

# HETEROGENEOUS CAUSAL EFFECTS WITH IMPERFECT COMPLIANCE: A BAYESIAN MACHINE LEARNING APPROACH

BY FALCO J. BARGAGLI-STOFFI<sup>1,\*</sup>, KRISTOF DE WITTE<sup>2</sup> AND GIORGIO GNECCO<sup>3</sup>

<sup>1</sup>Department of Biostatistics, Harvard T.H. Chan School of Public Health \*Corresponding author: fbargaglistoffi@hsph.harvard.edu.

<sup>2</sup>KU Leuven and Maastricht University

<sup>3</sup>IMT School for Advanced Studies

This paper introduces an innovative Bayesian machine learning algorithm to draw interpretable inference on heterogeneous causal effects in the presence of imperfect compliance (e.g., under an irregular assignment mechanism). We show, through Monte Carlo simulations, that the proposed Bayesian Causal Forest with Instrumental Variable (BCF-IV) methodology outperforms other machine learning techniques tailored for causal inference in discovering and estimating the heterogeneous causal effects while controlling for the familywise error rate (or – less stringently – for the false discovery rate) at leaves’ level. BCF-IV sheds a light on the heterogeneity of causal effects in instrumental variable scenarios and, in turn, provides the policy-makers with a relevant tool for targeted policies. Its empirical application evaluates the effects of additional funding on students’ performances. The results indicate that BCF-IV could be used to enhance the effectiveness of school funding on students’ performance.

## 1. Introduction

### 1.1. Motivation

Starting from the year 2002, the Flemish Ministry of Education promoted the “*Equal Educational Opportunities*” program (henceforth referred as EEO) to ensure equal educational opportunities to all students (OECD; 2017). The EEO program provides additional funding for secondary schools with a higher share of disadvantaged students. Proceeding from the seminal contributions of Coleman (1966) and Hanushek (2003) to recent contributions by Jackson et al. (2015) and Jackson (2018), the question on whether or not an increase in school spending affects students’ performances has been central in the social science literature. However, the bulk of studies on this topic have focused on average treatment effects (Hanushek et al.; 2016; Hanushek and Woessmann; 2017; D’Inverno et al.; 2021), often disregarding potential heterogeneities in how different subgroups of students and schools may be differentially affected by additional resources. While we acknowledge that the average treatment effect (ATE) is the most common starting point for any impact evaluation analysis, just focusing on it may mask potentially meaningful heterogeneities in the causal effects.

From a methodological point of view, the evaluation of the heterogeneous effects of additional school resources poses major challenges: (i) school funding is generally not randomly assigned to schools but its assignment correlates with schools’ characteristics and potential anticipations of the expected benefits (i.e., *confounding* issue) and; (ii) some schools may not comply with funds’ requirements and opt out even when eligible (i.e., *imperfect compliance* issue). When funding is not randomized but its assignment is based on the realized value of a variable, usually called the *forcing* (or *running*) variable, researchers can compare schools with very close values of the forcing variable – namely around the point where the discontinuity is observed – but with different levels of treatment. This comparison between units with different treatment levels, just above and below a given value of the forcing variable, is referred to as Regression Discontinuity (RD) design (Thistlethwaite and Campbell; 1960; Cook; 2008). RD designs are considered to be *quasi-experimental* set ups and lead to valid inference on causal effects of the treatment at the threshold. In the case of our application on EEO data, the forcing variable used to evaluate which schools are eligible for additional funding is the share of disadvantaged

---

*Keywords and phrases:* causal inference, instrumental variable, heterogeneous effects, interpretable machine learning, school funding, students’ performance.

students. Schools above an exogenously set threshold of 10% on the share of disadvantaged students are eligible to receive the funding, schools below are not. However, even when a quasi-experiment is set up, one cannot force individuals (or schools) to comply with the treatment assigned. This is the case of studies in which a certain number of units (e.g., individuals, groups of individuals, schools, companies and so on) are randomly assigned to receive a treatment (e.g., a drug, a training course, additional school funding, and so on), but not all the units that are assigned to receive it are actually treated (and vice versa for units assigned to the control branch). This issue is widely known as *imperfect compliance* problem (Angrist et al.; 1996; Balke and Pearl; 1997). For instance, in the case of our application, eligible schools need to comply with a minimal amount of additional teaching requirement to access the funding. In these cases, researchers can make use of a secondary treatment or *instrument* (i.e., being eligible for additional funding) to isolate the causal effects of the primary treatment (i.e., actually receiving the funding) on the outcome of interest. These designs are referred to as instrumental variable (IV) settings (Angrist et al.; 1996; Angrist and Krueger; 2001). In scenarios where the instrument is not randomized, but assigned based on threshold, and there is room for imperfect compliance between the treatment assigned and the treatment actually received, we are in the presence of a so-called *fuzzy RD* design (Trochim; 1984; Hahn et al.; 2001). Hence, fuzzy RD and IV methodologies are two sides of the same coin, and both techniques are tailored to draw causal inference in imperfect compliance settings (for further discussion on the similarities between fuzzy RD and IV see Lee and Lemieux; 2010). However, both these methodologies are not built to data-drivenly discover the subgroup with heterogeneous causal effects.

In recent years, various algorithms have been proposed to discover and estimate heterogeneous effects (see Dominici et al.; 2021, for a critical review). However, most of these techniques are tailored for drawing causal inference in settings where the treatment is randomly assigned to the units and do not address imperfect compliance issues. Nonetheless, in the real world, the implementation of policies or interventions often results in imperfect compliance which makes the policy evaluation complicated. Recently, some machine learning techniques have been proposed to discover heterogeneous effects while dealing with imperfect compliance using tailored reworks of tree-based algorithms (Bargagli-Stoffi and Gnecco; 2020; Bargagli Stoffi and Gnecco; 2018; Wang et al.; 2018; Johnson et al.; 2019), ensemble of trees algorithms (Athey et al.; 2019) or deep learning methodologies (Hartford et al.; 2017). However, these methods exhibit four principal limitations: (i) random forest-based algorithms for causal inference require large samples to converge to a good asymptotic behaviour for the estimation of causal effects, as shown in Hahn et al. (2019) and Wendling et al. (2018); (ii) deep learning-based algorithms require computationally expensive explorations of the space of possible hyper-parameters configurations (Zhang and Wallace; 2017), are extremely sensitive to tuning parameters (Novak et al.; 2018) and are often not able to tolerate significant sources of unmeasured variation (Prosperi et al.; 2020); (iii) results obtained from ensemble methods, such as forest-based algorithms and deep neural networks, are hard to interpret by human experts because their non-linear parametrizations of the covariate space are complex to probe (Lee et al.; 2020); (iv) more interpretable algorithms are based on single learning that performs worse as compared to multiple learning algorithms (i.e., ensemble methods).

## 1.2. Contributions

To address and accommodate the shortcomings introduced in the previous Sections, this paper innovates the literature in both a methodological and an empirical perspective.

First, we develop a machine learning algorithm tailored to draw causal inference in the presence of imperfect compliance. These situations, where the assignment depends on the observed and unobserved potential outcomes, are also referred to as *irregular assignment mechanisms* in the causal inference literature (Imbens and Rubin; 2015). The proposed method, Bayesian Instrumental Variable Causal Forest (BCF-IV), is an ensemble semi-parametric Bayesian regression model that directly builds on the Bayesian Additive Regression Trees (BART) algorithm (Chipman et al.; 2010). BCF-IV is tailored to discover and estimate the heterogeneous effects on the subgroup of units that comply with the treatment assignment (see Section 2.2), the so-called *compliers*. In this sense, the estimated effects can be seen as *doubly local effects* (namely, subgroup effects for the compliers subpopulation). In particular, BCF-IV discovers the heterogeneity in the form of an interpretable tree structure where each node of the tree corresponds to a discovered subgroup. For each leaf (final node) of the generated tree, BCF-IV allows to perform multiple hypotheses tests adjustments to control for Type I error rate (familywise error rate), or false discovery rate.

We evaluate the fit of the proposed algorithm by comparing it with two alternative machine learning methods explicitly developed to draw causal inference on the heterogeneous effects in the presence of irregular

assignment mechanisms: namely, the Generalized Random Forests (GRF) algorithm (Athey et al.; 2019) and the Honest Causal Trees with Instrumental Variables (HCT-IV) algorithm (Bargagli-Stoffi and Gnecco; 2020). Using Monte Carlo simulations, we compare and evaluate the algorithms with respect to three critical dimensions: (i) the ability of the algorithms to correctly detect the subgroups with heterogeneous causal effects; (ii) the overall performance in terms of estimation accuracy for the conditional causal effects within the correctly discovered subgroups and; (iii) the false positive rate. These dimensions are consistent with recent evaluations of various machine learning methods for causal inference that highlight the excellence of Bayesian algorithms for causal inference (see, e.g., Hahn et al.; 2019; Wendling et al.; 2018). We show that for each dimension, BCF-IV outperforms both GRF and HCT-IV in small samples and converges to an optimal large sample behaviour.

Second, in an empirical application, BCF-IV is used to evaluate the EEO policy. In particular, we employ BCF-IV to evaluate the heterogeneity using a unique administrative dataset on the universe of Flemish pupils in the first stage of education in the school year 2010/2011 (135,682 students). As the additional funding is allocated to schools based on being above or below an exogenously set threshold regarding the proportion of disadvantaged students, this provides us with a quasi-experimental identification strategy. There is also a second, exogenously set, eligibility criterion stating that schools have to generate a minimum number of teaching hours. This source of imperfect compliance, given by the fact that not all the schools fulfill both the criteria, enables us to exploit a fuzzy regression discontinuity design to draw causal effects. We focus on the effects of additional funding on students' performance: namely if a student gets the most favorable outcome (*A-certificate*). The results of our empirical application suggest that, although the effects of additional funding on the overall population of students are found to be not statistically significant, there is appreciable heterogeneity in the causal effects, and consistent sources of effect variation. The results are discussed in the application's Section.

The methodology proposed in this paper can be more widely applied to evaluations of the heterogeneous impact of an intervention in the presence of an irregular assignment mechanism in social and biomedical sciences. The remainder of this paper is organized as follows: in Section 2 we provide a general overview on causal inference and the applied machine learning frameworks, and we introduce our algorithm. In Section 3 we compare the performance of our algorithm with the performance of other methods already established in the literature. In Section 4 we depict the usage of our algorithm in an educational scenario to evaluate the heterogeneous causal effects of additional funding to schools. Section 5 discusses the results and highlights the further applications of heterogeneous causal effects discovery and targeted policies in education as well as in social and biomedical sciences. R code for the BCF-IV function can be found at <https://github.com/fbargaglistoffi/BCF-IV>.

## 2. Bayesian Instrumental Variable Causal Forest

### 2.1. Notation

This paper contributes to the causal inference literature by establishing a novel machine learning approach for the estimation of conditional causal effects in the presence of an irregular assignment mechanism.

