arXiv:1712.04802v4 [stat.ML] 3 Sep 2019

GENERIC MACHINE LEARNING INFERENCE ON HETEROGENOUS TREATMENT EFFECTS IN RANDOMIZED EXPERIMENTS

VICTOR CHERNOZHUKOV, MERT DEMIRER, ESTHER DUFLO, AND IVÁN FERNÁNDEZ-VAL

Abstract. We propose strategies to estimate and make inference on key features of heterogeneous effects in randomized experiments. These key features include *best linear predictors of the effects* using machine learning proxies, *average effects sorted by impact groups*, and *average characteristics of most and least impacted units*. The approach is valid in high dimensional settings, where the effects are proxied by machine learning methods. We post-process these proxies into the estimates of the key features. Our approach is generic, it can be used in conjunction with penalized methods, deep and shallow neural networks, canonical and new random forests, boosted trees, and ensemble methods. It does not rely on strong assumptions. In particular, we don't require conditions for consistency of the machine learning methods. Estimation and inference relies on repeated data splitting to avoid overfitting and achieve validity. For inference, we take medians of p-values and medians of confidence intervals, resulting from many different data splits, and then adjust their nominal level to guarantee uniform validity. This variational inference method is shown to be uniformly valid and quantifies the uncertainty coming from both parameter estimation and data splitting. We illustrate the use of the approach with two randomized experiments in development on the effects of microcredit and nudges to stimulate immunization demand.

Key words: Agnostic Inference, Machine Learning, Confidence Intervals, Causal Effects, Variational P-values and Confidence Intervals, Uniformly Valid Inference, Quantification of Uncertainty, Sample Splitting, Multiple Splitting, Assumption-Freeness, Microcredit, Immunization Incentives

JEL: C18, C21, D14, G21, O16

1. Introduction

Randomized experiments play an important role in the evaluation of social and economic programs and medical treatments (e.g., Imbens and Rubin (2015); Duflo et al. (2007)). Researchers and policy makers are often interested in features of the impact of the treatment that go beyond the simple average treatment effects. In particular, very often, they want to know whether treatment effect depends on covariates, such as gender, age, etc. It is essential to assess if the impact of the program would generalize to a different population with different characteristics, and for economists, to better understand the driving mechanism behind the effects of a particular program. In a review of 189 RCT published in top economic journals since 2006, we found that 76

Date: September 4, 2019.

We thank Susan Athey, Moshe Buchinsky, Denis Chetverikov, Siyi Luo, Max Kasy, Susan Murphy, Whitney Newey, and seminar participants at ASSA 2018, Barcelona GSE Summer Forum 2019, NYU, UCLA and Whitney Newey's Contributions to Econometrics conference for valuable comments. We gratefully acknowledge research support from the National Science Foundation.

(40%) report at least one subgroup analysis, wherein they report treatment effects in subgroups formed by baseline covariates.[1]

One issue with reporting treatment effects split by subgroups, however, is that there are often a large number of potential sample splits: choosing subgroups ex-post opens the possibility of overfitting. To solve this problem, medical journals and the FDA require pre-registering the sub-sample of interest in medical trials *in advance*. In economics, this approach has gained some traction, with the adoption of pre-analysis plans (which can be filed in the AEA registry for randomized experiments). Restricting heterogeneity analysis to pre-registered subgroups, however, amounts to throwing away a large amount of potentially valuable information, especially now that many researchers collect large baseline data sets. It should be possible to use the data to discover *ex post* whether there is any relevant heterogeneity in treatment effect by covariates.

To do this in a disciplined fashion and avoid the risk of overfitting, scholars have recently proposed using machine learning (ML) tools (see e.g. Athey and Imbens (2017) and below for a review). Indeed, ML tools seem to be ideal to explore heterogeneity of treatment effects, when researchers have access to a potentially large array of baseline variables to form subgroups, and little guiding principles on which of those are likely to be relevant. Several recent papers, which we review below, develop methods for detecting heterogeneity in treatment effects. Empirical researchers have taken notice.[2]

This paper develops a generic approach to use any of the ML tools to predict and make inference on heterogeneous treatment or policy effects. A core difficulty of applying ML tools to the estimation of heterogenous causal effects is that, while they are successful in prediction empirically, it is much more difficult to obtain uniformly valid inference. In fact, in high dimensional settings, absent strong assumptions, generic ML tools may not even produce consistent estimates of the *conditional average treatment effect* (CATE), the difference in the expected potential outcomes between treated and control groups conditional on covariates.

Previous attempts to solve this problem focus either on specific tools (for example the method proposed by Athey and Imbens (2016), which has become popular with applied researchers, and uses trees), or on situations where those assumptions might be satisfied. Our approach to resolve the fundamental impossibilities in non-parametric inference is different. Motivated by Genovese

[1]The papers were published in *Quarterly Journal of of Economics*, *American Economic Review*, *Review of Economics Studies*, *Econometrica* and *Journal of Political Economy*. We than Karthik Mularidharan, Mauricio Romero and Kaspar Wüthrich for sharing the list of papers they computed for another project.

[2]In the last few months alone, several new empirical papers in economics used ML methods to estimate heterogenous effects. E.g. Rigol et al. (2016) shows that villagers outperform the machine learning tools when they predict heterogeneity in returns to capital. Davis and Heller (2017) predicts who benefits the most from a summer internship projects. Deryugina et al. (Forthcoming) uses the methods developed in the present paper to evaluate the heterogeneity in the effect of air pollution on mortality. Crepon et al. (2019) also builds on the present paper to develop a methodology to determine if the impact of two different programs can be accounted for by different selection. The methodological papers reviewed later also contain a number of empirical applications.

and Wasserman (2008), instead of attempting to get consistent estimation and uniformly valid inference on the CATE itself, we focus on providing valid estimation and inference on *features* of CATE. We start by building a ML proxy predictor of CATE, and then develop valid inference on features of the CATE based on this proxy predictor. In particular, we develop valid inference on three objects, which are likely to be of interest to applied researchers and policy makers: First, the **Best Linear Predictor** (BLP) of the CATE based on the ML proxy predictor; second, the **Sorted Group Average Treatment Effects** (GATES) or average treatment effect by heterogeneity groups induced by the ML proxy predictor; and third, the **Classification Analysis** (CLAN) or the average characteristics of the most and least affected units defined in terms of the ML proxy predictor. Thus, we can find out if there is detectable heterogeneity in the treatment effect based on observables, and if there is any, what is the treatment effect for different bins. And finally we can describe which of the covariates is correlated with this heterogeneity.

There is a trade-off between more restrictive assumptions or tools and a more ambitious estimation. We chose a different approach to address this trade-off than previous papers: focus on coarser objects of the function rather than the function itself, but make as little assumptions as possible. This seems to be a worthwhile sacrifice: the objects for which we have developed inference appear to us at this point to be the most relevant, but in the future, one could easily use the same approach to develop methods to estimate other objects of interest.

The Model and Key Causal Functions. Let $Y(1)$ and $Y(0)$ be the potential outcomes in the treatment state 1 and the non-treatment state 0; see Neyman (1923) and Rubin (1974). Let Z be a vector of covariates that characterize the observational units. The main causal functions are the baseline conditional average (BCA):

$$b_0(Z) := \mathrm{E}[Y(0) \mid Z], \tag{1.1}$$

and the conditional average treatment effect (CATE):

$$s_0(Z) := \mathrm{E}[Y(1) \mid Z] - \mathrm{E}[Y(0) \mid Z]. \tag{1.2}$$

Suppose the binary treatment variable D is randomly assigned conditional on Z, with probability of assignment depending only on a subvector of stratifying variables $Z_1 \subseteq Z$, namely

$$D \perp\!\!\!\perp (Y(1), Y(0)) \mid Z, \tag{1.3}$$

and the propensity score is known and is given by

$$p(Z) := \mathrm{P}[D = 1 \mid Z] = \mathrm{P}[D = 1 \mid Z_1], \tag{1.4}$$

which we assume is bounded away from zero or one:

$$p(Z) \in [p_0, p_1] \subset (0, 1). \tag{1.5}$$

The observed outcome is $Y = DY(1) + (1 - D)Y(0)$. Under the stated assumption, the causal functions are identified by the components of the regression function of Y given D, Z:

$$Y = b_0(Z) + Ds_0(Z) + U, \;\; \mathrm{E}[U \mid Z, D] = 0,$$

that is,

$$b_0(Z) = \mathrm{E}[Y \mid D = 0, Z], \tag{1.6}$$

and

$$s_0(Z) = \mathrm{E}[Y \mid D = 1, Z] - \mathrm{E}[Y \mid D = 0, Z]. \tag{1.7}$$

We observe $\mathrm{Data} = (Y_i, Z_i, D_i)_{i=1}^N$, consisting of i.i.d. copies of the random vector (Y, Z, D) having probability law P. The expectation with respect to P is denoted by $\mathrm{E} = \mathrm{E}_P$. The probability law of the entire data is denoted by $\mathbb{P} = \mathbb{P}_P$ and the corresponding expectation is denoted by $\mathbb{E} = \mathbb{E}_P$.

Properties of Machine Learning Estimators of $s_0(Z)$ Motivating the Agnostic Approach. Machine learning (ML) is a name attached to a variety of new, constantly evolving statistical learning methods: Random Forest, Boosted Trees, Neural Networks, Penalized Regression, Ensembles, and Hybrids (see, e.g., Wasserman (2016) for a recent review, and Friedman et al. (2001) for a prominent textbook treatment). In modern high-dimensional settings, ML methods effectively explore the various forms of nonlinear structured sparsity to yield "good" approximations to $s_0(z)$ whenever such assumptions are valid, based on equations (1.6) and (1.7). As a result these methods often work much better than classical methods in high-dimensional settings, and have found widespread uses in industrial and academic applications.

Motivated by their practical predictive success, it is really tempting to apply ML methods directly to try to learn the CATE function $z \mapsto s_0(z)$ (by learning the two regression functions for treated and untreated and taking the difference). However, it is hard, if not impossible, to obtain uniformly valid inference on $z \mapsto s_0(z)$ using generic ML methods, under credible assumptions and practical tuning parameter choices. There are several fundamental reasons as well as huge gaps between theory and practice that are responsible for this.

One fundamental reason is that the ML methods might not even produce consistent estimators of $z \mapsto s_0(z)$ in high dimensional settings. For example, if z has dimension d and the target function $z \mapsto s_0(z)$ is assumed to have p continuous and bounded derivatives, then the worst case (minimax) lower bound on the rate of learning this function from a random sample of size N cannot be better than $N^{-p/(2p+d)}$ as $N \to \infty$, as shown by Stone Stone (1982). Hence if p is fixed and d is also small, but slowly increasing with N, such as $d \geqslant \log N$, then there exists no consistent estimator of $z \mapsto s_0(z)$ generally.

Hence, generic ML estimators cannot be regarded as consistent, unless further very strong assumptions are made. Examples of such assumptions include structured forms of linear and nonlinear sparsity and super-smoothness. While these (sometime believable and yet untestable) assumptions make consistent adaptive estimation possible (e.g.,Bickel et al. (2009)), inference remains a more difficult problem, as adaptive confidence sets do not exist even for low-dimensional nonparametric problems (Low et al. (1997); Genovese and Wasserman (2008)). Indeed, adaptive estimators (including modern ML methods) have biases of comparable or dominating order as compared to sampling error. Further assumptions such as "self-similarity" are needed to bound the biases and expand the confidence bands by the size of bias (see Giné and Nickl (2010); Chernozhukov et al. (2014)) to produce partly adaptive confidence bands. For more traditional statistical methods there are constructions in this vein that make use of either undersmoothing or bias-bounding arguments (Giné and Nickl (2010); Chernozhukov et al. (2014)). These methods, however, are not yet available for ML methods in high dimensions (see, however, Hansen et al. (2017) for a promising approach called "targeted undersmoothing" in sparse linear models).

Suppose we did decide to be optimistic (or panglossian) and imposed the strong assumptions, that made the theoretical versions of the ML methods provide us with high-quality consistent estimators of $z \mapsto s_0(z)$ and valid confidence bands based on them. This would still not give us a practical construction we would want for our applications. The reason is that there is often a gap between theoretical versions of the ML methods appearing in various theoretical papers and the practical versions (with the actual, data-driven tuning parameters) coded up in statistical computing packages used by practitioners.[3] The use of ML, for example, involves many tuning parameters with practical rules for choosing them, while theoretical work provides little guidance or backing for such practical rules; see e.g., the influential book Friedman et al. (2001) for many examples of such rules. Unfortunately, theoretical work often only provides existence results: there exist theoretical ranges of the tuning parameters that make the simple versions of the methods work for predictive purposes (under very strong assumptions), leaving no satisfactory guide to practice.

In this paper we take an agnostic view. We neither rely on any structured assumptions, which might be difficult to verify or believe in practice, nor impose conditions that make the ML estimators consistent. We simply treat ML as providing proxy predictors for the objects of interest.

Our Agnostic Approach. Here, we propose strategies for estimation and inference on

key features of $s_0(Z)$ rather than $s_0(Z)$ itself.

Because of this difference in focus we can avoid making strong assumptions about the properties of the ML estimators.

[3]There are cases where such gap does not exist, e.g., see Belloni et al. (2014, 2011) for the lasso. On the other hand, for example, even the wide use of K-fold cross-validation in high-dimensional settings for machine learning remains theoretically unjustified. There do exist, however, related subsample-based methods that achieve excellent performance for tuning selection (Wegkamp et al., 2003; Lecué and Mitchell, 2012).

Let (M, A) denote a random partition of the set of indices $\{1, \ldots, N\}$. The strategies that we consider rely on random splitting of

$$\text{Data} = (Y_i, D_i, Z_i)_{i=1}^N$$

into a main sample, denoted by $\text{Data}_M = (Y_i, D_i, Z_i)_{i \in M}$, and an auxiliary sample, denoted by $\text{Data}_A = (Y_i, D_i, Z_i)_{i \in A}$. We will sometimes refer to these samples as M and A. We assume that the main and auxiliary samples are approximately equal in size, though this is not required theoretically.

From the auxiliary sample A, we obtain ML estimators of the baseline and treatment effects, which we call the proxy predictors,

$$z \mapsto B(z) = B(z; \text{Data}_A) \text{ and } z \mapsto S(z) = S(z; \text{Data}_A).$$

These are possibly biased and noisy predictors of $b_0(z)$ and $s_0(z)$, and in principle, we do not even require that they are consistent for $b_0(z)$ and $s_0(z)$. We simply treat these estimates as proxies, which we post-process to estimate and make inference on the features of the CATE $z \mapsto s_0(z)$. We condition on the auxiliary sample Data_A, so we consider these maps as frozen, when working with the main sample.

Using the main sample and the proxies, we shall target and develop valid inference about *key features of* $s_0(Z)$ rather than $s_0(Z)$, which include

(1) **Best Linear Predictor** (BLP) of the CATE $s_0(Z)$ based on the ML proxy predictor $S(Z)$;

(2) **Sorted Group Average Treatment Effects** (GATES): average of $s_0(Z)$ (ATE) by heterogeneity groups induced by the ML proxy predictor $S(Z)$;

(3) **Classification Analysis** (CLAN): average characteristics of the most and least affected units defined in terms of the ML proxy predictor $S(Z)$.

Our approach is *generic* with respect to the ML method being used, and is *agnostic* about its formal properties.

We will make use of many splits of the data into main and auxiliary samples to produce robust estimates. Our estimation and inference will systematically account for two sources of uncertainty:

(I) **Estimation uncertainty** conditional on the auxiliary sample.

(II) **Splitting uncertainty** induced by random partitioning of the data into the main and auxiliary samples.

Because we account for the second source, we call the resulting collection of methods as variational estimation and inference methods (VEINs). For point estimates we report the median of the estimated key features over different random splits of the data. For the confidence intervals we take the medians of many random conditional confidence sets and we adjust their nominal confidence

level to reflect the splitting uncertainty. We construct p-values by taking medians of many random conditional p-values and adjust the nominal levels to reflect the splitting uncertainty. Note that considering many different splits and accounting for variability caused by splitting is very important. Indeed, with a single splitting practice, empiricists may unintentionally look for a "good" data split, which supports their prior beliefs about the likely results, thereby invalidating inference.[4]

Relationship to the Literature. We focus the review strictly on the literatures about estimation and inference on heterogeneous effects and inference using sample splitting.

We first mention work that uses linear and semiparametric regression methods. A semiparametric inference method for characterizing heterogeneity, called the sorted effects method, was given in Chernozhukov et al. (2015). This approach does provide a full set of inference tools, including simultaneous bands for percentiles of the CATE, but is strictly limited to the traditional semiparametric estimators for the regression and causal functions. Hansen et al. (2017) proposed a sparsity based method called "targeted undersmoothing" to perform inference on heterogeneous effects. This approach does allow for high-dimensional settings, but makes strong assumptions on sparsity as well as additional assumptions that enable the targeted undersmoothing. A related approach, which allows for simultaneous inference on many coefficients (for example, inference on the coefficients corresponding to the interaction of the treatment with other variables) was first given in Belloni et al. (2013) using a Z-estimation framework, where the number of interactions can be very large; see also Dezeure et al. (2016) for a more recent effort in this direction, focusing on de-biased lasso in mean regression problems. This approach, however, still relies on a strong form of sparsity assumptions. Zhao et al. (2017) proposed a post-selection inference framework within the high-dimensional linear sparse models for the heterogeneous effects. The approach is attractive because it allows for some misspecification of the model.

Next we discuss the use of tree-based and other methods. Imai and Ratkovic (2013) discussed the use of a heuristic support-vector-machine method with lasso penalization for classification of heterogeneous treatments into positive and negative ones. They used the Horvitz-Thompson transformation of the outcome (e.g., as in Hirano et al. (2003); Abadie (2005)) such that the new outcome becomes an unbiased, noisy version of CATE. Athey and Imbens (2016) made use of the Horvitz-Thompson transformation of the outcome variable to inform the process of building causal trees, with the main goal of predicting CATE. They also provide a valid inference result on average treatment effects for groups defined by the tree leaves, conditional on the data split in two subsamples: one used to build the tree leaves and the one to estimate the predicted values given the leaves. Like our methods, this approach is essentially assumption-free. The difference with our generic approach is that it is limited to trees and does not account for splitting uncertainty, which is important in practical settings. Wager and Athey (2017) provided a subsampling-based construction of a

[4]This problem is "solved" by fixing the Monte-Carlo seed and the entire data analysis algorithm before the empirical study. Even if such a huge commitment is really made and followed, there is a considerable risk that the resulting data-split may be non-typical. Our approach allows one to avoid taking this risk.

causal random forest, providing valid pointwise inference for CATE (see also the review in Wager and Athey (2017) on prior uses of random forests in causal settings) for the case when covariates are very low-dimensional (and essentially uniformly distributed).[5] Unfortunately, this condition rules out the typical high-dimensional settings that arise in many empirical problems, including the ones considered in this paper.

Our approach is different from these existing approaches, in that we are changing the target, and instead of hunting for CATE $z \mapsto s_0(z)$, we focus on key features of $z \mapsto s_0(z)$. We simply treat the ML methods as providing a proxy predictor $z \mapsto S(z)$, which we post-process to estimate and make inference on the key features of the CATE $z \mapsto s_0(z)$. Some of our strategies rely on Horvitz-Thompson transformations of outcome and some do not. The inspiration for our approach draws upon an observation in Genovese and Wasserman (2008), namely that some fundamental impossibilities in non-parametric inference could be avoided if we focus inference on coarser features of the non-parametric functions rather than the functions themselves.

Our inference approach is also of independent interest, and could be applied to many problems, where sample splitting is used to produce ML predictions, e.g. Abadie et al. (2017). Related references include Wasserman and Roeder (2009); Meinshausen et al. (2009), where the ideas are related but quite different in details, which we shall explain below. The premise is the same, however, as in Meinshausen et al. (2009); Rinaldo et al. (2016) – we should not rely on a single random split of the data and should adjust inference in some way. Our approach takes the medians of many conditional confidence intervals as the confidence interval and the median of many conditional p-values as the p-value, and adjusts their nominal levels to account for the splitting uncertainty. Our construction of p-values builds upon ideas in Benjamini and Hochberg (1995); Meinshausen et al. (2009), though what we propose is radically simpler, and our confidence intervals appear to be brand new. Of course sample splitting ideas are classical, going back to Hartigan (1969); Kish and Frankel (1974); Barnard (1974); Cox (1975); Mosteller and Tukey (1977), though having been mostly underdeveloped and overlooked for inference, as characterized by Rinaldo et al. (2016).

