_{32}		
	4	β_{41}	β_{42}	β_{43}	
	5	β_{51}	β_{52}	β_{53}	β_{54}

β_{ij} represents an estimated treatment effect on an outcome from wave i using surrogates from wave j and before

estimated treatment effect on an outcome from wave i using surrogates from wave j and before. For outcomes in the first wave with $i = 1$ (the first row), we cannot estimate a surrogate index as there are no earlier waves with shorter-term outcomes/surrogates. In the second wave, $i = 2$, we can only estimate a surrogate index using outcomes from the first wave as surrogates, $j = 1$. In the third wave, we can estimate two surrogate indexes. The first one is β_{31} where we just use the surrogates from the first wave to predict outcomes in the third wave. The second one is β_{32} where we use surrogates from both the first and the second wave to predict outcomes in the third wave. In the fourth wave, we can estimate three surrogate indexes, and in the fifth wave, we can estimate four surrogate indexes.

There are several ways to construct the observational dataset used for training the surrogate index prediction model. In the main article, we focus on what we call the same sample design. With this design, we construct the observational dataset from the control group. Meanwhile, the experimental dataset is constructed from both the control and treatment groups. This ensures that the experimental dataset contains variation in the treatment status.

The same sample design aims to minimise violations of the comparability of sample assumption by drawing the observational and experimental datasets from the same RCT sample. This design also mimics the idea that in practice, the observational dataset is likely to be constructed from a separate non-experimental sample where no treatment was administered. Finally, this design is always feasible, as there is always a control group in a randomised controlled trial. The drawback of this approach is that it might lead to overfitting of the surrogate index because the data used in the prediction and estimation stages of the estimator is partially overlapping.

In Appendix C we describe an alternative 'cross-arm' design, which can only be applied to RCTs with multiple treatment arms.

5.2 Implementations of surrogate index estimator

We now discuss how to use the observational dataset to predict the long-term outcome, thereby generating our surrogate index model. There are two key considerations here:

(1) Which prediction algorithm we use; and

(2) Which variables we use in that prediction algorithm.

For (1) we try five different algorithms: a kitchen sink linear regression using all surrogates, linear regression using a single surrogate, a lasso, XGBoost, and an algorithm we introduce

called the M-lasso. These algorithms are respectively detailed in sections 5.2.1 to 5.2.5 below.

For (2), we vary the sets of variables used in two ways. Firstly, we vary whether we only use surrogates, or else if we use surrogates *and* available baseline covariates. This lets us examine whether adding covariates helps us come closer to satisfying the surrogacy assumption, thereby reducing the bias in the surrogate index estimator. Recall from above that we also had five different prediction algorithms. This then gives us ten types of estimators overall: five without covariates and five with covariates.

Secondly, we vary the number of previous waves used to predict the long-term outcome, as discussed in table 2.[2] This allows us to assess how well the surrogate index method performs over different time horizons.

Once we have our predicted long-term outcomes (the surrogate index), we regress these on the treatment variable to obtain a surrogate estimate of the treatment effect. We also normalise all treatment effect estimates by dividing them by the standard deviation of the true long-term outcome in the control group. This converts everything into comparable effect sizes. As such, most of our figures do not come with units.

5.2.1 Kitchen sink linear regression

For the most basic approach, we include all surrogates in a kitchen-sink style linear regression to predict the long-term outcome. We do this both with only the surrogates (equation 6) and with the surrogates and the baseline covariates (equation 7).

$$Y_i^O = \beta S_i^O + \varepsilon_i \tag{6}$$

$$Y_i^O = \beta S_i^O + \gamma X_i^O + \varepsilon_i \tag{7}$$

5.2.2 Single surrogate linear regression

The next approach we follow uses only a single surrogate instead of all available surrogates in equations 6 and 7. We select this single surrogate by first correlating all available surrogates with the long-term outcome of interest. Then, we choose the surrogate with the highest absolute correlation. This is likely to give us the single most important surrogate. By comparing this approach with the kitchen sink approach, we can assess how much surrogate estimate bias comes from missing surrogates as opposed to missing surrogate-outcome confounders.

These first two approaches always produce predictions of the long-term outcome *even if* the surrogates are not predictive of the long-run outcome. By contrast, the approaches that follow utilise machine learning to select variables. As such, they might not select any surrogates if no surrogates are sufficiently predictive, meaning the surrogate estimate will be missing. This essentially means that only surrogate estimates above a certain 'quality' are produced. We argue that this generates a set of results that are more relevant to practitioners, as practitioners are unlikely to choose surrogates that are uncorrelated with their long-term outcome of interest.

5.2.3 Lasso

We now use lasso, a simple machine learning approach that adds variable selection and regularisation to the standard least squares regression to improve prediction accuracy. Whereas regular OLS solves the unconstrained problem, $\min_{\beta_0,\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 \right\}$, lasso instead solves the constrained problem:

[2] That is, we estimate all surrogate index estimates from β_{21} to β_{54} in table 2.

$$\min_{\beta_0,\beta}\left\{\sum_{i=1}^{N}(y_i-\beta_0-x_i^T\beta)^2\right\} \text{ subject to } \sum_{j=1}^{p}|\beta_j|\leq\lambda$$

.

This constraint on the sum of the absolute value of the regression coefficients shrinks the size of coefficients, forcing some of them to zero to avoid overfitting (Tibshirani, 1996). We further use the "rigorous" lasso introduced in Belloni et al. (2014) and implemented in the hdm R package (Chernozhukov et al., 2016). This method uses a data-driven approach for choosing the penalty level for the hyper-parameter λ. In particular, it sets:

$$\lambda = 2c\sqrt{n}\hat{\sigma}\Phi^{-1}(1-\gamma/2p),$$

where Φ is the cumulative standard normal distribution, $\hat{\sigma}$ is a preliminary estimate of $\sigma = \sqrt{\mathbb{E}\varepsilon^2}$, and c is a theoretical constant set to $c = 0.5$. Finally, γ is the probability level of mistakenly not removing X's when all of them have zero coefficients; this is set to $\gamma = 0.1$ (Chernozhukov et al., 2016).

5.2.4 M-lasso

Next, we introduce the *mediation lasso* (M-lasso), a lasso-based algorithm for selecting surrogates and covariates with similarities to the post double selection lasso of Belloni et al. (2014). This algorithm is a new contribution of this paper. Unlike the other prediction algorithms we test, it is specifically designed with the surrogate method in mind. In particular, it selects surrogates and covariates in a way that attempts to satisfy the surrogacy assumption as closely as possible.

The steps of the algorithm are as follows:

1. Using the observational dataset, run lasso of Y on S
2. Using the experimental dataset, run lasso of T on S
3. Take the union of the surrogates with non-zero coefficients from 1 and 2. Call this union $\tilde{S}$.
4. Using the observational dataset, run lasso of Y on X
5. Using the experimental dataset, run lasso of T on X
6. For each surrogate in $\tilde{S}$, run lasso of that surrogate on X
7. Take the union of the covariates with non-zero coefficients from 4, 5, and 6. Call this union $\tilde{X}$.
8. Run post-OLS of $Y = \beta\tilde{S} + \gamma\tilde{X} + \varepsilon$ to estimate the surrogate index

For the implementation of the M-lasso without baseline covariates, we simply skip steps 4-7 and run a post-OLS of $Y = \tilde{S} + \varepsilon$.

This approach aims to improve on the standard lasso approach in two ways. Firstly, by choosing surrogates that are predictive of treatment as well as the long-term outcome, we give ourselves a second chance at selecting mediators that lie on the causal path between treatment and the long-term outcome. This is analogous to how the post double selection lasso selects covariates that are predictive of either the treatment or the outcome.

Secondly, we now choose covariates that are predictive of the surrogates and long-term outcome, not just those that are predictive of the long-term outcome as in the standard lasso.

Recall from the causal graph in figure 2 that for the surrogacy assumption to be valid, we need to observe the confounders between the surrogates and the long-term outcome. By taking this approach, again similar to the post double selection lasso, we give ourselves two chances to select these potential confounders.

If the experiment is completely randomised, then in step 5 we should not find any covariates that are predictive of the randomised treatment. However, in cases where the assignment probability is different in different strata, this may select useful covariates.

5.2.5 XGBoost

XGBoost (eXtreme Gradient Boosting) is a supervised learning prediction algorithm based on boosted trees. This algorithm is known to perform well in supervised learning prediction competitions, so we include it to test if it also works well in causal inference questions. This algorithm uses boosting: it iteratively trains an ensemble of decision trees, with each iteration training a model to predict the residuals of the previous iteration's model. The final prediction is a weighted average of all the different trees' predictions. This process of boosting means that performance tends to improve in areas where previous iterations performed poorly.

We implement the XGBoost algorithm using the R package xgboost (Chen and Guestrin, 2016), using default parameters as far as possible. The one exception which has no default is the maximum number of boosting iterations (nrounds). To determine the optimal value of this parameter, we run a five-fold cross-validation where the maximum number of iterations is 1000, stopping once the performance does not improve for five consecutive rounds.

6 Results

We now present the results of our meta-analysis. In section 6.1 we summarise our meta-analysis on the average bias of each surrogate estimator. In subsequent sections we then focus on the performance of the M-lasso estimator, this method being a new contribution of our paper. We study the bias of the M-lasso estimator in sections 6.2 and 6.3. Section 6.4 further analyses the standard error of the M-lasso estimator. Finally, section 6.5 studies which surrogates were selected by the M-lasso procedure.

Further analysis on all estimators is available in Appendix B.

6.1 Overview of meta-analysis

Here we present our meta-analysis for the average bias of each surrogate estimator. We notate this bias as θ_i for causal effect i, and define it as follows:

$$\theta_i = \beta_i^{SI} - \beta_i \tag{8}$$

where β_i is the true long-term treatment effect, and β_i^{SI} is the limit of the surrogate estimator of this treatment effect.

However, in practice, we can only access feasible sample estimates of these objects. Thus, notating estimates with hats, we need to work with:

$$\hat{\theta}_i = \hat{\beta}_i^{SI} - \hat{\beta}_i^{RCT} \tag{9}$$

where $\hat{\beta}_i^{SI}$ is the surrogate estimate of the long-term treatment effect. Furthermore, $\hat{\beta}_i^{RCT}$ is an unbiased estimate of this treatment effect, which we compute using actual observed long-term outcomes from the RCT study. Finally, note that $\hat{\theta}_i$ is now the *estimated* bias of the surrogate estimator.

The meta-analysis allows us to deal with two issues. Firstly, the estimates we are working with have noise in them. To accommodate this, the meta-analysis lets us put more weight on more precise estimates. A second issue is that we have multiple estimates per study which are likely correlated with each other. This correlation can be incorporated into the way we set up the meta-analysis model.

The meta-analysis is structured as a study-level random-effects model:

$$\hat{\theta}_{ij} = \theta_{ij} + \varepsilon_{ij} \tag{10}$$

$$\theta_{ij} = \kappa_j + \varsigma_{(2)ij} \tag{11}$$

$$\kappa_j = \mu + \varsigma_{(3)j} \tag{12}$$

$\hat{\theta}_{ij}$ is the estimate of the true bias θ_{ij} for outcome i in study j. κ_j is the average bias in study j and μ is the overall average bias. $\varsigma_{(2)ij}$ is the within-study heterogeneity and $\varsigma_{(3)j}$ is the across-study heterogeneity. By substituting we can reduce the model to:

$$\hat{\theta}_{ij} = \mu + \varsigma_{(2)ij} + \varsigma_{(3)j} + \varepsilon_{ij} \tag{13}$$

We estimate this model with robust standard errors clustered at the study level. Results are contained in table 3. Interpreting this table, mean bias is the estimated average bias of the surrogate index approach, μ, from equation (13). SE are the standard errors on this mean. SD is the square root of the sum of the within and between variation, $Var(\varsigma_{(2)ij})$ and $Var(\varsigma_{(3)j})$ respectively. RMSE is the square root of the mean-squared error of the estimator. This analysis is done on the set of outcome-surrogate-treatment combinations for which all estimators selected at least one surrogate variable and produced a treatment effect estimate (n=743).

Note also that from here on out, when we refer to absolute (or mean) bias with no further qualifications, we are referring to μ by default.

There are three main results to highlight from table 3. First, we compute the predictive accuracy of each estimator by using the root mean squared error: $RMSE = \sqrt{\text{Mean bias}^2 + \text{Variance}}$. Against this metric, Lasso with baseline covariates performs best (RMSE = 0.101), followed very closely by M-lasso with no covariates (RMSE = 0.102).

A second main result from table 3 is the mean bias is always negative for all estimators. It is also often statistically significantly different from 0 at a 95% level of confidence, with the only exceptions being the 'Lasso' method of the surrogate index estimator. An average value of -0.057, when normalised by dividing by the standard deviation of the outcome variable in the control group, means that the surrogate index tends to underestimate the treatment effect from the long-term RCT by 5.7% of a standard deviation. The mean treatment effect from the RCTs in these contexts is 10.5% of a standard deviation. So, on average, the surrogate index estimator produces estimates that are little more than half of the true long-run effect. Although this is a large relative difference, 5.7% of a standard deviation is a only small absolute difference in most cases. However, this is just the *mean* bias. Given that the bias distribution has a standard deviation of roughly 0.13, some biases will be more negative and economically significant, while others will be 0 or positive. In section 6.2.2, we explore cases where the bias is large and negative and explain why this may occur.

The third main result is that for each estimator, the version *without* baseline covariates almost always performs better than the version with baseline covariates included. The mean bias, standard deviation, and RMSE are lower for almost every estimator in panel A relative to the same estimator in panel B. The only exception is the XGB estimator, whose standard deviation and RMSE are lower when baseline covariates are included. Regardless, the general result of the surrogate index method performing better when covariates are

Table 3: Meta-analysis of bias of different estimators

Estimator	(1) Mean bias	(2) SE	(3) SD	(4) RMSE
Panel A: No baseline covariates				
Single surrogate	-0.059*	0.030	0.093	0.110
Linear regression	-0.049	0.025	0.093	0.105
Lasso	-0.051	0.028	0.090	0.104
M-lasso	-0.050*	0.025	0.089	0.102
XGB	-0.057*	0.028	0.100	0.115
Panel B: With baseline covariates				
Single surrogate	-0.069*	0.028	0.114	0.134
Linear regression	-0.064*	0.025	0.114	0.131
Lasso	-0.054	0.028	0.085	0.101
M-lasso	-0.064*	0.025	0.110	0.127
XGB	-0.060*	0.028	0.095	0.113

Mean bias is the estimated average bias of the surrogate index approach, μ, from equation (13). SE are the standard errors on this mean. SD is the square root of the sum of the within and between variation, $Var(\varsigma_{(2)ij})$ (equation (11)) and $Var(\varsigma_{(3)j})$ (equation (12)) respectively. RMSE is the square root of the square of the mean bias plus the square of SD. An asterisk represents that the mean bias statistically significantly different from zero at a 95% level of confidence. This analysis is done on the set of outcome-surrogate-treatment combinations for which all estimators selected at least one surrogate and produced a treatment effect estimate (n=743). Across estimators, the mean bias is around -0.059, though this estimate is not always statistically significantly different from zero. For reference, the mean RCT treatment effect is 0.105 with standard error 0.049.

omitted is somewhat surprising. After all, according to the theory, we would expect the surrogacy approach to be violated by surrogate-outcome confounding. We explore this finding in further detail in section 6.2.3 and propose a potential explanation as to why it occurs.

In the rest of this section, we will focus on the M-lasso estimator as our representative example of the surrogate index method. We will therefore use the terms *M-lasso estimates* and *surrogate index estimates* interchangeably, unless otherwise specified. For completeness, however, the main graphs are replicated for all types of surrogate estimators in appendix B.

We focus on the M-lasso estimator because it is the only estimator specifically designed with the surrogate estimation procedure in mind, as well as being a new contribution that warrants further study. We specifically examine why there is bias even in the 'best-performing' version of the M-lasso in our meta-analysis, that is, the M-lasso *without* covariates. Our discussion therefore will relate to this form of the M-lasso from here on out, unless otherwise specified.

Based on the RMSEs in table 3, we can see that the M-lasso without covariates performs best across the estimators that exclude baseline covariates. However, the M-lasso with covariates does not stand out as much compared to other estimators that include baseline covariates.

6.2 Bias

In this section, we will explore the estimates underlying the meta-analysis in table 3.

Section 6.2.1 details the existence of bias in our surrogate estimators, while section 6.2.2 provides some potential explanations for this bias.

Section 6.2.3 then returns to the surprising result that covariate omission leads to smaller surrogate bias in our data. We propose a potential explanation: covariate omission may have generated a positive bias, cancelling out the negative bias from missing surrogates. This at least holds in our observed data, but we do not make strong claims about its generalisability.

6.2.1 Raw bias distribution

To begin, we look at the distribution of the raw estimated biases $\hat{\beta}^{SI} - \hat{\beta}^{RCT}$ in figure 3. These estimates include sampling error, so the observed variance of the estimated bias distribution will be greater than the true variance of the bias distribution.

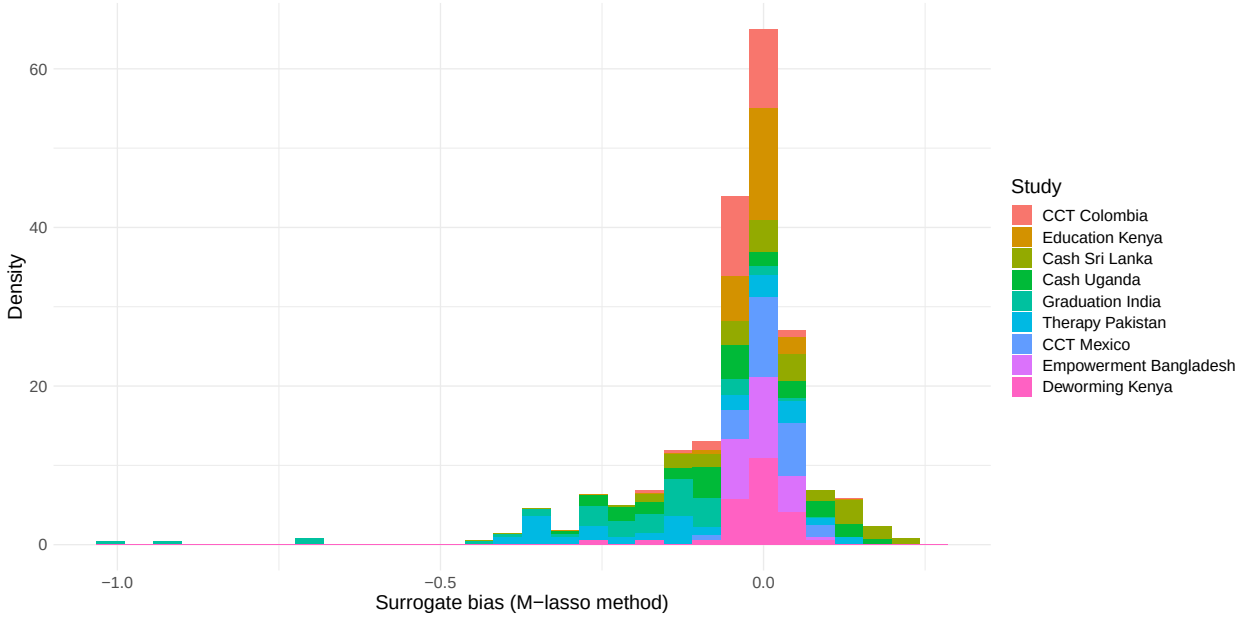
In figure 3, we can see that the bias distribution is centred slightly to the left of 0 and there is more mass on the negative, left-hand side of the distribution. There is a long tail in the negative part of the distribution. This indicates that the surrogate method is negatively biased.