We follow the standard notation of the Rubin's causal model (Rubin; 1974, 1978; Imbens and Rubin; 2015). Given a set of  $N$  units, indexed by  $i = 1, \dots, N$ , we denote with  $Y_i$  a generic outcome variable, with  $W_i$  a binary treatment indicator, with  $\mathbf{X}$  a  $N \times P$  matrix of  $P$  control variables, and with  $\mathbf{X}_i$  the  $i$ -th  $P$ -dimensional row vector of covariates. Given the Stable Unit Treatment Value Assumption (SUTVA), that excludes interference between the treatment assigned to one unit and the potential outcomes of another (Rubin; 1986), we can postulate the existence of a pair of potential outcomes:  $Y_i(W_i)$ . Specifically, the potential outcome for a unit  $i$  if assigned to the treatment is  $Y_i(W_i = 1) = Y_i(1)$ , and the potential outcome if assigned to the control is  $Y_i(W_i = 0) = Y_i(0)$ . We cannot observe for the same unit both the potential outcomes at the same time. However, we observe the potential outcome that corresponds to the assigned treatment:  $Y_i^{obs} = Y_i(1)W_i + Y_i(0)(1 - W_i)$ .

In order to draw proper causal inference in observational studies researchers need to assume *strong ignorability* to hold. This assumption states that:

$$(1) \quad Y_i(W_i) \perp\!\!\!\perp W_i \mid \mathbf{X}_i,$$

and

$$(2) \quad 0 < Pr(W_i = 1 \mid \mathbf{X}_i = x) < 1 \quad \forall x \in \mathcal{X},$$

where  $\mathcal{X}$  is the features space. The first assumption (*unconfoundedness*) rules out the presence of unmeasured confounders while the second condition (*common support*) needs to be invoked to be able to estimate the unbiased treatment effect on all the support of the covariates space. If these two conditions hold, we are in the presence of the so-called *regular assignment mechanism* (Imbens and Rubin; 2015). In such a scenario the Average Treatment Effect (ATE) can be expressed as:

$$(3) \quad \tau = \mathbb{E} \left[ Y_i^{obs} \mid W_i = 1 \right] - \mathbb{E} \left[ Y_i^{obs} \mid W_i = 0 \right],$$

and one can define, following Athey and Imbens (2016), the Conditional Average Treatment Effect (CATE) as:

$$(4) \quad \tau(x) = \mathbb{E} \left[ Y_i^{obs} \mid W_i = 1, \mathbf{X}_i = x \right] - \mathbb{E} \left[ Y_i^{obs} \mid W_i = 0, \mathbf{X}_i = x \right].$$

CATE is central for targeted policies as it enables the researcher to investigate the heterogeneity in causal effects. For instance, we may be interested in assessing how the effects of an intervention vary within different sub-populations.

In observational studies, the assignment to the treatment may be different from the reception of the treatment. In these scenarios, where one allows for non-compliance between the treatment assigned and the treatment received, one can assume that the assignment is unconfounded, wherein the receipt is confounded (Angrist et al.; 1996). In such cases, one can rely on an instrumental variable (IV),  $Z_i$ , to draw proper causal inference.<sup>1</sup>  $Z_i$  can be thought as a randomized assignment to the treatment, that affects the receipt of the treatment  $W_i$ , without directly affecting the outcome  $Y_i$  (*exclusion restriction*). Thus, one can then express the treatment received as a function of the treatment assigned:  $W_i(Z_i)$ .

In the following we assume the classical four IV assumptions (Angrist et al.; 1996) – *monotonicity, existence of compliers, unconfoundedness of the IV, exclusion restriction* – to hold. These assumptions need to hold to interpret the estimands proposed below as causal, and, in particular, researchers should devote attention to the necessary monotonicity assumption. Monotonicity states that for each subject, the level of the treatment that a subject would take if given a level of the IV is a monotonic increasing function of the level of the IV (Angrist et al.; 1996). Hence, this assumption, rules out the presence of so-called *defiers*: i.e., no unit who would be always defying from the treatment assigned, taking the treatment when assigned to the control and viceversa. Under the four IV assumptions, IV can then be used to identify the causal effect in the subset of units that comply with the treatment assigned (i.e., the compliers). We refer the reader to the Supplementary Material for a detailed discussion of the four assumptions and how they are assumed to hold in our application reported in Section 4. If the assumptions hold, one can get the causal effect of the treatment on the sub-population of compliers, the so-called Complier Average Causal Effect (CACE), that is:

$$(5) \quad \tau^{cace} = \frac{\mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0]}{\mathbb{E}[W_i \mid Z_i = 1] - \mathbb{E}[W_i \mid Z_i = 0]} = \frac{ITT_Y}{\pi_C},$$

where the numerator represents the average effect of the instrument, also referred to as Intention-To-Treat effect, and the denominator represents the overall proportion of units that comply with the treatment assignment, also referred to as proportion of compliers (Angrist et al.; 1996). CACE is also sometimes referred as Local Average Treatment Effects (see Angrist and Pischke; 2008) and represents the estimate of causal effect of the assignment to treatment on the principal outcome,  $Y_i$ , for the subpopulation of compliers (Imbens and Rubin; 2015). In this paper we consider the following conditional version of CATE. The conditional CACE,  $\tau^{cace}(x)$ , can be thought as the CACE for a sub-population of observations defined by a vector of characteristics  $x$ :

$$(6) \quad \tau^{cace}(x) = \frac{\mathbb{E}[Y_i \mid Z_i = 1, \mathbf{X}_i = x] - \mathbb{E}[Y_i \mid Z_i = 0, \mathbf{X}_i = x]}{\mathbb{E}[W_i \mid Z_i = 1, \mathbf{X}_i = x] - \mathbb{E}[W_i \mid Z_i = 0, \mathbf{X}_i = x]} = \frac{ITT_Y(x)}{\pi_C(x)},$$

where the numerator is the conditional intention to treat Intention-To-Treat effect, and the denominator the conditional proportion of compliers (Angrist et al.; 1996).

---

<sup>1</sup>Throughout the paper we assume the IV to be binary but, one could relax this assumption. However, as there are currently only a few studies that develop machine learning algorithms for the estimation of heterogeneous causal effects with a continuous treatment variable (see Woody et al.; 2020), we leave the investigation of these algorithms to further research.

## 2.2. Bayesian Instrumental Variable Causal Forest

In this paper, we propose an algorithm for the estimation of CATE in an irregular assignment mechanism scenario. In particular, we adapt the Bayesian Causal Forest (BCF) algorithm (Hahn et al.; 2020) for such a task. BCF was originally proposed for regular assignment mechanisms. This algorithm extends to a causal inference setting the Bayesian Additive Regression Trees (BART) algorithm (Chipman et al.; 2010), which in turn builds on the seminal Classification and Regression Trees (CART) algorithm (Breiman et al.; 1984). CART is a widely used algorithm for the construction of binary trees where each node is splitted into only two branches (see Zhang and Singer; 2010, for more details). The first node of the tree is called the root, its final nodes are referred to as leaves.

The accuracy of the predictions of binary trees can be dramatically improved by iteratively constructing the trees, and aggregating their results. BART, as well as BCF, are sum-of-trees ensemble algorithms, and their estimation approach relies on a fully Bayesian probability model (Kapelner and Bleich; 2013). In particular, the BART model can be expressed as:

$$(7) \quad Y_i = f(\mathbf{X}_i) + \epsilon_i \approx \mathcal{T}_1(\mathbf{X}_i) + \dots + \mathcal{T}_q(\mathbf{X}_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2),$$

where each of the  $q$  distinct binary trees is denoted by  $\mathcal{T}$ , where  $\mathcal{T}$  represents the entire tree: its structure, its nodes and its leaves (terminal nodes). The Bayesian component of the algorithm is incorporated in a set of three different priors on: (i) the structure of the trees (this prior is aimed at limiting the complexity of any single tree  $\mathcal{T}$  and works as a regularization device); (ii) the probability distribution of data in the nodes (this prior is aimed at shrinking the node predictions towards the center of the distribution of the response variable  $Y_i$ ); (iii) the error variance  $\sigma^2$  (which bounds away  $\sigma^2$  from very small values that would lead the algorithm to overfit the training data). The aim of these priors is to “regularize” the algorithm, preventing single trees to dominate the overall fit of the model (Kapelner and Bleich; 2013). For further details on the priors specifications we refer to Chipman et al. (2010).

The BCF algorithm proposed by Hahn et al. (2020) is a semi-parametric Bayesian regression model that directly builds on BART. It, however, introduces some significant changes in order to estimate heterogeneous treatment effects in regular assignment mechanisms (even in the presence of strong confounding). The principal novelties of this model are the expression of the conditional mean of the response variable as a sum of two functions and the introduction, in the BART model specification for causal inference, of an estimate of the propensity score,  $E[W_i = 1 | \mathbf{X}_i = x] = \pi(x)$ , in order to improve the estimation of heterogeneous treatment effects.<sup>2</sup> Results from empirical Monte Carlo simulations studies have shown an excellent performance of BART and BCF in causal inference tasks (Dorie et al.; 2019; Hahn et al.; 2019; Wendling et al.; 2018).

Here, we introduce the Bayesian Causal Forest with Instrumental Variable (BCF-IV) algorithm, which is tailored to data-drivenly discover and estimate heterogeneous causal effects in an interpretable way. Heterogeneous effects analyses are typically conducted for subgroups defined *a priori* to avoid potential cherry-picking problems connected to reporting the causal effects only for subgroups with extremely high or low treatment effects (Cook et al.; 2004). However, defining *a priori* these groups has two main shortcomings: (i) it requires a fairly good understanding of the treatment effect; (ii) researchers may miss unexpected subgroups. To overcome these limitations, we propose a data-driven approach to discover heterogeneous effects by using *honest* sample splitting (Athey and Imbens; 2016). Following a *honest* sample splitting approach we divide the data into two subsamples: one to build the tree for the discovery of the heterogeneous effects (*discover subsample*:  $\mathcal{I}^{dis}$ ) and another for making inference (*inference subsample*:  $\mathcal{I}^{inf}$ ). Accordingly to Athey and Imbens (2016), for both the simulations and empirical application, half of the sample is assigned to  $\mathcal{I}^{dis}$  and the other half to  $\mathcal{I}^{inf}$ .<sup>3</sup> As inference is a separated task from model selection, *honest* sample-splitting enables an *honest* inference for effect modification.<sup>4</sup> Algorithm (1) provides a general overview on the proposed BCF-IV algorithm.

<sup>2</sup>It is important to highlight that the propensity score is not used to estimate the causal effects but to moderate the distortive effects in treatment heterogeneity discovery due to strong confounding. Moreover, since BCF includes the entire predictors’ vector,  $\mathbf{X}_i$ , even if the propensity score is misspecified or poorly estimated, the model allows for the possibility that the response remains correctly specified (Hahn et al.; 2020). In the Supplementary Material, we show that even if the estimate  $\hat{\pi}(x)$  of the propensity score is incorrectly specified the results are still widely robust.

<sup>3</sup>As highlighted by Lee et al. (2020) different proportions of units could be assigned to the discovery and inference subsamples. In the R function implementing BCF-IV, we leave to the researcher the choice of the ratio between discovery and inference subsamples.

<sup>4</sup>Alternatively, *honest* inference could be obtained by techniques based on multiple testing and controlling for family wise error rate such as the ones proposed by Hsu et al. (2015) and Johnson et al. (2019).