2. Main Identification Results and Estimation Strategies

2.1. **BLP of CATE.** We consider two strategies for identifying and estimating the best linear predictor of $s_0(Z)$ using $S(Z)$:

$$\mathsf{BLP}[s_0(Z) \mid S(Z)] := \arg \min_{f(Z) \in \mathrm{Span}(1, S(Z))} \mathrm{E}[s_0(Z) - f(Z)]^2,$$

which, if exists, is defined by projecting $s_0(Z)$ on the linear span of 1 and $S(Z)$ in the space $L^2(P)$.

[5]The dimension d is fixed in Wager and Athey (2017); the analysis relies on the Stone's model with smoothness index $\beta = 1$, in which no consistent estimator exists once $d \geqslant \log n$. It'd be interesting to establish consistency properties and find valid inferential procedures for the random forest in high-dimensional ($d \propto n$ or $d \gg n$) *approximately sparse* cases, with continuous and categorical covariates, but we are not aware of any studies that cover such settings, which are of central importance to us.

BLP of CATE: The First Strategy. Here we shall identify the coefficients of the BLP from the weighted linear projection:

$$Y = \alpha' X_1 + \beta_1(D - p(Z)) + \beta_2(D - p(Z))(S - \mathrm{E}S) + \epsilon, \;\; \mathrm{E}[w(Z)\epsilon X] = 0, \tag{2.1}$$

where $S := S(Z)$,

$$w(Z) = \{p(Z)(1 - p(Z))\}^{-1}, \quad X := (X_1, X_2)$$
$$X_1 := X_1(Z), \quad \text{e.g.,} \quad X_1 = [1, B(Z)],$$
$$X_2 := [D - p(Z), (D - p(Z))(S - \mathrm{E}S)].$$

Note that the above equation uniquely pins down β_1 and β_2 under weak assumptions.

The interaction $(D - p(Z))(S - \mathrm{E}S)$ is orthogonal to $D - p(Z)$ under the weight $w(Z)$ and to all other regressors that are functions of Z under any Z-dependent weight.[6]

A consequence is our first main identification result, namely that

$$\beta_1 + \beta_2(S(Z) - \mathrm{E}S) = \mathsf{BLP}[s_0(Z) \mid S(Z)],$$

in particular $\beta_1 = \mathrm{E}s_0(Z)$ and $\beta_2 = \mathrm{Cov}(s_0(Z), S(Z))/\mathrm{Var}(S(Z))$.

Theorem 2.1 (BLP 1)**.** *Consider $z \mapsto S(z)$ and $z \mapsto B(z)$ as fixed maps. Assume that Y and X have finite second moments, $\mathrm{E}XX'$ is full rank, and $\mathrm{Var}(S(Z)) > 0$. Then, (β_1, β_2) defined in* (2.1) *also solves the best linear predictor/approximation problem for the target $s_0(Z)$:*

$$(\beta_1, \beta_2)' = \arg\min_{b_1, b_2} \mathrm{E}[s_0(Z) - b_1 - b_2 S(Z)]^2,$$

in particular $\beta_1 = \mathrm{E}S_0(Z)$ and $\beta_2 = \mathrm{Cov}(s_0(Z), S(Z))/\mathrm{Var}(S(Z))$.

The identification result is constructive. We can base the corresponding estimation strategy on the empirical analog:

$$Y_i = \widehat{\alpha}' X_{1i} + \widehat{\beta}_1(D_i - p(Z_i)) + \widehat{\beta}_2(D_i - p(Z_i))(S_i - \mathbb{E}_{N,M} S_i) + \widehat{\epsilon}_i, \quad i \in M,$$
$$\mathbb{E}_{N,M}[w(Z_i)\widehat{\epsilon}_i X_i] = 0,$$

where $\mathbb{E}_{N,M}$ denotes the empirical expectation with respect to the main sample, i.e.

$$\mathbb{E}_{N,M} g(Y_i, D_i, Z_i) := |M|^{-1} \sum_{i \in M} g(Y_i, D_i, Z_i).$$

[6]The orthogonalization ideas embedded in this strategy do have classical roots in econometrics (going back to at least Frisch and Waugh in the 30s), and similar strategies underlie the orthogonal or double machine learning approach (DML) in Chernozhukov et al. (2017). Our paper has different goals than DML, attacking the problem of inference on heterogeneous effects without rate and even consistency assumptions. The strategy here is more nuanced in that we are making it work under misspecification or inconsistent learning, which is likely to be true in very high-dimensional problems.

The properties of this estimator, conditional on the auxilliary data, are well known and follow as a special case of Lemma B.1 in the Appendix.

Comment 2.1 (Main Implications of the result)**.** If $S(Z)$ is a perfect proxy for $s_0(Z)$, then $\beta_2 = 1$. In general, $\beta_2 \neq 1$, correcting for noise in $S(Z)$. If $S(Z)$ is complete noise, uncorrelated to $s_0(Z)$, then $\beta_2 = 0$. Furthermore, if there is no heterogeneity, that is $s_0(Z) = s$, then $\beta_2 = 0$. Rejecting the hypothesis $\beta_2 = 0$ therefore means that there is both heterogeneity and $S(Z)$ is its relevant predictor. ■

Figure 1 provides two examples. The left panel shows a case without heterogeneity in the CATE where $s_0(Z) = 0$, whereas there right panel shows a case with strong heterogeneity in the CATE where $s_0(Z) = Z$. In both cases we evenly split 1000 observations between the auxiliary and main samples, Z is uniformly distributed in $(-1, 1)$, and the proxy predictor $S(Z)$ is estimated by random forest in the auxiliary sample following the standard implementation, see e.g. Friedman et al. (2001). When there is no heterogeneity, post-processing the ML estimates helps reducing sampling noise bringing the estimated BLP close to the true BLP; whereas under strong heterogeneity the signal in the ML estimates dominates the sampling noise and the post-processing has little effect.

Comment 2.2 (Digression: Naive Strategy that is not Quite Right)**.** It is tempting and "more natural" to estimate

$$Y = \tilde{\alpha}_1 + \tilde{\alpha}_2 B + \tilde{\beta}_1 D + \tilde{\beta}_2 D(S - \mathrm{E}S) + \epsilon, \quad \mathrm{E}[\epsilon \tilde{X}] = 0,$$

where $\tilde{X} = (1, B, D, D(S - \mathrm{E}S))$. This is a good strategy for predicting the conditional expectation of Y given Z and D. But, $\tilde{\beta}_2 \neq \beta_2$, and $\tilde{\beta}_1 + \tilde{\beta}_2(S - \mathrm{E}S)$ is not the best linear predictor of $s_0(Z)$. ■

BLP of CATE: The Second Strategy. The second strategy makes use of the Horvitz-Thompson transformation:

$$H = H(D, Z) = \frac{D - p(Z)}{p(Z)(1 - p(Z))}.$$

It is well known that the transformed response YH provides an unbiased signal about CATE:

$$\mathrm{E}[YH \mid Z] = s_0(Z)$$

and it follows that

$$\mathsf{BLP}[s_0(Z) \mid S(Z)] = \mathsf{BLP}[YH \mid S(Z)].$$

This simple strategy is completely fine for identification purposes, but can severely underperform in estimation and inference due to lack of precision. We can repair the deficiencies by considering, instead, the linear projection:

$$YH = \mu' X_1 H + \beta_1 + \beta_2(S - \mathrm{E}S) + \tilde{\epsilon}, \quad \mathrm{E}\tilde{\epsilon}\tilde{X} = 0, \tag{2.2}$$

where $B := B(Z)$, $S := S(Z)$, $\tilde{X} := (X_1' H, \tilde{X}_2')'$, $\tilde{X}_2 = (1, (S - \mathrm{E}S)')'$, and $X_1 = X_1(Z)$, e.g. $X_1 = B(Z)$ or $X_1 = (B(Z), S(Z), p(Z))'$. The terms X_1 are present in order to *reduce noise*.

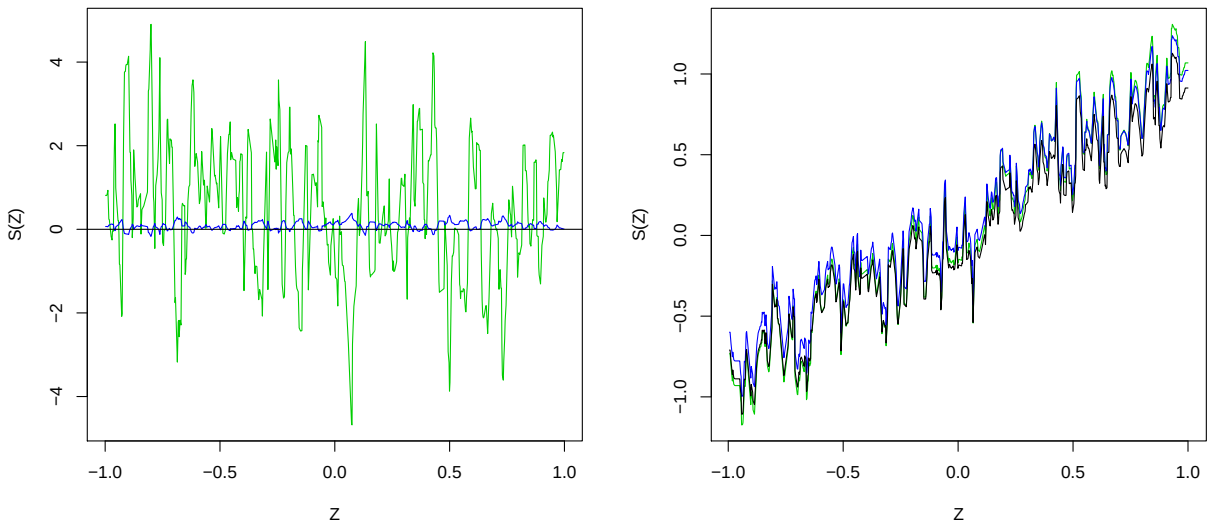

FIGURE 1. Example. In the left panel we have a homogeneous CATE $s_0(Z) = 0$; in the right panel we have heterogeneous CATE $s_0(Z) = Z$. The proxy predictor $S(Z)$ is produced by the Random Forest, shown by green line, the true BLP of CATE is shown by black line, and the estimated BLP of CATE is shown by blue line. The true and estimated BLP of CATE are more attenuated towards zero than the proxy predictor.

We show that, as a complementary main identification result,

$$\beta_1 + \beta_2(S - \mathrm{E}S) = \mathsf{BLP}[s_0(Z) \mid S(Z)].$$

Theorem 2.2 (BLP 2)**.** *Consider $z \mapsto S(z)$ and $z \mapsto B(z)$ as fixed maps. Assume that Y has finite second moments, $\tilde{X} = (X_1 H, 1, (S - \mathrm{E}S))$ is such that $\mathrm{E}\tilde{X}\tilde{X}'$ is finite and full rank, and $\mathrm{Var}(S(Z)) > 0$. Then, (β_1, β_2) defined in* (2.2) *solves the best linear predictor/approximation problem for the target $s_0(Z)$:*

$$(\beta_1, \beta_2)' = \arg\min_{b_1, b_2} \mathrm{E}[s_0(Z) - b_1 - b_2 S(Z)]^2,$$

in particular $\beta_1 = \mathrm{E}s_0(Z)$ and $\beta_2 = \mathrm{Cov}(s_0(Z), S(Z))/\mathrm{Var}(S(Z))$.

The corresponding estimator is defined through the empirical analog:

$$Y_i H_i = \widehat{\mu}' X_{1i} H_i + \widehat{\beta}_1 + \widehat{\beta}_2(S_i - \mathbb{E}_{N,M} S_i) + \widehat{\epsilon}_i, \quad \mathbb{E}_{N,M} \widehat{\epsilon}_i \tilde{X}_i = 0,$$

and the properties of this estimator, conditional on the auxiliary data, are well known and given in Lemma B.1.

Comment 2.3 (Comparison of Estimation Strategies)**.** A natural question that may arise is whether the two estimation strategies proposed can be ranked in terms of asymptotic efficiency. The answer is negative. We show in Appendix C that they produce estimators that have the same distribution in large samples.

2.2. **The Sorted Group ATE.** The target parameters are

$$\mathrm{E}[s_0(Z) \mid G],$$

where G is an indicator of group membership.

Comment 2.4. There are many possibilities for creating groups based upon ML tools applied to the auxiliary data. For example, one can group or cluster based upon predicted baseline response as in the "endogenous stratification" analysis (Abadie et al., 2017), or based upon actual predicted treatment effect S. We focus on the latter approach for defining groups, although our identification and inference ideas immediately apply to other ways of defining groups, and could be helpful in these contexts.

We build the groups to explain as much variation in $s_0(Z)$ as possible

$$G_k := \{S \in I_k\}, \quad k = 1, ..., K,$$

where $I_k = [\ell_{k-1}, \ell_k)$ are non-overlaping intervals that divide the support of S into regions $[\ell_{k-1}, \ell_k)$ with equal or unequal masses:

$$-\infty = \ell_0 < \ell_1 < \ldots < \ell_K = +\infty.$$

The parameters of interest are the Sorted Group Average Treatment Effects (GATES):

$$\mathrm{E}[s_0(Z) \mid G_k], \quad k = 1, \ldots, K.$$

Given the definition of groups, it is natural for us to impose the monotonicity restriction

$$\mathrm{E}[s_0(Z) \mid G_1] \leqslant ... \leqslant \mathrm{E}[s_0(Z) \mid G_K],$$

which holds asymptotically if $S(Z)$ is consistent for $s_0(Z)$ and the latter has an absolutely continuous distribution. Under the monotonicity condition, the estimates could be rearranged to obey the weak monotonicity condition, improving the precision of the estimator. The joint confidence intervals could also be improved by intersecting them with the set of monotone functions. Furthermore, as before, we can test for homogeneous effects, $s_0(Z) = s$, by testing whether,

$$\mathrm{E}[s_0(Z) \mid G_1] = ... = \mathrm{E}[s_0(Z) \mid G_K].$$

GATES: The First Strategy. Here we shall recover the GATES parameters from the weighted linear projection equation:

$$Y = \alpha' X_1 + \sum_{k=1}^{K} \gamma_k \cdot (D - p(Z)) \cdot 1(G_k) + \nu, \quad \mathrm{E}[w(Z)\nu W] = 0, \tag{2.3}$$

for $B := B(Z)$, $S := S(Z)$, $W = (X_1', W_2')'$,

$$W_2 = (\{(D - p(Z))1(G_k)\}_{k=1}^{K})'.$$

The presence of $D - p(Z)$ in the interaction $(D - p(Z))1(G_k)$ *orthogonalizes* this regressor relative to all other regressors that are functions of Z. The controls X_1, e.g. B, can be included to improve precision.

The second main identification result is that the projection coefficients γ_k are the GATES parameters:

$$\gamma = (\gamma_k)_{k=1}^{K} = (\mathrm{E}[s_0(Z) \mid G_k])_{k=1}^{K}.$$

Given the identification strategy, we can base the corresponding estimation strategy on the following empirical analog:

$$Y_i = \widehat{\alpha}' X_{1i} + \widehat{\gamma}' W_{2i} + \widehat{\nu}_i, \quad i \in M, \quad \mathbb{E}_{N,M}[w(Z_i)\widehat{\nu}_i W_i] = 0. \tag{2.4}$$

The properties of this estimator, conditional on the auxilliary data, are well known and stated as a special case of Lemma B.1.

A formal statement appears below, together with a complementary result.

Figure 2 provides two examples using the same designs as in fig. 1. Post-processing the ML estimates again has stronger effect when there is no heterogeneity, but in both cases help bring the estimated GATES close to the true GATES.

GATES: The Second Strategy. Here we employ linear projections on Horvitz-Thompson transformed variables:

$$YH = \mu' X_1 H + \sum_{k=1}^{K} \gamma_k \cdot 1(G_k) + \nu, \quad \mathrm{E}[\nu \tilde{W}] = 0, \tag{2.5}$$

for $B := B(Z)$, $S := S(Z)$, $\tilde{W} = (X_1' H, \tilde{W}_2')$, $\tilde{W}_2' = (\{1(G_k)\}_{k=1}^{K})$.

Again, we show that the projection parameters are GATES:

$$\gamma = (\gamma_k)_{k=1}^{K} = (\mathrm{E}[s_0(Z) \mid G_k])_{k=1}^{K}.$$

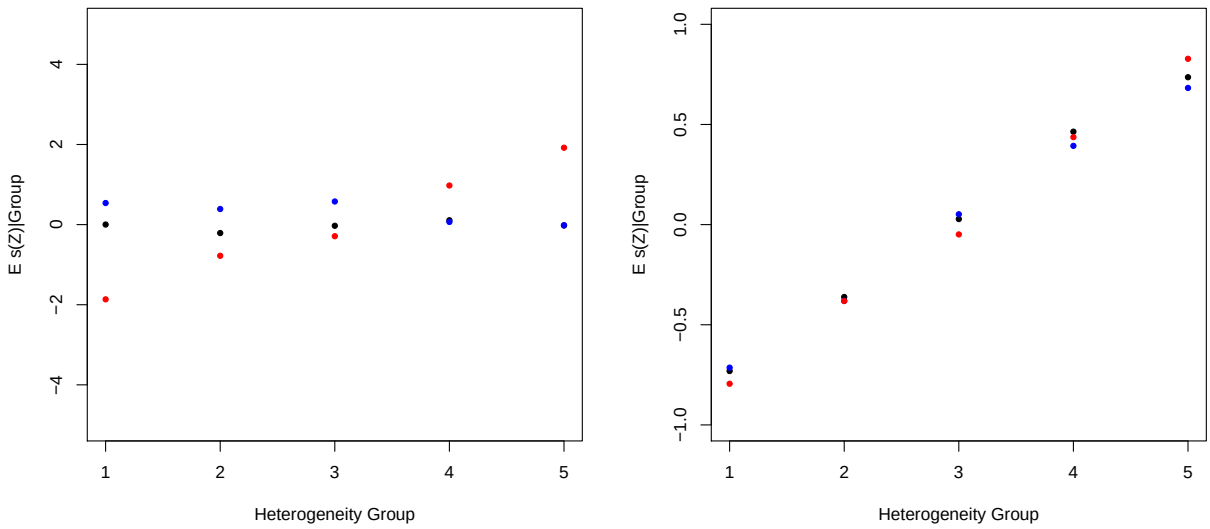

FIGURE 2. In the left panel we have the homogeneous CATE $s_0(Z) = 0$; in the right panel we have heterogeneous CATE $s_0(Z) = Z$. The proxy predictor $S(Z)$ for CATE is produced by the random forest, whose sorted averages by groups are shown as red dots, exhibiting large biases. These are the naive estimates. The true sorted group average treatment effects (GATES) $\mathrm{E}[s_0(Z) \mid G_k]$ are shown by black dots, and estimated GATES are shown by blue dots. The true and estimated GATES correct for the biases relative to the naive strategy shown in red. The estimated GATES shown by blue dots are always closer to the true GATEs shown by black dots than the naive estimates shown in red.

Given the identification strategy, we can base the corresponding estimation strategy on the following empirical analog:

$$Y_i H_i = \widehat{\mu}' X_{1i} H_i + \widehat{\gamma}' \tilde{W}_{2i} + \widehat{\nu}_i, \quad i \in M, \quad \mathbb{E}_{N,M}[\widehat{\nu}_i \tilde{W}_i] = 0. \tag{2.6}$$

The properties of this estimator, conditional on the auxiliary data, are well known and given in Lemma B.1. The resulting estimator has similar performance to the previous estimator, and under some conditions their first-order properties coincide.

The following is the formal statement of the identification result.

Theorem 2.3 (GATES). *Consider $z \mapsto S(z)$ and $z \mapsto B(z)$ as fixed maps. Assume that Y has finite second moments and the W's and $\tilde{W}$ defined above are such that $\mathrm{E}WW'$ and $\mathrm{E}\tilde{W}\tilde{W}'$ are finite and have full rank. Consider $\gamma = (\gamma_k)_{k=1}^K$ defined by the weighted regression equation (2.3) or by the regression equation (2.5). These parameters defined in two different ways are equivalent and are equal to the expectation of $s_0(Z)$*

conditional on the proxy group $\{S \in I_k\}$*:*

$$\gamma_k = \mathrm{E}[s_0(Z) \mid G_k].$$

2.3. **Classification Analysis (CLAN).** When the BLP and GATES analyses reveal substantial heterogeneity, it is interesting to know the properties of the subpopulations that are most and least affected. Here we focus on the "least affected group" G_1 and "most affect group" G_K. Under the monotonicity assumption, it is reasonable that the first and the last groups are the most and least affected, where the labels "most" and "least" can be swapped depending on the context.