We also show the same information in figure 4 by plotting the raw estimates from the M-lasso method against the raw RCT estimates that the M-lasso is trying to replicate. This is done separately for each dataset.

In figure 4, the blue line of best fit would ideally be on the black 45-degree line, as this would indicate that the surrogate index estimates were the same in expectation as the RCT estimates. However, in practice, we see that the line of best fit is never close to the 45-degree line. In 8 out of 9 studies, the line of best fit is still upwards-sloping, but shallower than the 45-degree line and attenuated towards zero. This suggests that the surrogate index estimates are themselves attenuated towards zero. To illustrate, note that the average slope of the lines is 0.3. This means that if we had a true positive treatment effect of (say) 0.3 standard deviations, in expectation the surrogate index estimates would be 0.09 standard deviations.

However, there could be an alternative explanation for why the line of best fit is attenuated. In particular, the slope of the line of best fit is the estimated OLS coefficient $\hat{\gamma}$ from a regression

Figure 3: Distribution of bias in raw surrogate index estimates (M-lasso method)



Note: Across nine RCT contexts, the bias distribution is centred slightly to the left of 0. There is more mass and a long tail in the negative, left-hand side of the distribution. This indicates that the surrogate method is negatively biased.

of $\hat{\beta}^{SI} = \alpha + \gamma\hat{\beta}^{RCT} + \varepsilon$. Yet $\hat{\beta}^{RCT}$ is only an estimate of the true β^{RCT}, and thus has random measurement error. Therefore, the OLS estimate $\hat{\gamma}$ exhibits attenuation bias relative to the true γ.[3] This might instead be responsible for our observed attenuation.

To rule out this possibility, in figure 5, we plot the raw estimated *bias* of the surrogate estimator against the raw RCT estimate of the treatment effect. In this graph, the line of best fit would ideally lie on the horizontal black line, indicating zero bias on average. Furthermore, if attenuation bias were just stemming from measurement error in $\hat{\beta}^{RCT}$, the slope of the line of best fit should be attenuated towards 0 in figure 5. This would bring the line of best fit closer to the ideal horizontal line, making the surrogacy approach look better than it is.[4]

In practice, however, the line of best fit is typically far from horizontal in figure 5. Instead, it is much closer to a downwards-sloping 45-degree line. Hence, figure 5 means we can be confident that our observed attenuation is not purely an artifact of measurement error. Note that the observed line of best fit implies that the more positive the RCT effect, the more negative the bias in the surrogate estimate. This also matches the observed negative mean bias in the meta-analysis.[5]

6.2.2 Why underestimates?

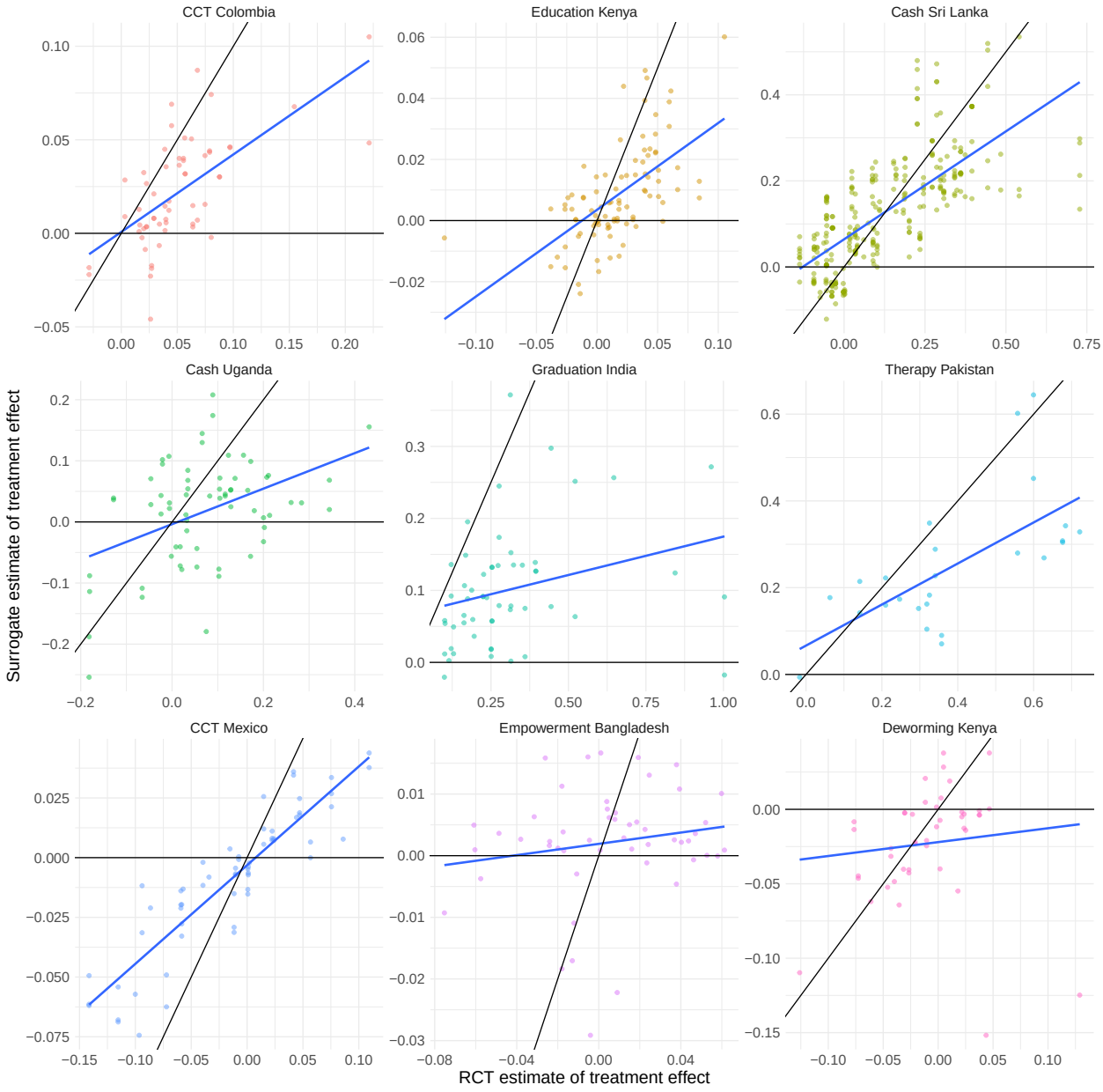
Why does the surrogate index approach tend to underestimate the true effect? One explanation is that we are missing surrogates, i.e. we are missing some causal pathways from treatment to long-term outcome, violating the surrogacy assumption. To further explain why the bias

[3]We also know that $\hat{\beta}^{SI}$ is an estimate of β^{SI} and so also has measurement error. However, measurement error in the LHS dependent variable only reduces the precision of the estimated coefficient; it does not similarly lead to attenuation bias.

[4]The opposite was previously true in figure 4, where attenuation bias pushed the line of best fit closer to horizontal and further away from the ideal 45-degree line.

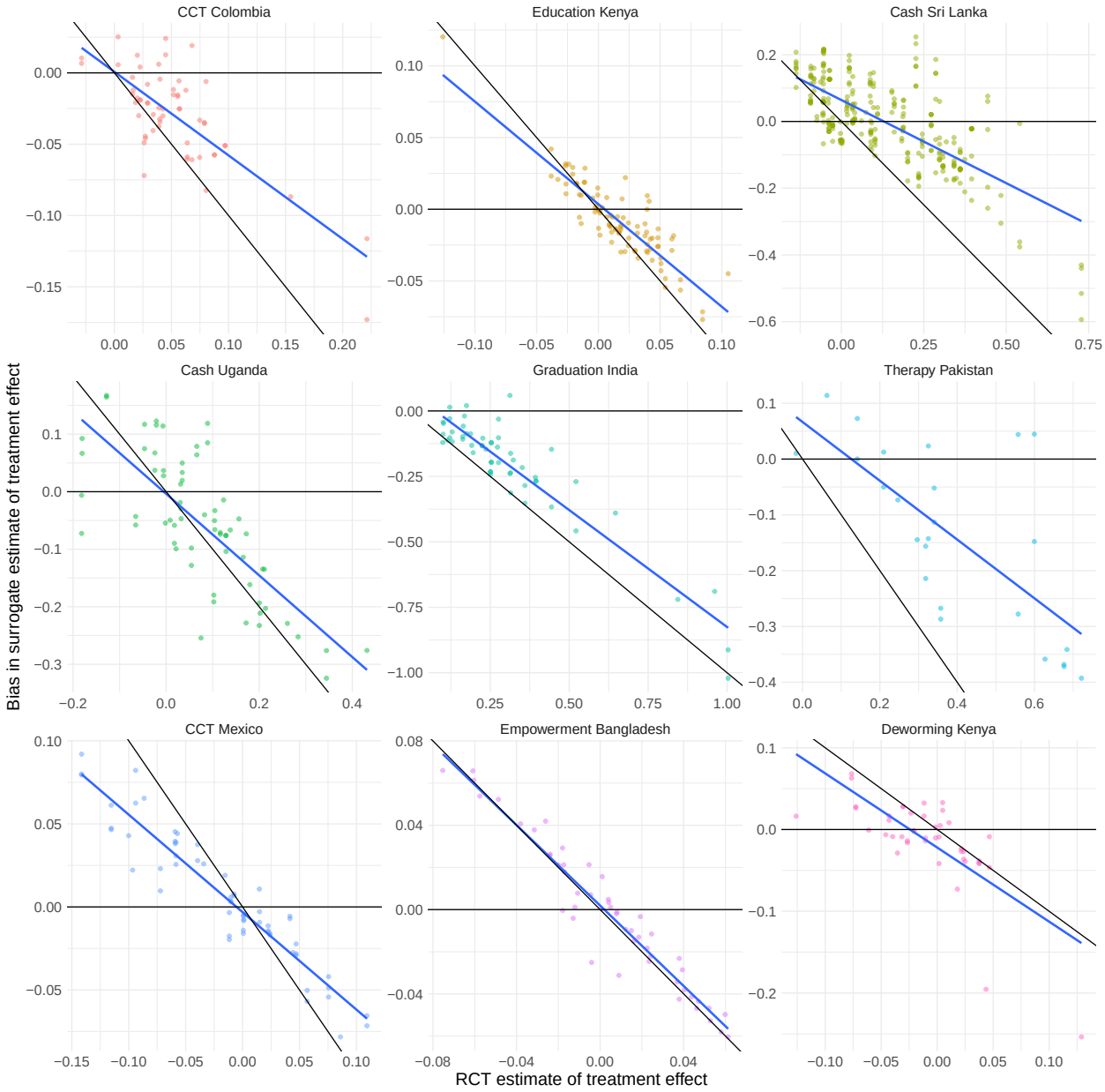
[5]The meta-analysis also deals with the issue of noise in the estimates; it does so by giving more weight to more precise estimates.

Figure 4: Normalised surrogate index raw estimates (M-lasso method) against normalised RCT raw estimates



Note: This graph plots surrogate estimates of the treatment effect (y-axis) against their corresponding RCT estimates (x-axis). Each RCT estimate is associated with multiple surrogate estimates because varying numbers of waves of surrogates were used for each treatment-outcome pair, as represented in table 2. The black line is the 45-degree line and the blue line is the line of best fit. All graphs share the same x and y-axis labels.

Figure 5: Normalised raw surrogate bias (M-lasso method) against normalised raw RCT treatment effect estimates



Note: The blue lines are the lines of best fit for surrogate bias against RCT treatment effect estimate across nine RCT contexts. The slope of this line is far from zero, indicating bias is not zero on average. The best fit is much closer to the 45-degree line, indicating increasing negative bias as RCT effects become more positive. Note that each RCT estimate is associated with multiple surrogate index biases because varying numbers of waves of surrogates were used to create surrogate index estimates, as represented in table 2. All graphs share the same x and y-axis labels.

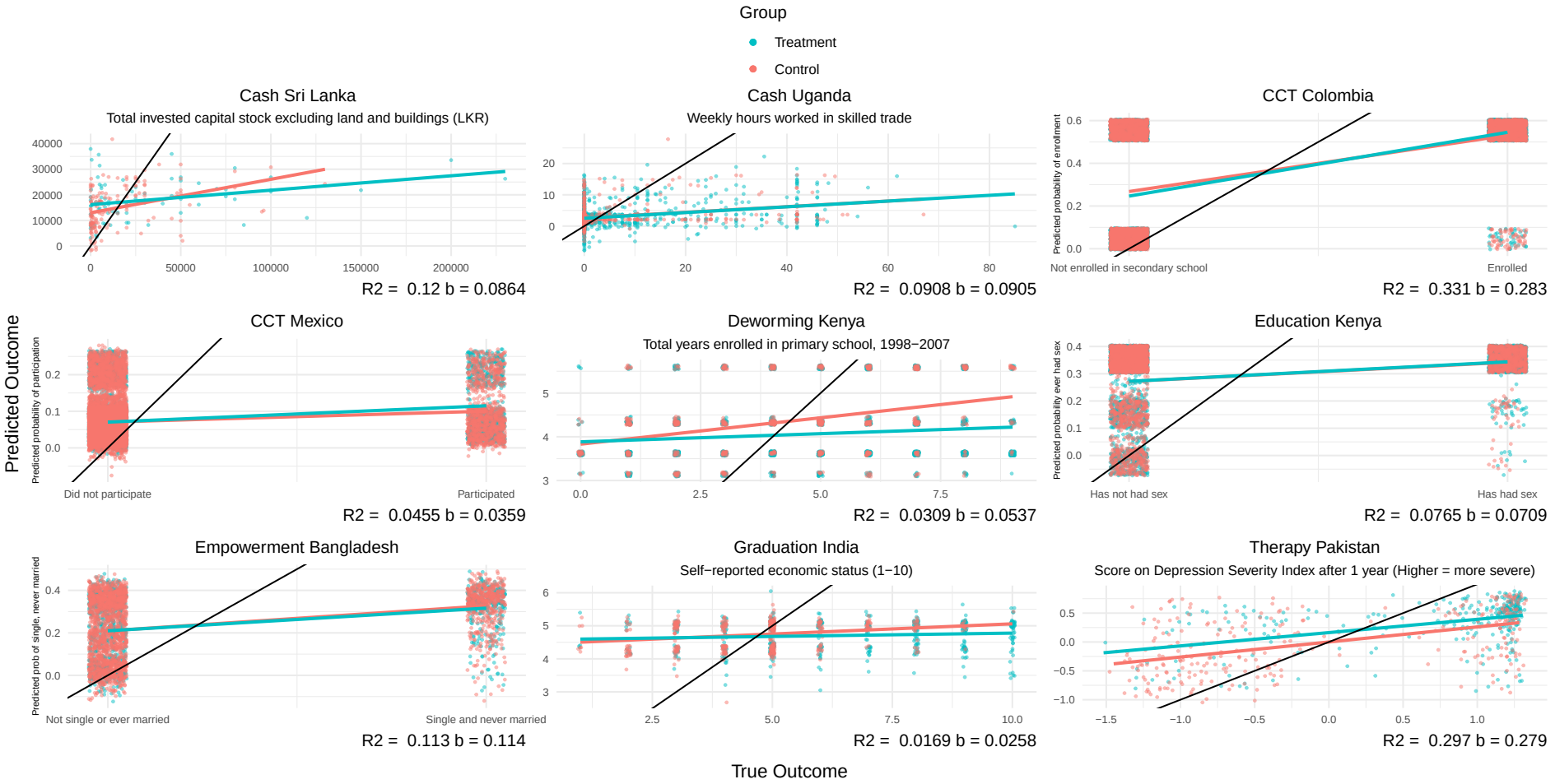
would be negative rather than positive, it must be the case that the signs of the missing causal pathways are positively correlated with the signs of the observed causal pathways.

Take the deworming study from Hamory et al. (2021) as an example. Suppose we had an observable surrogate, education, and an unobservable surrogate, socio-emotional skills. Further, suppose that we expect the observed effect of deworming on income via education to be positive, and we also expect the unobserved effect via socio-emotional skills to be positive. If so, then by omitting the unobserved surrogate (socio-emotional skills), we only observe part of the total effect on the long-term outcome.

However, we further suggest that surrogate estimates could be negatively biased *even if* all surrogates are observed, due to additional bias in the first-stage predictive model.

To illustrate, for each of the nine RCTs, we consider the least-biased uses of the surrogate index estimator using the M-lasso method. In figure 6, for continuous variables, we plot the predicted value of these outcomes against the true outcomes. For binary variables, we plot the prediction for the probability of the outcome against the true outcome. The figure shows that the predictions are attenuated towards the mean for both the treatment and control groups.

Figure 6: Predicted vs true long-run outcomes



Note: Each point on this graph plots an individual's true long-term outcome (x-axis), plus the corresponding predicted outcome from the first stage of the surrogate index approach (y-axis). R2 is the R^2 from a regression of the predicted outcome on the true outcome, and b is the coefficient on the true outcome in this regression. The black line is the 45-degree line; for points on this line, the true and predicted outcomes would be the same.

In particular, note that if the predictions were perfect, the lines of best fit for the treatment and control group would lie on the 45-degree line. However, we see that the slope of the line of best fit lies between 0.025 and 0.283 across the 9 outcomes under consideration, with a mean slope of 0.115. We interpret this to mean that on average, the predicted values or probabilities for the outcomes are approximately 12% of what they would be if the estimator was unbiased. The prediction algorithm fails to give unbiased predictions for the tails of the outcome distribution, instead pulling its predictions towards the mean. Note that this also reduces the variance in the predicted outcomes; the algorithm will trade off bias and variance in this way to minimise MSE loss.