**Algorithm 1** Overview of Bayesian Causal Forest with Instrumental Variable (BCF-IV)

**Inputs:**  $N$  units  $i$  ( $X_i, Z_i, W_i, Y_i$ ), where  $X_i$  is the feature vector,  $Z_i$  is treatment assignment (instrumental variable),  $W_i$  is the treatment receipt, and  $Y_i$  is the observed response.

**Outputs:** (1) a tree structure discovering the heterogeneity in the causal effects, and (2) estimates of the Complier Average Causal Effects within its leaves.

- The Honest Splitting Step:
  1. Randomly split the total sample into a discovery ( $\mathcal{I}^{dis}$ ) and an inference subsample  $\mathcal{I}^{inf}$ .
- The Discovery Step (performed on  $\mathcal{I}^{dis}$ ) (Section 2.2.1):
  2. Estimation of the Conditional CACE:
    - a) Estimate the conditional Intention-To-Treat:  $\widehat{ITT}(x)$ ;
    - b) Estimate the conditional proportion of compliers:  $\hat{\pi}_C(x)$ ;
    - c) Estimate the conditional CACE,  $\hat{\tau}^{cace}(x)$ , using the estimated values from (a) and (b) as in (15) (for a continuous outcome) or in (16) (for a binary outcome).
  3. Heterogeneous subpopulations discovery:
    - (d) Discover the heterogeneous effects by fitting a decision tree using the data  $(\hat{\tau}^{cace}(x), \mathbf{X}_i)$ .
- The Inference Step (performed on  $\mathcal{I}^{inf}$ ) (Section 2.2.2):
  4. Estimate the  $\hat{\tau}^{cace}(x)$  for all the discovered subpopulations (i.e., nodes and leaves) in the tree discovered in (d);
  5. Perform multiple hypotheses tests adjustments of the p-values to control for familywise error rate or – less stringently – for the false discovery rate;
  6. Run weak-instrument tests within every node and discard those nodes where a weak-instrument issue is detected.

**2.2.1. Discovery Step: Discovering Heterogeneity in the Conditional CACE**

After dividing the total sample into a discovery and an inference subsample, we perform the discovery of the heterogeneity in the conditional CACE on  $\mathcal{I}^{dis}$ . In particular, the BCF-IV algorithm starts from modifying (7) to adapt it for the estimation of the conditional Intention-To-Treat, by including the IV,  $Z_i$ :

$$(8) \quad Y_i = f(Z_i, \mathbf{X}_i) + \epsilon_i \approx \mathcal{T}_1(Z_i, \mathbf{X}_i) + \dots + \mathcal{T}_q(Z_i, \mathbf{X}_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2),$$

where, for simplicity, we assume the error to be a mean zero additive noise as in Hill (2011), Hahn et al. (2020), and Logan et al. (2019). The conditional expected value can of  $Y_i$  be expressed as:

$$(9) \quad \mathbb{E}[Y_i | Z_i = z, \mathbf{X}_i = x] = g(z, x),$$

and in turn the conditional intention-to-treat,  $ITT_Y(x)$ , is:

$$(10) \quad ITT_Y(x) = \mathbb{E}[Y_i | Z_i = 1, \mathbf{X}_i = x] - \mathbb{E}[Y_i | Z_i = 0, \mathbf{X}_i = x] = g(1, x) - g(0, x).$$

Then, adapting to an irregular assignment mechanism the model proposed by Hahn et al. (2020), we adopt the following functional form for (9):

$$(11) \quad \mathbb{E}[Y_i | Z_i = z, \mathbf{X}_i = x] = \mu(\pi(x), x) + ITT_Y(x)z$$

where  $\pi(x)$  is the propensity score for the IV:  $\pi(x) = E[Z_i = 1 | \mathbf{X}_i = x]$ . The expression of  $\mathbb{E}[Y_i | Z_i = z, \mathbf{X}_i = x]$  as a sum of two functions is central: the first component of the sum,  $\mu(\pi(x), x)$ , directly models the impact of the control variables on the conditional mean of the response (the component that is independent from the treatment effects) while the second component  $ITT_Y(x)z$  models directly the intention-to-treat effect as a nonlinear function of the observed characteristics (this second component captures the heterogeneity in the intention-to-treat). Both the functions  $\mu$  and  $ITT_Y$  are given independent priors. These priors are chosen in line with Hahn et al. (2020) to be for the first component the same priors of Chipman et al. (2010). However, for the second component the priors are changed in a way that allows for less deep, hence simpler trees.

The estimated propensity score is not used for the estimation of the effects but is included, as an additional covariate, in the first component of (11) to mitigate possible problems connected to *regularization induced confounding* (RIC)<sup>5</sup> and *targeted selection*.<sup>6</sup> Moreover, in scenarios where the IV is not randomized ex-ante, the

<sup>5</sup>RIC is analyzed in depth in Hahn et al. (2018). RIC issues rise when the ML algorithm used for regularizing the coefficient does not shrink to zero some coefficients due to a nonzero correlation between  $Z_i$  and  $\mathbf{X}_i$  resulting in an additional degree of bias that is not under the researcher's control.

<sup>6</sup>Targeted selection refers to settings where the treatment (or in an IV scenario the assignment to the treatment) is assigned based on an ex-ante prediction of the outcome conditional on some characteristics  $\mathbf{X}_i$ . We refer to Hahn et al. (2020) for a discussion of targeted selection problems.

inclusion of the estimated propensity score,  $\hat{\pi}(x)$ , leads to an improvement in the discovery of the heterogeneity in the causal effect (Hahn et al.; 2019).

Using a similar rework of BART, one can estimate the conditional proportion of compliers,  $\pi_C(x)$ . In particular, we propose to rework (7) as follows:

$$(12) \quad W_i = l(Z_i, \mathbf{X}_i) + \epsilon_i \approx \mathcal{T}_1(Z_i, \mathbf{X}_i) + \dots + \mathcal{T}_q(Z_i, \mathbf{X}_i) + \psi_i, \quad \psi_i \sim \mathcal{N}(0, \psi^2),$$

where, we assume again, the error  $\psi_i$  to be a mean zero additive noise. The conditional expected value of  $W_i$  can be expressed as:

$$(13) \quad \mathbb{E}[W_i | Z_i = z, \mathbf{X}_i = x] = \delta(z, x),$$

and the conditional proportion of compliers is:

$$(14) \quad \mathbb{E}[W_i | Z_i = 1, \mathbf{X}_i = x] - \mathbb{E}[W_i | Z_i = 0, \mathbf{X}_i = x] = \delta(1, x) - \delta(0, x),$$

where  $\delta(z, x)$  can be estimated, in the case of a binary  $Z_i$ , using the BART methodology for causal effects estimation proposed by Hill (2011) and implemented in R in the `bartCause` package (Dorie et al.; 2020).

Finally, the conditional CACE can be expressed, reworking (5), as the ratio between (11) and (14):

$$(15) \quad \tau^{cace}(x) = \frac{\mathbb{E}[Y_i | Z_i = z, \mathbf{X}_i = x]}{\mathbb{E}[W_i | Z_i = z, \mathbf{X}_i = x]} = \frac{\mu(\pi(x), x) + ITT_Y(x)z}{\delta(1, x) - \delta(0, x)}.$$

In the case of a binary outcome, the proposed methodology is implemented using again, instead of BCF, a suitable rework of the causal BART algorithm proposed by Hill (2011). In this case, the conditional CACE can be expressed as:

$$(16) \quad \tau^{cace}(x) = \frac{\mathbb{E}[Y_i | Z_i = z, \mathbf{X}_i = x]}{\mathbb{E}[W_i | Z_i = z, \mathbf{X}_i = x]} = \frac{g(1, x) - g(0, x)}{\delta(1, x) - \delta(0, x)},$$

where both  $g(\cdot, \cdot)$  and  $\delta(\cdot, \cdot)$  are estimated using BART using as an outcome the output  $Y_i$  and  $W_i$  respectively.

Once one estimated the conditional CACE as in (15), one can build a simple binary tree, using a CART (Breiman; 1984), on the fitted values ( $\hat{\tau}^{cace}(x)$ ) to discover, in an interpretable manner, the drivers of the heterogeneity (fit-the-fit approach). Indeed, as argued by Lee et al. (2021) and Lee et al. (2020), the subgroups discovered from a fit-the-fit approach using CART, are ideal to guarantee high levels of interpretability, where interpretability can be defined as the degree to which a human can understand the cause of a decision or consistently predict the results of the model (Miller; 2019; Kim et al.; 2016). The CART algorithm used for the detection of heterogeneous subgroups is implemented in R using the `rpart` package version 4.1-15 (Therneau et al.; 2015). The maximal depth of the tree can be set by the researcher. In our empirical example and in the simulations, the maximal depth is set to two to maintain a small level of complexity (and, in turn, enhance interpretability) and guarantee enough observations within each node (see Lee et al.; 2021, for further a discussion on tree depth, interpretability and subgroups sizes).

Alternatively, one could directly fit the CART on the estimated unit level ITT in (11). We call this alternative methodology BCF-ITT. BCF-ITT could be potentially helpful in scenarios where (i) one can imagine the heterogeneity to be driven by small values of  $\pi_C(x)$ , or (ii) one is solely interested in the detection of the heterogeneity in the ITT. BCF-ITT may suffer limitations when the ITT is relatively homogeneous while the proportion of compliers in various subgroups is not. In the simulations in the Supplementary Material, we discuss further the comparison between BCF-IV and BCF-ITT and how BCF-IV dominates BCF-ITT in the detection of the heterogeneity in CACE.

### 2.2.2. Estimation of Conditional CACE

Once the heterogeneous patterns in the intention-to-treat (ITT) are learned from the algorithm, one can estimate the conditional CACE,  $\tau^{cace}(x)$  on the inference subsample  $\mathcal{I}^{inf}$ . To do so, one can suitably estimate the method of moments in Equation (6) within all the different sub-populations that were detected in the previous step (see the Supplementary Material).

In the case of a binary instrument ( $Z_i \in \{0, 1\}$ ) and a binary treatment variable ( $W_i \in \{0, 1\}$ ), Angrist et al. (1996) and Imbens and Rubin (2015) revealed that the population versions of the method-of-moments estimator correspond to a Two Stage Least Squares (henceforth referred as 2SLS) estimator of  $\tau^{cace}$ , in the cases where

the four IV assumptions can be assumed to hold. Hence, since this case is analogous to our setting, one can apply the 2SLS method in every node  $\mathbb{X}_j$  of the discovered tree,  $\mathcal{T}$ , for the estimation of the effect on the compliers population, as it is presented by Imbens and Rubin (1997).

The two simultaneous equations of the 2SLS estimator are, in the population,

$$(17) \quad Y_i^{obs} = \alpha + \tau^{cace} W_i + \epsilon_i,$$

$$(18) \quad W_i = \pi_0 + \pi_C Z_i + \eta_i,$$

where  $\mathbb{E}(\epsilon_i) = \mathbb{E}(\eta_i) = 0$ , and  $\mathbb{E}(Z_i \eta_i) = 0$ .<sup>7</sup> In the econometric terminology, the explanatory variable  $W_i$  is *endogenous*, while the IV variable  $Z_i$  is *exogenous*. (17) is referred to as the *outcome equation* and (18) is referred to as the *treatment equation* (Lee and Lemieux; 2010).