Let $g(Y, Z)$ be a vector of characteristics of an observational unit. The parameters of interest are the average characteristics of the most and least affected groups:

$$\delta_1 = \mathrm{E}[g(Y, Z) \mid G_1] \quad \text{and} \quad \delta_K = \mathrm{E}[g(Y, Z) \mid G_K].$$

The parameters δ_K and δ_1 are identified because they are averages of variables that are directly observed. We can compare δ_K and δ_1 to quantify differences between the most and least affected groups. We call this type of comparisons as classification analysis or CLAN.

3. "Variational" Estimation and Inference Methods

3.1. **Estimation and Inference: The Generic Targets.** Let θ denote a generic target parameter or functional, for example,

- $\theta = \beta_2$ is the heterogeneity predictor loading parameter;
- $\theta = \beta_1 + \beta_2(S(z) - \mathrm{E}S)$ is the "personalized" prediction of $s_0(z)$;
- $\theta = \gamma_k$ is the expectation of $s_0(Z)$ for the group G_k;
- $\theta = \gamma_K - \gamma_1$ is the difference in the expectation of $s_0(Z)$ between the most and least affected groups;
- $\theta = \delta_K - \delta_1$ is the difference in the expectation of the characteristics of the most and least impacted groups.

3.2. **Quantification of Uncertainty: Two Sources.** There are two principal sources of sampling uncertainty:

(I) Estimation uncertainty regarding the parameter θ, conditional on the data split;

(II) Uncertainty or "variation" induced by the data splitting.

Conditional on the data split, quantification of estimation uncertainty is standard. To account for uncertainty with respect to the data splitting, it makes sense to examine the robustness and variability of the estimates/confidence intervals with respect to different random splits. One of our goals is to develop methods, which we call "variational estimation and inference" (VEIN) methods,

for quantifying this uncertainty. These methods can be of independent interest in many settings where the sample splitting is used.

Quantifying Source (I): Conditional Inference. We first recognize that the parameters implicitly depend on

$$\text{Data}_A := \{(Y_i, D_i, X_i)\}_{i \in A},$$

the auxiliary sample, used to create the ML proxies $B = B_A$ and $S = S_A$. Here we make the dependence explicit: $\theta = \theta_A$.

All of the examples admit an estimator $\widehat{\theta}_A$ such that under mild assumptions,

$$\widehat{\theta}_A \mid \text{Data}_A \sim_a N(\theta_A, \widehat{\sigma}_A^2),$$

in the sense that, as $|M| \to \infty$,

$$\mathbb{P}(\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta_A) \leqslant z \mid \text{Data}_A) \to_P \Phi(z).$$

Implicitly this requires the auxiliary data Data_A to be "sufficiently regular", and this should happen with high probability.

As a consequence, the confidence interval (CI)

$$[L_A, U_A] := [\widehat{\theta}_A \pm \Phi^{-1}(1 - \alpha/2)\widehat{\sigma}_A]$$

covers θ_A with approximate probability $1 - \alpha$:

$$\mathbb{P}[\theta_A \in [L_A, U_A] \mid \text{Data}_A] = 1 - \alpha - o_P(1).$$

This leads to straighforward conditional inference, which does not account for the sample splitting uncertainty.

Quantifying Source (II): "Variational" Inference. Different partitions (A, M) of $\{1, ..., N\}$ yield different targets θ_A. Conditional on the data, we treat θ_A as a random variable, since (A, M) are random sets that form random partitions of $\{1, \ldots, N\}$ into samples of size $|M|$ and $|A| = N - |M|$. Different partitions also yield different estimators $\widehat{\theta}_A$ and approximate distributions for these estimators. Hence we need a systematic way of treating the randomness in these estimators and their distributions.

Comment 3.1. In cases where the data sets are not large, it may be desirable to restrict attention to balanced partitions (A, M), where the proportion of treated units is equal to the designed propensity score.

We want to quantify the uncertainty induced by the random partitioning. Conditional on Data, the estimated $\widehat{\theta}_A$ is still a random variable, and the confidence band $[L_A, U_A]$ is a random set. For reporting purposes, we instead would like to report an estimator and confidence set, which are non-random conditional on the data.

Adjusted Point and Interval Estimators. Our proposal is as follows. As a point estimator, we shall report the median of $\widehat{\theta}_A$ as (A, M) vary (as random partitions):

$$\widehat{\theta} := \mathrm{Med}[\widehat{\theta}_A \mid \mathrm{Data}].$$

This estimator is more robust than the estimator based on a single split. To account for partition uncertainty, we propose to report the following confidence interval (CI) with the nominal confidence level $1 - 2\alpha$:

$$[l, u] := [\overline{\mathrm{Med}}[L_A \mid \mathrm{Data}], \underline{\mathrm{Med}}[U_A \mid \mathrm{Data}]].$$

Note that the price of splitting uncertainty is reflected in the discounting of the confidence level from $1 - \alpha$ to $1 - 2\alpha$. Alternatively, we can report the confidence interval based on inversion of a test based upon p-values, constructed below.

The above estimator and confidence set are non-random conditional on the data. The confidence set reflects the uncertainty created by the random partitioning of the data into the main and auxilliary data.

Comment 3.2. For a random variable X with law P_X we define

$$\underline{\mathrm{Med}}(X) := \inf\{x \in \mathbb{R} : \mathrm{P}_X(X \leqslant x) \geqslant 1/2\},$$

$$\overline{\mathrm{Med}}(X) := \sup\{x \in \mathbb{R} : \mathrm{P}_X(X \geqslant x) \geqslant 1/2\},$$

$$\mathrm{Med}(X) := (\underline{\mathrm{Med}}(X) + \overline{\mathrm{Med}}(X))/2.$$

Note that the lower median $\underline{\mathrm{Med}}(X)$ is the usual definition of the median. The upper median $\overline{\mathrm{Med}}(X)$ is the next distinct quantile of the random variable (or it is the usual median after reversing the order on $\mathbb{R}$). For example, when X is uniform on $\{1, 2, 3, 4\}$, then $\underline{\mathrm{Med}}(X) = 2$ and $\overline{\mathrm{Med}}(X) = 3$; and if X is uniform on $\{1, 2, 3\}$, then $\overline{\mathrm{Med}}(X) = \underline{\mathrm{Med}}(X) = 2$. For continuous random variables the upper and lower medians coincide. For discrete random variables they can differ, but the differences will be small for variables that are close to being continuous. ■

Suppose we are testing $H_0 : \theta_A = \theta_0$ against $H_1 : \theta_A < \theta_0$, conditional on the auxiliary data, then the p-value is given by

$$p_A = \Phi(\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta_0)).$$

The p-value for testing $H_0 : \theta_A = \theta_0$ against $H_1 : \theta_A > \theta_0$, is given by $p_A = 1 - \Phi(\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta_0))$.

Under the null hypothesis p_A is approximately distributed as the uniform variable, $p_A \sim U(0, 1)$, conditional on Data_A. Note that, conditional on Data, p_A still has randomness induced by random partitioning of the data, which we need to address.

Adjusted P-values. We say that testing the null hypothesis, based on the p-values p_A, that are random conditional on data, has significance level α if

$$\mathbb{P}(p_A \leqslant \alpha/2 \mid \text{Data}) \geqslant 1/2 \quad \text{or } p_{.5} = \underline{\text{Med}}(p_A \mid \text{Data}) \leqslant \alpha/2.$$

That is, for at least 50% of the random data splits, the realized p-value p_A falls below the level $\alpha/2$. Hence we can call $p = 2p_{.5}$ the *sample splitting-adjusted p-value*, and consider its small values as providing evidence against the null hypothesis.

Comment 3.3. Our construction of p-values builds upon the false-discovery-rate type adjustment ideas in Benjamini and Hochberg (1995); Meinshausen et al. (2009), though what we propose is much simpler, and is minimalistic for our problem, whereas the idea of our confidence intervals below appears to be new. ■

The main idea behind this construction is simple: the p-values are distributed as marginal uniform variables $\{U_j\}_{j\in J}$, and hence obey the following property.

Lemma 3.1 (A Property of Uniform Variables)**.** *Consider M, the (usual, lower) median of a sequence $\{U_j\}_{j\in J}$ of uniformly distributed variables, $U_j \sim U(0,1)$ for each $j \in J$, where variables are not necessarily independent. Then,*

$$\mathbb{P}(M \leqslant \alpha/2) \leqslant \alpha.$$

Proof. Let M denote the median of $\{U_j\}_{j\in J}$. Then $M \leqslant \alpha/2$ is equivalent to $|J|^{-1}\sum_{j\in J}[1(U_j \leqslant \alpha/2)] - 1/2 \geqslant 0$. So

$$\mathbb{P}[M \leqslant \alpha/2] = \mathbb{E}1\{|J|^{-1}\sum_{j\in J}[1(U_j \leqslant \alpha/2)] \geqslant 1/2\}.$$

By Markov inequality this is bounded by

$$2\mathbb{E}|J|^{-1}\sum_{j\in J}[1(U_j \leqslant \alpha/2)] \leqslant 2\mathbb{E}[1(U_j \leqslant \alpha/2)] \leqslant 2\alpha/2 = \alpha.$$

where the last inequality holds by the marginal uniformity. ■

Main Inference Result: Variational P-values and Confidence Intervals. We present a formal result on adjusted p-values using this condition:

PV. Suppose that $\mathcal{A}$ is a set of regular auxiliary data configurations such that for all $x \in [0,1]$, under the null hypothesis:

$$\sup_{P\in\mathcal{P}} |\mathbb{P}_P[p_A \leqslant x \mid \text{Data}_A \in \mathcal{A}] - x| \leqslant \delta = o(1),$$

and $\inf_{P\in\mathcal{P}} \mathbb{P}_P[\text{Data}_A \in \mathcal{A}] =: 1-\gamma = 1 - o(1)$. In particular, suppose that this holds for the p-values

$$p_A = \Phi(\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta_A)) \text{ and } p_A = 1 - \Phi(\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta_A)).$$

Lemma B.1 shows that this condition is plausible for the least squares estimators defined in the previous section under mild conditions.

Theorem 3.1 (Uniform Validity of Variational P-Value)**.** *Under condition* PV *and the null hypothesis holding,*

$$\mathbb{P}_P(p_{.5} \leqslant \alpha/2) \leqslant \alpha + 2(\delta + \gamma) = \alpha + o(1),$$

uniformly in $P \in \mathcal{P}$.

In order to establish the properties of the confidence interval $[l, u]$, we first consider the properties of the related confidence interval, which is based on the inversion of the p-value based tests:

$$\text{CI} := \{\theta \in \mathbb{R} : p_u(\theta) > \alpha/2,\ p_l(\theta) > \alpha/2\}, \tag{3.1}$$

for $\alpha < .25$, where, for $\widehat{\sigma}_A > 0$,

$$p_l(\theta) \quad := \quad \underline{\text{Med}}(1 - \Phi[\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta)] \mid \text{Data}), \tag{3.2}$$

$$p_u(\theta) \quad := \quad \underline{\text{Med}}(\Phi[\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta)] \mid \text{Data}). \tag{3.3}$$

The confidence interval CI has the following representation in terms of the medians of t-statistics implied by the proof Theorem 3.2 stated below:

$$\text{CI} = \left\{ \theta \in \mathbb{R} : \begin{array}{l} \overline{\text{Med}} \left[\frac{\theta - \widehat{\theta}_A}{\widehat{\sigma}_A} - \Phi^{-1}(1 - \alpha/2) \mid \text{Data} \right] < 0 \\ \underline{\text{Med}} \left[\frac{\theta - \widehat{\theta}_A}{\widehat{\sigma}_A} + \Phi^{-1}(1 - \alpha/2) \mid \text{Data} \right] > 0 \end{array} \right\}. \tag{3.4}$$

This CI can be (slightly) tighter than $[l, u]$, while the latter is much simpler to construct.

The following theorem establishes that both confidence sets maintain the approximate confidence level $1 - 2\alpha$.

Theorem 3.2 (Uniform Validity of Variational Confidence Intervals)**.** CI *can be represented as (3.4) and* $\text{CI} \subseteq [l, u]$*, and under condition* PV*,*

$$\mathbb{P}_P(\theta_A \in \text{CI}) \geqslant 1 - 2\alpha - 2(\delta + \gamma) = 1 - 2\alpha - o(1),$$

uniformly in $P \in \mathcal{P}$.

4. Other Considerations and Extensions

1. Choosing the Best ML Method Targeting CATE in Stage 1. There are several options. The best ML method can be chosen using the auxiliary sample, based on either (a) the ability to predict YH using BH and S or (b) the ability to predict Y using B and $(D - p(Z))(S - \text{E}(S))$ under the weight $w(Z)$ (as in the first type of strategies we developed earlier). To be specific, we can solve either of the following problems:

(a) minimize the errors in the prediction of YH on BH and S:

$$(B, S) = \arg \min_{B \in \mathcal{B}, S \in \mathcal{S}} \sum_{i \in A} [Y_i H_i - B(Z_i) H_i - S(Z_i)]^2,$$

where $\mathcal{B}$ and $\mathcal{S}$ are parameter spaces for $z \mapsto B(z)$ and $z \mapsto S(z)$; or

(b) minimize the errors in the weighted prediction of Y on B and $(D - p(Z))(S - \mathrm{E}(S))$:

$$(B, S) = \arg \min_{B \in \mathcal{B}, S \in \mathcal{S}} \sum_{i \in A} w(Z_i)[Y_i - B(Z_i) - (D_i - p(Z_i))\{S(Z_i) - \bar{S}(Z_i)\}]^2,$$

where $\bar{S}(Z_i) = |A|^{-1} \sum_{i \in A} S(Z_i)$ and $\mathcal{B}$ and $\mathcal{S}$ are parameter spaces for $z \mapsto B(z)$ and $z \mapsto S(z)$.

This idea improves over simple but inefficient strategy of predicting YH just using S, which have been suggested before for causal inference. It also improves over the simple strategy that predicts Y using B and DS (which chooses the best predictor for $\mathrm{E}[Y \mid D, Z]$ in a given class but not necessarily the best predictor for CATE $s_0(Z)$). Note that this idea is new and is of major independent interest.

2. Choosing the Best ML Method BLP Targeting CATE in Stage 2. The best ML method can also be chosen in the main sample by maximizing

$$\Lambda := |\beta_2|^2 \mathrm{Var}(S(Z)) = \mathrm{Corr}^2(s_0(Z), S(Z)) \mathrm{Var}(s_0(Z)). \tag{4.1}$$

Maximizing Λ is equivalent to maximizing the correlation between the ML proxy predictor $S(Z)$ and the true score $s_0(Z)$, or equivalent to maximizing the R^2 in the regression of $s_0(Z)$ on $S(Z)$.

3. Choosing the Best ML Method GATES Targeting CATE in Stage 2. Analogously, for GATES the best ML method can also be chosen in the main sample by maximizing

$$\bar{\Lambda} = \mathrm{E}\left(\sum_{k=1}^{K} \gamma_k 1(S \in I_k)\right)^2 = \sum_{k=1}^{K} \gamma_k^2 \mathrm{P}(S \in I_k). \tag{4.2}$$

This is the part of variation $\mathrm{E}s_0^2(Z)$ of $s_0(z)$ explained by $\bar{S}(Z) = \sum_{k=1}^{K} \gamma_k 1(S(Z) \in I_k)$. Hence choosing the ML proxy $S(Z)$ to maximize $\bar{\Lambda}$ is equivalent to maximizing the R^2 in the regression of $s_0(Z)$ on $\bar{S}(Z)$ (without a constant). If the groups $G_k = \{S \in I_k\}$ have equal size, namely $\mathrm{P}(S(Z) \in I_k) = 1/K$ for each $k = 1, ..., K$, then

$$\bar{\Lambda} = \frac{1}{K} \sum_{k=1}^{K} \gamma_k^2.$$

4. Stratified Splitting. The idea is to balance the proportions of treated and untreated in both A and M samples, so that the proportion of treated is equal to the experiment's propensity scores across strata. This formally requires us to replace the i.i.d. assumption by the i.n.i.d. assumption

(independent but not identically distributed observations) when accounting for estimation uncertainty, conditional on the auxiliary sample. This makes the notation more complicated, but the results in Lemma B.1 still go through with notational modifications.

5. When Proxies have Little Variation. The analysis may generate proxy predictors S that have little variation, so we can think of them as "weak", which makes the parameter β_2 weakly identified. We can either add small noise to the proxies (jittering), so that inference results go through, or we may switch to testing rather than estimation. For practical reasons, we prefer the jittering approach.

5. Further Potential Applications to Prediction and Causal Inference Problems

Our inference approach generalizes to any problem of the following sort.

Generalization. Suppose we can construct an *unbiased signal* $\tilde{Y}$ such that

$$\mathrm{E}[\tilde{Y} \mid Z] = s_0(Z),$$

where $s_0(Z)$ is now a generic target function. Let $S(Z)$ denote an ML proxy for $s_0(Z)$. Then, using previous arguments, we immediately can generate the following conclusions:

(1) The projection of $\tilde{Y}$ on the ML proxy $S(Z)$ identifies the BLP of $s_0(Z)$ using $S(Z)$.
(2) The grouped average of the target (GAT) $\mathrm{E}[s_0(Z) \mid G_k]$ is identified by $\mathrm{E}[\tilde{Y} \mid G_k]$.
(3) Using ML tools we can train proxy predictors $S(Z)$ to predict $\tilde{Y}$ in auxiliary samples.
(4) We post-process $S(Z)$ in the main sample, by estimating the BLP and GATs.
(5) We apply variational inference on functionals of the BLP and GATs.

The noise reduction strategies, like the ones we used in the context of H-transformed outcomes, can be useful in these cases, but their construction could depend on the context.

Example 1. Forecasting or Predicting Regression Functions using ML proxies. This is the most common type of the problem arising in forecasting. Here the target is the best predictor of Y using Z, namely $s_0(Z) = \mathrm{E}[Y \mid Z]$, and $\tilde{Y} = Y$ trivially serves as the unbiased signal. The interesting part here is the use of variational inference tools developed in this paper for constructing confidence intervals for the predicted values produced by the estimated BLP of $s_0(Z)$ using $S(Z)$.

Example 2. Predicting Structural Derivatives using ML proxies. Suppose we are interested in best predicting the conditional average partial derivative $s_0(z) = \mathrm{E}[g'(X, Z) \mid Z = z]$, where $g'(x, z) = \partial g(x, z)/\partial x$ and $g(x, z) = \mathrm{E}[Y \mid X = x, Z = z]$. In the context of demand analysis, Y is the log of individual demand, X is the log-price of a product, and Z includes prices of other products and characteristics of individuals. Then, the unbiased signal is given by $\tilde{Y} = -Y[\partial \log p(X \mid Z)/\partial x]$, where $p(\cdot \mid \cdot)$ is the conditional density function of X given Z. That is, $\mathrm{E}[\tilde{Y} \mid Z] = s_0(Z)$ under mild conditions on the density using the integration by parts formula.

6. Empirical Applications and Implementation Algorithms

To illustrate the methods developed in this paper, we consider two empirical examples. The first example is an RCT conducted in Morocco, which investigates the effect of microfinance access on several outcomes. The second example analyzes a randomized intervention program in India to improve immunization. We conclude this section by providing the implementation algorithm.

6.1. **Heterogeneity in the Effect of Microcredit Availability.** We analyze a randomized experiment designed to evaluate the impact of microcredit availability on borrowing and self-employment activities, which was previously studied in Crépon et al. (2015). The experiment was conducted in 162 villages in Morocco, divided into 81 pairs of villages with similar observable characteristics (number of households, accessibility to the center of the community, existing infrastructure, type of activities carried out by the households, and type of agriculture activities). One of the villages in each pair was randomly assigned to treatment and the other to control. Between 2006 and 2007 a microfinance institution started operating in the treated villages. Two years after the intervention an endline household survey was conducted with 5,551 households, which constitute our sample. There was no other microcredit penetration in these villages, before and for the duration of the study. Therefore, we interpret the treatment as the availability of microcredit.

Recent randomized evaluations of access to microcredit at the community level have found limited impacts of microcredit.[7] Despite evidence that access to microfinance leads to an increase in borrowing (Angelucci et al. (2015), Banerjee et al. (2015b), Tarozzi et al. (2015)) and business creation or expansion (Angelucci et al. (2015), Attanasio et al. (2015), Banerjee et al. (2015b), Tarozzi et al. (2015)), most studies have found that this does not translate into an increase in economic outcomes such as profit, income, labor supply and consumption (Angelucci et al. (2015), Banerjee et al. (2015b), Crépon et al. (2015)). Moreover, there is also no evidence of substantial gains in human development outcomes, such as education and health (Banerjee et al. (2015b),Tarozzi et al. (2015)). Studies which estimate the impact of microfinance by randomizing microcredit at the individual level confirm these findings (Augsburg et al. (2012), Karlan and Zinman (2009), Karlan and Zinman (2011)).