Figure 7 helps illuminate why this predictive bias leads to a negative bias in the treatment effect estimation. To start with, observe the treatment and control group distributions of the true outcomes in figure 7. We can see that in each case, the true treatment group distribution is shifted to the right of the control group distribution. In particular, take the dotted blue line, the mean of the treatment group distribution. In the true outcomes, this is always to the right of the dotted red line, the mean of the control group distribution. This makes sense as we expect the treatment to result in better outcomes for the treatment group.

However, *predicted* outcomes for treated individuals are pulled to the center of the overall distribution. Contrast the means of the predicted outcomes of the treatment and control groups. The predicted control outcomes tend to have a similar mean to the true control outcomes (represented by red dotted lines). However, the predicted *treatment* group outcome means are smaller than in the true outcome histograms (represented by a blue dotted line), often pulled towards the mean outcome of the joint (treatment *and* control) sample. Sometimes, it is even pulled completely to the left of the mean predicted outcome of the control group. This negative bias in predicted treatment outcomes then leads to a negative bias in the overall treatment effect estimate.

We expect the negative bias in these figures to be something of an optimistic indication of just how biased the M-lasso method can be. This is partially because we selected the least biased versions of the method, suggesting that every other choice of outcome variable in each RCT context will yield even greater bias. Moreover, in these examples, we used the control group data to train the prediction algorithm. As such, the mean of the predicted control group outcome is the same as the mean of the true control group outcome. Had we instead used another independent data source to train the prediction algorithm, we would also expect the control group predictions to be pulled to the center of the overall distribution. This would raise the mean predicted outcome for the control group, further increasing the negative bias of the surrogate index approach.

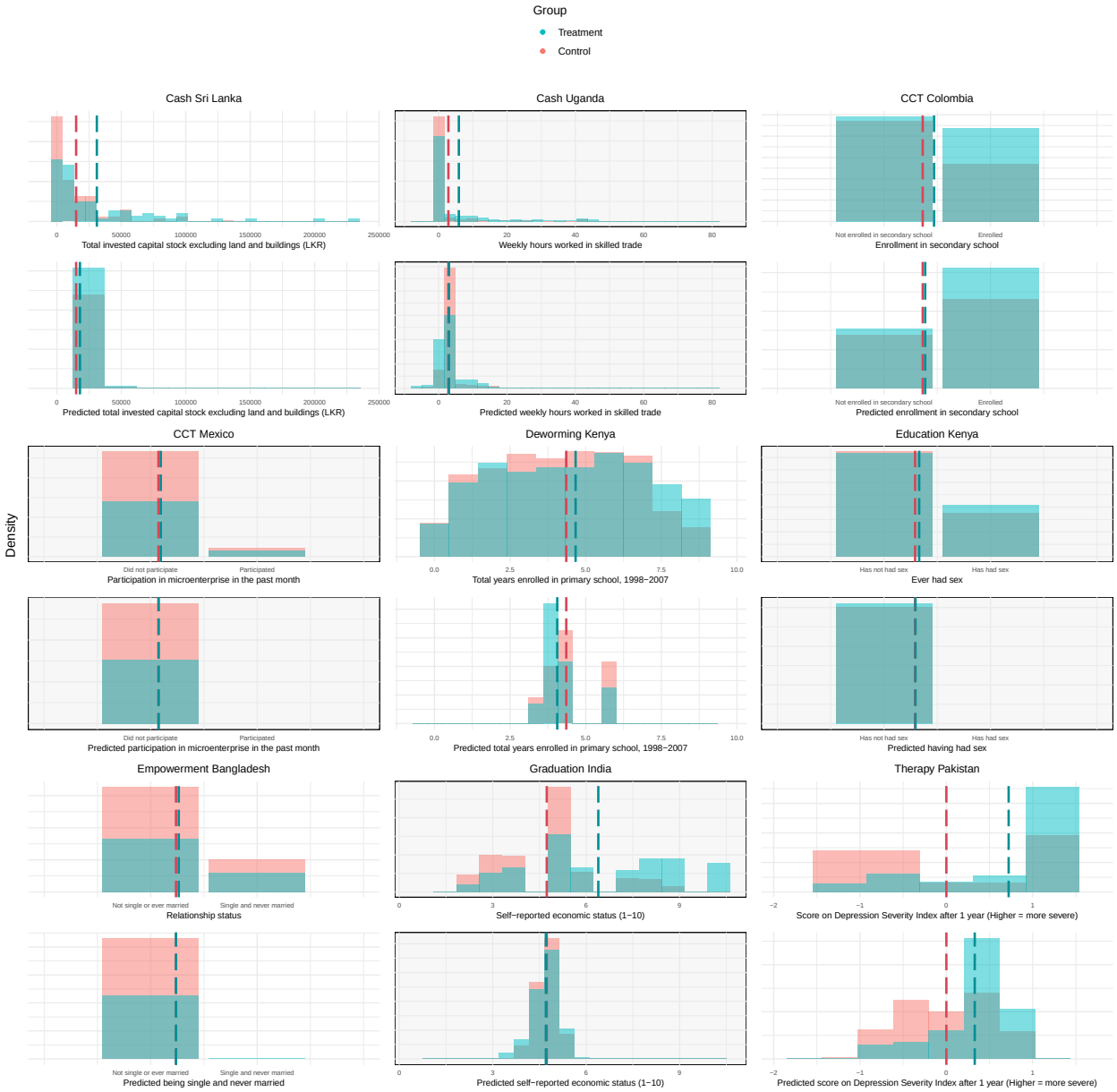
6.2.3 With or without baseline covariates?

We now explore the third main result from the meta-analysis in table 3, which is that the versions of the estimators without baseline covariates systematically do better than those with covariates. To re-examine this, we run a meta-analysis on the M-lasso mean biases and absolute biases, summarised in table 4. Here, the absolute bias is defined as $|\hat{\beta}_i^{SI} - \hat{\beta}_i^{RCT}|$. With the absolute bias, positive and negative biases do not cancel out, so this gives different information to the mean bias.

The meta-analysis in table 4 is computed for estimators both with and without covariates. We also include an indicator for whether the biases come from a version of the M-lasso estimates with covariates included. This is incorporated as a moderator in the meta-analysis.

To start with, we replicate the mean bias of the M-lasso from table 3. As before, including covariates makes the bias more negative: the coefficient on the indicator for covariate inclusion (-0.015) is negative and statistically significant. This result holds whether we include study

Figure 7: Examples of predicted vs true distributions for long-run outcomes



Note: For each of the nine RCT contexts, the true treatment effect appears more positive than the predicted treatment effect. Though it is harder to see in the cases with binary outcome variables, the distributions of predicted outcomes for both the control and treatment groups are closer to the centre of the joint distribution. As such, the predicted treatment effects are smaller than the true treatment effects. We speculate that these trends occur because predicting outcomes closer to the mean reduces prediction error on average. All graphs have density for their y-axis.

Table 4: Meta-analysis of bias with and without baseline covariates

	(1) Bias	(2) Bias	(3) Absolute Bias	(4) Absolute Bias
Intercept	-0.050*		0.081*	
standard error	(0.025)		(0.023)	
Covariates included	-0.015*	-0.015*	0.020	0.020
standard error	(0.007)	(0.007)	(0.021)	(0.021)
Study FEs	No	Yes	No	Yes

Our meta-analysis suggests that including covariates weakly increases the amount of bias we observe. The asterisks represent statistical significance at the 5% level. The relationships between biases and including covariates are not affected by the inclusion or exclusion of study fixed effects in the meta analysis.

fixed effects or not.

Next, we meta-analyse the absolute bias. We find that, on average, when covariates are included, the absolute bias is 0.02 higher, a 25% increase. However, this difference is not statistically significant, with or without study fixed effects.

All in all, our results suggest that across our RCTs, deliberately omitting the observable baseline covariates tends to add positive bias to the surrogate estimates. Since these surrogate estimates are negatively biased on average, this has the effect of cancelling out some of that negative bias. Positive bias from covariate omission also seems intuitively plausible. As an illustrative example, consider the setting from Barrera-Osorio et al. (2019), an RCT on conditional cash transfers that we included in our meta-analysis. Here our long-term outcome was tertiary enrollment, a surrogate was high school enrollment, and a (observable) baseline covariate was household wealth.

Now, consider two possible regressions:

$$tertiary\_enrollment = \alpha_1 + \beta_1\ high\_school\_enrollment + \varepsilon_1$$
$$tertiary\_enrollment = \alpha_2 + \beta_2\ high\_school\_enrollment + \gamma\ household\_wealth + \varepsilon_2$$

The former regression deliberately omits the covariate relating to household wealth. However, we expect household wealth to be positively correlated with both high school enrollment and tertiary enrollment. Therefore, we expect β_1 to have positive omitted variable bias, such that $\beta_1 > \beta_2$.

As such, if we used β_1 rather than β_2 to make our predictions of the long-term outcome, we would overestimate the relationship between our surrogate and our long-term outcome. This would lead to positive omitted variable bias in the surrogate estimate of the long-term treatment effect. In turn, this could cancel out some existing negative surrogate bias, decreasing the magnitude of the overall surrogate bias.

Note that we cannot say for sure whether this finding of a positive omitted variable bias would generalise to:

1. *Unobserved* covariates in our studies, which we are forced to omit across all specifications. By definition, we have no information on unobserved covariates, so we cannot identify the extent of the bias by omitting them.

2. Other studies that we did not cover in this meta-analysis.

This is essentially a question of external validity. However, the generalisability of our finding might be improved by the fact that our meta-analysis covered a wide variety of different RCTs.

As such, suppose we expected there to be some existing negative bias in our surrogate estimates - due to missing surrogates, or bias in the first-stage predictive model. One could then deliberately leave out covariates when estimating the surrogate index, hoping to induce a positive bias that partially cancels out the existing negative bias. However, this approach of relying on two biases of opposite directions to cancel out is likely to be fragile in practice. We do not recommend it.

6.3 Determinants of bias

Next, we look at what influences the size of the surrogate bias. In particular, sections 6.3.1 to 6.3.3 below consider the following factors respectively:

1. The time horizon between the surrogates and the long-term outcome;
2. The strength of the correlation between the surrogate index and the true long-term outcome;
3. The complexity of the intervention.

6.3.1 Time horizon

In this section we study an *a priori* key determinant of the bias: the time in years that elapses between the surrogates and the long-term outcome. The more time there is between the surrogates and the long-term outcome, the less likely the surrogates are to cover the key causal pathways between the treatment and the outcome, and the worse the prediction of the long-term outcome is likely to be. As such, our prior is that an increasing time horizon between the surrogates and the long-term outcome leads to a larger violation of the surrogacy assumption, hence larger surrogate bias.

For this analysis, we take advantage of the fact that we have many post-treatment waves, as shown in table 2. We look at the final wave of outcomes from each RCT study and use every possible cumulative combination of waves to derive our sets of surrogates (i.e. we compute all the estimates from row 5 in table 2: β_{51}, β_{52}, β_{53}, and β_{54}). This lets us examine the relationship between time horizon and bias. Table 5 presents the results.

In the table, Horizon, included as a moderator, represents the gap in years between the final outcome and the latest surrogates used in that estimation. Experiment FEs are fixed effects for each combination of treatment and outcome variables, for each implementation of the surrogate index method. Including these fixed effects allows us to better isolate the relationship between time horizon and bias. Both mean bias and absolute mean bias are reported, though we focus on interpreting the absolute mean bias.

We find that the coefficients on Horizon in table 5 are positive but not statistically significant at the 5% level. This evidence is therefore somewhat inconclusive and does not strongly support our prior. Further graphs of this relationship are available in appendix E.

Interpretation of this result is made somewhat more difficult by the selection of studies in which we have long-run results available. Typically, long-run follow-ups are run if researchers find significant short-run effects which are also expected to persist over time. However, such characteristics would also make it easier to predict long-term outcomes using short-term data.

Table 5: Meta-analysis of bias and time horizon

	(1) Bias	(2) Bias	(3) Absolute Bias	(4) Absolute Bias
Intercept	0.016		0.010	
standard error	(0.020)		(0.022)	
Horizon (Years)	-0.008	-0.008	0.009	0.006
standard error	(0.005)	(0.006)	(0.005)	(0.005)
Experiment FEs	No	Yes	No	Yes

This meta-analysis studies the surrogate index method's bias for treatment effect estimates on the longest-run outcomes of nine RCTs. Absolute bias seems to increase as the years between the observation of the latest surrogate variable and the final outcomes. Experimental fixed effects, which capture the fixed effects for each combination of treatment and outcome variables in each implementation of the surrogate index method, appear to have no effect. However, no estimates are statistically significant at the 5% level.

If these kinds of studies are over-represented in our meta-analysis, we may be overstating the effectiveness of the first-stage predictive model. In turn, this would lead us to understate the extent to which larger time horizons increase surrogate bias for the typical RCT.

6.3.2 First-stage prediction accuracy

Next, we examine the relationship between the predictive accuracy of the first-stage prediction model and bias in the final surrogate estimate of the treatment effect.

To measure first-stage predictive accuracy, we compute the correlation between the predicted and true long-run outcomes in the observational sample. We henceforth refer to this as the *observational correlation*.

Next, we compute the correlation between both types of bias. We find that observational correlations vary widely in our sample, with a minimum of 0.016 and a maximum of 0.92. The mean and median correlations are both around 0.45. The interquartile range is 0.27 to 0.62.

In table 6, we then meta-analyse the relationship between this first-stage observational correlation and bias in the final surrogate estimator. Overall, we find that all of the coefficients on the first-stage correlation are small and insignificant. This holds whether we look at the mean or absolute surrogate bias and whether we include study fixed effects. This suggests that the predictive accuracy of the first stage does not matter much for the usefulness of the surrogate index method for estimating long-term effects.

To understand this result, recall that the key focus of the first stage of the surrogate method is to capture variation in the long-term outcome that is influenced by treatment. That is, the correlation between the predicted and actual outcomes in the observational sample is only beneficial to the extent that it reflects causal pathways from treatment to outcome. If the first-stage predictive model selects only variables that are orthogonal to treatment, then we would expect the first stage correlation to be high but the second-stage accuracy to be low. The performance of the surrogate index approach ultimately depends on whether it can capture the causal pathways from treatment to outcome, and this cannot be

Table 6: Meta-analysis of bias and first stage predictive accuracy

	(1) Bias	(2) Bias	(3) Absolute Bias	(4) Absolute Bias
Intercept	-0.058		0.071*	
standard error	(0.035)		(0.031)	
Correlation	0.023	-0.006	0.027	-0.003
standard error	(0.064)	(0.013)	(0.038)	(0.015)
Study FEs	No	Yes	No	Yes

Notes: The relationship between first stage predictive accuracy and bias is small and statistically insignificant at the 5% level, regardless whether study fixed effects are accounted for. First stage predictive accuracy is measured by the correlation, in the observational sample, between the true long-term outcome and the surrogate index's prediction of the long-term outcome. An asterisk represents statistical significance at 5%.

assessed with a simple correlation.

6.3.3 Intervention complexity

We now look at whether the complexity of an intervention affects the bias of the surrogate index approach. Simple interventions are defined as those that have one component that is the same for all participants. Complex interventions are those that either have multiple components, or components that are not standard across individuals. We test one plausible prior, which is that simple interventions have fewer causal pathways from treatment to long-run outcome.[6] As such, the surrogate index approach should be able to capture a larger fraction of the causal pathways and therefore perform better. By contrast, complex interventions are likely to have many causal pathways, more of which are likely to be unobserved, resulting in worse performance. However, this may be offset to some extent by the fact that RCTs with more complex treatments may collect a wider variety of outcomes to capture the more complex effects. Furthermore, as complexity only varies at the intervention level, we only have 16 effective data points here. Hence, our analysis is likely to be underpowered.

Following the above definitions, we classify the 16 treatments from the 9 different RCTs as simple or complex. These are enumerated in table 7. We then meta-analyse the surrogate biases in table 8, including complexity as a moderator.

In column (3) of table 8, we find that simple interventions have an average absolute bias of 0.059, whereas the average absolute bias of complex interventions is 0.048 higher, totalling 0.107. Meanwhile, in column (1) of the table, simple interventions have a smaller-magnitude average mean bias of -0.016, and adding complexity increases the magnitude of the bias to -0.093. However, none of these differences in bias are statistically significant, likely due to the limited number of distinct interventions in the analysis sample.

[6]We think this prior is a reasonable starting point for most cases, though we also acknowledge that there could be exceptions to this. For instance, cash transfers are a relatively simple intervention, but they might operate through many channels.

[7]Either with or without the free uniform grant.

[8]Either with or without the financial incentive.

Table 7: Treatments categorised by simple or complex

Simple interventions
Unconditional cash transfers from Cash Sri Lanka and Cash Uganda
Conditional cash transfer from CCT Mexico
Basic conditional cash transfer from CCT Colombia
Free uniform grant from Education Kenya
Financial incentive to delay marriage from Empowerment Bangladesh
Deworming pills from Deworming Kenya
Complex interventions
Savings and incentive conditional cash transfers from CCT Colombia
HIV education program from Education Kenya[7]
Graduation program from Graduation India
Cognitive behavioural psychotherapy from Therapy Pakistan
Empowerment program from Empowerment Bangladesh[8]

Table 8:
Meta-analysis of bias and intervention complexity

	(1)	(2)	(3)	(4)
			Absolute	Absolute
	Bias	Bias	Bias	Bias
Intercept	-0.016		0.059*	
	(0.011)		(0.017)	
Complex	-0.077	0.007	0.048	-0.003
	(0.042)	(0.008)	(0.042)	(0.015)
Study FEs	No	Yes	No	Yes

Notes: The relationship between bias and intervention complexity is small and statistically insignificant at the 5% level, regardless of whether study fixed effects are accounted for. Complex interventions are those that either have multiple components, or components that are not standard across individuals. An asterisk represents statistical significance at 5%.

We also see that when we include study fixed effects, our coefficients drop to approximately 0. In this case, the identifying variation comes from Empowerment Bangladesh and Education Kenya, the only studies with at least one simple and one complex intervention. These preliminary findings point to more complex studies having larger biases, but the limited sample size of interventions warrants further investigation with a broader range of studies and interventions.

6.4 Standard errors

In this section, we test whether the surrogate estimates in our meta-analysis have greater accuracy than long-run RCT estimates. The intuition here is as follows. If the surrogacy assumption is true, then the observed surrogates explain all the variation in the long-term outcome caused by the randomised treatment. However, the long-term outcome also contains variation from other sources. This random noise is irrelevant for estimating the treatment effect and merely reduces precision.