We can express the 2SLS equations, conditional on a subpopulation of a node  $\mathbb{X}_j$ , as

$$(19) \quad Y_{i,\mathbb{X}_j}^{obs} = \alpha_{\mathbb{X}_j} + \tau_{\mathbb{X}_j}^{cace} W_{i,\mathbb{X}_j} + \epsilon_{i,\mathbb{X}_j},$$

$$(20) \quad W_{i,\mathbb{X}_j} = \pi_{0,\mathbb{X}_j} + \pi_{C,\mathbb{X}_j} Z_{i,\mathbb{X}_j} + \eta_{i,\mathbb{X}_j},$$

where  $\mathbb{E}(\epsilon_{i,\mathbb{X}_j}) = \mathbb{E}(\eta_{i,\mathbb{X}_j}) = 0$ , and  $\mathbb{E}(Z_{i,\mathbb{X}_j} \eta_{i,\mathbb{X}_j}) = 0$ .

Moreover, the following reduced equation (obtained plugging (20) into (19)) holds:

$$(21) \quad \begin{aligned} Y_{i,\mathbb{X}_j}^{obs} &= \left( \alpha_{\mathbb{X}_j} + \tau_{\mathbb{X}_j}^{cace} \pi_{0,\mathbb{X}_j} \right) + \left( \tau_{\mathbb{X}_j}^{cace} \pi_{C,\mathbb{X}_j} \right) Z_{i,\mathbb{X}_j} + \left( \epsilon_{i,\mathbb{X}_j} + \tau_{\mathbb{X}_j}^{cace} \eta_{i,\mathbb{X}_j} \right) \\ &= \bar{\alpha}_{\mathbb{X}_j} + \gamma_{\mathbb{X}_j} Z_{i,\mathbb{X}_j} + \psi_{i,\mathbb{X}_j}. \end{aligned}$$

In the case of a single instrument, the logic of IV regression is that one can estimate the respective parameters  $\pi_{C,\mathbb{X}_j}$  and  $\gamma_{\mathbb{X}_j} = \tau_{\mathbb{X}_j}^{cace} \pi_{C,\mathbb{X}_j}$  of the regressions (20) and (21) above by least squares, when the observations in each node are independent and identically distributed, then obtaining an estimate of the parameter  $\tau_{\mathbb{X}_j}^{cace}$  in (19).

In particular, for every element  $\mathbf{X}_i$  of a node  $\mathbb{X}_j$ , one can estimate  $\tau^{CACE}(\mathbf{X}_i) = \tau_{\mathbb{X}_j}^{cace}$  through 2SLS, as the following ratio (Imbens and Rubin; 2015):

$$(22) \quad \hat{\tau}^{CACE}(\mathbf{X}_i) \equiv \hat{\tau}_{\mathbb{X}_j}^{2SLS} = \frac{\hat{\gamma}_{\mathbb{X}_j}}{\hat{\pi}_{C,\mathbb{X}_j}}.$$

The theoretical properties of these conditional estimators are reported in the Supplementary Material.

Within each subgroup  $\mathbb{X}_j$ , we evaluate whether  $\hat{\tau}_{\mathbb{X}_j}^{2SLS}$  is significantly different from zero, and we run weak-instrument tests for the estimated effects and we discard those nodes where a weak-instrument issue is detected. This is done to avoid problems connected to the discovery of heterogeneous effects driven by small proportion of compliers in the subgroup (aka potential weak-instrument issue).

Furthermore, for each leaf of the generated tree, we implement a set of multiple hypotheses tests adjustments to control for Type I error rate (aka familywise error rate – e.g., the probability of making a Type I error among a specified group, or *family*, of tests), or false discovery rate (e.g., the expected proportion of false discoveries amongst the rejected hypotheses). Indeed, while the main focus of the BCF-IV algorithm is to discover heterogeneous effects in the form of an interpretable tree structure and then estimate the subgroup effects, it may be of interest to assess the significance of the discovered heterogeneity at leaves-level while controlling for potential spurious heterogeneity discovery. In particular, the adjustment methods that are implemented in BCF-IV are the Bonferroni correction and the corrections proposed by Holm (1979), Hochberg (1988), Hommel (1988), Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001). The first four methods are designed to control for familywise error rate, while the last two corrections control for the false discovery rate. The false discovery rate is a less stringent condition than the family-wise error rate, so the latter two methods are more powerful than the others. By default, BCF-IV implements Holm’s method, which is more stringent, dominates the Bonferroni correction, and is valid under arbitrary assumptions, while Hochberg’s and Hommel’s methods are valid when the hypothesis tests are independent or when they are non-negatively associated (Sarkar and Chang; 1997). However, depending on the specific setting of application and the specific research interest, any of the other methods can be implemented instead.

<sup>7</sup>The latter comes from the fact that (17) is assumed to represent the linear projection of  $W_i$  onto  $Z_i$ .



### 3. Monte Carlo Simulations

The overall goal of this Section is to provide insights on how discovering and estimating the causal effects in the presence of imperfect compliance is a complex task that depends on multiple factors such as the structure and magnitude of heterogeneity in the causal effects, the size of the available data, the strength of the IV used and the heterogeneity of compliance rate among different subgroups. To do so, a set of Monte Carlo simulations' scenarios is designed to analyse the performance of the proposed methodologies as a function of: (1) the structure and magnitude of the heterogeneous effects within the different subgroups; (2) the size of the data used in the analysis. In the Supplementary Material, we provide a number of robustness checks of the Monte Carlo simulation. In particular, we focus on what happens to the fit of the three algorithms when one introduces: (3) the different compliance rates, ranging from a strong instrument (high compliance) to a weak one (low compliance); and (4) the variation in the compliance rates among different subgroups; (5) confounding in the generation of the IV; (6) covariance in the covariates matrix; and (7) misspecification in the propensity score. The performance of BCF-IV does not significantly deteriorate, as compared to the baseline models introduced in the following simulations.

The following simulations were designed to closely mimic the empirical example discussed later in Section 4 in order to furnish a useful guidance on the reliability of the empirical results. In particular, the generated Monte Carlo data reproduce an imperfect compliance setting under the assumption of one-sided non-compliance (i.e., units that are not assigned to the treatment will never be able to get it) with potential heterogeneity in the causal effects and in the compliance rate. Sample sizes (4,000) and compliance rates (0.75) are simulated to be slightly smaller than the ones in the empirical application on EEO data reported later in Section 4 in order to provide conservative guidance on the algorithms' performances.<sup>8</sup>

To evaluate the performance of the BCF-IV and BCF-ITT algorithms, we contrast them with other two methods tailored for drawing causal inference in irregular assignment mechanism scenarios: the Honest Causal Trees with Instrumental Variable (HCT-IV) algorithm (Bargagli-Stoffi and Gnecco; 2020) and the Generalized Random Forests (GRF) algorithm (Athey et al.; 2019). Both these algorithms were shown to outperform other causal machine learning methodologies in irregular assignment mechanisms. Since the foremost focus of this paper is on discovery of the heterogeneity in the effects in terms of an interpretable tree structure and the estimation of this heterogeneity in all the detected subgroups, we compare and evaluate the algorithms on three critical dimensions: (i) the ability of the algorithms to correctly detect the subgroups (at leaves level) with heterogeneous causal effects, (ii) the capacity of BCF-IV to control for false positive discovery at leaves level through p-value adjustment, and (iii) the overall performance in terms of estimation accuracy for the conditional causal effects within the correctly discovered subgroups at leaves level (namely, the ones with heterogeneous effects). With respect to the first dimension, we compare BCF-IV and BCF-ITT with HCT-IV as these three methods are able to identify, in an interpretable manner, the groups with highest levels of heterogeneity in the causal effect. In this case, we leave out the comparison with GRF as this technique is not providing interpretable guidance on the subgroups with highest levels of heterogeneity. For the last dimension, we compare the estimation ability of BCF-IV with the one of GRF.<sup>9</sup>

We start by generating the set of potential outcomes  $Y_i(Z_i)$ , potential treatments  $W_i(Z_i)$ , and covariates  $\mathbf{X}_i$  for each unit  $i$ , where the observed treatment and outcome are based on the value of the instrument  $Z_i$ . For each data-generating process, we generate the covariate matrix  $\mathbf{X}$  with 10 binary covariates from  $X_{i1}$  to  $X_{i10}$  where each covariate is sampled from a binomial distribution with success probability equal to 0.5:  $X_{ip} \sim \text{Binom}(0.5)$ . The binary instrument  $Z_i$  is also drawn from a binomial distribution,  $Z_i \sim \text{Binom}(0.5)$ . Given one-sided non-compliance, for units not assigned to the treatment their potential treatment is always 0, namely  $W_i(0) = 0$ . The potential treatment is a Bernoulli trial with a success rate (or compliance rate) of  $\pi_{X_{i1}, X_{i2}}$ . This means that the compliance rate depends on the two covariates  $X_{i1}$  and  $X_{i2}$  mimicking a setting where people assigned to the treatment may decide to actually receive it or not based on the values of some observed covariates. The potential outcomes of units having not received the instrument are sampled from a standard normal distribution,  $Y_i(0) \sim \mathcal{N}(0, 1)$ , while the potential outcomes of units having received the instrument are a function of the treatment value and  $\tau^{cace}(\mathbf{X}_i)$ :  $Y_i(1) = Y_i(0) + W_i(1)\tau^{cace}(\mathbf{X}_i)$ .  $\tau^{cace}(\mathbf{X}_i)$  indicates the heterogeneous conditional CACE, which is a function of the set of covariates levels of each unit. To simplify, we make  $\tau^{cace}(\mathbf{X}_i)$  dependent on  $X_{i1}$  and  $X_{i2}$  and we generate two scenarios:

<sup>8</sup>The smallest learning sample for the EEO has 4,300 observations and an overall compliance rate of roughly 0.80.

<sup>9</sup>Since BCF-IV and BCF-ITT are based on the same estimator we implement just BCF-IV for this contrast.

1. *strong heterogeneity scenario*:

$$\tau^{cace}(\mathbf{X}_i) = \begin{cases} k & \text{if } \mathbf{X}_i \in \ell_1 = \{\mathbf{X}_i : X_{i1} = 0, X_{i2} = 0\}; \\ -k & \text{if } \mathbf{X}_i \in \ell_2 = \{\mathbf{X}_i : X_{i1} = 1, X_{i2} = 1\}; \\ 0 & \text{otherwise;} \end{cases}$$

2. *slight heterogeneity scenario*:

$$\tau^{cace}(\mathbf{X}_i) = \begin{cases} k & \text{if } \mathbf{X}_i \in \ell_1 = \{\mathbf{X}_i : X_{i1} = 0, X_{i2} = 0\}; \\ -k & \text{if } \mathbf{X}_i \in \ell_2 = \{\mathbf{X}_i : X_{i1} = 1, X_{i2} = 1\}; \\ 0.5k & \text{if } \mathbf{X}_i \in \ell_3 = \{\mathbf{X}_i : X_{i1} = 1, X_{i2} = 0\}; \\ -0.5k & \text{if } \mathbf{X}_i \in \ell_4 = \{\mathbf{X}_i : X_{i1} = 0, X_{i2} = 1\}; \end{cases}$$

where  $k$  is a positive number and, to simplify the notation, we refer to  $\ell_1$  for  $X_{i1} = 0, X_{i2} = 0$ ,  $\ell_2$  for  $X_{i1} = 1, X_{i2} = 1$ ,  $\ell_3$  for  $X_{i1} = 1, X_{i2} = 0$ , and  $\ell_4$  for  $X_{i1} = 0, X_{i2} = 1$ .