One question that remains elusive is whether the lack of evidence on the average effects masks heterogeneity, in which there are potential winners and losers of the microcredit expansion. Understanding this heterogeneity can have important implications for evaluating the welfare effects of microcredit, designing policies and targeting the groups that would benefit from microfinance. Indeed, the idea that there might be heterogeneity in the impact of microcredit has been a common theme among RCTs evaluating microfinance programs. Having found mostly positive but insignificant coefficients, the papers cited above attempt to explore heterogeneous treatment effects, mostly using quantile treatment effects. For profits, most studies seem to find positive impact at the higher quantiles (and in the data set we study here, Crépon et al. (2015) actually find *negative* impacts at

[7]See Banerjee (2013) for a summary of the recent literature

lower level). Using Bayesian hierarchical methods to aggregate the evidence across studies, Meager (2017) cautions that these results on quantiles may not be generalizable: the profit variables seems to have too much noise to lend itself to quantile estimation.

A number of recent papers also consider heterogeneous treatment effects by studying the effect of microfinance on subpopulations. In a follow-up study of Banerjee et al. (2015b), Banerjee et al. (2015a) investigates whether the heterogeneity is persistent six years after the microfinance was introduced. They find that credit has a much bigger impact on the business outcomes of those who started a business before microfinance entered than of those without prior businesses. Using the same dataset as in this application, Crépon et al. (2015) classifies households into three categories in terms of their probability to borrow before the intervention and finds that microcredit access has a significant impact on investment and profit, but still no impact on income and consumption among those who are most likely to borrow. It is worth noting that the original strategy for this study was to construct groups which, *ex ante* had different probability to borrow, in order to separately estimate the direct effect of microcredit on those most likely to borrow, and the indirect effect on those very unlikely to borrow. The researchers initially tried to predict the probability to borrow fitting a model to a first group of villages for which they had collected a short survey. However they ended up predicting the probability to borrow *ex-post* because the model proved to have low predictive power. This ex-post classification may lead to overfitting. One cause for concern in this case is that different variables predict the probability to borrow in different waves, which makes it less likely that those variables reflect true structural relationships.

The strategy developed in this paper provides several advantages in studying heterogeneity in the treatment effects of microfinance. First, contrary to the literature, which relies on ad hoc subgroup analysis across a few baseline characteristics, we are agnostic about the source of heterogeneity. While the variable "had a prior business" has proven to be a robust and generalizable predictor of differences in treatment effect (Meager (2017)) and could therefore be pre-specified in future pre-analysis plans, we have little idea about what else predicts heterogeneity. Second, our approach is valid in high dimensional settings, allowing us to include a rich set of characteristics in an unspecified functional form. Finally, using the CLAN estimation we are able to identify the characteristics of the most and least affected subpopulations, which could be an important input for a welfare analysis or targeting households who are likely to benefit from access to microfinance.

We focus on heterogeneity in treatment effects on four household outcome variables, Y: the amount of money borrowed, the output from self-employment activities, profit from self-employment activities, and monthly consumption. The treatment variable, D, is an indicator for the household residing in a treated village. The covariates, Z, include some baseline household characteristics such as number of members, number of adults, head age, indicators for households doing animal husbandry, doing other non-agricultural activity, having an outstanding loan over the past 12 months, household spouse responded to the survey, another household member (excluding the household head) responded to the survey, and 81 village pair fixed effects (these are the variables

that are available for all households). We also include indicators for missing observation at baseline as controls. Table 1 shows some descriptive statistics for the variables used in the analysis (all monetary variables are expressed in Moroccan Dirams, or MAD). Treated and control households have similar characteristics and the unconditional average treatment effect on loans, output, profit and consumption are respectively 1,128, 5,237, 1,844 and -31.

Table 1. Descriptive Statistics of Households

	All	Treated	Control
Outcome Variables			
Total Amount of Loans	2,359	2,930	1,802
Total output from self-employment activities (past 12 months)	32,499	35,148	29,911
Total profit from self-employment activities (past 12 months)	10,102	11,035	9,191
Total monthly consumption	3,012	2,996	3,027
Baseline Covariates			
Number of Household Members	3.879	3.872	3.886
Number of Members 16 Years Old or Older	2.604	2.601	2.607
Head Age	35.976	35.937	36.014
Declared Animal Husbandry Self-employment Activity	0.415	0.426	0.404
Declared Non-agricultural Self-employment Activity	0.146	0.129	0.164
Borrowed from Any Source	0.210	0.224	0.196
Spouse of Head Responded to Self-employment Section	0.067	0.074	0.061
Member Responded to Self-employment Section	0.044	0.048	0.041

We implement our methods using the algorithm and ML methods described in Section 6.3. By design the propensity score $p(Z_i) = 1/2$ for all the households. Table 2 compares the four ML methods for producing the proxy predictors $S(Z_i)$ considered in Stage 1. We find that the Random Forest and Elastic Net outperform the Boosted Tree and Neural Network across all outcome variables for both metrics. Accordingly, we focus on these two methods for the rest of the analysis.[8]

Table 3 presents results of the BLP of CATE using the ML proxies $S(Z)$ for the four outcome variables. We report estimates of the coefficients β_1 and β_2, which correspond to the ATE and heterogeneity loading (HET) parameters in the BLP, respectively. In parentheses, we report confidence intervals adjusted for variability across the sample splits using the median method; and in brackets, we report adjusted p-values for the hypothesis that the parameter is equal to zero. The estimated ATEs of microfinance availability are consistent with the findings of Crépon et al. (2015) and are similar to the unconditional ATE, as expected by virtue of the randomization. The ATE on the amount of loans and output are positive and statistically significant at least at the 10% level

[8]The results obtained using Boosted Tree and Neural Network are similar to the results reported, but they are slightly less precise. These results are not reported but are available from the authors upon request.

Table 2. Comparison of ML Methods: Microfinance Availability

	Elastic Net	Boosting	Neural Network	Random Forest
Amount of Loans				
Best BLP (Λ)	2,808,960	1,919,609	2,175,872	2,753,511
Best GATES ($\bar{\Lambda}$)	875	283	568	1290
Output				
Best BLP (Λ)	142,021,759	81,927,950	72,908,917	123,485,223
Best GATES ($\bar{\Lambda}$)	8,677	3,625	4,986	5,123
Profit				
Best BLP (Λ)	32,462,874	16,674,642	13,411,383	43,184,732
Best GATES ($\bar{\Lambda}$)	4,595	2,167	1,447	4,344
Consumption				
Best BLP (Λ)	45,084	26,158	38,578	37,507
Best GATES ($\bar{\Lambda}$)	101	69	85	109

Notes: Medians over 100 splits in half.

Table 3. BLP of Microfinance Availability

	Elastic Net		Random Forest	
	ATE (β_1)	HET (β_2)	ATE (β_1)	HET (β_2)
Amount of Loans	1,163	0.238	1,180	0.390
	(545, 1,737)	(0.021,0.448)	(546, 1,770)	(0.037, 0.779)
	[0.000]	[0.060]	[0.001]	[0.062]
Output	5,096	0.262	4,854	0.190
	(230, 10,027)	(0.084, 0.431)	(-167, 9,982)	(-0.099, 0.498)
	[0.079]	[0.008]	[0.116]	[0.385]
Profit	1,554	0.243	1,625	0.275
	(-1,344, 4,388)	(0.079, 0.416)	(-1,332, 4,576)	(0.036,0.510)
	[0.584]	[0.008]	[0.577]	[0.045]
Consumption	-59.2	0.154	-58.5	0.183
	(-161.4, 43.9)	(-0.054, 0.382)	(-167.0, 45.9)	(-0.177, 0.565)
	[0.513]	[0.270]	[0.494]	[0.617]

Notes: Medians over 100 splits. 90% confidence interval in parenthesis.
P-values for the hypothesis that the parameter is equal to zero in brackets.

with both ML methods. Microfinance availability does not have a significant impact on profit and consumption.

TABLE 4. GATES of 20% Most and Least Affected Groups

	Elastic Net			Random Forest		
	20% Most (γ_5)	20% Least (γ_1)	Difference ($\gamma_5 - \gamma_1$)	20% Most (γ_5	20% Least (γ_1)	Difference ($\gamma_5 - \gamma_1$)
Amount of Loans	2,678	-197	2,995	2,883	70	2,942
	(1,298, 4,076)	(-1,835, 1,308)	(946, 5,104)	(1,141, 4,695)	(-1,630, 1,594)	(551, 5,355)
	[0.000]	[1.000]	[0.008]	[0.002]	[1.000]	[0.034]
Output	22,070	-2,882	2,531	21,551	690	21,790
	(7,343, 36,960)	(-12,602, 6,920)	(7,201, 42,649)	(6,764, 37,498)	(-12,457, 13,840)	(-313.6, 42,831)
	[0.007]	[1.000]	[0.012]	[0.011]	[1.000]	[0.108]
Profit	10,707	-1,227	11,768	12,000	-2,130	14,056
	(1,628, 19,032)	(-7,273, 5,003)	(1,186, 22,485)	(2,911, 20,638)	(-9,135, 4,853)	(2,292, 25,698)
	[0.028]	[1.000]	[0.059]	[0.018]	[1.000]	[0.035]
Consumption	60	-342	378	56	-309	313
	(-174, 281)	(-686, -0.32)	(-66, 808)	(-252, 360)	(-691, 59)	(-211, 813)
	[1.000]	[0.100]	[0.189]	[1.000]	[0.222]	[0.522]

Notes: Medians over 100 splits. 90% confidence interval in parenthesis.
P-values for the hypothesis that the parameter is equal to zero in brackets.

Turning to the heterogeneity results, we reject the hypothesis that HET is zero at the 10% level for the amount of loans, output and profit with the elastic net method, suggesting the presence of heterogeneity in the effect of microfinance availability. The results are consistent across both ML methods except for output, for which HET coefficient on the Random Forest proxy is not significantly different from zero at the 10% level. Finally, the BLP analysis does not reveal any significant heterogeneity in the effect on consumption. Overall, these results suggest that microfinance availability has heterogenous impacts on business-related outcomes that do not seem to translate into a detectable contemporaneous effect on the standard of living as represented by consumption, even for the most positively affected households. One possible explanation is that households that are most likely to borrow and get higher profits from microfinance compensate by reducing their labor supply: this is the finding in Crépon et al. (2015).

We next estimate the GATES. We divide the households into $K = 5$ groups based on the quintiles of the ML proxy predictor $S(Z)$ and estimate the average effect for each group. Figures 3-6 presents the estimated GATES coefficients $\gamma_1 - \gamma_5$ along with joint confidence bands. We also report the ATE and its confidence interval that were obtained in the BLP analysis for comparison. The GATES provide a richer understanding of the heterogeneity. In particular, the figures reveal that there are groups of winners, the most affected groups, for which the GATES on amount of loans, output and profit are significantly different from zero. These groups are likely to drive the heterogeneity in the treatment effect that we find in the BLP analysis. We further investigate the GATES by comparing the most and least affected groups in Table 4. We find that the difference of GATES of these two

Table 5. Predictive Power of Covariates for Treatment Effect Heterogeneity

	Elastic Net	Random Forest
Pair Fixed Effects		
Amount of Loans	0.94	0.81
Output	0.95	0.72
Profit	0.98	0.73
Baseline Covariates		
Amount of Loans	0.35	0.28
Output	0.26	0.15
Profit	0.16	0.08

groups is significantly different from zero at least at the 10% level on amount of loans, output and profit, whereas we fail to reject the hypothesis that this difference is zero at conventional levels on consumption. The results are robust to the ML method used. Looking at the least affected group, it is reassuring to see that we have no evidence of negative impact on profit and income, mitigating the concerns that there are adversely affected households. However, there is negative and insignificant effect on consumption for the same group. A possible explanation for this result is that investment is lumpy and some households cut back consumption to increase investment.

After presenting evidence on the heterogeneity of treatment effects for three outcomes we examine what drives this heterogeneity in the data using CLAN. We omit the results for consumption as we do not detect significant heterogeneity for this outcome. Remember that in our estimation, we used two sets of covariates to predict heterogeneity: baseline household characteristics and village pair fixed effects. In the original design, similar villages were paired based both on the fact that they were under the catchment area on the same branch, and on some observable characteristics such as the number of households, accessibility to the center of the community, existing infrastructure, type of activities carried out by the households, and type of agriculture activities. However, our dataset does not contain these village-level characteristics. Thus, we can view the village pair fixed effects as a rich set of proxy variables for both village-level characteristics that are unobservables (to us), and also the dynamism of the branch manager in recruiting clients in these new villages. It is important to distinguish whether any heterogeneity appears to be driven mainly by household level covariates or by village level fixed effect for several reasons. First, if the household level covariates account for a significant part of the heterogeneity, we can relate it to household-level decision. Second, the original empirical strategy of Crépon et al. (2015) to estimate any spillover effect on non borrowing households was to identify a set of households that, based on original covariates, was unlikely to borrow, and then to estimate heterogenous effect based on this predicted probability to borrow. This is by definition a *within village strategy* and will only be robust if the heterogeneity in loan take up is related to baseline covariates.

In order to quantify the relative importance of the two set of covariates, we look at their predictive power for heterogeneity in treatment effects. For this purpose, we create an indicator variable which equals one if an individual belongs to the most affected group and zero if she belongs to the least affected group, defined by the quintiles of the CATE proxy $S(Z)$. Then we estimate what fraction of the variation in this variable is due to the baseline household characteristic and village pair fixed effects. In particular, we regress this indicator variable on the village and household-level covariates separately and report the R-squares from these regressions.

Table 6. CLAN of Microfinance Availability

	Elastic Net			Random Forest		
	20% Most (δ_5)	20% Least (δ_1)	Difference ($\delta_5 - \delta_1$)	20% Most (δ_5)	20% Least (δ_1)	Difference ($\delta_5 - \delta_1$)
Amount of Loans						
Head Age	30.5	39.0	-8.4	24.5	38.1	-13.5
	(28.4, 32.6)	(36.8, 40.9)	(-11.3, -5.4)	(22.4, 26.6)	(36.0, 40.2)	(-16.5, -10.4)
	-	-	[0.000]	-	-	[0.000]
Number of Household Members	3.26	4.54	-1.17	2.64	4.47	-1.85
	(2.98, 3.55)	(4.27, 4.83)	(-1.56, -0.79)	(2.36, 2.91)	(4.19, 4.75)	(-2.25, -1.45)
	-	-	[0.000]	-	-	[0.000]
Number of Members over 16	2.37	2.68	-0.29	1.84	2.82	-1.04
	(2.17, 2.57)	(2.48, 2.88)	(-0.57, -0.01)	(1.65, 2.04)	(2.62, 3.02)	(-1.32, -0.75)
	-	-	[0.081]	-	-	[0.000]
Output						
Non-agricultural self-emp.	0.277	0.051	0.228	0.249	0.098	0.150
	(0.247, 0.306)	(0.021, 0.081)	(0.186, 0.269)	(0.217, 0.281)	(0.067, 0.128)	(0.105, 0.195)
	-	-	[0.000]	-	-	[0.000]
Number of Members over 16	2.92	2.33	0.60	2.77	2.27	0.43
	(2.72, 3.12)	(2.13, 2.53)	(0.32, 0.88)	(2.55, 2.98)	(2.06, 2.48)	(0.13, 0.74)
	-	-	[0.000]	-	-	[0.009]
Number of Household Members	4.10	3.74	0.43	3.86	3.49	0.41
	(3.82, 4.37)	(3.46, 4.023)	(0.04, 0.81)	(3.56, 4.17)	(3.19, 3.79)	(-0.02, 0.82)
	-	-	[0.059]	-	-	[0.120]
Profit						
Non-agricultural self-emp.	0.198	0.103	0.086	0.186	0.108	0.074
	(0.169, 0.227)	(0.073, 0.132)	(0.046, 0.127)	(0.156, 0.215)	(0.079, 0.138)	(0.033, 0.115)
	-	-	[0.000]	-	-	[0.001]
Animal Husbandry self-emp.	0.321	0.570	-0.243	0.378	0.483	-0.113
	(0.280, 0.361)	(0.529, 0.610)	(-0.300, -0.186)	(0.336, 0.419)	(0.442, 0.525)	(-0.171, -0.054)
	-	-	[0.000]	-	-	[0.000]
Head Age	34.11	39.99	-6.08	31.83	35.77	-4.20
	(32.06, 36.18)	(37.90, 42.06)	(-9.05, -3.10)	(29.56, 34.14)	(33.52, 37.99)	(-7.29, -1.10)
	-	-	[0.000]	-	-	[0.017]

Notes: Medians over 100 splits. 90% confidence interval in parenthesis.
P-values for the hypothesis that the parameter is equal to zero in brackets.

The results presented in Table 5 suggest that village pair fixed effects have much more predictive power for treatment effect heterogeneity than the baseline household covariates. When we use elastic net to estimate the most/least affected groups, the village pair fixed effects explain close to 100% of the variation in heterogeneity in all outcomes, whereas individual-level covariates explain only between 16-35% of the variation. With the random forest proxy, results are similar but R-squares are slightly lower for both set of covariates. From this analysis, we conclude that village-level covariates explain a significant part of the heterogeneity in treatment effects. Potential explanations for this observation include unobserved manager quality, heterogeneity in spillovers, and general equilibrium effects that occur within a village. While it is not possible to learn what causes heterogeneity from the CLAN, this evidence can still be useful. On a negative level, it suggests that it is not possible to use the heterogeneity in microfinance take up to say much about spillover effects, since any apparent individual-level heterogeneity seems to be a result of overfitting. It also suggest that it is very difficult to predict individually who will take up or benefit from microfinance. On a more positive level, it suggests that more work can be done in identifying village-level driver in the success of microfinance.

We conclude by looking at the average baseline characteristics of the most and least affected groups. This is illustrative, since the previous analysis suggests that they do not have as much predictive power as the village pair fixed effects. Still, they account for some part of the heterogeneity. Furthermore, unlike the village pair fixed effects, they can be interpreted. We focus on three characteristics for each outcome that are most correlated with the heterogeneity score $S(Z)$, after dropping ones with a correlation less than 0.01 in absolute value. For the selected characteristics, Table 6 reports the CLAN for the 20% least and most affected groups defined by the quintiles of the CATE proxy $S(Z)$ as well as the difference between the two. We find that households with young heads, fewer number of households members and fewer adults are more likely to borrow more from the microfinance institution. For output and profits, the main finding is that households with non-agricultural self employment at baseline are much more likely to be in the group with the large impact (for example, 28% of the households in the top quintile of impact for output had a prior non agricultural business, versus 5% in the bottom quintile). This is a very interesting finding, because the majority of studies on microfinance report larger positive effect on outputs and profits for households that already had a non-agricultural business before microfinance. Meager (2017) finds this differential effect to be robust and generalizable across studies. Banerjee et al. (2015a) shows that the long term effects of microfinance are radically different for people who had a prior business and those who did not. It is reassuring that the one individual variable that is robustly discovered to empirically drive heterogeneity in the CLAN is precisely the one that empirical researchers had identified as relevant.