The first stage of the surrogate estimator then predicts the long-term outcome using the surrogates. This isolates variation in the outcome that is related to the surrogates, discarding any further random noise. As such, the variance of the predicted long-term outcome is less than the variance of the true long-term outcome (c.f. figure 7). In turn, this would make the surrogate index estimate more precise than the long-term RCT estimate.

Of course, if the surrogacy assumption is not valid, then we risk introducing some bias by incorrectly assuming it to be true. In this case, the surrogates would only predict *part* of the variation in the long-term outcome caused by the treatment. Some further relevant variation would be discarded, along with the random noise.

In figure 8, we plot the standard error from the long-term RCT against the standard error from the M-Lasso surrogate index approach. Almost all points lie below the black 45-degree line, implying that the surrogate index standard errors are typically lower than the RCT standard errors. Indeed, across all outcomes, the surrogate index standard error is on average 52% of the RCT standard error. This suggests that there are substantial precision gains to be made from using the surrogate index approach in cases when the surrogacy assumption is satisfied. Such precision gains could motivate a researcher to use the surrogate approach even if they already have access to long-term outcomes.

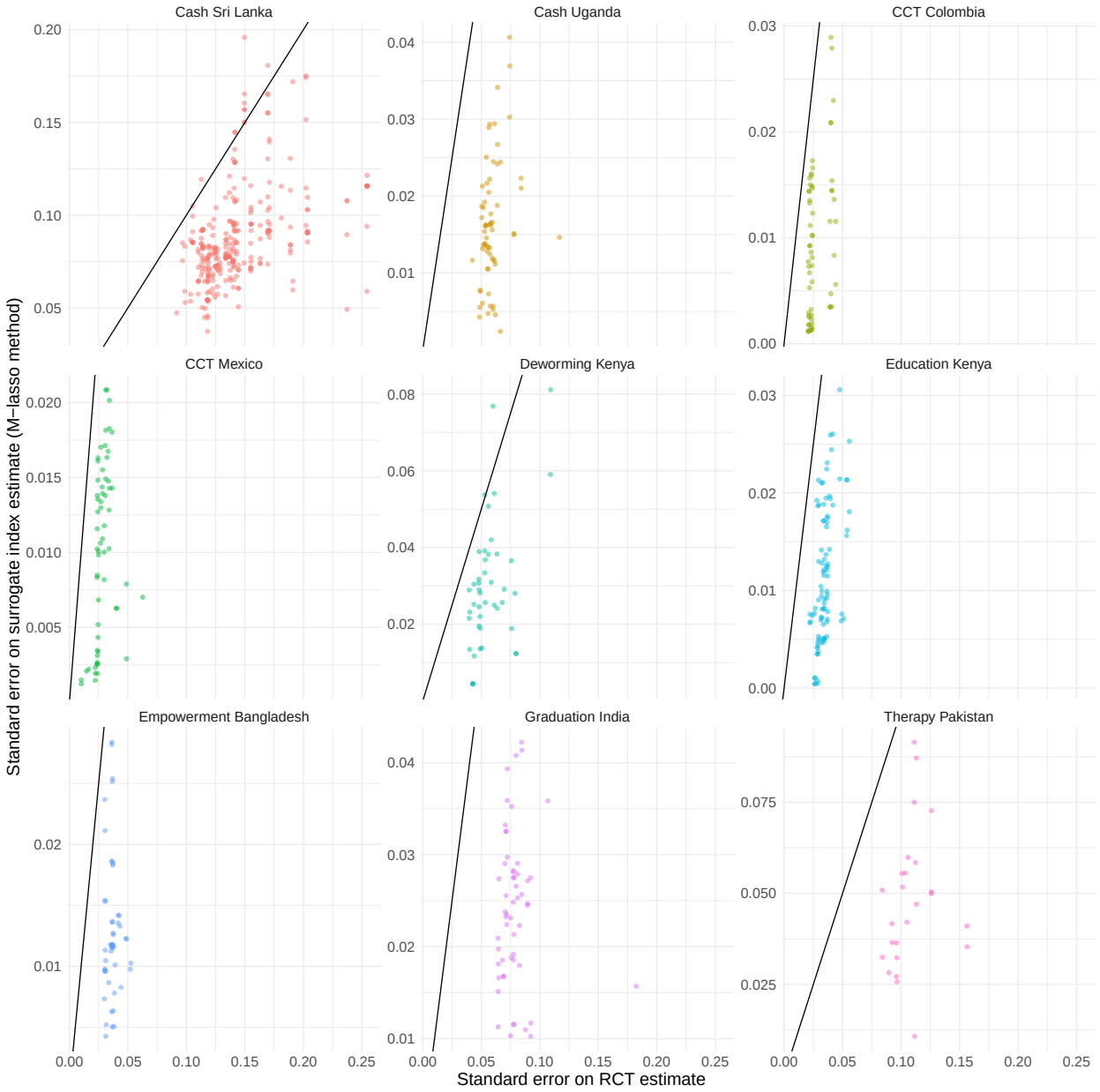
However, in practice, the surrogacy assumption is rarely satisfied exactly. If so, using the surrogacy approach introduces a negative bias. This means that the researcher faces a bias-variance trade-off: they could improve their precision by using a predicted long-term outcome instead of the true long-term outcome, but this will also generate some bias. Whether this is desirable depends on the researcher's exact loss function. For instance, this might be acceptable if the researcher is minimising an MSE criterion. On the other hand, this is less likely to be acceptable in settings where the focus is on obtaining unbiased estimates of treatment effects.

Note that the case for the surrogate approach could also be strengthened if the researcher is happy to subsequently 'debias' their surrogate index estimates - that is, to subtract the empirically estimated mean bias of the surrogate index estimator (-0.05). However, this is only feasible if the researcher is willing to make an exchangeability assumption between their study and the studies in this analysis.

6.5 Surrogates selected by M-lasso

This section describes the characteristics of surrogates the M-lasso method picks for the primary long-term outcome of each study, for each treatment. We combine these findings

Figure 8: Normalised surrogate index standard errors (M-lasso method) against normalised RCT standard errors



Notes: Each point plots the standard error of an RCT treatment effect estimate (x-axis) and a corresponding standard error of a surrogate index treatment effect estimate (y-axis). Each RCT standard error is associated with multiple surrogate index standard errors because we use varying numbers of waves of surrogates in each surrogate index estimator, as represented in table 2. The black line is the 45-degree line. Surrogate index standard errors are systemically smaller than RCT standard errors.

with the previous results to guide practitioners interested in implementing these methods. The primary outcomes are selected from either the pre-analysis plan of each long-run RCT or the abstract of the associated paper. In most cases, the primary outcome is a welfare measure such as income, consumption, or profit. However, for CCT Colombia it is whether students enrolled in tertiary education; for Education Kenya, it is the number of grades completed; and for Therapy Pakistan it is depression severity.

The full table of primary outcomes, treatments, and selected surrogates for each study appears in appendix F.

Several patterns emerge in the selected surrogates. In some cases, the selected surrogates are mechanically related to the primary outcome. This pattern is particularly evident in the CCT Colombia study, where the long-term outcome of interest is tertiary enrollment 12 years after treatment. The surrogates selected include enrollment in high school and taking the high school exit exam, both of which are necessary prerequisites for tertiary enrollment in many educational systems. This suggests that in cases where there are clear steps from one outcome to the next, it is useful to select surrogates that appear earlier in the process.

Another recurring pattern is to select a lagged version of the primary outcome as a surrogate. One example of this is the Cash Uganda study, where the income index at the 4-year mark is used as a surrogate for the primary outcome of the income index at 9 years. Using lagged versions of the primary outcome as a surrogate also allows us to incorporate information about the trajectory of the primary outcome, which could increase the precision and accuracy of the surrogate index. This strategy is likely to be especially effective when the treatment's impact is expected to persist or evolve in consistently over the relevant period.

In studies with multiple treatments, the selected surrogates tend to be consistent across different treatments. This is evident in the four studies with multiple treatments (CCT Colombia, Cash Sri Lanka, Education Kenya, and Empowerment Bangladesh) where there is significant overlap in the surrogates selected for each treatment. The consistency of surrogate selection across treatments suggests that these short-term indicators are broadly relevant to the long-term outcomes of interest, regardless of the specific treatment administered. This suggests that these surrogates may be generalisable to other contexts and treatments.

Finally, missing data indicators are often selected as surrogates too. We follow the standard practice of replacing missing values with a constant and creating a dummy variable to indicate missingness. We do this to avoid losing observations due to missingness in one surrogate, but the selection of these missingness indicators suggests that missing information can itself be informative about the primary outcome.

7 Conclusion

Our findings suggest that while the surrogate index is a powerful tool for estimating long-term treatment effects, it is not without shortcomings. Analysing data from nine long-term RCTs, we found that the surrogate index approach consistently underestimates positive long-term treatment effects, irrespective of the estimation method employed. This leads to a bias that is, on average, 0.05 standard deviations. Such bias could come from omitted surrogates, as well as bias in the first-stage predictive model.

We do not necessarily believe that this negative bias should dissuade researchers from the surrogate approach. After all, in practice, there may still be no better alternative for estimating unobservable long-term effects. However, this may suggest that we think of surrogate index treatment effect estimates as a lower bound on the benefits from treatment.

We observed that, contrary to theoretical expectations, the deliberate exclusion of (observable) covariates improved the method's performance in our meta-analysis. It appears

that excluding these covariates introduced a positive bias that offset the negative bias from missing surrogates. However, we cannot say for sure how far this result generalises, so we do not recommend deliberately excluding covariates in practice.

Despite the inherent biases in the surrogate index method, it offers considerable precision gains, with standard errors being approximately 52% the size of those from the long-term RCT. Therefore, even if researchers have access to long-term outcomes, they may be more interested in the surrogate method if they are willing to trade off a small amount of bias for a large reduction in variance.

Finally, our paper introduced the M-lasso algorithm. This is specifically designed with the surrogate method in mind and can be used to construct the first-stage predictive model.

Our analysis comes with some caveats. In particular, the long-term datasets used in our meta-analysis are likely selectively drawn from studies expected to have significant short-term and long-term effects. The surrogate index might perform especially well in these settings, as large short-run effects could provide a stronger basis for predicting long-term outcomes. As such, our analysis might overstate the performance of the surrogate index, relative to settings with weaker short-term effects.

On the other hand, the RCTs in our meta-analysis were designed without consideration for using the surrogate index methodology. As such, the original researchers likely focused on collecting baseline covariates that are predictive of long-term outcomes, to improve power. For researchers designing studies where they plan to use the surrogate index approach, they should focus more on collecting information on mediators of the long-run causal effect (and potentially mediator-outcome confounders). This might *improve* performance of these methods relative to our meta-analysis findings.

More work remains to be done in this field. Future research could expand our analysis to include more datasets, explore other domains outside of development economics, and investigate alternative methods for estimating the surrogate index. Further exploration into the causes of negative bias is also recommended. These efforts will give us greater confidence in using surrogate indexes and other related approaches throughout the social sciences.

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*. Accessed: 2019-04-09.

Athey, S., Chetty, R., Imbens, G. W., and Kang, H. (2019). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Working Paper 26463, National Bureau of Economic Research.

Baird, S., Hicks, J. H., Kremer, M., and Miguel, E. (2016). Worms at work: Long-run impacts of a child health investment. *The Quarterly Journal of Economics*, 131(4):1637–1680.

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*.

Banerjee, A., Duflo, E., and Sharma, G. (2021). Long-term effects of the targeting the ultra poor program. *American Economic Review: Insights*, 3(4):471–86.

Baranov, V., Bhalotra, S., Biroli, P., and Maselko, J. (2020). Maternal depression, women's empowerment, and parental investment: Evidence from a randomized controlled trial. *American Economic Review*, 110(3):824–59.

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.

Blattman, C., Fiala, N., and Martinez, S. (2020). The long-term impacts of grants on poverty: Nine-year evidence from uganda's youth opportunities program. *American Economic Review: Insights*, 2(3):287–304.

Bouguen, A., Huang, Y., Kremer, M., and Miguel, E. (2019). Using randomized controlled trials to estimate long-run impacts in development economics. *Annual Review of Economics*, 11:523–561.

Buchmann, N., Field, E., Glennerster, R., Nazneen, S., and Wang, X. Y. (2023). A signal to end child marriage: Theory and experimental evidence from bangladesh. *American Economic Review*, 113(10):2645–88.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 785–794.

Chernozhukov, V., Hansen, C., and Spindler, M. (2016). hdm: High-Dimensional Metrics. *The R Journal*, 8(2):185–199.

De Mel, S., McKenzie, D., and Woodruff, C. (2012). One-time transfers of cash or capital have long-lasting effects on microenterprises in sri lanka. *Science*, 335(6071):962–966.

Duflo, E., Dupas, P., and Kremer, M. (2015). Education, hiv, and early fertility: Experimental evidence from kenya. *American Economic Review*, 105(9):2757–97.

Dynarski, S., Libassi, C., Michelmore, K., and Owen, S. (2021). Closing the gap: The effect of reducing complexity and uncertainty in college pricing on the choices of low-income students. *American Economic Review*, 111(6):1721–56.

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.

Gertler, P. J., Martinez, S. W., and Rubio-Codina, M. (2012). Investing cash transfers to raise long-term living standards. *American Economic Journal: Applied Economics*, 4(1):164–92.

Guzman, J., Oh, J. J., and Sen, A. (2020). How do (green) innovators respond to climate change scenarios? evidence from a field experiment.

Hamory, J., Miguel, E., Walker, M., Kremer, M., and Baird, S. (2021). Twenty-year economic impacts of deworming. *Proceedings of the National Academy of Sciences*, 118(14):e2023185118.

Imbens, G. W. (2020). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, 58(4):1129–1179.

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

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, 76(4):604–620.

McKenzie, D. (2020). Using a surrogate index to estimate long-term treatment impacts from a short-term follow-up. Accessed: 2023-09-14.

Miguel, E. and Kremer, M. (2004). Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, 72(1):159–217.

Otero, S., Barahona, N., and Dobbin, C. (2021). Affirmative action in centralized college admission systems: Evidence from brazil. Accessed: 2023-09-14.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

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

A Appendix - Description of RCTs used in meta-analysis

In this appendix, we provide a detailed description of each RCT employed in our meta-analysis. For a shorter overview of these RCTs, see section 4.

A.1 Conditional transfers in Colombia

Barrera-Osorio et al. (2019) use two experiments to study the impact of three different forms of conditional cash transfer (CCT) on educational outcomes in Bogota, Colombia. 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. This is paid over 10 months of the year for a total of $150. The second treatment ('savings') is similar but gives only $20 every two months, transferring the remaining $50 at the time of enrollment the following year. This is intended to reduce liquidity constraints at the time of enrollment, which is when students typically incur additional expenses such as uniforms and school supplies. Therefore, this treatment should encourage greater enrollment. Both of these treatments were studied in the San Cristobal locality relative to one control group.

A third treatment ('incentive') was studied in a separate experiment in the Suba locality 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. Note that San Cristobal and Suba are two of the poorest localities in Bogota.

In San Cristobal, 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).

The authors study effects on enrollment and graduation for secondary and tertiary education, from 2 to 12 years after the experiment. The medium-term outcomes come from annual secondary school enrollment data, administrative data from the institute that organises secondary school exit examinations, and panel data which tracks students who enrol in college. This is where we observe the long-term outcomes measured in 2016, as well as information on whether students had enrolled in tertiary education by 2012. We also utilise 17 baseline variables including measures of household wealth, income, education, and sizes as well as the students' age and gender.

A.2 Education subsidies and HIV education in Kenya

Duflo et al. (2015) is a seven-year cross-randomised RCT studying an HIV prevention education programme as well as education subsidies in the form of providing students uniforms. 328 schools in Kenya were randomised to receive either the free uniform programme (83 schools, 4,764 pupils), the HIV education programme (83 schools, 4,936 pupils), both programmes jointly (80 schools, 4,652 pupils), or the control group which received neither programme (82 schools, 4,927 pupils). The authors also randomised an add-on component to the HIV education programme which focused on condoms, but we do not use this cross-randomisation in our analysis. Students enrolled in sixth grade in 2003 formed the study sample. The authors followed up with these students three, five, and seven years after treatment.

The initial outcomes focus on whether students were present and enrolled in school, plus whether they were married, had children, or had ever been pregnant. A long-run follow-up was done in 2010 which included the same questions as well as measuring biomarkers for HIV

and herpes. The authors analyse boys and girls separately throughout the original paper, but I merge the groups in my analysis. The education subsidy is found to be effective for reducing school dropout, teen marriage, and pregnancy. However, the HIV programme has mostly null effects. Long-run effects are mostly small and null in this analysis, except for the effect of the education subsidy on the number of grades completed.

A.3 Cash grants in Sri Lanka

De Mel et al. (2012) is a five-year RCT studying the impact of cash grants to microenterprises with no employees in Sri Lanka. The grants were randomly either of size $100 or $200. The authors further randomised half the grants to be in the form of cash and the other half as in-kind purchases of equipment, but we do not use this variation in my analysis. This RCT is significantly smaller than the others in our analysis, with only 408 microenterprises taking part. The authors focus instead on surveying the microenterprises multiple times, resulting in there being 12 post-treatment survey waves. For the first two years, the owners are surveyed quarterly, and in year 3 they are surveyed twice at six-month intervals. The authors then return two years later to survey the microenterprises twice in year 5, again at six-month intervals.

The primary outcomes are monthly profits, capital stock, and the labor supply of the owner. These outcomes are observed in each of the 12 post-treatment survey waves, except labor supply, which is not observed in the final two surveys in year 5. The headline result in the paper is that effects on business survival and profits are large and persistent for men and consistently null for women.

A.4 Cash grants in Uganda

Blattman et al. (2020) is a nine-year follow-up of an RCT evaluating the Youth Opportunity Program in Uganda. In this RCT, groups of young people submitted applications for grants to help them start skilled trades and businesses. The program was oversubscribed, so for groups who had sufficiently good applications, the authors randomised them to receive the grants or not. 535 groups of approximately 12,000 members submitted eligible applications and 265 of the groups received the grant. The authors randomly sampled around 5 people per group resulting in a final sample size of 2,677.

Two-year, four-year, and nine-year follow-up surveys were done after grant disbursement in 2008. The key outcomes concern the participants' income, capital, employment hours, and whether they worked in a skilled trade. Effects were significant two and four years after the program, but most effects had faded to zero by the time of the nine-year endline, with the control group catching up to the treatment group.

A.5 Graduation program in India

Banerjee et al. (2021) is a ten-year follow-up of a graduation or big-push program which provided a large asset transfer, consumption support and training to ultra-poor Indian households. Households had to meet several criteria to be eligible for the intervention. Those who were eligible were individually randomised with stratification at the hamlet level. However, only 266 of the 514 who were assigned to receive the intervention accepted. Like the original authors, we focus on intent-to-treat effects.