Half of the learning sample was assigned to the discovery subsample and the other half to the inference subsample.

The performance of the BCF-IV algorithm is evaluated using a number of goodness-of-fit measures, averaged over  $M = 500$  generated data sets. With respect to the first two dimensions – the ability to correctly discover the subgroups with heterogeneous effects at leaves level and the capacity of BCF-IV to control for false positive discovery at leaves level – we evaluated the performance of the algorithms using a suitable rework of common performance measures from the classification literature. With respect to the first dimension, we assess the True Positive Rate (TPR) as the rate at which the algorithm is able to correctly detect the subgroups with true heterogeneity at leaves level ( $\ell_1$  and  $\ell_2$  in the case of strong heterogeneity, and  $\ell_1, \ell_2, \ell_3$  and  $\ell_4$  in the case of slight heterogeneity). TPR ranges between zero (worst performance) and one (best performance) and is computed as the ratio  $\frac{TP}{TP+FN}$  where TP stands for the number of True Positive, and FN stands for the number of False Negatives. Regarding the second dimension, the performance of the algorithm to control for false positive discovery is evaluated using the False Positive Rate (FPR). FPR ranges from zero (best performance) to one (worst performance) and is the ratio  $\frac{FP}{FP+TN}$ , where FP stands for the number of False Positives and TN for the number of True Negatives. We refer to Table 1 for additional details. The evaluation is performed on leaves level. The intuition is that if the algorithm is able to detect/discard heterogeneous subgroup at the final level (read leaves) it will also be able to detect it at any intermediate level (read nodes).<sup>10</sup> Finally, the overall performance in terms of estimation accuracy for the conditional effects within the correctly discovered subgroups is evaluated using a number of performance measures for all the correctly discovered heterogeneous subgroups at leaves level. In particular, we measured for the heterogeneous subgroups the Monte Carlo estimated bias, the Monte Carlo Mean Squared Error (MSE), and the Monte Carlo coverage. These measures are introduced in more detail in the Supplementary Material.

Building on the data generating process described above, we introduce variations in the effect size of the heterogeneous causal effects, in the structure of the heterogeneity and in the sample sizes. These features are important for various reasons. Firstly, as the size of the effect within the heterogeneous subgroups and its difference between various such subgroups are growing the algorithms should depict a higher ability to correctly discover these subgroups. Secondly, it is important to assess the ability of the algorithms to discover the correct subgroups in the presence of various structures of the heterogeneity. Finally, both the discovery and the estimation ability of the algorithms are expected to vary based on the size of the learning sample.

In the proposed simulation setting, we consider three varying factors: (i) the effect size  $k$  from 0 (no heterogeneity) to 2 (greater level of heterogeneity); (ii) the structure of the heterogeneity: *slight heterogeneity scenario* vs *strong heterogeneity scenario* and; (iii) the sample size with  $N = 1,000$ , or  $4,000$ . These sample sizes refer to the size of data used for either the discovery or the inference step. The compliance rate is set to be 0.75 (strong instrument scenario) and the covariance between the covariates to be zero. The results for TPR are depicted in

<sup>10</sup>We want to highlight that, while the TPR is evaluated comparing BCF-IV (and BCF-ITT) with HCT-IV, FPR is evaluated just for BCF-IV. Indeed, as HCT-IV is not able to correct for false positive (or false discovery) rate the advantage of BCF-IV are manifest in this scenario. Hence – to build the fairest comparison – we employed p-values correction just in the evaluation of the FPR for BCF-IV. The significance level is set to 0.05.

Figures 1. Figure 2 reports the results for the FDR. The results for the overall performance in terms of estimation accuracy within the subgroups are shown in Tables 2 and 3. Figure 1 depicts the results for the TPR in the *strong* and *slight heterogeneity* scenario both with 1,000 and 4,000 observations. We compare the results for the subgroups identified with BCF-IV (blue line) with the ones of BCF-ITT (green line) and HCT-IV (red line). In both scenarios, BCF-IV and BCF-ITT are able to converge to the highest possible TPR as the effect size grows larger. This convergence is faster as the sample sizes increases. Moreover, their performance is very similar, and they always outperform HCT-IV. Figure 2 depicts the results for the FDR for BCF-IV in the same scenarios. These results complement the information from the previous figure on TPR. While the former results focus on the ability of BCF-IV to correctly discover the heterogeneous subgroups, the latter results evaluate the ability of the algorithm to control for false discoveries (FPR). In both *strong* and *slight heterogeneity* scenarios, BCF-IV is able to keep a FPR very close to zero by discarding false discoveries through p-values adjustments controlling for the familywise error rate.

Tables 2 and 3 show the results for the estimated conditional effects within the subgroups with heterogeneous effects for the case of *strong heterogeneity* with 1,000 and 4,000 data points respectively.<sup>11</sup> Also with respect to the estimation of the conditional CACEs, BCF-IV seems to perform in a very good way irrespective of the effect size. Indeed, the Monte Carlo estimated MSE and bias are smaller than the ones of GRF, and the Monte Carlo coverage approaches the value of 0.95.

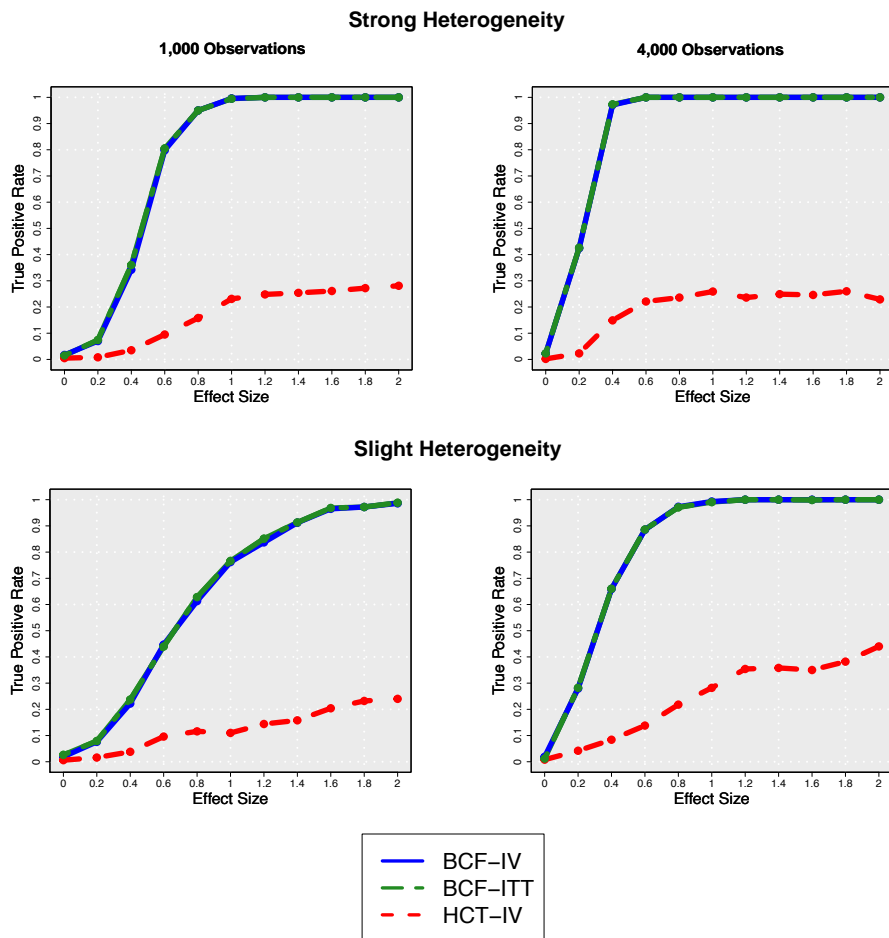


Fig 1: TPR for the *strong heterogeneity* and the *slight heterogeneity* scenarios.

<sup>11</sup>The estimation results for the case of *slight heterogeneity* mimic the results for the case of *strong heterogeneity*, as these two designs do not differently affect the ability of the algorithms to precisely estimate the conditional effects.

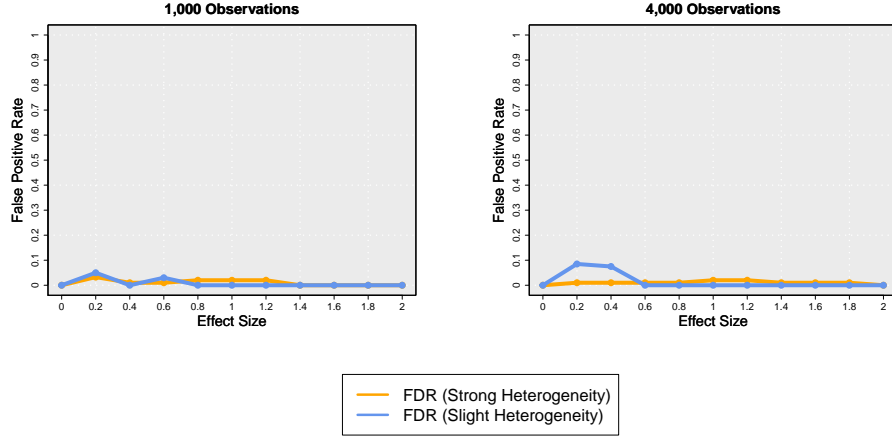


Fig 2: FDR for the *strong heterogeneity* and the *slight heterogeneity* scenarios.

		True Condition	
		HES	Not HES
Prediction	Predicted as HES	True positive $TP$ (Correct)	False positive $FP$ (Incorrect)
	Not Predicted as HES	False negative $FN$ (Incorrect)	True negative $TN$ (Correct)

TABLE 1

Classification for heterogeneous subgroups detection, where HES refers to Heterogeneous Effect Subgroup at leaves level.