6.2. **Heterogeneity in the Effect of Immunization Incentives.** In the Morocco microcredit example, most of the heterogeneity we detected was not easily interpretable because it was dominated by the village pair fixed effects. We worked out other examples, omitted for brevity, where there was

TABLE 7. Selected Descriptive Statistics of Villages

	All	Treated	Control
Outcome Variables *(Village-Month Level)*			
Number of children who completed the immunization schedule	7.458	9.090	6.640
Baseline Covariates–Demographic Variables *(Village Level)*			
Household financial status (on 1-10 scale)	3.479	3.17	3.627
Fraction Scheduled Caste-Scheduled Tribe (SC/ST)	0.191	0.199	0.188
Fraction Other Backward Caste (OBC)	0.222	0.207	0.23
Fraction Hindu	0.911	0.851	0.939
Fraction Muslim	0.059	0.109	0.035
Fraction Christian	0.001	0.003	0
Fraction Literate	0.797	0.786	0.802
Fraction Single	0.053	0.052	0.053
Fraction Married (living with spouse)	0.517	0.499	0.526
Fraction Married (not living with spouse)	0.003	0.003	0.003
Fraction Divorced or Separated	0.002	0.005	0
Fraction Widow or Widower	0.04	0.037	0.041
Fraction who received Nursery level education or less	0.152	0.154	0.151
Fraction who received Class 4 level education	0.081	0.08	0.082
Fraction who received Class 9 education	0.157	0.162	0.154
Fraction who received Class 12 education	0.246	0.223	0.257
Fraction who received Graduate or Other Diploma level education	0.085	0.078	0.088
Baseline Covariates–Immunization History of Older Cohort *(Village Level)*			
Number of vaccines administered to pregnant mother	2.276	2.211	2.307
Number of vaccines administered to child since birth	4.485	4.398	4.527
Fraction of children who received polio drops	0.999	1	0.999
Number of polio drops administered to child	2.982	2.985	2.98
Fraction of children who received an immunization card	0.913	0.871	0.933
Number of Observations			
Villages	103	25	78
Village-Months	1321	320	1001

"apparent" heterogeneity when splitting the sample by covariates, but we ultimately found out no detectable heterogeneity using the strategy in this paper. In these instances, the naive approach of reporting some *ad hoc* split likely led to spurious findings. This underscores the importance of a systematic approach.

We now discuss an interesting example where we discover heterogeneity associated with baseline village-level variables, which leads to actionable policy recommendations. It is based on an

RCT aimed at increasing demand for vaccines in India (The interventions are described and analyzed in Banerjee et al. (2019a) and Banerjee et al. (2019b)). The experiment was conducted in 2017 in collaboration with the government of the state of Haryana, where the immunization baseline levels were particularly low. The government health system rolled out an e-health platform designed by a research team, in which nurses collected data on which child was given which shot at each immunization camp. The platform was implemented in over 2,000 villages in seven districts, and provides excellent quality administrative data on immunization coverage.[9] Prior to the launch of the interventions, survey data were collected in 912 of those villages using a sample of 15 households with children aged 1-3 per village. The baseline data covers demographic and socio-economic variables as well as immunization history of these children, who were too old to be included in the intervention. In these 912 villages, three different interventions (and their variants) were cross-randomized:

(1) Small incentives for immunization: parents/caregivers receive mobile phone credit upon bringing children for vaccinations.
(2) Social network intervention: information about immunization camps was diffused through key members of a social network.
(3) Reminders: a fraction of parents/caregivers who had come at least one time received phone reminders for pending vaccinations of the children.

For each of these interventions, there were several possible variants: incentives were either low or high, and either flat or increasing with each shot; the key members of the social network were identified to be either information central using the "gossip" methodology developed by Banerjee et al. (Forthcoming), a trusted person, or both; and reminders were sent to either 33% or 66% of the people concerned. Moreover, each of the interventions were cross-cut, generating a large number of cells of possible treatment combination. Banerjee et al. (2019a) use the method developed by Andrews et al. (2019) to identify the most effective policy to increase the number of children completing the full course of immunization at the village level, and estimate its effects. They find that the combination of a information-central seed ("gossip"), the presence of reminders (we pool 33% and 66% reminders cells for simplicity), and increasing incentives (regardless of levels) is the most effective policy. This is also the most expensive package, so the government was interested in prioritizing villages: where should they scale up the full package?

For this illustration, we focus on evaluating the heterogeneity of the effect of the most effective policy. In particular, we compare 25 villages where the policy was implemented (treatment group) with 78 villages that received neither sloped incentives, nor any social network intervention, nor reminder (control group). Our data constitute an approximately balanced monthly panel of the 103 treated and control villages for 12 months (the duration of the intervention). The outcome variable, Y, is the number of children in a given month in a given village that receive the measles shot

[9] Banerjee et al. (2019b) discusses validation data from random checks conducted by independent surveyors.

Table 8. Comparison of ML Methods: Immunization Incentives

	Elastic Net	Boosting	Neural Network	Random Forest
Best BLP (Λ)	55.830	24.860	35.670	15.830
Best GATES ($\bar{\Lambda}$)	7.164	4.634	5.276	3.767

Notes: Medians over 100 splits in half.

Table 9. BLP of Immunization Incentives

Elastic Net		Neural Network	
ATE (β_1)	HET (β_2)	ATE (β_1)	HET (β_2)
3.069	1.085	1.903	0.916
(1.789, 4.303)	(0.872, 1.293)	(0.750, 3.016)	(0.732, 1.111)
[0.000]	[0.000]	[0.003]	[0.000]

Notes: Medians over 100 splits. 90% confidence interval in parenthesis.
P-values for the hypothesis that the parameter is equal to zero in brackets.

(the last vaccine in the sequence, and thus the completion of the course). The treatment variable, D, is an indicator of the household being in a village that receives the policy. The covariates, Z, include 36 baseline village-level characteristics, including religion, caste, financial status, marriage and family status, education, and baseline immunization. The propensity score is constant. Table 7 shows sample averages in the control and treated groups for some of the variables used in the analysis weighted by village population, as the rest of the analysis. Treatment and control villages have similar baseline characteristics (in particular, the immunization status of the older cohort was similar). During the course of the intervention, on average 6.64 children completed the immunization sequence in control villages, and 9.09 did in treatment villages. This is a raw difference of 2.49, or 37% of the baseline mean.[10] The combined treatment was very effective on average.

The implementation details for the heterogeneity analysis are the same as in the microfinance example, with three differences due to the design: we weight village-level estimations by village population, include district–time fixed effects, and cluster standard errors at the village level. Table 8 compares the ML methods based on Stage 1 proxy predictors. We find that elastic net, as in the previous example, outperforms the other methods, but the second best method is neural network, differently from the previous example. Table 9 presents results of the BLP of CATE using the ML proxies. The ATE estimates in column 1 and 3 indicate that the package treatment increases the number of immunized children by 3 based on elastic net and by 2 based on neural network. Reassuringly these estimates are on either side of the raw difference in means (2.49). Focusing on

[10]The baseline survey suggest that about 40% of children aged 1-3 were fully immunized at baseline. These estimate imply that the rates would jump to about 55%.

the HET estimates, we find strong heterogeneity in treatment effects, as indicated by the statistically significant estimates. Moreover, the estimates are close to 1, suggesting that the ML proxies are good predictors of the CATE.

Next, we estimate the GATES by quintiles of the ML proxies. Figure 7 present the estimated GATES coefficients $\gamma_1 - \gamma_5$ along with joint confidence bands and the ATE estimates. In Table 10 we present the result from the hypothesis test that the difference of the ATE for the most and least affected groups is statistically significant. We find that this difference is 20.3 and 15 based on elastic net and neural network methods, respectively, and statistically significant. Given that the ATE estimates in the whole population are between 2 and 3, these results suggest a large and potentially policy-relevant heterogeneity. Importantly, the impact is an increase by at least 12 in the number of fully immunized children in the most affected group, and a *negative* and significant effect in the least affected group. In some context, it looks like the combined package of small incentives, reminder, and persuasion by members of the social network actually put people off immunization.

The government of Haryana was interested in scaling up this program, but faced a budget constraint. Understandably, they might want to carry out the expansion in the villages with the lowest immunization rate. A natural question is whether there exists a trade-off between this desire of equity, and maximizing the effectiveness of the dollars spent on the policy. To answer this question, we can explore what variables are associated with the heterogeneity detected in BLP and GATES via CLAN. Table 11 reports the CLAN estimates for a selected set of covariates and Tables 14–15 in the appendix for the rest of covariates. Regardless of the method used, the estimates of differences in means between most and least affected groups for the number of vaccines to pregnant mother, number of vaccines to kids since birth, and kids receiving immunization card YN are negative and statistically significant. These results suggest that the villages with low levels of pretreatment immunization are the most affected by the incentives. These are in fact the only variables that consistently pop up. Thus, in this instance, the policy that is preferred ex-ante by the government also happens to be the most effective.

While the heterogeneity associated with the baseline immunization rates cannot be causally interpreted (it could always be proxying for other things), it still sheds interesting light on the negative effect we find for the least affected group. In these villages, immunization rates were higher to start with. Perhaps villagers were intrinsically motivated to get immunized. The nudging with small incentives and mild social pressure may have backfired, by crowding out intrinsic motivation without providing a strong enough extrinsic motivation to act as in Gneezy and Rustichini (2000).

TABLE 10. GATES of 20% Most and Least Affected Groups

Elastic Net			Nnet		
20% Most (γ_5)	20% Least (γ_1)	Difference ($\gamma_5 - \gamma_1$)	20% Most (γ_5	20% Least (γ_1)	Difference ($\gamma_5 - \gamma_1$)
12.310	-7.962	20.320	8.718	-6.342	15.040
(8.434, 16.00)	(-12.03, -3.756)	(14.08, 26.35)	(6.379, 11.15)	(-9.069, -3.560)	(11.18, 18.73)
[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]

Notes: Medians over 100 splits. 90% confidence interval in parenthesis.
P-values for the hypothesis that the parameter is equal to zero in brackets.

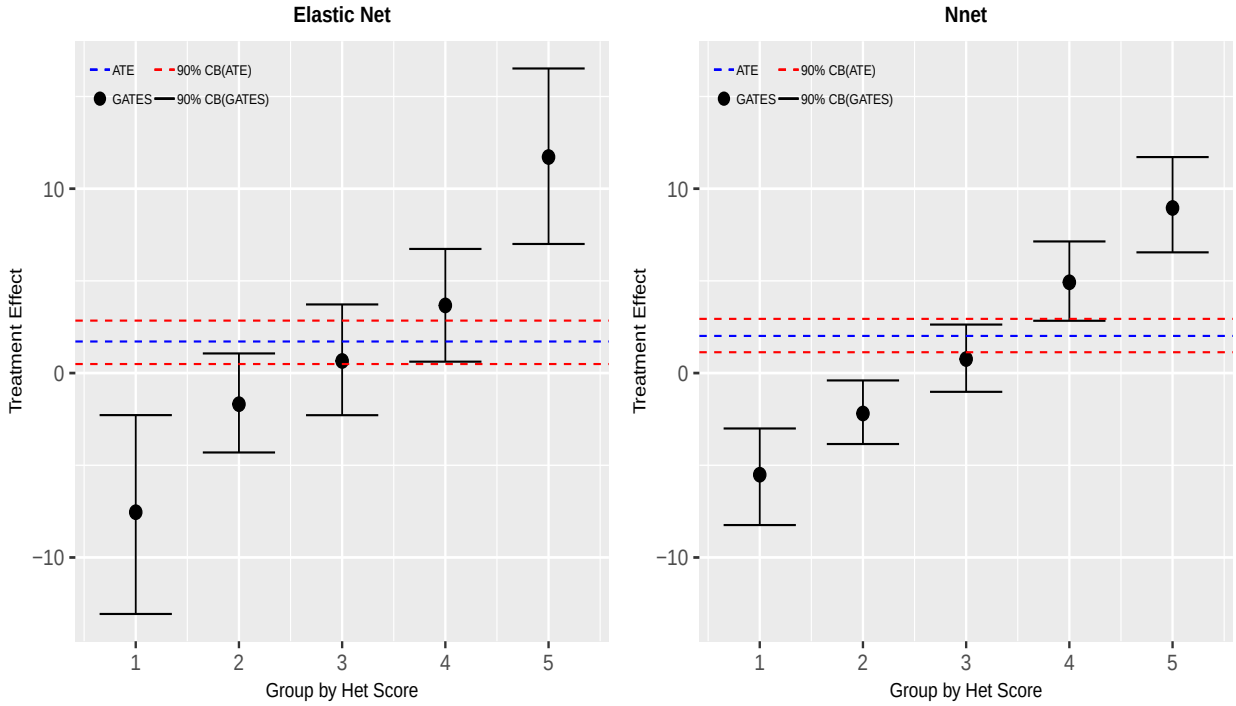

FIGURE 7. GATES of Immunization Incentives. Point estimates and 90% adjusted confidence intervals uniform across groups based on 100 random splits in half

6.3. **Implementation Algorithm.** In this section we describe an algorithm based on the first identification strategy and provide some specific implementation details for the empirical examples.

Algorithm 1 (Inference Algorithm). The inputs are given by the data on units $i \in [N] = \{1, ..., N\}$.

Step 0. Fix the number of splits S and the significance level α, e.g. $S = 100$ and $\alpha = 0.05$.

Step 1. Compute the propensity scores $p(Z_i)$ for $i \in [N]$.

TABLE 11. CLAN of Immunization Incentives

	Elastic Net			Nnet		
	20% Most (δ_5)	20% Least (δ_1)	Difference ($\delta_5 - \delta_1$)	20% Most (δ_5)	20% Least (δ_1)	Difference ($\delta_5 - \delta_1$)
Number of vaccines to pregnant mother	2.199	2.310	-0.102	2.196	2.287	-0.092
	(2.154, 2.247)	(2.264, 2.352)	(-0.166, -0.038)	(2.154, 2.237)	(2.248, 2.326)	(-0.149, -0.035)
	-	-	[0.003]	-	-	[0.003]
Number of vaccines to child since birth	4.111	4.645	-0.513	4.328	4.696	-0.368
	(3.972, 4.251)	(4.524, 4.775)	(-0.698, -0.319)	(4.215, 4.435)	(4.583, 4.813)	(-0.534, -0.215)
	-	-	[0.000]	-	-	[0.000]
Fraction of children received polio drops	0.998	1.000	-0.002	1.000	1.000	0.000
	(0.996, 1.000)	(0.998, 1.002)	(-0.004, 0.001)	(1.000, 1.000)	(1.000, 1.000)	(0.000, 0.000)
	-	-	[0.261]	-	-	[0.000]
Number of polio drops to child	2.945	2.994	-0.049	2.957	3.000	-0.041
	(2.932, 2.959)	(2.983, 3.007)	(-0.067, -0.031)	(2.947, 2.967)	(2.989, 3.008)	(-0.055, -0.027)
	-	-	[0.000]	-	-	[0.000]
Fraction of children received immunized card	0.806	0.926	-0.120	0.906	0.928	-0.028
	(0.776, 0.837)	(0.899, 0.951)	(-0.162, -0.077)	(0.886, 0.922)	(0.910, 0.946)	(-0.053, -0.007)
	-	-	[0.000]	-	-	[0.019]

Notes: Medians over 100 splits. 90% confidence interval in parenthesis.
Notes: P-values for the hypothesis that the parameter is equal to zero in brackets.

Step 2. Consider S splits in half of the indices $i \in \{1, ..., N\}$ into the main sample, M, and the auxiliary sample, A. Over each split $s = 1, .., S$, apply the following steps:

a. Tune and train each ML method separately to learn $B(\cdot)$ and $S(\cdot)$ using A. For each $i \in M$, compute the predicted baseline effect $B(Z_i)$ and predicted treatment effect $S(Z_i)$. If there is zero variation in $B(Z_i)$ and $S(Z_i)$ add Gaussian noise with a variance of 0.1 to the proxies.
b. Estimate the BLP parameters by weighted OLS in M, i.e.,

$$Y_i = \widehat{\alpha}' X_{1i} + \widehat{\beta}_1(D_i - p(Z_i)) + \widehat{\beta}_2(D_i - p(Z_i))(S_i - \mathbb{E}_{N,M} S_i) + \widehat{\epsilon}_i, \quad i \in M$$

such that $\mathbb{E}_{N,M}[w(Z_i)\widehat{\epsilon}_i X_i] = 0$ for $X_i = [X_{1i}', D_i - p(Z_i), (D_i - p(Z_i))(S_i - \mathbb{E}_{N,M} S_i)]'$, where $w(Z_i) = \{p(Z_i)(1 - p(Z_i))\}^{-1}$ and X_{1i} includes a constant, $B(Z_i)$ and $S(Z_i)$.
c. Estimate the GATES parameters by weighted OLS in M, i.e.,

$$Y_i = \widehat{\alpha}' X_{1i} + \sum_{k=1}^{K} \widehat{\gamma}_k \cdot (D_i - p(Z_i)) \cdot 1(S_i \in I_k) + \widehat{\nu}_i, \quad i \in M,$$

such that $\mathbb{E}_{N,M}[w(Z_i)\widehat{\nu}_i W_i] = 0$ for $W_i = [X_{i1}', \{(D_i - p(Z_i))1(S_i \in I_k)\}_{k=1}^{K}]'$, where $w(Z_i) = \{p(Z_i)(1 - p(Z_i))\}^{-1}$, X_{1i} includes a constant, $B(Z_i)$ and $S(Z_i)$, $I_k = [\ell_{k-1}, \ell_k)$, and ℓ_k is the (k/K)-quantile of $\{S_i\}_{i \in M}$.
d. Estimate the CLAN parameters in M by

$$\widehat{\delta}_1 = \mathbb{E}_{N,M}[g(Y_i, Z_i) \mid S_i \in I_1] \quad \text{and} \quad \widehat{\delta}_K = \mathbb{E}_{N,M}[g(Y_i, Z_i) \mid S_i \in I_K],$$

where $I_k = [\ell_{k-1}, \ell_k)$ and ℓ_k is the (k/K)-quantile of $\{S_i\}_{i \in M}$.

e. Compute the two performance measures for the ML methods

$$\widehat{\Lambda} = |\widehat{\beta}_2|^2 \widehat{\mathrm{Var}}(S(Z)) \qquad \widehat{\bar{\Lambda}} = \frac{1}{K}\sum_{k=1}^{K} \widehat{\gamma}_k^2.$$

Step 3: Choose the best ML methods based on the medians of $\widehat{\Lambda}$ and $\widehat{\bar{\Lambda}}$ over the splits.

Step 4: Compute the estimates, $(1-\alpha)$-level conditional confidence intervals and conditional p-values for all the parameters of interest. Monotonize the confidence intervals if needed. For example, construct a $(1-\alpha)$ joint confidence interval for the GATES as

$$\{\widehat{\gamma}_k \pm \widehat{c}(1-\alpha)\widehat{\sigma}_k, \;\; k = 1, \ldots, K\}, \tag{6.1}$$

where $\widehat{c}(1-\alpha)$ is a consistent estimator of the $(1-\alpha)$-quantile of $\max_{k\in 1,\ldots,K} |\widehat{\gamma}_k - \gamma_k|/\widehat{\sigma}_k$ and $\widehat{\sigma}_k$ is the standard error of $\widehat{\gamma}_k$ conditional on the data split. Monotonize the band (6.1) with respect to k using the rearrangement method of Chernozhukov et al. (2009).

Step 5: Compute the adjusted $(1-2\alpha)$-confidence intervals and adjusted p-values using the VEIN methods described in Section 3.

Comment 6.1 (ML Methods)**.** We consider four ML methods to estimate the proxy predictors: elastic net, boosted trees, neural network with feature extraction, and random forest. The ML methods are implemented in R using the package caret (Kuhn, 2008). The names of the elastic net, boosted tree, neural network with feature extraction, and random forest methods in caret are glmnet, gbm, pcaNNet and rf, respectively. For each split of the data, we choose the tuning parameters separately for $B(z)$ and $S(z)$ based on mean squared error estimates of repeated 2-fold cross-validation, except for random forest, for which we use the default tuning parameters to reduce the computational time.[11] In tuning and training the ML methods we use only the auxiliary sample. In all the methods we rescale the outcomes and covariates to be between 0 and 1 before training.

Comment 6.2 (Microfinance Application)**.** We adopt two strategies to improve precision, and to adapt our strategy to the experimental design. First, since the stratification was conducted within pairs, the linear projections of the BLP and GATES control for village pair fixed effects along with the predicted baseline effect, $B(z)$ and predicted treatment effect, $S(z)$. Second, as suggested in Section 4, we use stratified sample splitting where the strata are village pairs. We cluster the standard errors at the village level to account for potential correlated shocks within each village. All reported results are medians over $S = 100$ splits and $\alpha = 0.05$.

[11]We have the following tuning parameters for each method: Elastic Net: alpha (Mixing Percentage), lambda (Regularization Parameter), Boosted trees: n.trees (Number of Boosting Iterations), interaction.depth (Max Tree Depth), shrinkage (Shrinkage), n.minobsinnode (Min. Terminal Node Size), size (Number of Hidden Units) , decay (Weight Decay), mtry (Number of Randomly Selected Predictors).

7. Concluding Remarks

We propose to focus inference on key features of heterogeneous effects in randomized experiments, and develop the corresponding methods. These key features include best linear predictors of the effects and average effects sorted by groups, as well as average characteristics of most and least affected units. Our new approach is valid in high dimensional settings, where the effects are estimated by machine learning methods. The main advantage of our approach is its credibility: the approach is agnostic about the properties of the machine learning estimators, and does not rely on incredible or hard-to-verify assumptions. Estimation and inference relies on data splitting, where the latter allows us to avoid overfitting and all kinds of non-regularities. Our inference quantifies uncertainty coming from both parameter estimation and the data splitting, and could be of independent interest. Two empirical applications illustrate the practical uses of the approach.