The key outcomes of interest in this study were consumption, food security, income, and health. The authors found positive long-term effects on all these outcomes: consumption increased by 0.6 standard deviations (SD), food security by 0.1 SD, income by 0.3 SD, and

health by 0.2 SD. These effects grew for the first seven years following the transfer and persisted until year 10. An important channel for these persistent effects was the opportunity for treated households to diversify into more lucrative wage employment, notably through migration.

A.6 Psychotherapy in Pakistan

Baranov et al. (2020) study the seven-year effects of a cognitive behavioural psychotherapy intervention for prenatally depressed mothers on women's mental health, financial empowerment, and parenting decisions. Randomisation was done at the community level with 20 communities being assigned to treatment and 20 serving as controls. There were 903 women in total in the trial and 585 of them were identified at the seven-year endline, making this one of the world's largest studies of psychotherapy intervention.

After one year the intervention reduced depression by 39 percentage points. However, after seven years the effect was five percentage points. Additionally, the intervention resulted in improved financial empowerment for women and led to an increase in both time- and money-intensive parental investments. These improvements were measured as being between 0.2 and 0.3 standard deviations.

A.7 Conditional cash transfers in Mexico

Gertler et al. (2012) study the five-year impacts of the Progresa / Oportunidades conditional cash transfer program for poor households in rural Mexico. Communities are randomised into receiving the cash transfers, but only eligible households classified as low income by a proxy means test are eligible for the program.

The study found that households invested part of their cash transfers in productive assets, which increased their agricultural income by almost 10% after 18 months of receiving benefits. This effect persisted, with a 5.6% increase in consumption after five and a half years. For each peso transferred, households consumed 74 cents and invested the rest, permanently increasing long-term consumption by about 1.6 cents. These results suggest that cash transfers can lead to long-term increases in consumption through investment in productive activities, allowing beneficiary households to achieve higher living standards that are sustained even after transitioning off the program.

A.8 Female empowerment in Bangladesh

Buchmann et al. (2023) study two programs in Bangladesh, aimed at increasing girls' education while reducing teenage marriage or childbearing. The first treatment is a six-month empowerment program, the second treatment is a financial incentive to delay marriage, and the third treatment provides both programs simultaneously. Cluster randomisation is done at the community level and girls aged 10-19 are the target program participants.

The authors find that after 4.5 years, the empowerment program did not affect child marriage or teenage childbearing, although it did increase school enrollment. It also increased an income-generating activities index by 0.5 SDs. By contrast, the incentive program had effects on marriage, giving birth *and* school enrollment: eligible girls were 8.9 percentage points less likely to be married, 4.8 percentage points less likely to have given birth under 20, and 7.0 percentage points more likely to be in school.

A.9 Deworming in Kenya

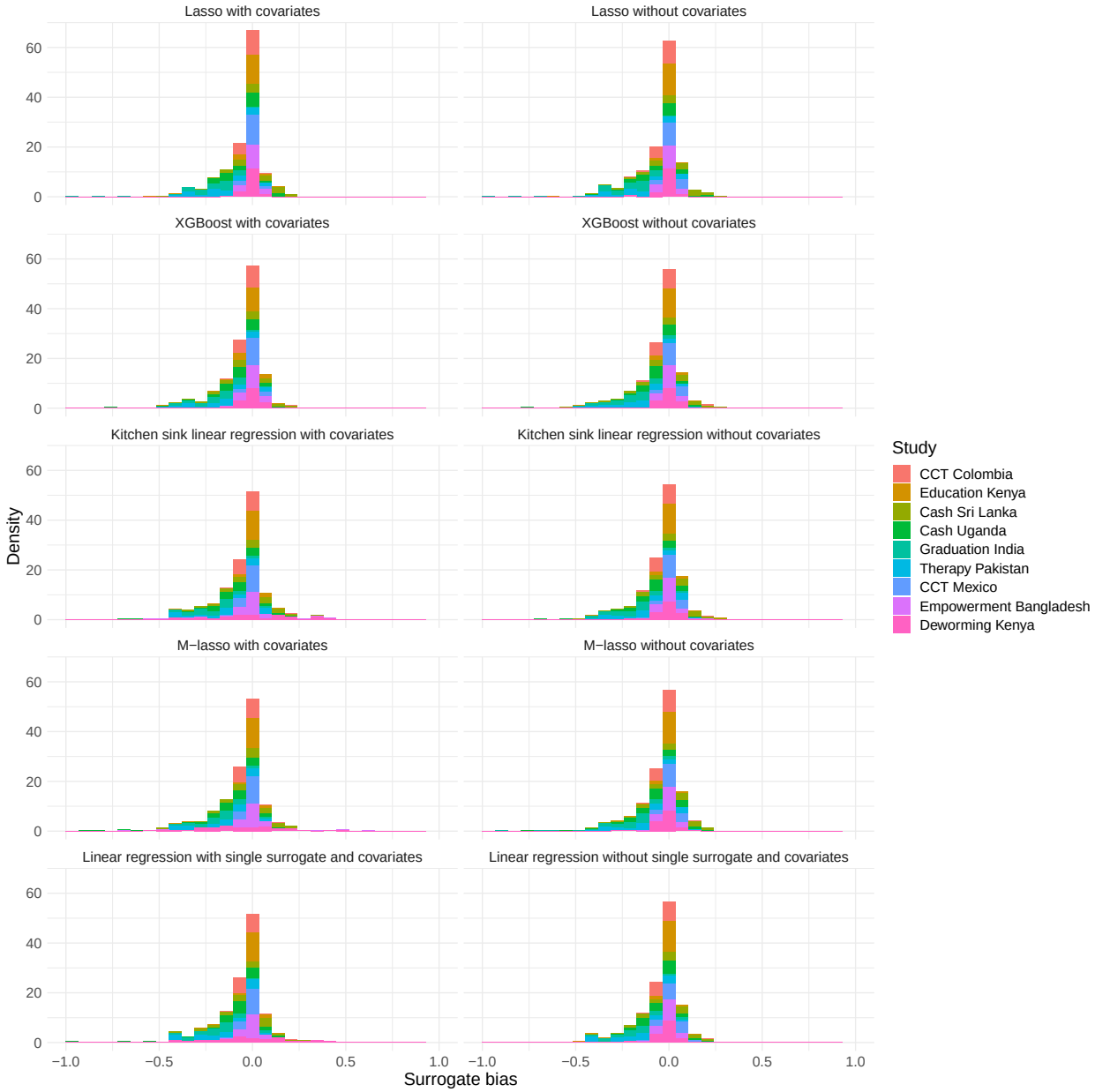
Hamory et al. (2021) study the twenty-year effects of deworming in Kenya. Schools were randomised to start receiving deworming pills at different times such that some children received treatment two to three years earlier. Participants were surveyed one to two years after the start of the program (Miguel and Kremer, 2004), ten years (Baird et al., 2016), fifteen and twenty years (Hamory et al., 2021).

Miguel and Kremer (2004) show that the program improved health and school participation and had positive externalities on untreated children in treated and neighbouring schools. Baird et al. (2016) study the medium-run effects and show that there were positive effects on school enrollment for males and females, as well as positive labour market impacts in terms of time spent working and income for men.

Hamory et al. (2021) show that these effects persist twenty years after treatment, with those who received additional, earlier deworming having 14% greater consumption expenditures and 13% higher hourly earnings. They also work more in nonagricultural sectors and are more likely to live in urban areas.

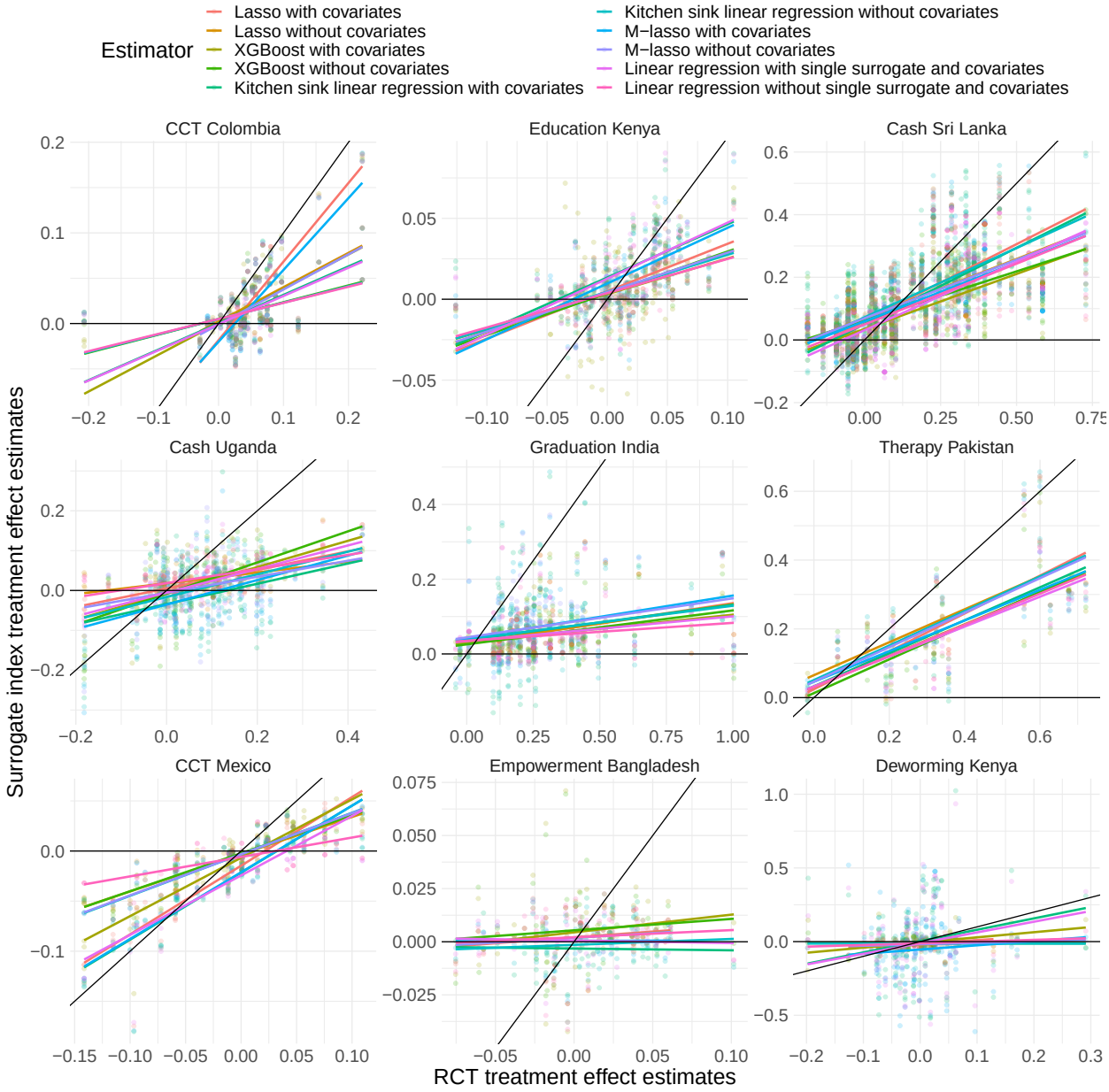
B Appendix - Graphs with all estimators

Figure 9: Distribution of the bias in raw normalised surrogate index estimates, all methods



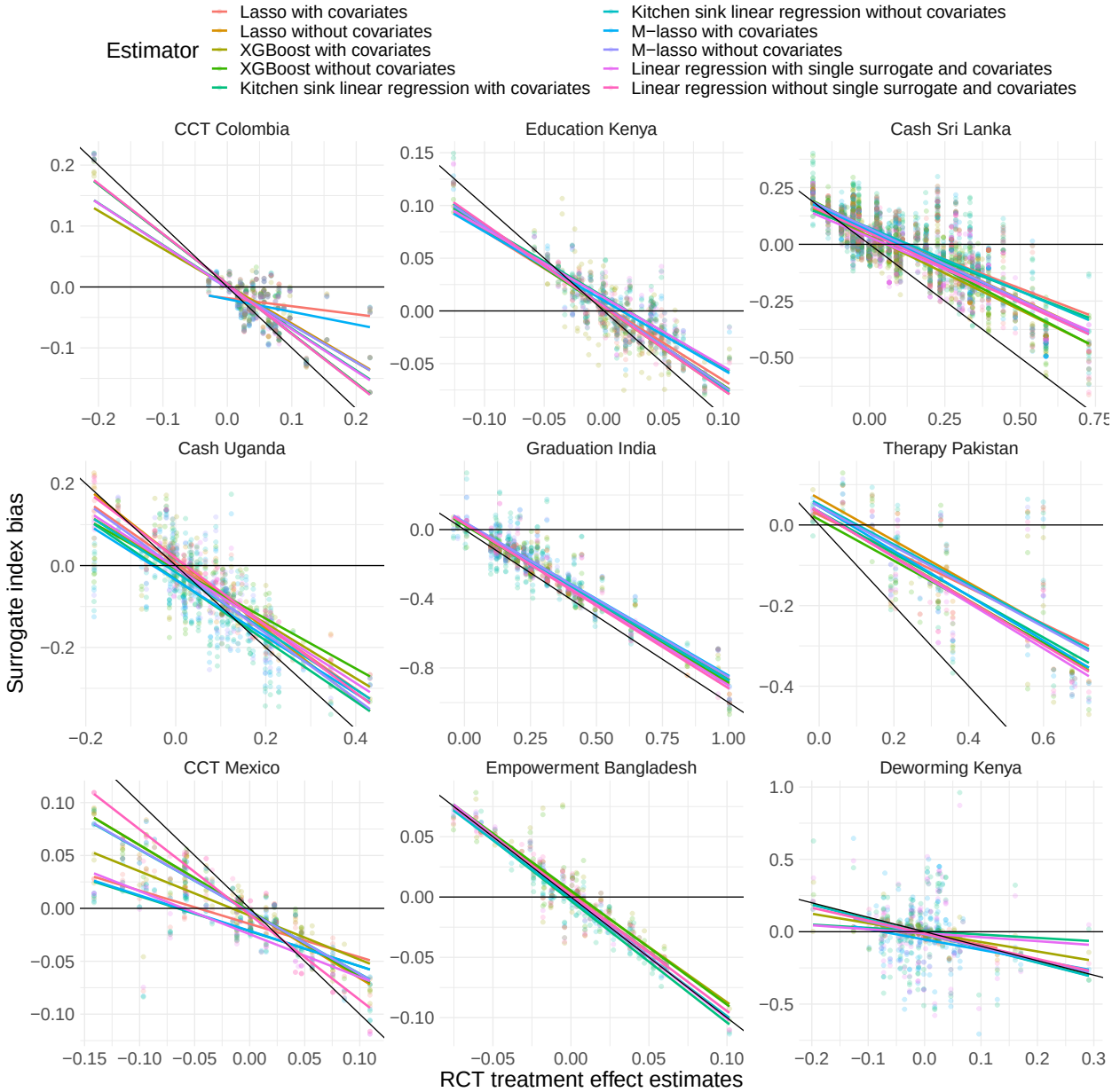
Notes: Across nine RCT contexts, for all methods of implementing the surrogate index estimator, the bias distribution has a left tail and is centred slightly left of zero. This indicates the surrogate method is negatively biased across all of our implementation methods.

Figure 10: Normalised surrogate index treatment effect point estimates (all methods) against normalised RCT treatment effect point estimates



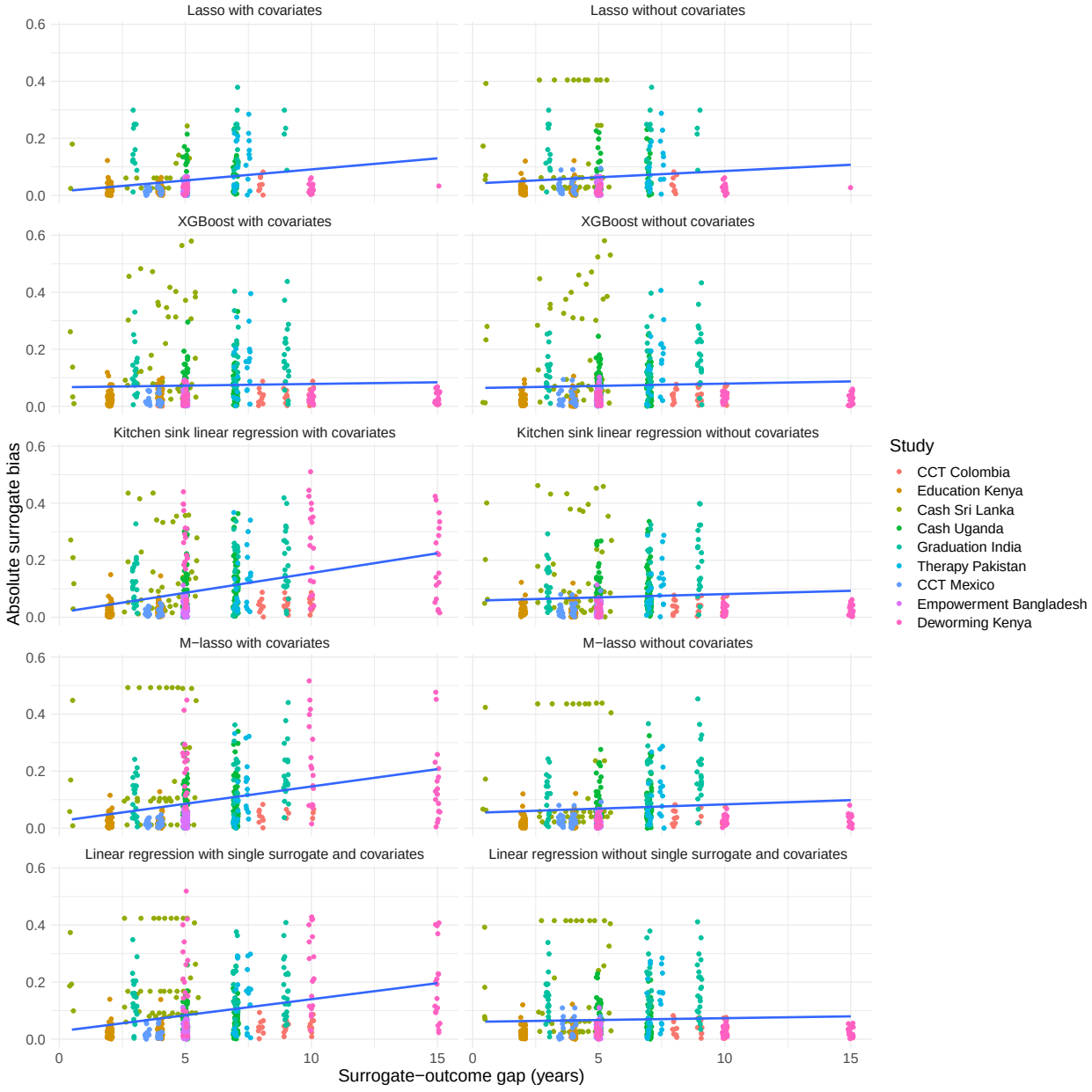
Notes: Across nine RCT contexts, for each surrogate index estimation method, we plot the normalised surrogate estimates of treatment effects (y-axis) against their corresponding normalised RCT estimates (x-axis). Each RCT estimate is associated with multiple surrogate estimates because varying numbers of waves of surrogates were used for each treatment-outcome pair, as represented in table 2. The black line is the 45-degree line and the coloured lines are the lines of best fit. For all estimation methods, the surrogate index best fit line has a smaller positive slope than the 45-degree line, suggesting that the surrogate index method overestimates smaller treatment effects and underestimates larger ones.