Effect Size	$MSE(\hat{\tau}_{l_1}^{cace})$	$Bias(\hat{\tau}_{l_1}^{cace})$	$Coverage(\hat{\tau}_{l_1}^{cace})$	$MSE(\hat{\tau}_{l_2}^{cace})$	$Bias(\hat{\tau}_{l_2}^{cace})$	$Coverage(\hat{\tau}_{l_2}^{cace})$
BCF-IV						
0	0.030	0.138	0.952	0.031	0.140	0.938
0.1	0.021	0.020	0.974	0.019	0.018	0.980
0.2	0.028	-0.004	0.960	0.027	-0.014	0.968
0.3	0.027	0.008	0.956	0.031	0.010	0.938
0.4	0.028	-0.013	0.962	0.031	-0.022	0.950
0.5	0.029	0.007	0.954	0.026	-0.014	0.964
0.6	0.026	-0.004	0.960	0.031	-0.016	0.944
0.7	0.028	0.003	0.958	0.032	-0.012	0.944
0.8	0.030	-0.003	0.944	0.029	-0.002	0.952
0.9	0.031	0.010	0.936	0.029	0.002	0.944
1	0.028	0.004	0.952	0.029	0.004	0.950
GRF						
0	0.018	0.107	0.998	0.017	0.103	0.998
0.1	0.015	-0.072	1.000	0.015	-0.075	1.000
0.2	0.070	-0.235	0.962	0.066	-0.222	0.954
0.3	0.135	-0.328	0.732	0.136	-0.329	0.708
0.4	0.197	-0.398	0.646	0.197	-0.399	0.648
0.5	0.235	-0.427	0.652	0.234	-0.431	0.658
0.6	0.249	-0.439	0.684	0.264	-0.445	0.672
0.7	0.232	-0.394	0.752	0.234	-0.408	0.752
0.8	0.216	-0.367	0.798	0.224	-0.384	0.782
0.9	0.208	-0.344	0.838	0.219	-0.357	0.824
1	0.186	-0.318	0.862	0.181	-0.309	0.876

TABLE 2

Simulation results for 1,000 data points and strong instrument.

<i>Effect Size</i>	$MSE(\hat{\tau}_{l_1}^{cace})$	$Bias(\hat{\tau}_{l_1}^{cace})$	$Coverage(\hat{\tau}_{l_1}^{cace})$	$MSE(\hat{\tau}_{l_2}^{cace})$	$Bias(\hat{\tau}_{l_2}^{cace})$	$Coverage(\hat{\tau}_{l_2}^{cace})$
BCF-IV						
0	0.007	0.068	0.950	0.008	0.071	0.944
0.1	0.007	0.004	0.960	0.007	0.007	0.944
0.2	0.007	-0.004	0.950	0.007	-0.003	0.942
0.3	0.007	0.001	0.940	0.007	-0.003	0.960
0.4	0.007	0.002	0.962	0.007	0.002	0.940
0.5	0.007	-0.008	0.958	0.007	0.003	0.956
0.6	0.007	-0.006	0.940	0.008	-0.004	0.938
0.7	0.006	0.002	0.968	0.007	-0.004	0.948
0.8	0.008	0.006	0.942	0.007	-0.001	0.952
0.9	0.007	0.001	0.960	0.008	0.002	0.942
1	0.008	-0.005	0.938	0.006	0.007	0.966
GRF						
0	0.006	0.063	1.000	0.006	0.063	1.000
0.1	0.013	-0.080	1.000	0.012	-0.079	1.000
0.2	0.032	-0.143	0.968	0.030	-0.135	0.976
0.3	0.033	-0.126	0.972	0.033	-0.129	0.980
0.4	0.031	-0.102	0.992	0.028	-0.097	0.978
0.5	0.030	-0.093	0.984	0.025	-0.078	0.986
0.6	0.023	-0.051	0.994	0.025	-0.055	0.996
0.7	0.017	-0.026	1.000	0.019	-0.032	0.998
0.8	0.018	-0.014	1.000	0.016	-0.015	0.998
0.9	0.017	-0.011	1.000	0.019	-0.005	1.000
1	0.016	-0.008	0.996	0.014	-0.004	0.998

TABLE 3

*Simulation results for 4,000 data points and strong instrument.*

## 4. Heterogeneous Causal Effects of Education Funding

There is a wide consensus that education positively influences labor market outcomes (see the review by Psacharopoulos and Patrinos; 2018). Moreover, the question on whether or not school spending affects students' performances has been central in the economic literature (Coleman; 1966; Hanushek; 2003; Jackson et al.; 2015; Jackson; 2018). Students' performance can be driven by multiple factors connected with students' characteristics and environmental characteristics. However, to the best of our knowledge, this is the first paper to study the heterogeneous impact of additional school funding on students' performance using machine learning techniques tailored for causal inference. In this Section we apply the BCF-IV algorithm to evaluate the impact and estimate the heterogeneity in the effects of additional funding to schools with disadvantaged students on students' performance. First, we describe the data used for this application. Next, we depict the identification strategy. Finally, we describe the results obtained and their relevance in the economics of education literature.

### 4.1. Data

The EEO program, promoted by the Flemish Ministry of Education to encourage "Equal Educational Opportunities", provides additional funding for secondary schools with a significant share of disadvantaged students. Owing to the funding schools can hire additional teachers and increase the number of teaching hours. Pupils are considered to be disadvantaged on the basis of five different indicators: (i) the pupil lives outside the family; (ii) the pupil does not speak Dutch as a native language; (iii) the mother of the pupil does not have a secondary education degree; (iv) the pupil receives educational grant guaranteed for low income families; and (v) one of the parents is part of the travelling population. In order for a school to be eligible for the EEO funding, it needs to satisfy two conditions: the first condition is that the share of students with at least one of the five characteristics has to exceed an exogenously set threshold; to avoid fragmentation of resources, the second condition requires that the additional resources should be at least larger than six teaching hours a week. The exogenous threshold is, for students in the first two years of secondary education (first stage students), a minimum share of 10% disadvantaged students.

The Flemish Ministry of Education provided us with data on the universe of pupils in the first stage of education in the school year 2010/2011 (135,682 students). In particular, we have data on student level characteristics and school level characteristics. The student level characteristics cover the gender of the pupil (*gender*), the

grade retention in primary school (*retention*) and the inclusion of the pupil in the special needs student population in primary school (which serves as a proxy of student's low cognitive skills). The school level characteristics include both the teacher characteristics, such as the teachers' age, seniority and education, in addition to principal characteristics, such as the principals' age and seniority. Teacher and principal seniority measures the level of experience of the teachers and principals, respectively. These variables assume values in the range of 1 to 7, where the teachers (and principals) with a seniority level of 1 are the least experienced (0-5 years of experience) and teachers (and principals) with a seniority level of 7 are the most experienced (more than 30 years of experience).<sup>12</sup> Similarly, the ages of teacher and principal are reported as categorical variables that range from 1 to 8, where teachers/principals in the first category are the youngest (less than 30 years old) and teachers/principals in the last category are the oldest (more than 60 years old).<sup>13</sup> Teachers' education records whether or not the teacher holds a pedagogical training (in the following we will refer to it as "teacher training"). All these variables are aggregated at school level in the form of averages (for age and seniority) and shares (for teachers' education) and assigned to each student with respect to the school where he/she is enrolled.

The outcome variable is a dummy variables defined as follows: the variable *A-certificate* assumes value 1 if the student gets an "A-certificate" at the end of the school year (which is the most favorable outcome) and 0 if not. Since we do not have data on standardized test scores for Flemish students, *A-certificate* is a good, available proxy of student performance. Every year, each student performs a final test and gets a ranking from "A" to "C". Students that get an "A" can progress school without any restriction, while the students that get either "B" or "C" can progress school but only in specific programs or have some grade retention. Furthermore, we additionally analyse – in the Supplementary Material – the results that would be obtained using the variable *progress school* as outcome. This variable assumes value 1 if the student progresses to the following year without any grade retention and 0 if not (this variable is a complement of school retention). Both these outcome variables are proxies for different levels of students' performance: a positive *A-certificate* proxies for a higher level of performance than a positive *progress school*. In principle, the target of a policy-maker could be to have the highest possible share of students getting "A-certificates" and the lowest share of students not progressing through school.

## 4.2. Identification Strategy

To evaluate the impact of the policy on students' performance we apply the BCF-IV within a regression discontinuity design (Trochim; 1984; Hahn et al.; 2001). Regression Discontinuity Design (RDD) is a method that aims at evaluating the causal effects of interventions in settings where the assignment to the treatment is determined (at least partly) by the values of an observed covariate lying on either sides of a threshold point. The idea is that subjects just above and below this threshold are very similar and one can assume a quasi-randomization around the threshold (Mealli and Rampichini; 2012). RDDs are categorized in sharp RDDs and fuzzy RDDs.

In sharp RDDs, the central assumption is that, around the threshold, there is a sharp discontinuity (from 0 to 1) in the probability of being treated. This is due to the fact that in sharp RDDs there is no room for imperfect compliance. In many real world scenarios, however, thresholds are not strictly implemented, as in the case of our application. To deal with these situations, one can use fuzzy RDDs, which are applicable when around the threshold the probability of being actually treated changes discontinuously, but not sharply from 0 to 1 (i.e., the jump in the probability of being treated is less than 1).

In our application of the fuzzy RDD technique, we exploit two cutoffs around the 10% share of disadvantaged students in the first stage of secondary education. The students in schools just below the threshold are assigned to the control group ( $Z_i = 0$ ), while the students in schools just above the threshold are assigned to the treated group ( $Z_i = 1$ ). The bandwidth around the threshold (from which one obtains the two cutoffs) is determined using the data-driven methodology proposed by Calonico et al. (2014) and implemented in the `rdrobust` package in R (Calonico et al.; 2015). This methodology is used as a stand-alone bandwidth selector, and inference on the treatment effect is performed, on the inference subsample, using the estimators introduced in Section 2.2.2. Following the indication of Imbens and Lemieux (2008), we focus on the outcome equation to

<sup>12</sup>Teachers and principals' seniority classes are the following: class 1: between 0 and 5 years of experience; class 2: between 6 and 10; class 3: between 11 and 15; class 4: between 16 and 20; class 5: between 20 and 25; class 6: between 26 and 30; class 7: more than 30.

<sup>13</sup>Teachers and principals' age classes are the following: class 1: less than 30 years old; class 2: between 30 and 34; class 3: between 35 and 39; class 4: between 40 and 44; class 5: between 45 and 49; class 6: between 50 and 54; class 7: between 55 and 60; class 8: more than 60.

select the bandwidth and then we use the same bandwidth for the estimation of the treatment equation. Moreover, as pointed out by Lee and Lemieux (2010), the usage of the same bandwidth guarantees the validity of the 2SLS estimators used to estimate the causal effects.

The selected bandwidth around the threshold is 3.5% for the outcome variables *A-certificate*. To further validate this bandwidth, selected using the method of Calonico et al. (2014), we run additional analyses implementing the Bayesian methods proposed by Li et al. (2015) and by Mattei and Mealli (2016). In particular, we run a series of robustness checks for the selection of the bandwidth around the threshold implementing a hierarchical Bayesian model for assessing the balance of the covariates between the groups of observations assigned to the treatment and the ones assigned to the control. For the selected bandwidth, the probability of the pre-assignment variables being well-balanced is high for the subpopulations defined by values of the cutoff strictly lower than 3.5%. Indeed, these probabilities are larger than or close to 0.8, indicating that the covariates are balanced in the two groups.

Moreover, to guarantee an equal representation to all the schools, and to avoid biases related to the over-representation of biggest schools' students, we sample 50 pupils from each school. In turn, this leads to a higher balance among the averages between the observations assigned to the treatment and the observations assigned to the control, as shown in panel (a) of Figure C.1 in the Supplementary Material.