A researcher might be concerned about the application of our method to detect heterogeneity due to the possible power loss induced by sample splitting. We argue that this power loss is the price to pay when the researcher is not certain or willing to fully specify the form of the heterogeneity prior to conducting the experiment. Thus, if the researcher has a well-defined pre-analysis plan that spells out a small number of heterogeneity groups in advance, then there is no need of splitting the sample.[12] However, this situation is not common. In general, the researcher might not be able to fully specify the form of the heterogeneity due to lack of information, economic theory, or willingness to take a stand at the early stages of the analysis. She might also face data limitations that preclude the availability of the desired covariates. Here we recommend the use of our method to avoid overfitting and p-hacking, and impose discipline to the heterogeneity analysis at the cost of some power loss due to sample splitting. This loss is difficult to quantify as we are not aware of any alternative method that works at the same level of agnosticism as ours. In Appendix D we provide a numerical example using a simple parametric model where standard methods are available. We find that the extent of the power loss for not using the parametric form of the heterogeneity roughly corresponds to reducing the sample size by half in a test for the presence of heterogeneity, although the exact comparison depends on features of the data generating process.

Appendix A. Proofs

Proof of Theorem 2.1. The subset of the normal equations, which correspond to $\beta := (\beta_1, \beta_2)'$, are given by $\mathrm{E}[w(Z)(Y - \alpha' X_1 - \beta' X_2)X_2] = 0$. Substituting $Y = b_0(Z) + s_0(Z)D + U$, and using the definition $X_2 = X_2(Z, D) = [D - p(Z), (D - p(Z)(S - \mathrm{E}S)]'$, $X_1 = X_1(Z)$, and the law of iterated

[12]More generally, the plan needs to specify a parametric form for the heterogeneity as a low dimensional function of prespecified covariates (e.g., Chernozhukov et al., 2015). In this case, we still recommend the use of ML tools to efficiently estimate the CATEs in the presence of control variables (Belloni et al., 2017; Chernozhukov et al., 2017).

expectations, we notice that:

$$\begin{aligned}
&\mathrm{E}[w(Z)b_0(Z)X_2] = \mathrm{E}[w(Z)b_0(Z)\underbrace{\mathrm{E}[X_2 \mid Z]}_{=0}] = 0,\\
&\mathrm{E}[w(Z)UX_2] = \mathrm{E}[w(Z)\underbrace{\mathrm{E}[U \mid Z,D]}_{0}X_2(Z,D)] = 0,\\
&\mathrm{E}[w(Z)X_1X_2] = \mathrm{E}[w(Z)X_1(Z)\underbrace{\mathrm{E}[X_2(Z,D) \mid Z]}_{=0}] = 0.
\end{aligned}$$

Hence the normal equations simplify to: $\mathrm{E}[w(Z)(s_0(Z)D - \beta' X_2)X_2] = 0$. Since

$$\mathrm{E}[\{D - p(Z)\}\{D - p(Z)\} \mid Z] = p(Z)(1 - p(Z)) = w^{-1}(Z),$$

and $S = S(Z)$, the components of X_2 are orthogonal by the law of iterated expectations:

$$\mathrm{E}w(Z)(D - p(Z))(D - p(Z))(S - \mathrm{E}S) = \mathrm{E}(S - \mathrm{E}S) = 0.$$

Hence the normal equations above further simplify to

$$\begin{aligned}
&\mathrm{E}[w(Z)\{s_0(Z)D - \beta_1(D - p(Z))\}(D - p(Z))] = 0,\\
&\mathrm{E}[w(Z)\{s_0(Z)D - \beta_2(D - p(Z))(S - \mathrm{E}S)\}(D - p(Z))(S - \mathrm{E}S)] = 0.
\end{aligned}$$

Solving these equations and using the law of iterated expectations, we obtain

$$\beta_1 = \frac{\mathrm{E}w(Z)\{s_0(Z)D(D - p(Z))\}}{\mathrm{E}w(Z)(D - p(Z))^2} = \frac{\mathrm{E}w(Z)s_0(Z)w^{-1}(Z)}{\mathrm{E}w(Z)w^{-1}(Z)} = \mathrm{E}s_0(Z),$$

$$\begin{aligned}
\beta_2 &= \frac{\mathrm{E}w(Z)\{s_0(Z)D(D - p(Z))(S - \mathrm{E}S)\}}{\mathrm{E}w(Z)(D - p(Z))^2(S - \mathrm{E}S)^2}\\
&= \frac{\mathrm{E}w(Z)s_0(Z)w^{-1}(Z)(S - \mathrm{E}S)}{\mathrm{E}w(Z)w^{-1}(Z)(S - \mathrm{E}S)^2} = \mathrm{Cov}(s_0(Z), S)/\mathrm{Var}(S).
\end{aligned}$$

The conclusion follows by noting that these coefficients also solve the normal equations

$$\mathrm{E}\{[s_0(Z) - \beta_1 - \beta_2(S - \mathrm{E}S)][1, (S - \mathrm{E}S)]'\} = 0,$$

which characterize the optimum in the problem of best linear approximation/prediction of $s_0(Z)$ using S. ■

Proof of Theorem 2.2. The normal equations defining $\beta = (\beta_1, \beta_2)'$ are given by $\mathrm{E}[(YH - \mu' X_1 H - \beta' \tilde{X}_2)\tilde{X}_2] = 0$. Substituting $Y = b_0(Z) + s_0(Z)D + U$, and using the definition $\tilde{X}_2 = \tilde{X}_2(Z) = [1, (S(Z) - \mathrm{E}S(Z))]'$, $X_1 = X_1(Z)$, and the law of iterated expectations, we notice that:

$$\begin{aligned}
&\mathrm{E}[b_0(Z)H\tilde{X}_2(Z)] = \mathrm{E}[b_0(Z)\underbrace{\mathrm{E}[H \mid Z]}_{=0}\tilde{X}_2(Z)] = 0,\\
&\mathrm{E}[UH\tilde{X}_2(Z)] = \mathrm{E}[\underbrace{\mathrm{E}[U \mid Z,D]}_{0}H(D,Z)\tilde{X}_2(Z)] = 0,\\
&\mathrm{E}[X_1(Z)H\tilde{X}_2(Z)] = \mathrm{E}[X_1(Z)\underbrace{\mathrm{E}[H \mid Z]}_{=0}\tilde{X}_2(Z)] = 0.
\end{aligned}$$

Hence the normal equations simplify to:

$$\mathrm{E}[(s_0(Z)DH - \beta' \tilde{X}_2)\tilde{X}_2] = 0.$$

Since 1 and $S - \mathrm{E}S$ are orthogonal, the normal equations above further simplify to

$$\begin{aligned}&\mathrm{E}\{s_0(Z)DH - \beta_1\} = 0,\\&\mathrm{E}[\{s_0(Z)DH - \beta_2(S - \mathrm{E}S)\}(S - \mathrm{E}S)] = 0.\end{aligned}$$

Using that

$$\mathrm{E}[DH \mid Z] = [p(Z)(1 - p(Z))]/[p(Z)(1 - p(Z))] = 1,$$

$S = S(Z)$, and the law of iterated expectations, the equations simplify to

$$\begin{aligned}&\mathrm{E}\{s_0(Z) - \beta_1\} = 0,\\&\mathrm{E}\{s_0(Z) - \beta_2(S - \mathrm{E}S)\}(S - \mathrm{E}S) = 0.\end{aligned}$$

These are normal equations that characterize the optimum in the problem of best linear approximation/prediction of $s_0(Z)$ using S. Solving these equations gives the expressions for β_1 and β_2 stated in the theorem. ■

Proof of Theorem 2.3. The proof is similar to the proof of Theorem 2.1- 2.2. Moreover, since the proofs for the two strategies are similar, we will only demonstrate the proof for the second strategy.

The subset of the normal equations, which correspond to $\gamma := (\gamma_k)_{k=1}^K$, are given by $\mathrm{E}[(YH - \mu'\tilde{W}_1 - \gamma'\tilde{W}_2)\tilde{W}_2] = 0$. Substituting $Y = b_0(Z) + s_0(Z)D + U$, and using the definition $\tilde{W}_2 = \tilde{W}_2(Z) = [1(S \in I_k)_{k=1}^K]'$, $\tilde{W}_1 = X_1(Z)H$, and the law of iterated expectations, we notice that:

$$\begin{aligned}&\mathrm{E}[b_0(Z)H\tilde{W}_2(Z)] = \mathrm{E}[b_0(Z)\underbrace{\mathrm{E}[H \mid Z]}_{=0}\tilde{W}_2(Z)] = 0,\\&\mathrm{E}[UH\tilde{W}_2(Z)] = \mathrm{E}[\underbrace{\mathrm{E}[U \mid Z, D]}_{0}H(D, Z)\tilde{W}_2(Z)] = 0,\\&\mathrm{E}[\tilde{W}_1\tilde{W}_2(Z)] = \mathrm{E}[X_1(Z)\underbrace{\mathrm{E}[H \mid Z]}_{=0}\tilde{W}_2(Z)] = 0.\end{aligned}$$

Hence the normal equations simplify to:

$$\mathrm{E}[\{s_0(Z)DH - \gamma'\tilde{W}_2\}\tilde{W}_2] = 0.$$

Since components of $\tilde{W}_2 = \tilde{W}_2(Z) = [1(G_k)_{k=1}^K]'$ are orthogonal, the normal equations above further simplify to

$$\mathrm{E}[\{s_0(Z)DH - \gamma_k 1(G_k)\}1(G_k)] = 0.$$

Using that

$$\mathrm{E}[DH \mid Z] = [p(Z)\{1 - p(Z)\}]/[p(Z)\{1 - p(Z)\}] = 1,$$

$S = S(Z)$, and the law of iterated expectations, the equations simplify to

$$\mathrm{E}[\{s_0(Z) - \gamma_k 1(G_k)\}1(G_k)] = 0 \iff \gamma_k = \mathrm{E}s_0(Z)1(G_k)/\mathrm{E}[1(G_k)] = \mathrm{E}[s_0(Z) \mid G_k].$$

■

Proof of Theorem 3.1. We have that $p_{.5} \leqslant \alpha/2$ is equivalent to $\mathbb{E}_P[1(p_A \leqslant \alpha/2) \mid \text{Data}] \geqslant 1/2$. So

$$\mathbb{P}_P[p_{.5} \leqslant \alpha/2] = \mathbb{E}_P 1\{\mathbb{E}_P[1(p_A \leqslant \alpha/2) \mid \text{Data}] \geqslant 1/2\}.$$

By Markov inequality,

$$\mathbb{E}_P 1\{\mathbb{E}_P[1(p_A \leqslant \alpha/2) \mid \text{Data}] \geqslant 1/2\} \leqslant 2\mathbb{P}_P[p_A \leqslant \alpha/2].$$

Moreover,

$$\mathbb{P}_P(p_A \leqslant \alpha/2) \leqslant \mathbb{E}_P[\mathbb{P}_P[p_A \leqslant \alpha/2 \mid \text{Data}_A \in \mathcal{A}] + \gamma] \leqslant \alpha/2 + \delta + \gamma.$$

■

Proof of Theorem 3.2. To show the second claim, we note that

$$\begin{aligned}\mathbb{P}_P(\theta_A \notin \text{CI}) &= \mathbb{P}_P(p_l(\theta_A) \leqslant \alpha/2) + \mathbb{P}_P(p_u(\theta_A) \leqslant \alpha/2) \\ &\leqslant \alpha + \delta + \gamma + \alpha + \delta + \gamma,\end{aligned}$$

where the inequality holds by Theorem 3.1 on the p-values. The last bound is upper bounded by $2\alpha + o(1)$ by the regularity condition PV for the p-values, uniformly in $P \in \mathcal{P}$.

To show the first claim, we need to show the following inequalities:

$$\sup\{\theta \in \mathbb{R} : p_u(\theta) > \alpha/2\} \leqslant u, \quad \inf\{\theta \in \mathbb{R} : p_l(\theta) > \alpha/2\} \geqslant l.$$

We demonstrate the first inequality, and the second follows similarly.

We have that

$$\begin{aligned}\{\theta \in \mathbb{R} : p_u(\theta) > \alpha/2\} &= \{\theta \in \mathbb{R} : \underline{\text{Med}}[\Phi\{\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta)\} \mid \text{Data}] > \alpha/2\} \\ &= \{\theta \in \mathbb{R} : \Phi\{\underline{\text{Med}}[\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta) \mid \text{Data}]\} > \alpha/2\} \\ &= \{\theta \in \mathbb{R} : \underline{\text{Med}}[\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta) \mid \text{Data}] > \Phi^{-1}(\alpha/2)\} \\ &= \{\theta \in \mathbb{R} : \overline{\text{Med}}[\widehat{\sigma}_A^{-1}(\theta - \widehat{\theta}_A) \mid \text{Data}] < \Phi^{-1}(1-\alpha/2)\} \\ &= \left\{\theta \in \mathbb{R} : \overline{\text{Med}}\left[\frac{\theta - \widehat{\theta}_A}{\widehat{\sigma}_A} - \Phi^{-1}(1-\alpha/2) \mid \text{Data}\right] < 0\right\},\end{aligned}$$

where we have used the equivariance of $\overline{\text{Med}}$ and $\underline{\text{Med}}$ to monotone transformations, implied from their definition. We claim that by the definition of

$$u := \underline{\text{Med}}[\widehat{\theta}_A + \widehat{\sigma}_A \Phi^{-1}(1-\alpha/2) \mid \text{Data}],$$

we have

$$\overline{\text{Med}}\left[\frac{u - \widehat{\theta}_A}{\widehat{\sigma}_A} - \Phi^{-1}(1-\alpha/2) \mid \text{Data}\right] \geqslant 0.$$

Indeed, by the definition of u,

$$\mathbb{E}\left(1(u - \widehat{\theta}_A - \widehat{\sigma}_A \Phi^{-1}(1-\alpha/2) \geqslant 0) \mid \text{Data}\right) \geqslant 1/2.$$

Since $\widehat{\sigma}_A > 0$ by assumption,

$$1(u - \widehat{\theta}_A - \widehat{\sigma}_A \Phi^{-1}(1-\alpha/2) \geqslant 0) = 1\left(\frac{u - \widehat{\theta}_A}{\widehat{\sigma}_A} - \Phi^{-1}(1-\alpha/2) \geqslant 0\right),$$

and it follows that

$$\mathbb{P}\Big(\frac{u - \widehat{\theta}_A}{\widehat{\sigma}_A} - \Phi^{-1}(1-\alpha/2) \geqslant 0 \mid \text{Data}\Big) \geqslant 1/2.$$

The claimed inequality $\sup\{\theta \in \mathbb{R} : p_u(\theta) > \alpha/2\} \leqslant u$ follows. ■

Appendix B. A Lemma on Uniform in P Conditional Inference

Lemma B.1. *Fix two positive constants c and C, and a small constant $\delta > 0$. Let $\tilde{Y}$ and X denote a generic outcome and a generic d-vector of regressors, whose use and definition may differ in different places of the paper. Assume that for each $P \in \mathcal{P}$, $\mathrm{E}_P|\tilde{Y}|^{4+\delta} < C$ and let $0 < \underline{w} \leqslant w(Z) \leqslant \overline{w} < \infty$ denote a generic weight, and that $\{(\tilde{Y}_i, Z_i, D_i)\}_{i=1}^N$ are i.i.d. copies of $(\tilde{Y}, Z, D)$. Let $\{\text{Data}_A \in \mathcal{A}_N\}$ be the event such that the ML algorithm, operating only on Data_A, produces a vector $X_A = X(Z, D; Data_A)$ that obeys, for $\epsilon_A = \tilde{Y} - X'\beta_A$ defined by: $\mathbb{E}_P[\epsilon_A w(Z) X_A \mid \text{Data}_A] = 0$, the following inequalities, uniformly in $P \in \mathcal{P}$*

$$\mathbb{E}_P[\|X_A\|^{4+\delta} \mid \text{Data}_A] \leqslant C, \;\; \textit{mineig}\; \mathbb{E}_P[X_A X_A' \mid \text{Data}_A] > c, \;\; \textit{mineig}\; \mathbb{E}_P[\epsilon_A^2 X_A X_A' \mid \text{Data}_A] > c.$$

Suppose that $\mathbb{P}_P\{\text{Data}_A \in \mathcal{A}_N\} \geqslant 1 - \gamma \to 1$ uniformly in $P \in \mathcal{P}$, as $N \to \infty$. Let $\widehat{\beta}_A$ be defined by:

$$\mathbb{E}_{N,M}[w(Z) X_A \widehat{\epsilon}_A] = 0, \quad \widehat{\epsilon}_A = Y_A - X'\widehat{\beta}_A.$$

Let $\widehat{V}_{N,A} := (\mathbb{E}_{N,M} X_A X_A')^{-1} \mathbb{E}_{N,M} \widehat{\epsilon}_A^2 X_A X_A' (\mathbb{E}_{N,M} X_A X_A')^{-1}$ be an estimator of

$$V_{N,A} = (\mathbb{E}_P[X_A X_A' \mid \text{Data}_A])^{-1} \mathbb{E}_P[\epsilon_A^2 X_A X_A' \mid \text{Data}_A] (\mathbb{E}_P[X_A X_A' \mid \text{Data}_A])^{-1}.$$

Let I_d denote the identify matrix of order d. Then for any convex set R in $\mathbb{R}^d$, we have that uniformly in $P \in \mathcal{P}$:

$$\mathbb{P}_P[\widehat{V}_{N,A}^{-1/2}(\widehat{\beta}_A - \beta_A) \in R \mid \text{Data}_A] \to_P \mathbb{P}(N(0, I_d) \in R),$$
$$\mathbb{P}_P[\widehat{V}_{N,A}^{-1/2}(\widehat{\beta}_A - \beta_A) \in R \mid \{\text{Data}_A \in \mathcal{A}_N\}] \to \mathbb{P}(N(0, I_d) \in R),$$

and the same results hold with $\widehat{V}_{N,A}$ replaced by $V_{N,A}$.

Proof. It suffices to demonstrate the argument for an arbitrary sequence $\{P_n\}$ in $\mathcal{P}$. Let $z \mapsto \tilde{X}_{A,N}(z)$ be a deterministic map such that the following inequalities hold, for $\tilde{e}_A$ defined by

$$\mathrm{E}_{P_n}[\tilde{e}_A w(Z) \tilde{X}_{A,N}(Z)] = 0$$

and $\tilde{X}_{A,N} = \tilde{X}_{A,N}(Z)$:

$$\mathrm{E}_{P_n}[\|\tilde{X}_{A,N}\|^4] < C, \;\; \text{mineig}\; \mathrm{E}_{P_n}[\tilde{X}_{A,N} \tilde{X}_{A,N}'] > c, \;\; \text{mineig}\; \mathrm{E}_{P_n}[\tilde{e}_A^2 \tilde{X}_{A,N} \tilde{X}_{A,N}'] > c.$$

Then we have that (abusing notation):

$$B_N := \sup_{\tilde{X}_{A,N}} \sup_{h \in \mathrm{BL}_1(\mathbb{R}^d)} |\mathbb{E}_{P_n} h(\tilde{V}_{N,A}^{-1/2}(\widehat{\beta}_A - \beta_A) \mid \tilde{X}_{A,N}) - \mathbb{E} h(N(0, I_d))| \to 0,$$

by the standard argument for asymptotic normality of the least squares estimator, which utilizes the Lindeberg-Feller Central Limit Theorem. Here

$$\tilde{V}_{N,A} := (\mathrm{E}\tilde{X}_A\tilde{X}_A')^{-1}\mathrm{E}\tilde{\epsilon}_A^2\tilde{X}_A\tilde{X}_A'(\mathrm{E}\tilde{X}_A\tilde{X}_A')^{-1},$$

and $\mathrm{BL}_1(\mathbb{R}^d)$ denotes the set of Lipschitz maps $h : \mathbb{R}^d \to [0,1]$ with the Lipschitz coefficient bounded by 1.