Figure 11: Normalised surrogate index bias (all methods) against normalised RCT treatment effect point estimates



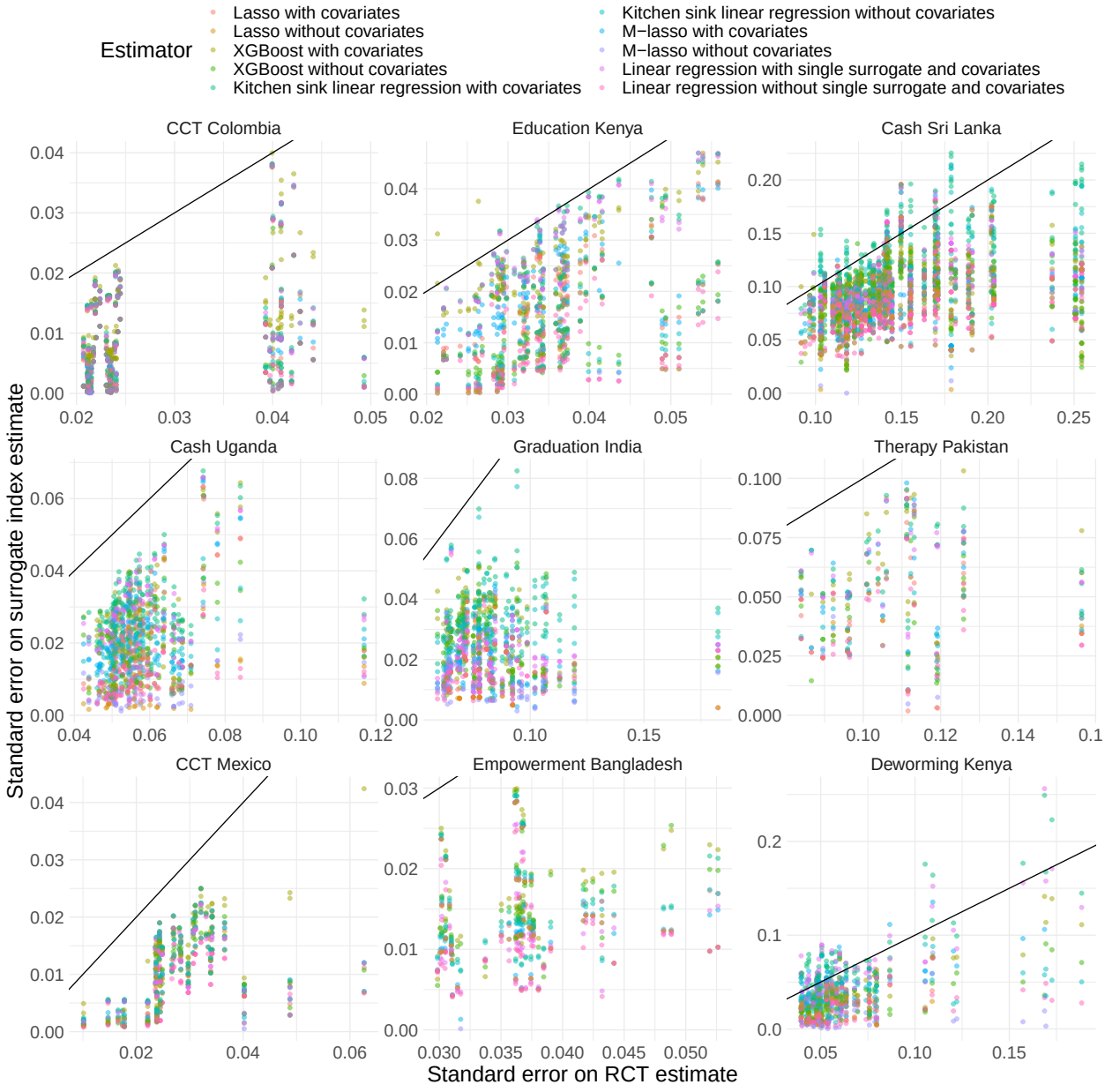
Notes: The coloured lines are the lines of best fit for the surrogate bias of each estimation method against the RCT treatment effect estimate across nine RCT contexts. Though the kitchen sink linear regression method without covariates does comparatively well, the coloured lines are different from the horizontal black line representing zero bias. Often, the slopes of these best fit lines are far from zero, closer to the 45-degree line, indicating increasing negative bias as RCT effects become more positive. Note that each RCT estimate is associated with multiple surrogate index biases because varying numbers of waves of surrogates were used to create surrogate index estimates, as represented in table 2. All graphs share the same x and y-axis labels.

Figure 12: Normalised absolute bias of all estimators against time horizon



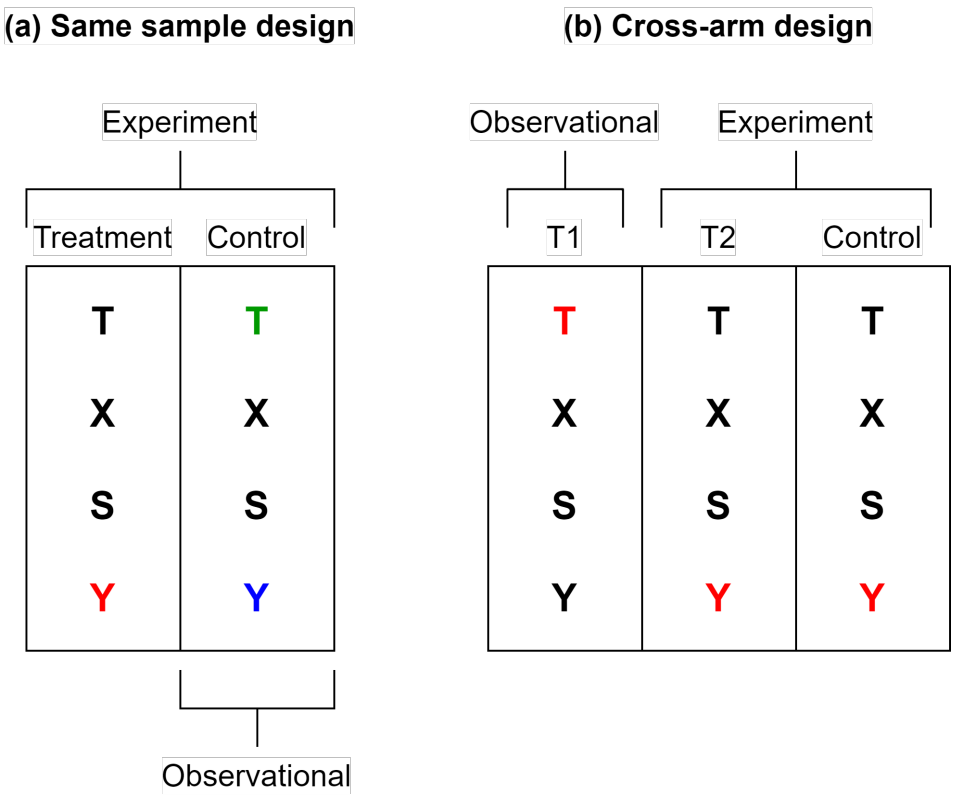
This graph plots biases against time horizons for all experiments and colour-codes them by the study from which they derive. In general, absolute bias tends to weakly increase with time horizon.

Figure 13: Normalised standard errors on treatment effects: surrogate index results (all methods; cross arm design) against RCT results



Notes: For the four studies for which cross-arm design is possible, we show the relationship between RCT-estimated standard errors on treatment effects with the same for surrogate index-estimated standard errors. Each RCT standard error is associated with multiple surrogate index standard errors because we use varying numbers of waves of surrogates in each surrogate index estimator, as represented in Table 2. The black line is the 45-degree line. Points are colour-coded by the estimation method. Surrogate index standard errors tend to be systemically smaller than RCT standard errors. The Cash Sri Lanka study has some expectations, with surrogate index estimators producing larger standard errors than the RCT method does, particularly for the 'kitchen sink linear regression with covariates' estimation method.

Figure 14: Two designs for testing surrogacy approach



Note: The red Ys and Ts represent that we do not use this variable when implementing the surrogate index approach. The blue Y represents that we do not use this variable when using the data as experimental, but we do when using it as observational. The green T represents that we do not use this variable when using the data as observational, but we do when using it as experimental.

C Appendix - Cross-arm design

Recall that the meta-analysis in the main article is constructed using a 'same sample' design. With this design, for every RCT we construct an observational dataset from the control group. Meanwhile, an experimental dataset is constructed from both the control and treatment group – this ensures that the experimental dataset contains variation in the treatment status. This is shown in figure 14 (a).

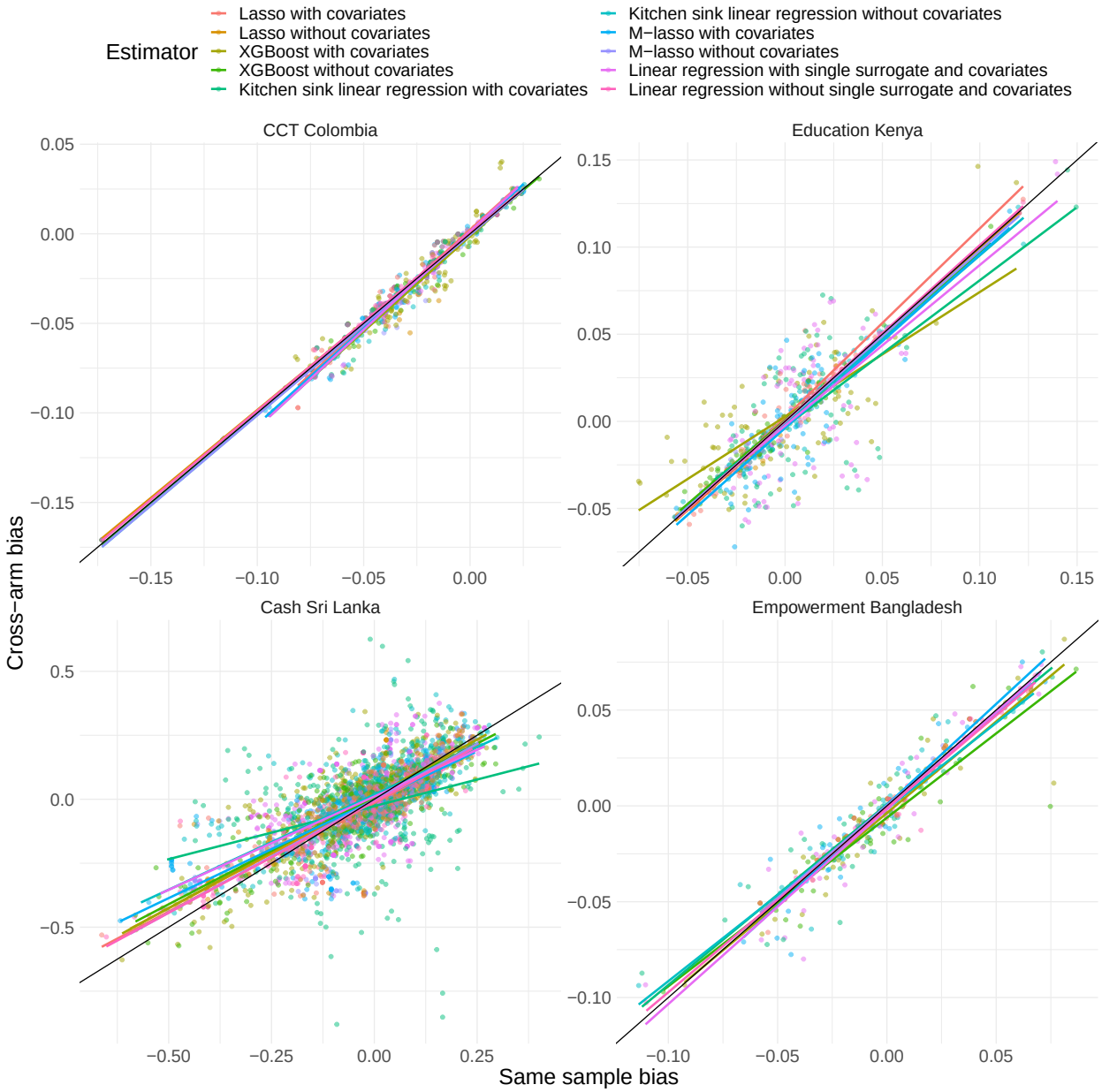
For RCTs containing at least two treatment arms and one control, we can also consider an alternative design that we call the *cross-arm* design. Here we use one treatment arm as the observational dataset. We then use all the other arms (including the control) to construct the experimental dataset. This reduces concerns of overfitting since the observational and experimental datasets no longer overlap. However, it introduces the *possibility* for the comparability of samples assumption to be violated, seeing as treatment status is now guaranteed to be different in the observational dataset versus the experimental dataset. This problem is mitigated by the fact that the observational and experimental datasets still come from the same RCT, such that both datasets will have the same distribution of baseline characteristics in expectation.

As noted above, a downside of the cross-arm approach is that it is only feasible in RCTs where there are at least two treatment arms and one control arm. Only four out of the nine RCTs we use satisfy this requirement, so we focus on the same sample design for the main

analysis. However, we show in appendix D that the main graphical results are robust to using the cross-arm design in the six RCTs which have more than one treatment arm. In particular, (1) we show that the bias of the surrogate index is similar for the same-sample and cross-arm approaches in the studies where we can estimate both, and (2) we replicate the main figures of the paper but with the cross-arm design.

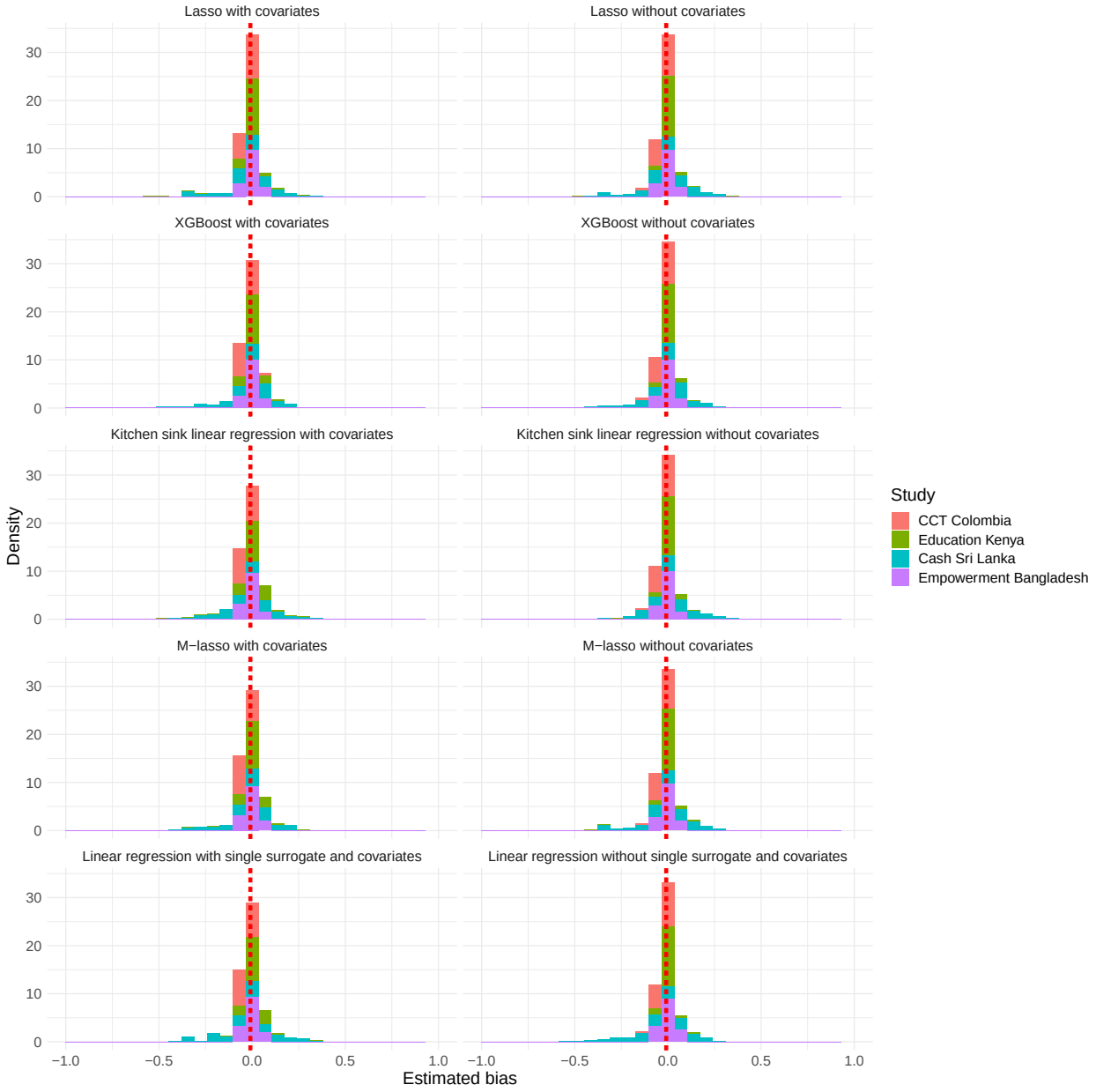
D Appendix - Robustness to cross-arm design

Figure 15: Normalised cross-arm bias against equivalent normalised same-sample bias