In the Supplementary Material, we run a series of tests to show that the RDD (Regression Discontinuity Design) is valid for this application. Moreover, as a robustness check we sample a higher number of students according to the size of the smallest school (62 pupils) from every school. We show the balance in the samples of units assigned to the treatment and to the control in the second scenario.

### 4.3. Results

This Section assesses the effects of the additional funding on students' performances and highlights the main drivers of the heterogeneity in causal effects. These analyses are made for the outcome variable *A-certificate*. In the Supplementary Material, we report the results for a *progress school*. It is important to highlight that the results for both the outcomes, considered separately, in terms of effects and heterogeneity drivers, remain roughly the same when we (i) widen the sample of units included in the analysis, and (ii) perform the heterogeneity discovery on the ITT (results are reported in the Supplementary Material).

The variable *A-certificate* serves as a proxy for positive performance. In our sample, the students that got an "A-certificate" are the 91.73% of the total population. In Figure 3, the heterogeneous Complier Average Causal Effects (CACE) estimated using the proposed model are depicted. The darker the shade of blue in the node the higher the causal effect.

Although positive, the overall effect of the additional funding is not significant. This finding is in line with the recent literature on school spending and students' performance in a cross-country scenario (Hanushek et al.; 2016; Hanushek and Woessmann; 2017) and in the Flanders, in particular (D'Inverno et al.; 2021). However, it is compelling to observe the main drivers of the heterogeneity in the causal effect. It is worth highlighting that we did not implement any leaves trimming using p-values corrections as the aim of this application was to discover an interpretable representation of potential heterogeneity in the effects, even in the absence of significant results.

The first driver in the heterogeneity of the effects is the variable *principal age*: for students in schools with younger principals, the effects of funding are larger. These results, even if not significant, show that the treatment effects are higher for students that are in schools with principals younger than 55 years old. The second driver of heterogeneity is the seniority of the teachers: students in schools with younger principals and with less senior teachers – namely, teachers with less than 11 years of experience – have an increase in their performance. The conditional effect is 0.051, meaning that being treated leads to an increase of 5.1% in the probability of getting the highest possible grade. On the opposite side, students in schools with older principals and with more senior teachers – namely, teachers with more than 11 years of experience – have a decrease in their performance. The conditional effect is -0.039, meaning that being treated leads to a decrease of 3.9% in the probability of getting the highest possible grade.

Both these heterogeneity drivers, namely, the seniority and age of the teachers and principals, are particularly appreciable, as there are evidences in the education literature that connect teachers' seniority (Rice; 2010; Harris and Sass; 2011), teachers' age (Holmlund and Sund; 2008) to their teaching performance, and in turn teaching performance to students' positive achievements (Goldhaber and Hansen; 2010). Moreover, there is a compelling evidence in the literature regarding the role of principals in driving higher students' achievements (Eberts and

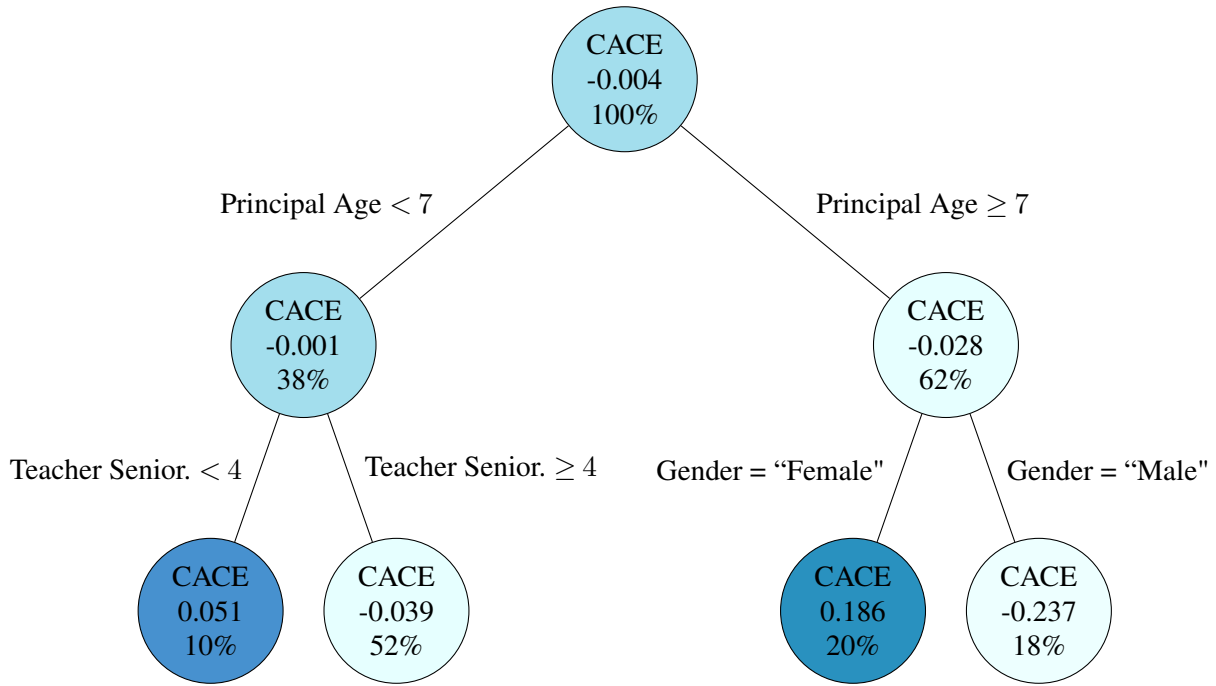


Fig 3: Visualization of the heterogeneous Complier Average Causal Effects (CACE) of additional funding on *A-certificate* estimated using the proposed BCF-IV model. The overall learning sample size is 4,300. The tree is a summarizing classification tree fit to posterior point estimates of individual treatment effects.

The significance level is \* for a significance level of 0.1, \*\* for a significance level of 0.05 and \*\*\* for a significance level of 0.01.

Stone; 1988; Gentilucci and Muto; 2007), however this is, to the extent of our knowledge, the first research that highlights the role of principal's age and seniority as drivers of treatment effect variations. Interestingly, another source of treatment variation is student's gender. Female students in schools with older principals, depict a positive effect while male students a negative effect. Yet, the effects for both subgroups are not significant.

This evidence can be interpreted in the following way: the additional funding has a negative, but not statistically significant, effect in the performance of students in the overall population, but it increases its effect in a notable way for those students in schools with less senior teachers and younger principals. These results are in line with the evidence that additional school funding does not boost the performance of the overall population of students (Hanushek et al.; 2016; Hanushek and Woessmann; 2017; D'Inverno et al.; 2021) and with the literature that connects students' achievements with principals (Eberts and Stone; 1988; Gentilucci and Muto; 2007) and teachers performance (Holmlund and Sund; 2008; Rice; 2010; Harris and Sass; 2011; Goldhaber and Hansen; 2010). It is important to highlight that even if we find some evidence of treatment effects variation connected to teachers' seniority, principal age, and students gender the conditional causal effects are not significantly different from zero. These results are robust when the outcome variable considered is *progress school* (see Supplementary Material).

## 5. Conclusion and Discussion

This paper developed a novel Bayesian machine learning technique, BCF-IV, to draw causal inference in scenarios with imperfect compliance. By investigating the heterogeneity in the causal effects, the technique expedites targeted policies. We manifested that the BCF-IV technique outperforms other machine learning techniques tailored for causal inference in precisely estimating the causal effects and converges to an optimal large sample performance in identifying the subgroups with heterogeneous effects. Moreover, using Monte-Carlo simulations, we showed that the competitive advantages of using BCF-IV, as compared to GRF or HCT-IV, are substantial.

In our application, we evaluated the effects of additional funding on students' performances. While the overall effects are negative but not significant, there are significant differences among different sub-populations of students. Indeed, for students in schools with less senior and younger principals (principals younger than 55 years old and with less than 30 years of experience) the effects of the policy are greater. We want to highlight that



age and seniority of principals as treatment variation drivers are robust to different definitions of the outcome variable (see results in Section 4), to variations in the algorithm used (BCF-IV vs BCF-ITT), in the size and definition of the learning sample and, in turn, in the balance between the group of treated and control units (see the Supplementary Material).

On one hand, as an underlying mechanism, the need for additional funds can be higher in schools with younger and less senior principals, who are more often observed in the most disadvantaged schools. This phenomenon arises as senior principals select themselves out of the most disadvantaged schools and more into advantaged schools, thereby creating relatively more vacancies in disadvantaged schools. Therefore, on average, younger and less senior principals lack a real choice but to start working in the most disadvantaged schools. Moreover, we can think of the motivation for principals to decrease as they grow older and this, in turn, have an impact on their performance, and their ability to effectively allocate the additional funds. To the best of our knowledge, this is the first study that investigates the effects of age and seniority of principals on enhancing the effectiveness of school funding on students' performance. The investigation of the true causal channel is beyond the goals of this paper and is left to further investigation where more granular teachers' and principals' characteristics are available.

These results are relevant to the policy as they furnish the instruments to policy-makers to enhance the effects of additional funding on students' performance. Indeed, on one side policy-makers could target just students in school with positive, effects reducing the overall costs of the policy and using the savings to experiment more effective policies in the other schools. On the other side, policy-makers could analyze the reason of lack of the effectiveness of funding in schools with certain characteristics and implement policies to boost the effects of future funding. Furthermore, the added value of our algorithm is that it could enable policy-makers to target just those units that benefit the most from the treatment and it provides an insight on possible inefficiencies in the allocation and/or usage of funding. From our analysis it seems that there is room for policies that support less senior principals since students in their schools show higher returns in terms of performance from additional funding.

The extension of these methods to other fields of economic investigation and the development of novel machine learning algorithms for targeted policies and welfare maximization can form the future scope of further research. In particular, the development of an algorithm that could deal with welfare maximization in the context of multiple outcomes is of interest. As a further extension, it may be worth exploring hierarchical testing procedures to control for familywise error rate (see, e.g., Marcus et al.; 1976) or for false discovery rate (see, e.g., Yekutieli; 2008) at each node of the tree discovered by BCF-IV and not just in the terminal nodes. Such algorithms have been explored in the context of matching procedures by Johnson et al. (2019) and Lee et al. (2021). Moreover, the "usual" Bayesian way for estimating CACE is via a data augmentation scheme, (e.g., imputing compliance status and estimating impacts among estimated compliers, as in Imbens and Rubin; 2015). In our algorithm we do not implement such a methodology, however, as a further line of research, it could be extended including a data augmentation scheme.