Then, for the stochastic sequence $X_{A,N} = X_{A,N}(\mathrm{Data}_A)$,

$$\sup_{h\in \mathrm{BL}_1(\mathbb{R}^d)} |\mathbb{E}_{P_n}[h(V_{N,A}^{-1/2}(\widehat{\beta}_A - \beta_A)) \mid X_{A,N}] - \mathbb{E}[h(N(0,I_d))]| \leqslant B_N + 2(1 - 1\{\mathrm{Data}_A \in \mathcal{A}_N\}) \to_{P_n} 0.$$

Since under the stated bounds on moments, $\widehat{V}_{N,A}^{1/2}V_{N,A}^{-1/2} \to_{P_n} I_d$ by the standard argument for consistency of the Eicker-Huber-White sandwich, we further notice that

$$\sup_{h\in \mathrm{BL}_1(\mathbb{R}^d)} |\mathbb{E}_{P_n}[h(\widehat{V}_{N,A}^{-1/2}(\widehat{\beta}_A - \beta_A)) \mid X_{A,N}] - \mathbb{E}_{P_n}[h(V_{N,A}^{-1/2}(\widehat{\beta}_A - \beta_A)) \mid X_{A,N}]|$$

$$\leqslant \mathbb{E}_{P_n}[\|\widehat{V}_{N,A}^{-1/2}V_{N,A}^{1/2} - I_d\| \wedge 1 \cdot \|V_{N,A}^{-1/2}(\widehat{\beta}_A - \beta_A)\| \wedge 1 \mid X_{A,N}] \to_{P_n} \mathbb{E}[0 \wedge 1 \cdot \|N(0,I_d)\| \wedge 1] = 0,$$

in order to conclude that

$$\sup_{h\in \mathrm{BL}_1(\mathbb{R}^d)} \mathbb{E}_{P_n}[h(\widehat{V}_{N,A}^{-1/2}(\widehat{\beta}_A - \beta_A)) \mid X_{A,N}] - \mathbb{E}[h(N(0,I_d))] \to_P 0.$$

Moreover, since $\mathbb{E}_{P_n}[h(\widehat{V}_{N,A}^{-1/2}(\widehat{\beta}_A - \beta_A)) \mid X_{A,N}] = \mathbb{E}_{P_n}[h(\widehat{V}_{N,A}^{-1/2}(\widehat{\beta}_A - \beta_A)) \mid \mathrm{Data}_A]$, the first conclusion follows: $\mathbb{P}_{P_n}[\widehat{V}_{N,A}^{-1/2}(\widehat{\beta}_A - \beta_A) \in R \mid \mathrm{Data}_A] \to_{P_n} \mathbb{P}(N(0,I_d) \in R)$, by the conventional smoothing argument (where we approximate the indicator of a convex region by a smooth map with finite Lipschitz coefficient). The second conclusion

$$\mathbb{P}_{P_n}[\widehat{V}_{N,A}^{-1/2}(\widehat{\beta}_A - \beta_A) \in R \mid \mathrm{Data}_A \in \mathcal{A}_N] \to \mathbb{P}(N(0,I_d) \in R)$$

follows from the first by

$$\mathbb{P}_{P_n}[\widehat{V}_{N,A}^{-1/2}(\widehat{\beta}_A - \beta_A) \in R \mid \mathrm{Data}_A \in \mathcal{A}_N] =$$

$$= \mathbb{E}_{P_n}[\mathbb{P}_{P_n}[\widehat{V}_{N,A}^{-1/2}(\widehat{\beta}_A - \beta_A) \in R \mid \mathrm{Data}_A]1(\{\mathrm{Data}_A \in \mathcal{A}_N\})/\mathbb{P}_{P_n}\{\mathrm{Data}_A \in \mathcal{A}_N\}]$$

$$\to \mathrm{E}[\mathbb{P}(N(0,I_d) \in R) \cdot 1],$$

using the definition of the weak convergence, implied by the convergence to the constants in probability. ■

Appendix C. Comparison of Two Estimation Strategies

We focus on the estimation of the BLP. The analysis can be extended to the GATES using analogous arguments.

Let $X_{2i} = (1, S_i - \mathbb{E}_{N,M} S_i)'$ and $\widehat{\beta} = (\widehat{\beta}_1, \widehat{\beta}_2)'$. In the first strategy, we run the weighted linear regression

$$Y_i = X_{1i}'\widehat{\alpha} + (D_i - p(Z_i))X_{2i}'\widehat{\beta} + \widehat{\epsilon}_i, \quad i \in M,$$

$$\mathbb{E}_{N,M}[w(Z_i)\widehat{\epsilon}_i X_i] = 0, \;\; w(Z) = \{p(Z)(1-p(Z))\}^{-1}, \;\; X_i = [X_{1i}', (D_i - p(Z_i))X_{2i}']'.$$

Let $\widehat{\theta} := (\widehat{\alpha}', \widehat{\beta}')'$. Then, this estimator is

$$\widehat{\theta} = \left(\mathbb{E}_{N,M}[w(Z_i)X_i X_i']\right)^{-1} \mathbb{E}_{N,M}[w(Z_i)X_i Y_i].$$

Let $X = [X_1', (D - p(Z))X_2']'$ with $X_2 = (1, S - \mathrm{E}S)'$. By standard properties of the least squares estimator and the central limit theorem

$$\widehat{\theta} = \left(\mathrm{E}[w(Z)XX']\right)^{-1} \mathbb{E}_{N,M}[w(Z_i)X_i Y_i] + o_P(M^{-1/2}),$$

where

$$\mathrm{E}[w(Z)XX'] = \begin{pmatrix} \mathrm{E}w(Z)X_1X_1' & 0 \\ 0 & \mathrm{E}X_2X_2' \end{pmatrix}.$$

In the previous expression we use that $\mathrm{E}w(Z)(D-p(Z))X_1X_2' = 0$ and $\mathrm{E}w(Z)(D-p(Z))^2X_2X_2' = \mathrm{E}X_2X_2'$ by iterated expectations. Then,

$$\widehat{\beta} = \left(\mathrm{E}X_2X_2'\right)^{-1} \mathbb{E}_{N,M}[w(Z_i)(D_i - p(Z_i))X_{2i}Y_i] + o_P(M^{-1/2}),$$

using that $\mathrm{E}[w(Z)XX']$ is block-diagonal between $\widehat{\alpha}$ and $\widehat{\beta}$.

In the second strategy, we run the linear regression

$$H_iY_i = H_iX_{1i}'\tilde{\alpha} + X_{2i}'\tilde{\beta} + \tilde{\epsilon}_i, \quad \mathbb{E}_{N,M}\tilde{\epsilon}_i\tilde{X}_i = 0, \quad H_i = (D_i - p(Z_i))w(Z_i), \quad \tilde{X}_i = [H_iX_{1i}', X_{2i}]',$$

which yields the estimator, for $\tilde{\theta} = (\tilde{\alpha}', \tilde{\beta}')'$,

$$\tilde{\beta} = \left(\mathbb{E}_{N,M}[\tilde{X}_i\tilde{X}_i']\right)^{-1} \mathbb{E}_{N,M}[H_i\tilde{X}_iY_i].$$

Let $\tilde{X} = [HX_1', X_2']'$ with $X_2 = (1, S - \mathrm{E}S)'$. By standard properties of the least squares estimator and the central limit theorem

$$\tilde{\theta} = \left(\mathrm{E}[\tilde{X}\tilde{X}']\right)^{-1} \mathbb{E}_{N,M}[H_i\tilde{X}_iY_i] + o_P(M^{-1/2}),$$

where

$$\mathrm{E}[\tilde{X}\tilde{X}'] = \begin{pmatrix} \mathrm{E}w(Z)X_1X_1' & 0 \\ 0 & \mathrm{E}X_2X_2' \end{pmatrix} = \mathrm{E}[w(Z)XX'].$$

In the previous expression we use that $\mathrm{E}HX_1X_2' = 0$ and $\mathrm{E}H^2X_1X_1' = \mathrm{E}w(Z)X_1X_1'$ by iterated expectations. Hence,

$$\tilde{\beta} = \left(\mathrm{E}[X_2X_2']\right)^{-1} \mathbb{E}_{N,M}[w(Z_i)(D_i - p(Z_i))X_{2i}Y_i] + o_P(M^{-1/2}),$$

where we use that $\mathrm{E}[\tilde{X}\tilde{X}']$ is block-diagonal between $\widehat{\alpha}$ and $\widehat{\beta}$, and $\mathbb{E}_{N,M}[H_i X_{2i} Y_i] = \mathbb{E}_{N,M}[w(Z_i)(D_i - p(Z_i))X_{2i}Y_i]$.

We conclude that $\widehat{\beta}$ and $\tilde{\beta}$ have the same asymptotic distribution because they have the same first order representation.

Appendix D. Power Calculations

We conduct a numerical simulation to compare the power of the proposed method with the available standard methods to detect heterogeneity. The comparison is complicated because the existing methods do not apply to the general class of models that we consider. We therefore focus on a parametric low dimensional setting for which there are standard methods available. The design is a linear interactive model:

$$Y = \alpha_0 + \alpha_1 Z + \alpha_2 D + \beta Z D + \sigma\varepsilon, \tag{D.1}$$

where Z is standard normal, D is Bernouilli with probability 0.5, ε is standard normal, $\alpha_0 = \alpha_1 = 0$, and $\sigma = 1$. The parameter β determines whether there is heterogeneity in the CATE, $s_0(Z) = \alpha_2 + \beta Z$. We vary its value across the simulations from no heterogeneity $\beta = 0$ to increasing levels of heterogeneity $\beta \in \{.1, .2, .3, .4, .6, .8\}$. The benchmark of comparison is a t-test of $\beta = 0$ based on the least squares estimator with heteroskedasticity-robust standard errors in (D.1) using the entire sample.[13] We implement our test that the BLP is equal to zero using sample splitting. In the first stage we estimate the proxies of the CATE by least squares in the linear interactive model (D.1) using half of the sample. In the second stage we run the adjusted linear regression of strategy 1 using the other half of the sample. We repeat the procedure for 100 splits and use the median p-value multiplied by 2 to carry out the test. The nominal level of the test for both the standard and proposed method is 5%. We consider several sample sizes, $n \in \{100, 200, 300, 400, 600, 800\}$, to study how the power scales with n. All the results are based on $5,000$ replications.

Tables 12 and 13 report the empirical size and power for the standard and proposed test, respectively. One might conjecture that the standard test is as powerful as the proposed test with double the sample size due to sample splitting. The results roughly agree with this conjecture, but the power comparison depends nonlinearly on the heterogeneity coefficient β. Thus, the standard test is more powerful than the proposed test with double the sample size for low values of β, but the proposed test is more powerful than the standard test with half of the sample size for high values of β. We also note that the proposed test is conservative in this design.

[13] Note that this method is only applicable when researcher is willing to specify a parametric model for the expectation of Y conditional on D and Z.

TABLE 12. Empirical Size and Power of Standard Test by Sample Size

	β=0	β=.1	β=.2	β=.3	β=.4	β=.6	β=.8
n =100	0.07	0.10	0.19	0.35	0.53	0.84	0.97
n =200	0.06	0.12	0.30	0.58	0.81	0.98	1.00
n =300	0.05	0.15	0.42	0.74	0.93	1.00	1.00
n =400	0.05	0.18	0.52	0.86	0.98	1.00	1.00
n =600	0.06	0.24	0.69	0.95	1.00	1.00	1.00
n =800	0.05	0.29	0.81	0.99	1.00	1.00	1.00

Notes: Nominal level is 5%. $5,000$ simulations.

TABLE 13. Empirical Size and Power of Proposed Test by Sample Size

	β=0	β=.1	β=.2	β=.3	β=.4	β=.6	β=.8
n =100	0.00	0.01	0.03	0.08	0.17	0.48	0.80
n =200	0.00	0.01	0.05	0.17	0.40	0.85	0.99
n =300	0.00	0.02	0.09	0.30	0.63	0.97	1.00
n =400	0.00	0.02	0.14	0.45	0.79	1.00	1.00
n =600	0.00	0.03	0.24	0.70	0.96	1.00	1.00
n =800	0.00	0.04	0.38	0.86	0.99	1.00	1.00

Notes: Nominal level is 5%. 100 sample splits in half. $5,000$ simulations.

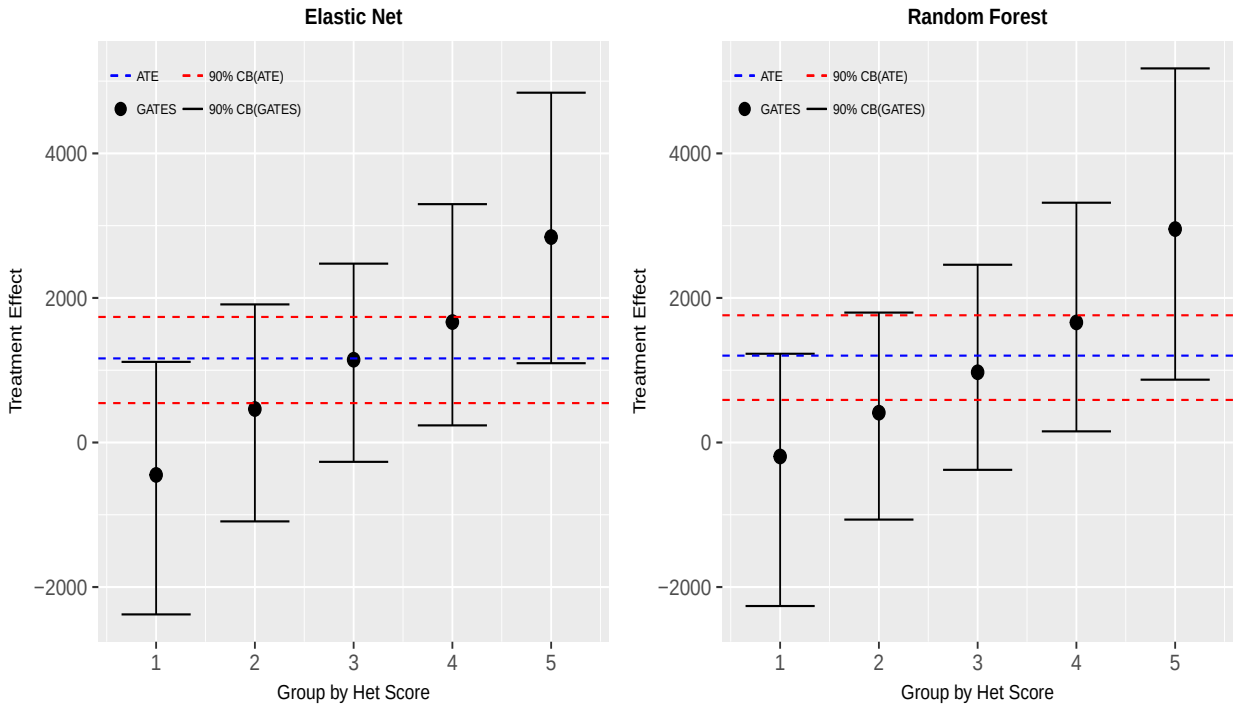

FIGURE 3. GATES of Microfinance Availability: Amount of Loans. Point estimates and 90% adjusted confidence intervals uniform across groups based on 100 random splits in half

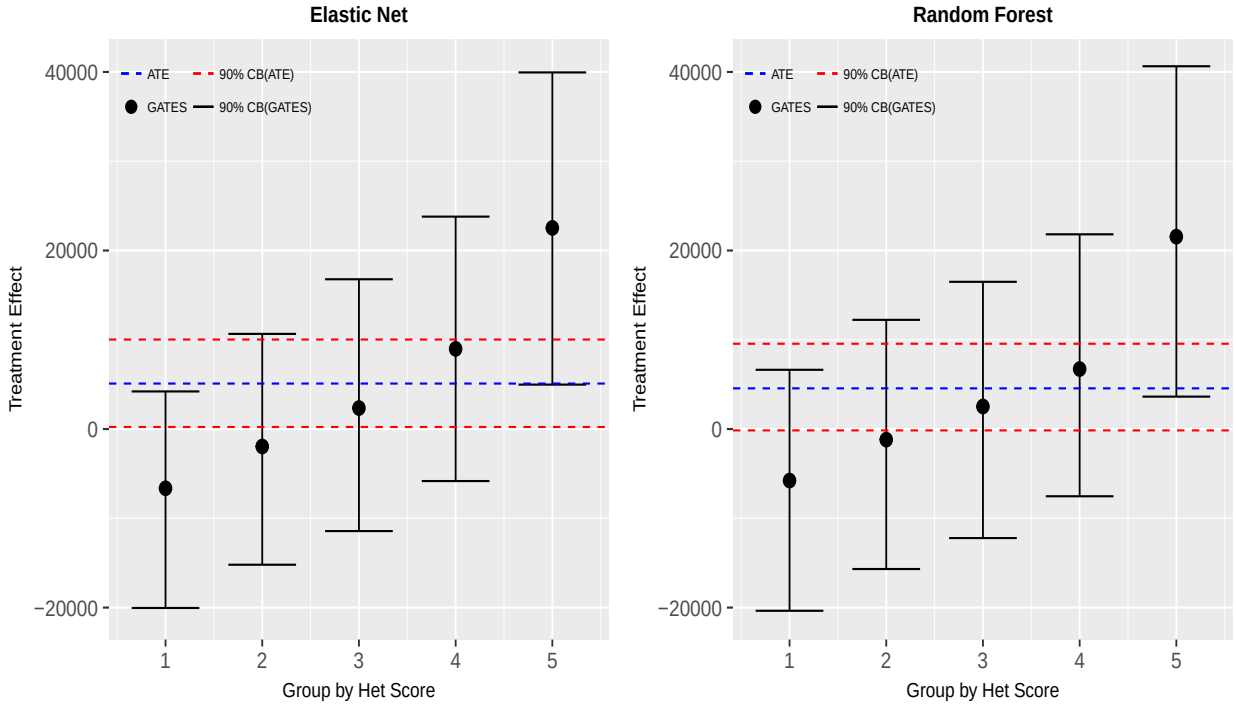

FIGURE 4. GATES of Microfinance Availability: Output. Point estimates and 90% adjusted confidence intervals uniform across groups based on 100 random splits in

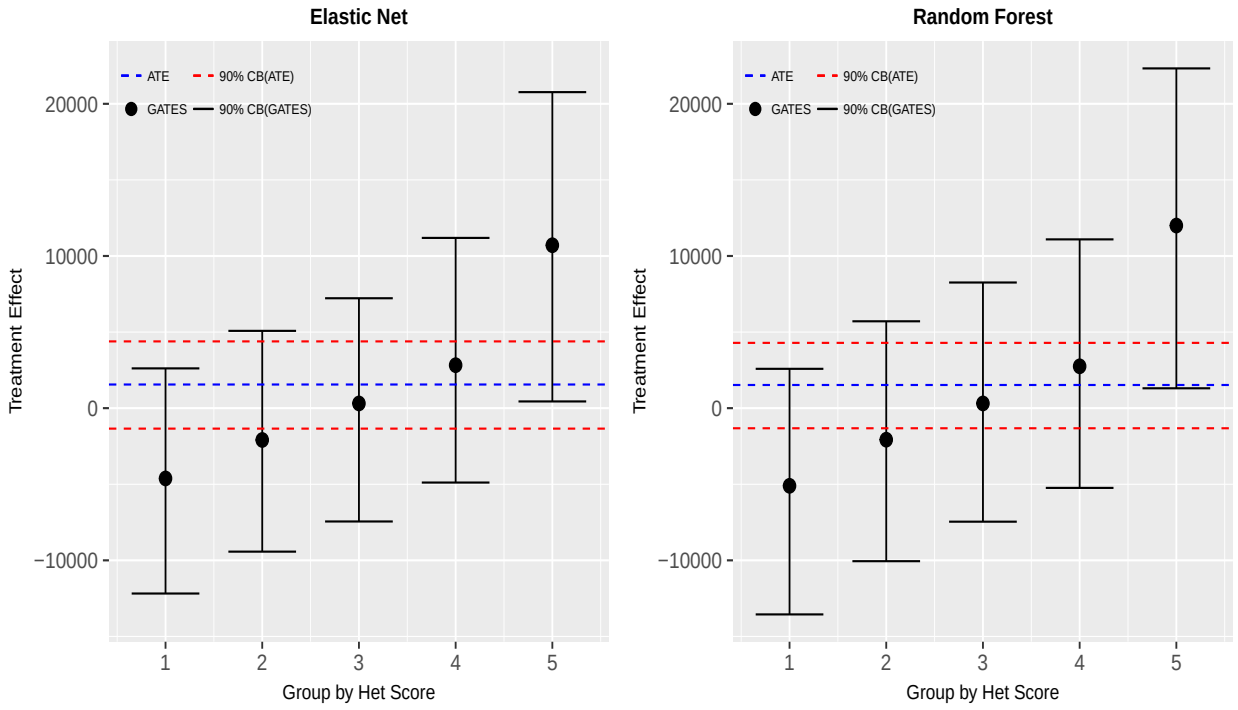

FIGURE 5. GATES of Microfinance Availability: Profit. Point estimates and 90% adjusted confidence intervals uniform across groups based on 100 random splits in half