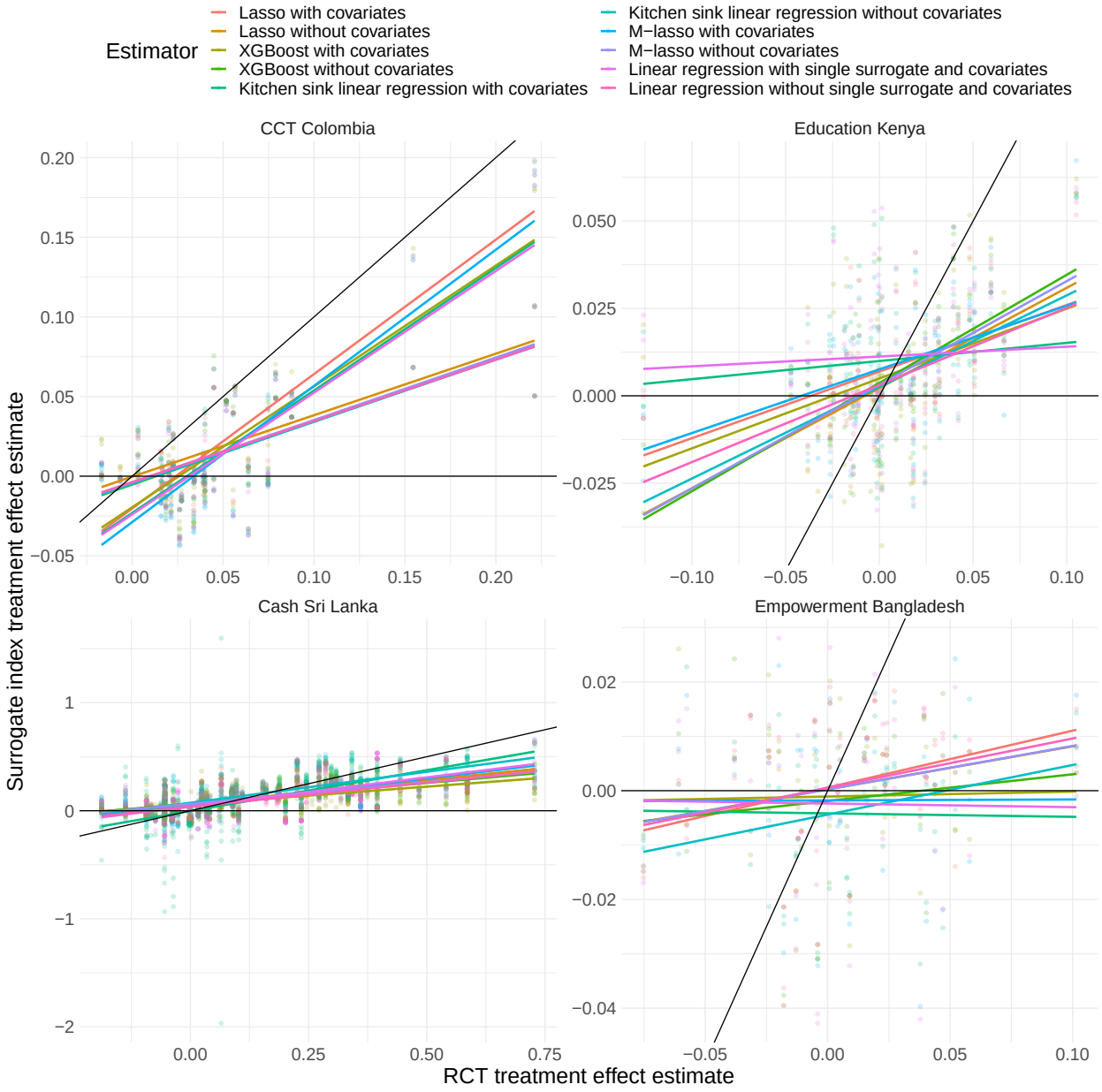
The biases of the surrogate index are similar for the same-sample and cross-arm approaches in the studies where both can be estimated.

Figure 16: Distribution of the bias in raw normalised surrogate index estimates, all methods, cross-arm design



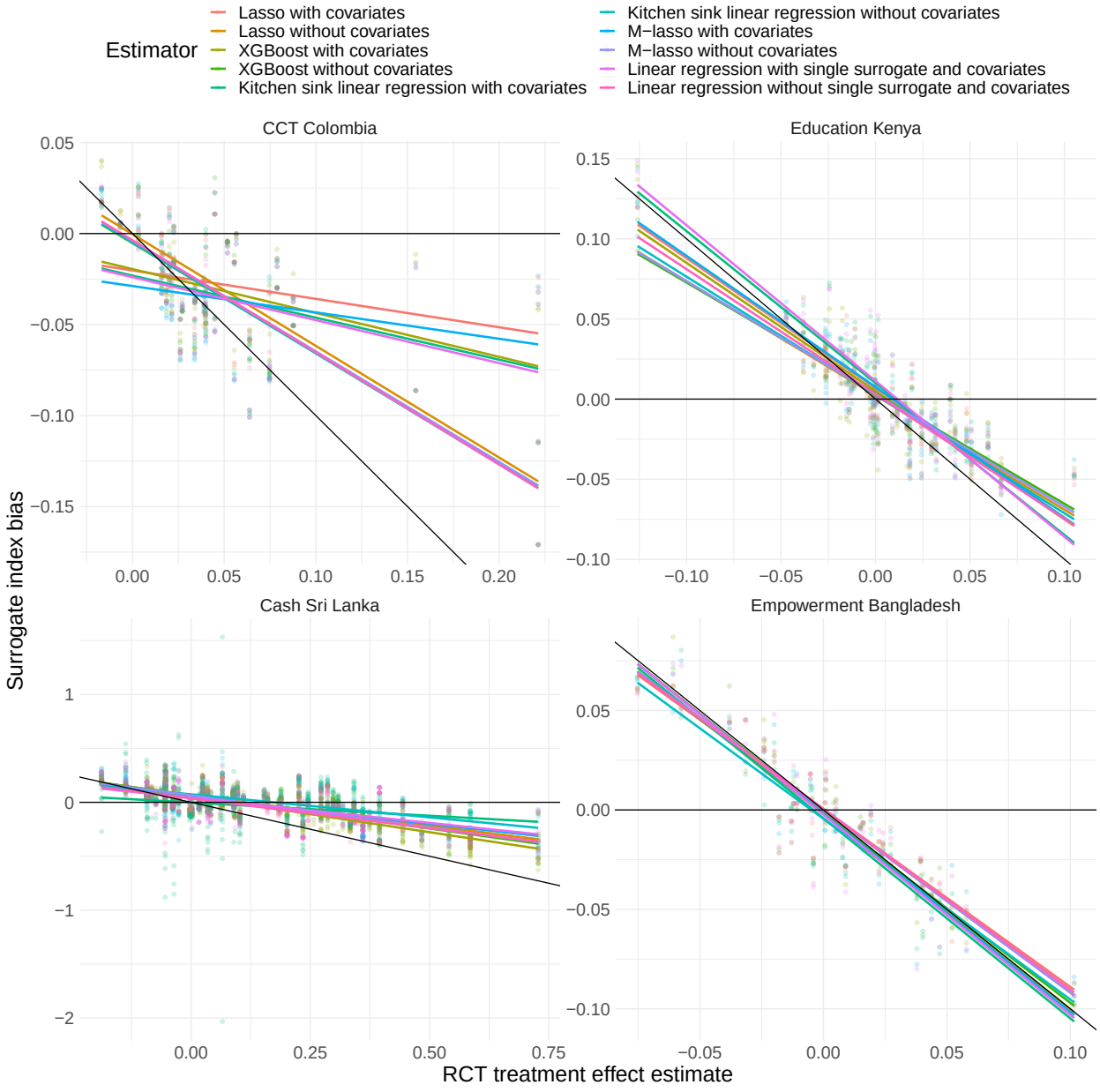
Across the four RCT contexts for which the cross-arm design is possible, for all methods of implementing the surrogate index estimator, the bias distribution is centred very slightly left of zero. This indicates the surrogate method could be negatively biased for the cross-arm design.

Figure 17: Normalised surrogate index treatment effect point estimates (all methods) against normalised RCT treatment effect point estimates, cross-arm design



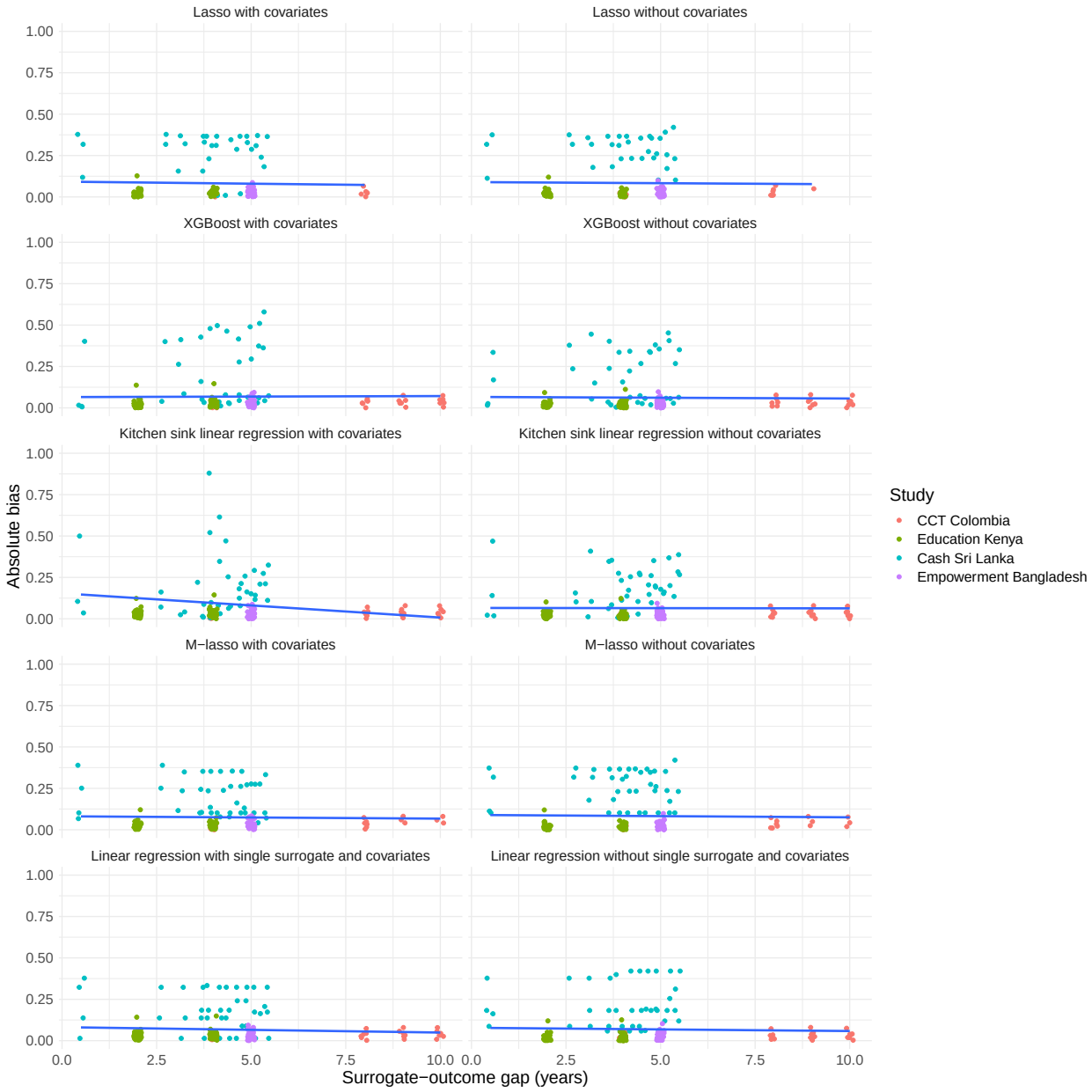
Across the four RCT contexts for which the cross-arm design is possible, we plot the surrogate estimates against their corresponding RCT estimates. Each RCT estimate is associated with multiple surrogate estimates because varying numbers of waves of surrogates were used for each treatment-outcome pair, as represented in table 2. The black line is the 45-degree line and the coloured lines are the lines of best fit. For all estimation methods, the surrogate index best fit line has a smaller positive slope than the 45-degree line, suggesting that the surrogate index method using the cross-arm design overestimates smaller treatment effects and underestimates larger ones across all of our estimation methods.

Figure 18: Normalised surrogate index bias (all methods) against normalised RCT treatment effect point estimates, cross-arm design



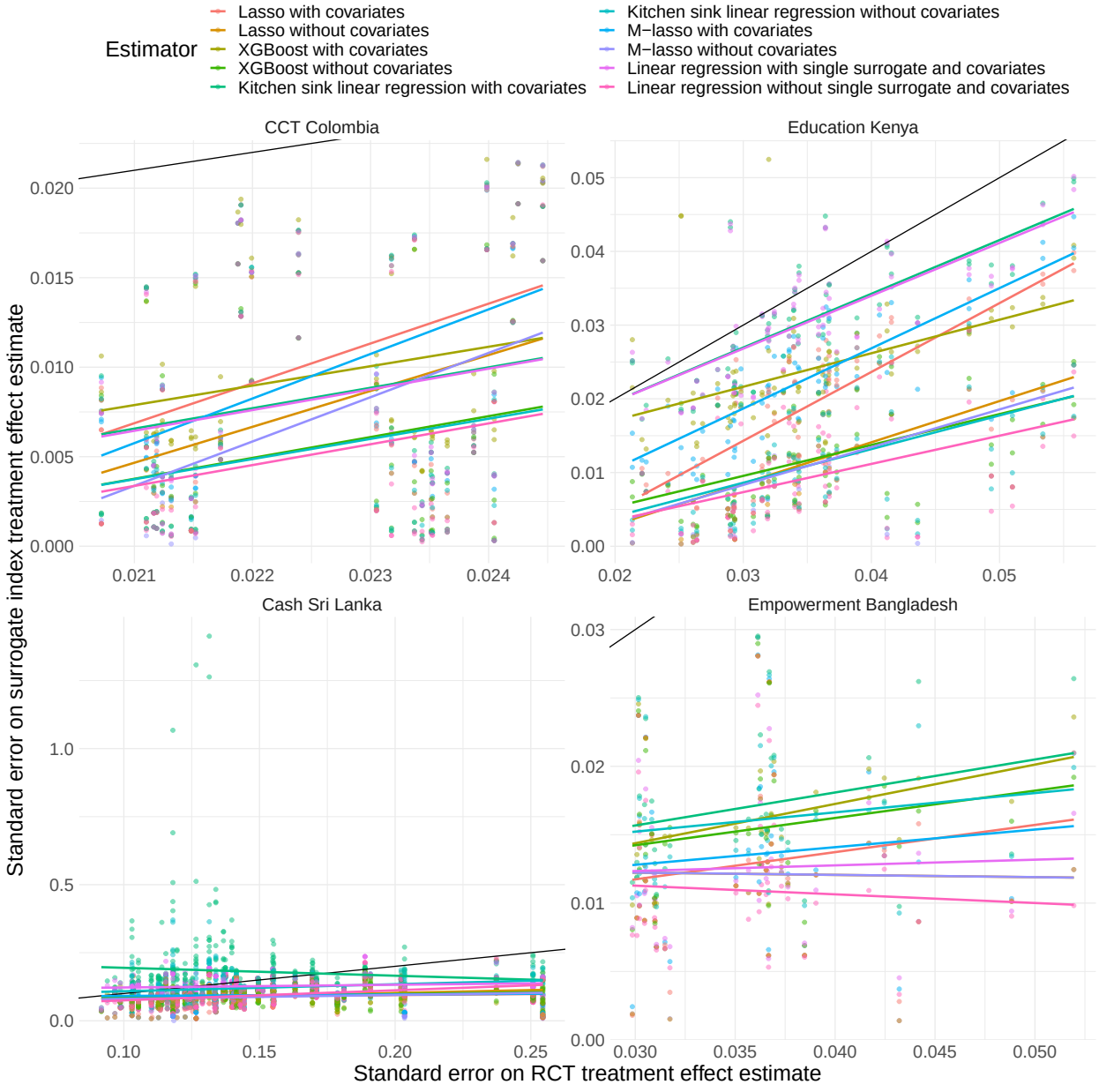
The coloured lines are the lines of best fit for the surrogate bias of each estimation method against the RCT treatment effect estimate across the four RCT contexts for which the cross-arm design is possible. These lines are often different from the horizontal black line representing zero bias. Sometimes, their slope is much closer to the 45-degree line in black. This indicates that, using cross-arm design, surrogate bias is not zero on average across all estimation methods, and that bias grows more negative as RCT effects become more positive. Note that each RCT estimate is associated with multiple surrogate index biases because varying numbers of waves of surrogates were used to create surrogate index estimates, as represented in table 2. All graphs share the same x and y-axis labels.

Figure 19: Normalised absolute bias of all cross-arm estimators against time horizon



This graph plots biases against time horizons for all experiments for which cross-arm design is possible. Experiments are colour-coded by the study from which they were derived. In general, the absolute bias appears to weakly decrease with time horizon. This opposes the result from the same-sample design. However, there are very few observations and study heterogeneity makes us suspect the trend is not robust.