## REFERENCES

- Andrews, I., Stock, J. H. and Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice, *Annual Review of Economics* **11**: 727–753.
- Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996). Identification of causal effects using instrumental variables, *Journal of the American Statistical Association* **91**(434): 444–455.
- Angrist, J. D. and Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments, *Journal of Economic perspectives* **15**(4): 69–85.
- Angrist, J. D. and Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*, Princeton University Press.
- Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects, *Proceedings of the National Academy of Sciences* **113**(27): 7353–7360.
- Athey, S., Tibshirani, J., Wager, S. et al. (2019). Generalized random forests, *The Annals of Statistics* **47**(2): 1148–1178.
- Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance, *Journal of the American Statistical Association* **92**(439): 1171–1176.
- Bargagli Stoffi, F. J. and Gnecco, G. (2018). Estimating heterogeneous causal effects in the presence of irregular assignment mechanisms, *In Proceedings of the 5th IEEE Conference in Data Science and Advanced Analytics* pp. 1–10.
- Bargagli-Stoffi, F. J. and Gnecco, G. (2020). Causal tree with instrumental variable: an extension of the causal tree framework to irregular assignment mechanisms, *International Journal of Data Science and Analytics* **9**(3): 315–337.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal statistical society: series B (Methodological)* **57**(1): 289–300.

- Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency, *Annals of statistics* pp. 1165–1188.
- Breiman, L. (1984). *Classification and regression trees*, Routledge, New York, New York.
- Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and regression trees*, Belmont, CA: Wadsworth & Brooks .
- Calonico, S., Cattaneo, M. D. and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs, *Econometrica* **82**(6): 2295–2326.
- Calonico, S., Cattaneo, M. D. and Titiunik, R. (2015). rdrobust: An r package for robust nonparametric inference in regression-discontinuity designs, *R Journal* **7**(1): 38–51.
- Chipman, H. A., George, E. I., McCulloch, R. E. et al. (2010). Bart: Bayesian additive regression trees, *The Annals of Applied Statistics* **4**(1): 266–298.
- Coleman, J. S. (1966). Equality of educational opportunity, *Washington DC: US Government Printing Office* pp. 1–32.
- Cook, D. I., Gebski, V. J. and Keech, A. C. (2004). Subgroup analysis in clinical trials, *Medical Journal of Australia* **180**(6): 289–291.
- Cook, T. D. (2008). “waiting for life to arrive”: a history of the regression-discontinuity design in psychology, statistics and economics, *Journal of Econometrics* **142**(2): 636–654.
- Ding, P., Feller, A. and Miratrix, L. (2019). Decomposing treatment effect variation, *Journal of the American Statistical Association* **114**(525): 304–317.
- Dominici, F., Bargagli-Stoffi, F. J. and Mealli, F. (2021). From controlled to undisciplined data: Estimating causal effects in the era of data science using a potential outcome framework, *Harvard Data Science Review* .
- Dorie, V., Hill, J. and Dorie, M. V. (2020). Package ‘bartcause’.
- Dorie, V., Hill, J., Shalit, U., Scott, M. and Cervone, D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition, *Statistical Science* **34**(1): 43–68.
- D’Inverno, G., Smet, M. and De Witte, K. (2021). Impact evaluation in a multi-input multi-output setting: Evidence on the effect of additional resources for schools, *European Journal of Operational Research* **290**(3): 1111–1124.
- Eberts, R. W. and Stone, J. A. (1988). Student achievement in public schools: Do principals make a difference?, *Economics of Education Review* **7**(3): 291–299.
- Gentilucci, J. L. and Muto, C. C. (2007). Principals’ influence on academic achievement: The student perspective, *NASSP bulletin* **91**(3): 219–236.
- Goldhaber, D. and Hansen, M. (2010). Using performance on the job to inform teacher tenure decisions, *American Economic Review* **100**(2): 250–55.
- Hahn, J., Todd, P. and Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design, *Econometrica* **69**(1): 201–209.
- Hahn, P. R., Carvalho, C. M., Puelz, D., He, J. et al. (2018). Regularization and confounding in linear regression for treatment effect estimation, *Bayesian Analysis* **13**(1): 163–182.
- Hahn, P. R., Dorie, V. and Murray, J. S. (2019). Atlantic causal inference conference (acic) data analysis challenge 2017, *arXiv preprint arXiv:1905.09515* .
- Hahn, P. R., Murray, J. S. and Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion), *Bayesian Anal.* **15**(3): 965–1056.
- Hanushek, E. A. (2003). The failure of input-based schooling policies, *The Economic Journal* **113**(485): F64–F98.
- Hanushek, E. A., Machin, S. J. and Woessmann, L. (2016). *Handbook of the Economics of Education*, Elsevier.
- Hanushek, E. A. and Woessmann, L. (2017). School resources and student achievement: A review of cross-country economic research, *Cognitive abilities and educational outcomes*, Springer, pp. 149–171.
- Harris, D. N. and Sass, T. R. (2011). Teacher training, teacher quality and student achievement, *Journal of Public Economics* **95**(7–8): 798–812.
- Hartford, J., Lewis, G., Leyton-Brown, K. and Taddy, M. (2017). Deep iv: A flexible approach for counterfactual prediction, *International Conference on Machine Learning*, PMLR, pp. 1414–1423.
- Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference, *Journal of Computational and Graphical Statistics* **20**(1): 217–240.
- Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance, *Biometrika* **75**(4): 800–802.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure, *Scandinavian journal of statistics* pp. 65–70.
- Holmlund, H. and Sund, K. (2008). Is the gender gap in school performance affected by the sex of the teacher?, *Labour Economics* **15**(1): 37–53.
- Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified bonferroni test, *Biometrika* **75**(2): 383–386.
- Hsu, J. Y., Zubizarreta, J. R., Small, D. S. and Rosenbaum, P. R. (2015). Strong control of the familywise error rate in observational studies that discover effect modification by exploratory methods, *Biometrika* **102**(4): 767–782.
- Imbens, G. W. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice, *Journal of Econometrics* **142**(2): 615–635.
- Imbens, G. W. and Rubin, D. B. (1997). Estimating outcome distributions for compliers in instrumental variables models, *The Review of Economic Studies* **64**(4): 555–574.
- Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press.
- Jackson, C. K. (2018). Does school spending matter? the new literature on an old question, *Technical report*, National Bureau of Economic Research.
- Jackson, C. K., Johnson, R. C. and Persico, C. (2015). The effects of school spending on educational and economic outcomes: Evidence from school finance reforms, *The Quarterly Journal of Economics* **131**(1): 157–218.

- Johnson, M., Cao, J. and Kang, H. (2019). Detecting heterogeneous treatment effect with instrumental variables, *arXiv preprint arXiv:1908.03652*.
- Kapelner, A. and Bleich, J. (2013). Bartmachine: Machine learning with bayesian additive regression trees, *arXiv preprint, arXiv:1312.2171*.
- Kim, B., Khanna, R. and Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability, *Advances in Neural Information Processing Systems*, pp. 2280–2288.
- Lee, D. S. and Lemieux, T. (2010). Regression discontinuity designs in economics, *Journal of Economic Literature* **48**(2): 281–355.
- Lee, K., Bargagli-Stoffi, F. J. and Dominici, F. (2020). Causal rule ensemble: Interpretable inference of heterogeneous treatment effects, *arXiv preprint arXiv:2009.09036*.
- Lee, K., Small, D. S. and Dominici, F. (2021). Discovering heterogeneous exposure effects using randomization inference in air pollution studies, *Journal of the American Statistical Association* pp. 1–12.
- Li, F., Mattei, A. and Mealli, F. (2015). Evaluating the causal effect of university grants on student dropout: evidence from a regression discontinuity design using principal stratification, *The Annals of Applied Statistics* pp. 1906–1931.
- Logan, B. R., Sparapani, R., McCulloch, R. E. and Laud, P. W. (2019). Decision making and uncertainty quantification for individualized treatments using bayesian additive regression trees, *Statistical Methods in Medical Research* **28**(4): 1079–1093.
- Marcus, R., Eric, P. and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance, *Biometrika* **63**(3): 655–660.
- Mattei, A. and Mealli, F. (2016). Regression discontinuity designs as local randomized experiments, *Observational Studies* **66**: 156–173.
- McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test, *Journal of Econometrics* **142**(2): 698–714.
- Mealli, F. and Rampichini, C. (2012). Evaluating the effects of university grants by using regression discontinuity designs, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **175**(3): 775–798.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences, *Artificial Intelligence* **267**: 1–38.
- Mullainathan, S. and Spiess, J. (2017). Machine learning: an applied econometric approach, *Journal of Economic Perspectives* **31**(2): 87–106.
- Nelson, C. R. and Startz, R. (1990). The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one, *Journal of Business* pp. S125–S140.
- Nelson, C. R. and Stratz, R. (1990). Some further results on the exact small sample properties of the instrumental variable estimator, *Econometrica* **58**(4): 967–976.
- Novak, R., Bahri, Y., Abolafia, D. A., Pennington, J. and Sohl-Dickstein, J. (2018). Sensitivity and generalization in neural networks: an empirical study, *arXiv preprint arXiv:1802.08760*.
- OECD (2017). Educational opportunity for all: Overcoming inequality throughout the life course.
- Prosperi, M., Guo, Y., Sperrin, M., Koopman, J. S., Min, J. S., He, X., Rich, S., Wang, M., Buchan, I. E. and Bian, J. (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare, *Nature Machine Intelligence* **2**(7): 369–375.
- Psacharopoulos, G. and Patrinos, H. A. (2018). Returns to investment in education: a decennial review of the global literature, *Education Economics* **26**(5): 445–458.
- Rice, J. K. (2010). The impact of teacher experience: Examining the evidence and policy implications, *National Center for Analysis of Longitudinal Data in Education Research*.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies., *Journal of Educational Psychology* **66**(5): 688.
- Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization, *The Annals of Statistics* pp. 34–58.
- Rubin, D. B. (1986). Comment: Which ifs have causal answers, *Journal of the American Statistical Association* **81**(396): 961–962.
- Sarkar, S. K. and Chang, C.-K. (1997). The simes method for multiple hypothesis testing with positively dependent test statistics, *Journal of the American Statistical Association* **92**(440): 1601–1608.
- Therneau, T., Atkinson, B., Ripley, B. and Ripley, M. B. (2015). Package ‘rpart’ package version 4.1-15, Available online: <https://cran.r-project.org/web/packages/rpart/rpart.pdf>.
- Thistlethwaite, D. L. and Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment., *Journal of Educational Psychology* **51**(6): 309.
- Trochim, W. M. (1984). *Research design for program evaluation: The regression-discontinuity approach*, Vol. 6, SAGE Publications, Inc.
- Wang, G., Li, J. and Hopp, W. J. (2018). An instrumental variable tree approach for detecting heterogeneous treatment effects in observational studies, Available at SSRN 3045327.
- Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N. and Gallego, B. (2018). Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases, *Statistics in Medicine* **37**(23): 3309–3324.
- Woody, S., Carvalho, C. M., Hahn, P. R. and Murray, J. S. (2020). Estimating heterogeneous effects of continuous exposures using bayesian tree ensembles: revisiting the impact of abortion rates on crime, *arXiv preprint arXiv:2007.09845*.
- Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*, MIT Press, Cambridge, Massachusetts.
- Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*, Nelson Education.
- Yekutieli, D. (2008). Hierarchical false discovery rate–controlling methodology, *Journal of the American Statistical Association* **103**(481): 309–316.
- Zhang, H. and Singer, B. H. (2010). *Recursive partitioning and applications*, Springer Science & Business Media.
- Zhang, Y. and Wallace, B. C. (2017). A sensitivity analysis of (and practitioners’ guide to) convolutional neural networks for sentence classification, *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 253–263.