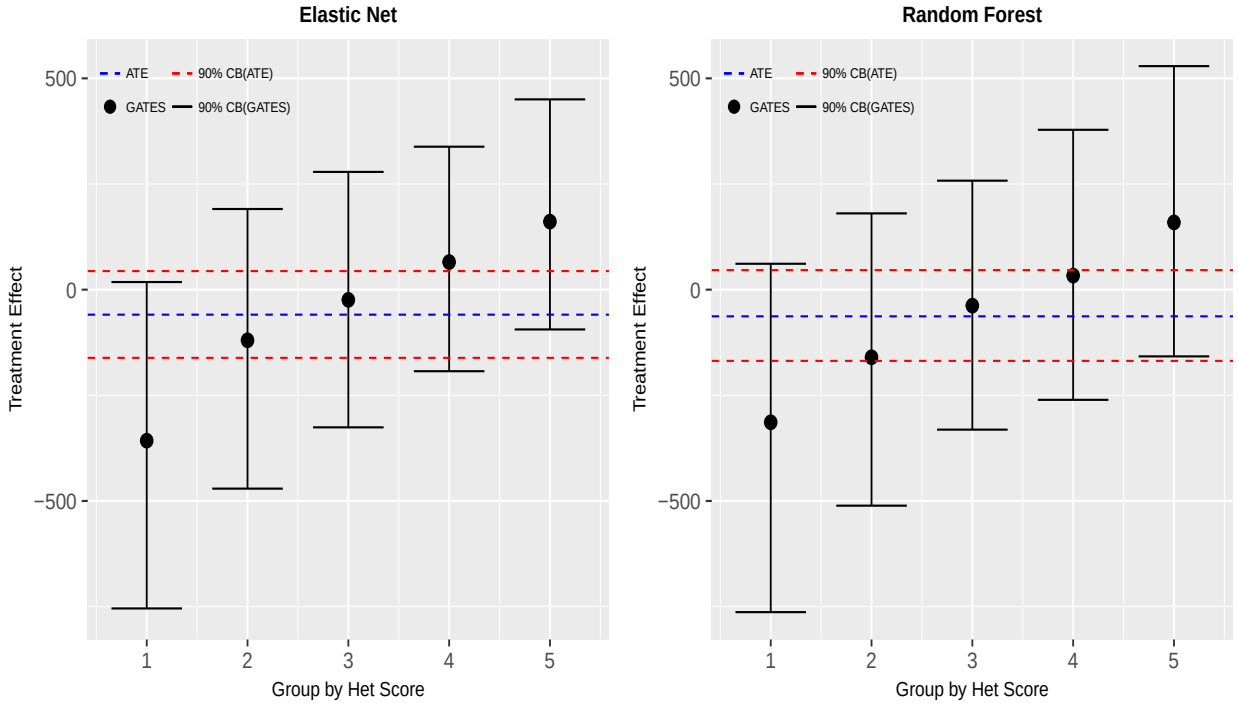

FIGURE 6. GATES of Microfinance Availability: Consumption. Point estimates and 90% adjusted confidence intervals uniform across groups based on 100 random

TABLE 14. CLAN of Immunization Incentives

	Elastic Net			Nnet		
	20% Most (δ_5)	20% Least (δ_1)	Difference ($\delta_5 - \delta_1$)	20% Most (δ_5)	20% Least (δ_1)	Difference ($\delta_5 - \delta_1$)
Fraction participating in Employment Generating Schemes	0.124	0.032	0.089	0.074	0.027	0.045
	(0.107, 0.141)	(0.017, 0.048)	(0.066, 0.113)	(0.060, 0.090)	(0.014, 0.040)	(0.027, 0.066)
	-	-	[0.000]	-	-	[0.000]
Fraction Below Poverty Line (BPL)	0.206	0.183	0.021	0.179	0.174	0.004
	(0.171, 0.241)	(0.148, 0.217)	(-0.027, 0.069)	(0.150, 0.211)	(0.144, 0.201)	(-0.038, 0.046)
	-	-	[0.670]	-	-	[1.000]
Fraction Scheduled Caste-Scheduled Tribes (SC/ST)	0.174	0.126	0.050	0.189	0.139	0.047
	(0.148, 0.201)	(0.100, 0.151)	(0.013, 0.087)	(0.162, 0.217)	(0.114, 0.164)	(0.011, 0.086)
	-	-	[0.014]	-	-	[0.023]
Fraction Other Backward Caste (OBC)	0.276	0.154	0.124	0.335	0.168	0.169
	(0.243, 0.309)	(0.123, 0.185)	(0.078, 0.170)	(0.305, 0.367)	(0.139, 0.196)	(0.126, 0.212)
	-	-	[0.000]	-	-	[0.000]
Fraction Minority Caste	0.007	0.008	-0.001	0.004	0.004	0.000
	(0.001, 0.013)	(0.001, 0.014)	(-0.010, 0.008)	(0.001, 0.010)	(0.001, 0.008)	(-0.004, 0.005)
	-	-	[1.000]	-	-	[1.000]
Fraction General Caste	0.202	0.537	-0.332	0.228	0.505	-0.274
	(0.160, 0.244)	(0.497, 0.578)	(-0.391, -0.276)	(0.188, 0.267)	(0.463, 0.546)	(-0.333, -0.215)
	-	-	[0.000]	-	-	[0.000]
Fraction No Caste	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)
	-	-	[1.000]	-	-	[1.000]
Fraction Other caste	0.000	0.000	0.000	0.001	0.000	0.001
	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.002)	(-0.001, 0.001)	(0.000, 0.002)
	-	-	[1.000]	-	-	[0.292]
Fraction Dont know caste	0.326	0.167	0.155	0.236	0.179	0.052
	(0.288, 0.366)	(0.128, 0.205)	(0.098, 0.212)	(0.202, 0.272)	(0.145, 0.211)	(0.007, 0.099)
	-	-	[0.000]	-	-	[0.047]
Fraction Hindu	0.806	0.940	-0.130	0.959	0.945	0.006
	(0.754, 0.854)	(0.898, 0.985)	(-0.199, -0.062)	(0.936, 0.979)	(0.915, 0.971)	(-0.017, 0.029)
	-	-	[0.000]	-	-	[1.000]
Fraction Muslim	0.165	0.026	0.135	0.020	0.020	0.005
	(0.119, 0.210)	(-0.014, 0.066)	(0.071, 0.198)	(0.009, 0.037)	(0.003, 0.046)	(-0.010, 0.020)
	-	-	[0.000]	-	-	[1.000]
Fraction Christian	0.000	0.004	-0.004	0.000	0.004	-0.004
	(-0.005, 0.005)	(-0.001, 0.009)	(-0.012, 0.003)	(-0.005, 0.005)	(-0.001, 0.009)	(-0.011, 0.003)
	-	-	[0.537]	-	-	[0.524]
Fraction Buddhist	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)
	-	-	[1.000]	-	-	[1.000]
Fraction Sikh	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)
	-	-	[1.000]	-	-	[1.000]
Fraction Jain	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)
	-	-	[1.000]	-	-	[1.000]
Fraction Other Religion	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)
	-	-	[1.000]	-	-	[1.000]

Notes: Medians over 100 splits. 90% confidence interval in parenthesis.P-values for the hypothesis that the parameter is equal to zero in brackets.

Table 15. CLAN of Immunization Incentives-2

	Elastic Net			Nnet		
	20% Most (δ_5)	20% Least (δ_1)	Difference ($\delta_5 - \delta_1$)	20% Most (δ_5)	20% Least (δ_1)	Difference ($\delta_5 - \delta_1$)
Fraction Don't Know Religion	0.032	0.027	0.005	0.018	0.029	-0.013
	(0.020, 0.044)	(0.016, 0.038)	(-0.010, 0.020)	(0.008, 0.029)	(0.019, 0.039)	(-0.027, 0.001)
	-	-	[1.000]	-	-	[0.152]
Fraction Literate	0.782	0.781	0.002	0.819	0.783	0.034
	(0.769, 0.796)	(0.769, 0.794)	(-0.016, 0.021)	(0.809, 0.829)	(0.773, 0.794)	(0.020, 0.048)
	-	-	[1.000]	-	-	[0.000]
Fraction unmarried	0.053	0.049	0.003	0.054	0.046	0.008
	(0.049, 0.057)	(0.045, 0.054)	(-0.003, 0.010)	(0.049, 0.058)	(0.042, 0.050)	(0.001, 0.015)
	-	-	[0.610]	-	-	[0.038]
Fraction of adults Married (living with spouse)	0.491	0.515	-0.022	0.517	0.517	0.001
	(0.482, 0.501)	(0.507, 0.524)	(-0.035, -0.009)	(0.510, 0.525)	(0.509, 0.524)	(-0.010, 0.011)
	-	-	[0.002]	-	-	[1.000]
Fraction of adults Married (not living with spouse)	0.003	0.005	-0.002	0.004	0.003	0.001
	(0.001, 0.005)	(0.003, 0.006)	(-0.004, 0.001)	(0.002, 0.005)	(0.002, 0.004)	(-0.001, 0.002)
	-	-	[0.314]	-	-	[0.784]
Fraction of adults Divorced or Seperated	0.006	0.001	0.005	0.004	0.001	0.003
	(0.005, 0.007)	(0.000, 0.002)	(0.004, 0.007)	(0.003, 0.006)	(0.000, 0.002)	(0.002, 0.005)
	-	-	[0.000]	-	-	[0.000]
Fraction Widow or Widower	0.035	0.037	-0.002	0.036	0.039	-0.004
	(0.031, 0.038)	(0.034, 0.040)	(-0.006, 0.003)	(0.033, 0.039)	(0.036, 0.042)	(-0.008, 0.001)
	-	-	[0.847]	-	-	[0.200]
Fraction Marriage Status Unknown	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)
	-	-	[1.000]	-	-	[1.000]
Fraction Marriage status NA"	0.412	0.393	0.018	0.384	0.392	-0.007
	(0.400, 0.424)	(0.382, 0.404)	(0.001, 0.034)	(0.375, 0.393)	(0.383, 0.402)	(-0.019, 0.005)
	-	-	[0.083]	-	-	[0.481]
Fraction who received Nursery level education or less	0.152	0.170	-0.017	0.133	0.168	-0.033
	(0.144, 0.162)	(0.162, 0.178)	(-0.029, -0.005)	(0.127, 0.140)	(0.162, 0.175)	(-0.043, -0.024)
	-	-	[0.014]	-	-	[0.000]
Fraction who received Class 4 level education	0.079	0.090	-0.011	0.079	0.090	-0.011
	(0.074, 0.084)	(0.085, 0.095)	(-0.018, -0.005)	(0.074, 0.084)	(0.085, 0.094)	(-0.018, -0.004)
	-	-	[0.002]	-	-	[0.004]
Fraction who received Class 9 level education	0.171	0.160	0.010	0.161	0.155	0.006
	(0.164, 0.179)	(0.153, 0.167)	(0.000, 0.021)	(0.155, 0.168)	(0.148, 0.162)	(-0.004, 0.015)
	-	-	[0.095]	-	-	[0.451]
Fraction who received Class 12 level education	0.208	0.224	-0.013	0.246	0.225	0.020
	(0.195, 0.220)	(0.213, 0.235)	(-0.029, 0.003)	(0.235, 0.257)	(0.215, 0.235)	(0.006, 0.034)
	-	-	[0.219]	-	-	[0.009]
Fraction who received Graduate or Other Diploma level education	0.077	0.089	-0.013	0.088	0.093	-0.005
	(0.068, 0.085)	(0.081, 0.098)	(-0.026, -0.001)	(0.079, 0.096)	(0.085, 0.101)	(-0.016, 0.006)
	-	-	[0.080]	-	-	[0.801]

Notes: Medians over 100 splits. 90% confidence interval in parenthesis.P-values for the hypothesis that the parameter is equal to zero in brackets.

References

Alberto Abadie. Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1):1–19, 2005.

Alberto Abadie, Matthew M Chingos, and Martin R West. Endogenous stratification in randomized experiments. Technical report, National Bureau of Economic Research, 2017.

Isaiah Andrews, Toru Kitagawa, and Adam McCloskey. Inference on winners, 2019.

Manuela Angelucci, Dean Karlan, and Jonathan Zinman. Microcredit impacts: Evidence from a randomized microcredit program placement experiment by compartamos banco. *American Economic Journal: Applied Economics*, 7(1):151–182, 2015.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Susan Athey and Guido W Imbens. The econometrics of randomized experiments. *Handbook of Economic Field Experiments*, 1:73–140, 2017.

Orazio Attanasio, Britta Augsburg, Ralph De Haas, Emla Fitzsimons, and Heike Harmgart. The impacts of microfinance: Evidence from joint-liability lending in mongolia. *American Economic Journal: Applied Economics*, 7(1):90–122, 2015.

Britta Augsburg, Ralph De Haas, Heike Harmgart, and Costas Meghir. Microfinance, poverty and education. 2012.

Abhijit Banerjee, Emily Breza, Esther Duflo, and Cynthia Kinnan. Do credit constraints limit entrepreneurship? heterogeneity in the returns to microfinance. *Evanston, USA: Department of Economics Northwestern University*, 2015a.

Abhijit Banerjee, Esther Duflo, Rachel Glennerster, and Cynthia Kinnan. The miracle of microfinance? evidence from a randomized evaluation. *American Economic Journal: Applied Economics*, 7(1):22–53, 2015b.

Abhijit Banerjee, Arun Chandrasekhar, Esther Duflo, Suresh Dalpath, John Floretta, Matthew Jackson, Loza Francine, Harini Kannan, and Anna Schrimpf. Improving full immunization rates in haryana, india: Evaluating incentives and communication methods. 2019a.

Abhijit Banerjee, Arun Chandrasekhar, Esther Duflo, Suresh Dalpath, John Floretta, Matthew Jackson, Harini Kannan, Anna Schrimpf, and Mahesh Shrestha. Leveraging the social network amplifies the effectiveness of interventions to stimulate take up of immunization. 2019b.

Abhijit Banerjee, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson. Using gossips to spread information: Theory and evidence from two randomized controlled trials. *The Review of Economic Studies*, Forthcoming.

Abhijit Vinayak Banerjee. Microcredit under the microscope: what have we learned in the past two decades, and what do we need to know? *Annu. Rev. Econ.*, 5(1):487–519, 2013.

G. Barnard. Discussion of "Cross-validatory choice and assessment of statistical predictions" by Stone. *Journal of the Royal Statistical Society. Series B (Methodological)*, page 133?135, 1974.

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

A. Belloni, V. Chernozhukov, I. Fernández-Val, and C. Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017. doi: 10.3982/ECTA12723. URL `https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA12723`.

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Pivotal estimation of nonparametric functions via square-root lasso. *arXiv preprint arXiv:1105.1475*, 2011.

Alexandre Belloni, Victor Chernozhukov, and Kengo Kato. Uniform post selection inference for lad regression models. *arXiv preprint arXiv:1304.0282*, 2013.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

V. Chernozhukov, I. Fernández-Val, and A. Galichon. Improving point and interval estimators of monotone functions by rearrangement. *Biometrika*, 96(3):559–575, 2009. ISSN 0006-3444. doi: 10.1093/biomet/asp030. URL `http://dx.doi.org/10.1093/biomet/asp030`.

V. Chernozhukov, I. Fernandez-Val, and Y. Luo. The Sorted Effects Method: Discovering Heterogeneous Effects Beyond Their Averages. *ArXiv e-prints*, December 2015.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics*, 42(5):1787–1818, 2014.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 2017.

DR Cox. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441–444, 1975.

Bruno Crépon, Florencia Devoto, Esther Duflo, and William Parienté. Estimating the impact of microcredit on those who take it up: Evidence from a randomized experiment in morocco. *American Economic Journal: Applied Economics*, 7(1):123–150, 2015.

Bruno Crepon, Esther Duflo, Huillery Elisa, William Pariente, Juliette Seban, and Paul-Armand Veillon. Cream skimming and the comparison between social interventions evidence from entrepreneurship programs for at-risk youth in france. 2019. Mimeo.

Jonathan Davis and Sara B Heller. Rethinking the benefits of youth employment programs: The heterogeneous effects of summer jobs. Technical report, National Bureau of Economic Research, 2017.

Tatyana Deryugina, Garth Heutel, Nolan H Miller, David Molitor, and Julian Reif. The mortality and medical costs of air pollution: Evidence from changes in wind direction. *The American Economic Review*, Forthcoming.

Ruben Dezeure, Peter Bühlmann, and Cun-Hui Zhang. High-dimensional simultaneous inference with the bootstrap. *arXiv preprint arXiv:1606.03940*, 2016.

Esther Duflo, Rachel Glennerster, and Michael Kremer. Using randomization in development economics research: A toolkit. *Handbook of development economics*, 4:3895–3962, 2007.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Christopher Genovese and Larry Wasserman. Adaptive confidence bands. *The Annals of Statistics*, pages 875–905, 2008.

Evarist Giné and Richard Nickl. Confidence bands in density estimation. *The Annals of Statistics*, 38(2):1122–1170, 2010.

Uri Gneezy and Aldo Rustichini. A fine is a price. *The Journal of Legal Studies*, 29(1):1–17, 2000.

Christian Hansen, Damian Kozbur, and Sanjog Misra. Targeted undersmoothing. *arXiv preprint arXiv:1706.07328*, 2017.

John A Hartigan. Using subsample values as typical values. *Journal of the American Statistical Association*, 64(328):1303–1317, 1969.

Keisuke Hirano, Guido W. Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003. ISSN 0012-9682. doi: 10.1111/1468-0262.00442. URL `http://dx.doi.org/10.1111/1468-0262.00442`.

Kosuke Imai and Marc Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

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

Dean Karlan and Jonathan Zinman. Expanding credit access: Using randomized supply decisions to estimate the impacts. *The Review of Financial Studies*, 23(1):433–464, 2009.

Dean Karlan and Jonathan Zinman. Microcredit in theory and practice: Using randomized credit scoring for impact evaluation. *Science*, 332(6035):1278–1284, 2011.

Leslie Kish and Martin Richard Frankel. Inference from complex samples. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–37, 1974.

Max Kuhn. Caret package. *Journal of Statistical Software*, 28(5):1–26, 2008.

Guillaume Lecué and Charles Mitchell. Oracle inequalities for cross-validation type procedures. *Electronic Journal of Statistics*, 6:1803–1837, 2012.

Mark G Low et al. On nonparametric confidence intervals. *The Annals of Statistics*, 25(6):2547–2554, 1997.

Rachael Meager. Aggregating distributional treatment effects: A bayesian hierarchical analysis of the microcredit literature. *Working Paper*, 2017.

Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009.

Frederick Mosteller and John Wilder Tukey. Data analysis and regression: a second course in statistics. *Addison-Wesley Series in Behavioral Science: Quantitative Methods*, 1977.

J Neyman. Sur les applications de la theorie des probabilites aux experiences agricoles: essai des principes (masters thesis); justification of applications of the calculus of probabilities to the solutions of certain questions in agricultural experimentation. excerpts english translation (reprinted). *Stat Sci*, 5:463–472, 1923.

Natalia Rigol, Reshmaan Hussam, and Benjamin Roth. Targeting high ability entrepreneurs using community information: Mechanism design in the field, 2016.

Alessandro Rinaldo, Larry Wasserman, Max G'Sell, Jing Lei, and Ryan Tibshirani. Bootstrapping and sample splitting for high-dimensional, assumption-free inference. *arXiv preprint arXiv:1611.05401*, 2016.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10(4):1040–1053, 1982. ISSN 0090-5364. URL `http://links.jstor.org/sici?sici=0090-5364(198212)10:4<1040:OGROCF>2.0.CO;2-2&origin=MSN`.

Alessandro Tarozzi, Jaikishan Desai, and Kristin Johnson. The impacts of microcredit: Evidence from ethiopia. *American Economic Journal: Applied Economics*, 7(1):54–89, 2015.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (just-accepted), 2017.

Larry Wasserman. Machine learning overview. In *Becker-Friedman Institute, Conference on ML in Economics*, 2016.

Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009.

Marten Wegkamp et al. Model selection in nonparametric regression. *The Annals of Statistics*, 31 (1):252–273, 2003.

Qingyuan Zhao, Dylan S Small, and Ashkan Ertefaie. Selective inference for effect modification via the lasso. *arXiv preprint arXiv:1705.08020*, 2017.