Figure 20: Standard errors of the treatment effects estimated by cross-arm estimators

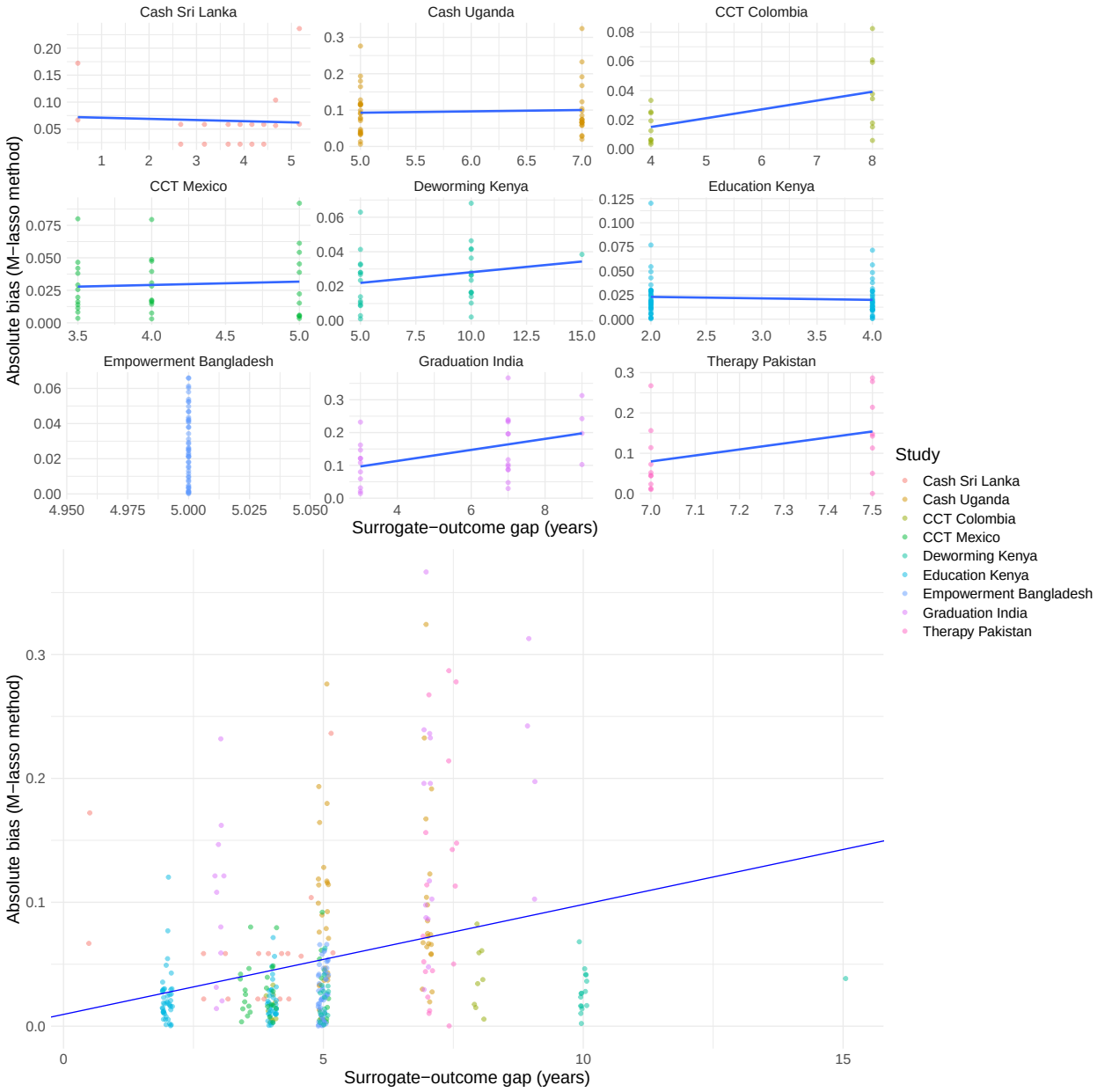


E Appendix - Plotting time horizon and bias

Recall that in section 6.3.1 we found only weak evidence of a positive relationship between time horizon and surrogate bias. We expand on this relationship further with figure 21.

The top half of figure 21 plots the absolute bias for each study separately. The bottom half of figure 21 plots absolute bias for all studies jointly, plus the horizon-absolute bias relationship estimated in the meta-analysis. In general, figure 21 reports the same positive relationship between time horizon and bias within most individual studies. The studies Education Kenya and Cash Transfers Sri Lanka are exceptions. Along with CCT Mexico, these two studies have the shortest time horizon gaps between surrogates and outcomes.

Figure 21: Absolute surrogate bias (M-lasso method) over time horizon



Note: In general, there is only weak evidence that absolute bias increases with time horizon. This relationship is shown for each study separately in the top half of the figure, with y-axes scaled differently for each study. Empowerment Bangladesh only has two post-treatment waves, so only one set of surrogate index estimates is available. The bottom half of the figure combines the nine individual plots to represent the absolute bias and surrogate-outcome gap relationship estimated in the meta-analysis in table 5

.

F Appendix - Table of selected surrogates for primary outcomes

This appendix identifies the surrogates picked by the M-lasso method for the primary long-term outcome of each study, for each treatment.

Table 9: Selected surrogates for primary outcomes

Study	Primary outcome (years after treatment)	Treatment	Selected surrogates (years after treatment)
CCT Colombia	Tertiary enrollment (12)	Basic conditional cash transfer	Enrolled in high school (4), Took high school exit exam (8), Tertiary enrollment (8)
CCT Colombia	Tertiary enrollment (12)	Savings conditional cash transfer	Took high school exit exam (8), Tertiary enrollment (8)
CCT Colombia	Tertiary enrollment (12)	Incentive conditional cash transfer	Took high school exit exam (8), Tertiary enrollment (8), Unreported type of tertiary education (8)
Cash Sri Lanka	Real profits (5)	$100 cash grant	Real profits (4.5), Missing profits (0.25)
Cash Sri Lanka	Real profits (5)	$200 cash grant	Real profits (4.5), Missing hours worked (0.25), Missing profits (1)
Cash Uganda	Income index (9)	Cash grant	Enrolled in vocational training (2), Hours of training received (2), Business assets (2), Durable assets (2), Missing enrolled in vocational training (2), Income index (4), Durable assets (4), Skilled trade hours per week (4)
Education Kenya	Grades completed (7)	Free uniform	Ever dropped out of primary (3), Present in school (3), Ever dropped out of primary (5), Ever married (5), Ever pregnant (5), Ever married and not pregnant (5), Missing ever married (5)
Education Kenya	Grades completed (7)	HIV Education	Ever dropped out of primary (3), Present in school (3), Ever dropped out of primary (5), Ever married (5), Ever pregnant (5), Ever married and not pregnant (5), Missing ever married (5)
Education Kenya	Grades completed (7)	Uniform and HIV Education	Ever dropped out of primary (3), Present in school (3), Ever dropped out of primary (5), Ever married (5), Ever pregnant (5), Ever married and not pregnant (5)
Graduation India	Consumption (7)	Graduation program	Food consumption (2), Livestock revenue (2), Missing asset index (2), Productive asset index (4), Missing asset index (4), Livestock revenue (8), Productive asset index (8), Self-reported economics status (8)

Continued on next page

Table 9 – continued from previous page

Study	Primary outcome (years after treatment)	Treatment	Selected surrogates (years after treatment)
Therapy Pakistan	Depression severity (7)	Cognitive behavioural therapy	Depressed (0.5), Brief Disability Questionnaire score (0.5), Brief Disability Questionnaire score (1), Perceived social support score (1), Depression index (1)
CCT Mexico	Consumption (6)	Conditional cash transfer	Using land (1), Value of production animals (1), Consumption (1), Home production (1), Money received from migrants and friends (1), Missing using land (1), Missing hectares of land(1), Missing microenterprise (1), Value of draft animals (2), Value of production animals (2), Hectares of land (2), Microenterprise (2), Consumption(2), Home production (2), Missing owns farm animals (2), Missing hectares (2), Owns production animals (2.5), Value of draft animals (2.5), Value of production animals (2.5) Hectares of land (2.5)
Empowerment Bangladesh	Income (10)	Financial incentive	Used withdrawal method (5), Hours per day spent on income generating activity (5), Missing age started income generating activity (5)
Empowerment Bangladesh	Income (10)	Empowerment program	Used withdrawal method (5), Hours per day spent on income generating activity (5), Missing age started income generating activity (5)
Empowerment Bangladesh	Income (10)	Financial incentive and empowerment program	Used withdrawal method (5), Hours per day spent on income generating activity (5), Missing age started income generating activity (5)
Deworming Kenya	Consumption (20)	Deworming pills	Missing cleanliness (1), Missing wearing shoes (1), Miscarriage (10), Highest education level (10), Missing moderate to heavy worm infection (10), Lives in an urban area (15), Farm hours worked (15)

The number of years between the treatment and the year the outcome/surrogate was observed is in brackets. Surrogates that start with "Missing" are binary indicators for whether the value of the associated variable is missing.

G Appendix - Formal overview of surrogate index

Here we give a formal overview of the surrogate index, summarising results from Athey et al. (2019). The surrogate index requires two data samples: a short-run experimental sample ($P_i = E$) with N_E units, and an observational sample ($P_i = O$) with N_O units. In the experiment, we are interested in the impact τ of a binary treatment $T_i \in \{0, 1\}$ on the long-term outcome Y_i. However, this long-run outcome Y_i cannot be observed in the short-run experimental sample.[9] Instead, we observe many short-term outcomes or surrogates, S_i, as well as pre-treatment covariates that are unaffected by treatment, X_i.

The observational sample is made up of a separate group of people. The observational units do not have to be exposed to any treatment and we do not need to know their treatment status. In the observational sample, we observe the same pre-treatment covariates and surrogates as we observe in the experimental sample and we also observe the long-term outcome of interest, (X_i, S_i, Y_i). The data requirements are shown in figure 1.

We follow the potential outcomes framework and are interested in the average effect of the treatment on the outcome, that is, $\tau = \mathbb{E}[Y_i(1) - Y_i(0) \mid P_i = E]$. As we cannot observe both potential outcomes for any given individual, we focus on the average treatment effect across the sample. Note that the surrogates also have two potential outcomes, $S_i(1)$ and $S_i(0)$, and we can similarly define an average treatment effect on the surrogates $\tau_S = E[S_i(1) - S_i(0)]$, although we are not specifically interested in this treatment effect.

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

Definition 1. Propensity score
$e(x) = Pr(T_i = 1 \mid X_i = x, P_i = E)$

Assumption 1. Ignorable treatment assignment
(i) $T_i \perp\!\!\!\perp (Y_i(0), Y_i(1), S_i(0), S_i(1)) \mid X_i, P_i = E$
(ii) $0 < e(x) < 1$ for all $x \in \mathbb{X}$

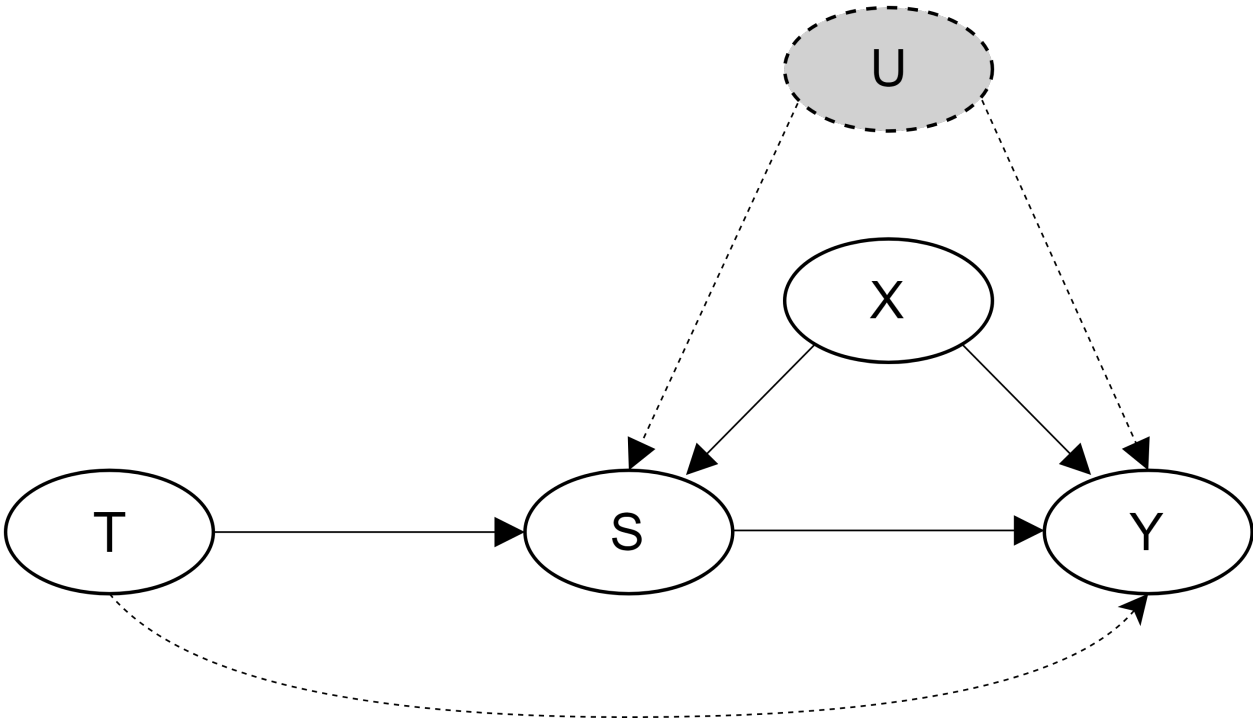
where $\mathbb{X}$ is the support of x. As we use RCTs for the analysis, this assumption is true by design as randomisation ensures the independence of treatment status and potential outcomes. This assumption implies that the average treatment effects (ATE) on the short-run outcomes in the experiment are identified. Furthermore, if we did observe the long-term outcome Y_i in the experimental sample, this assumption would be sufficient for identifying the ATE on the long-term outcome. However, we do not observe Y_i in the experimental sample. Therefore, we must rely on the surrogates, and make further assumptions to estimate the treatment effect of T_i on Y_i in the experimental sample.

Assumption 2. Surrogacy
$T_i \perp\!\!\!\perp Y_i \mid S_i, X_i, P_i = E$

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 22. This assumption implies two sub-assumptions. Firstly, there cannot be any direct effect of the treatment on the long-term outcome that is not mediated through the surrogates. We can express this as $Y_i(t, s) = Y_i(t', s)$, i.e. if we hold the values of the surrogates S_i fixed at s, then changing the

[9] Note that Y_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. We focus on the case where Y_i is unobserved because it occurs in the future, but the analysis would be identical for contemporaneous outcomes.

Figure 22: Directed Acyclic Graph showing potential violations of surrogacy assumption



Notes: Treatment is represented by T, surrogates are S, the long-term outcome is Y, observed surrogate-outcome confounders are X, and unobserved surrogate-outcome confounders are U. Dotted lines are causal paths ruled out by the surrogacy assumption.

value of treatment T_i from t to t' does not affect the value of Y_i. In the language of causal graphs, there must be no direct path $T \rightarrow Y$. This assumption is analogous to an exclusion restriction in the more familiar instrumental variables setting.

The second sub-assumption is that there are no unobserved confounders of the surrogates and the long-term outcome, which can be expressed as $S_i \perp\!\!\!\perp Y_i \mid X_i$. In the above DAG, this assumption requires that there are no unobserved mediator-outcome confounders U outside the set of observed covariates X.

A graphical representation of this 'no unobserved confounders' assumption is that there is no open backdoor path from T to Y. The backdoor paths which we are concerned by are $T \rightarrow S \leftarrow X \rightarrow Y$ and $T \rightarrow S \leftarrow U \rightarrow Y$. Both paths are initially closed as S is a collider on both paths. However, when we make T and Y independent by conditioning on S to close the $T \rightarrow S \rightarrow Y$ path, we open the two backdoor paths as we condition on the only collider on those paths. We can close the $T \rightarrow S \leftarrow X \rightarrow Y$ path by additional conditioning on X. However, as we do not observe U, we cannot condition on it, and so we cannot close the $T \rightarrow S \leftarrow U \rightarrow Y$ path if one exists.

The surrogacy assumption is recognised as critical, but previous work has only allowed for one surrogate. If either (1) this 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 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. If there are many causal paths or mechanisms between the treatment and the long-term outcome, observing more short-run outcomes increases the number of

mechanisms we can account for. Furthermore, we can control for observed confounders to further reduce the risk of the surrogacy assumption being violated by surrogate-outcome confounding.

Admittedly, the surrogacy assumption is very strong and it is unlikely that it will be exactly satisfied in most contexts. Nonetheless, 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 $\tau = \mathbb{E}_E[Y^1_{E,i} - Y^0_{E,i}]$ [the treatment effect on the long-run outcome in the experiment] 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.

This paper tests in which applications and contexts, and with what set of surrogates, we might expect the surrogacy assumption to be (approximately) satisfied.

To move from the single surrogate to the multiple surrogate case, Athey et al. (2019) introduce a new concept called the *surrogate index*. The surrogate index is the conditional expectation of the primary outcome, given the covariates and surrogates in the observational sample:

Definition 2. The surrogate index
$h_O(s, x) = \mathbb{E}[Y_i \mid S_i = s, X_i = x, P_i = O]$

Since we observe Y_i in the observational sample, $h_O(s, x)$ is estimable. We can similarly define the same conditional expectation within the experiment, $h_E(s, x) = \mathbb{E}[Y_i \mid S_i = s, X_i = x, P_i = E]$, and even more precisely, within each treatment arm of the experiment $\mu_E(s, x, t) = \mathbb{E}[Y_i \mid S_i = s, X_i = x, T_i = t, P_i = E]$. However, as we do not observe Y_i in the experiment, these conditional expectations are not possible to estimate directly. The question that now arises is: under what conditions can we use the estimable $h_O(s, x)$ to approximate the inestimable $h_E(s, x)$ and $\mu_E(s, x, t)$? Athey et al. (2019) introduce the *comparability of samples* assumption to tackle this issue.

Assumption 3. Comparability of samples
$Y_i \mid S_i, X_i, P_i = O \sim Y_i \mid S_i, X_i, P_i = E$
and $\mathbb{X}_E \in \mathbb{X}_O$, and $\mathbb{S}_E \in \mathbb{S}_O$.

The comparability of samples assumption requires that the conditional distribution of Y_i given (S_i, X_i) in the experimental dataset is the same as the conditional distribution of Y_i given (S_i, X_i) in the observational dataset. It must also be the case that the support of X_i and S_i in the experiment ($\mathbb{X}_E$ and $\mathbb{S}_E$ respectively) is contained within the support of X_i and S_i in the observational dataset ($\mathbb{X}_O$ and $\mathbb{S}_O$ respectively). This avoids making out of sample extrapolations.

Under comparability and surrogacy, $h_O(s, x)$, $h_E(s, x)$ and $\mu(s, x, t)$ are all equivalent which allows us to transfer the model we estimate in the observational data to the experiment to impute the missing long-run outcomes.

Proposition 1. Surrogate index
(i) Under surrogacy (assumption 2) we have

$\mu_E(s, x, t) = h_E(s, x)$, for all $s \in \mathbb{S}$, $x \in \mathbb{X}$ and $t \in \mathbb{T}$

(ii) Under comparability (assumption 3) we have

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

(iii) Under surrogacy and comparability we have $\mu_E(s, x, t) = h_O(s, x)$, for all $s \in \mathbb{S}$, $x \in \mathbb{X}$ and $t \in \mathbb{T}$

(i) says that if the surrogacy assumption is true, the conditional expectation of treated and untreated people within the experiment is the same, conditional on surrogates and baseline covariates. This is because the surrogacy assumption rules out the treatment affecting the long-term outcome independently of the surrogates, as well as ruling out the treatment modifying the relationship between the surrogates and the long-run outcome. (ii) adds that if the comparability of samples assumption is true, then the conditional expectation of the long-term outcomes conditional on the short-term outcomes and covariates is the same in the experimental dataset and the observational dataset. We can then combine these two results to get (iii), which shows that the inestimable conditional expectation of the long-term outcome in each treatment arm is equivalent to the estimable conditional expectation of the long-term outcome in the observational dataset.