GENERIC MACHINE LEARNING INFERENCE ON HETEROGENOUS TREATMENT EFFECTS IN RANDOMIZED EXPERIMENTS, WITH AN APPLICATION TO IMMUNIZATION IN INDIA

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* on 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. Estimation and inference are based 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, which quantifies the uncertainty coming from both parameter estimation and data splitting, is shown to be uniformly valid for a large class of data generating processes. We illustrate the use of the approach with a randomized field experiment that evaluated a combination of nudges to stimulate demand for immunization in India.

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, Heterogeneous Effects, Immunization incentives, Nudges.

JEL: C18, C21, D14, G21, O16

1. Introduction

Randomized Controlled Trials (RCT) and Machine Learning (ML) are arguably two of the most important developments in data analysis methods for applied researchers. RCT play an important role in the evaluation of social and economic programs, medical treatments and marketing (e.g., Imbens and Rubin, 2015; Duflo et al., 2007). ML is a name attached to a variety of new, constantly evolving statistical learning methods including Random Forest, Boosted Trees, Neural Networks, Penalized Regression, Ensembles, and Hybrids; see, e.g., Wasserman (2016) for a recent review, and

Date: December 27, 2020.

The authors respectively from MIT, MIT, MIT and BU. This paper was delivered (virtually) at the Fischer-Shultz Lecture of the Econometric Society World Congress, 2020. We thank the organizing committee for including us in the program. We thank Susan Athey, Moshe Buchinsky, Denis Chetverikov, Guido Imbens, Steven Lehrer, Siyi Luo, Max Kasy, Susan Murphy, Whitney Newey, Patrick Power and seminar participants at ASSA 2018, Barcelona GSE Summer Forum 2019, Brazilian Econometric Society Meeting 2019, BU, Lancaster, NBER summer institute 2018, NYU, UCLA, Whitney Newey's Contributions to Econometrics conference, and York for valuable comments. We gratefully acknowledge research support from the National Science Foundation, AFD, USAID, and 3ie.

Friedman et al. (2001) for a prominent textbook treatment. ML has become a key tool for prediction and pattern recognition problems, surpassing classical methods in high dimensional settings.

At first blush, those two sets of methods may seem to have very different applications: in the most basic randomized controlled experiment, there is a sample with a single treatment and a single outcome. Covariates are not necessary and even linear regression is not the best way to analyze the data (Imbens and Rubin, 2015). In practice however, applied researchers are often confronted with more complex experiments. For example, there might be accidental imbalances in the sample, which require select control variables in a principled way. ML tools, such as the lasso method proposed in Belloni et al. (2014, 2017) or the double machine learning method proposed in Chernozhukov et al. (2017), have proven useful for this purpose. Moreover, some complex RCT designs have so many treatments combinations that ML methods may be useful to select the few treatments that actually work and pool the rest with the control groups (Banerjee et al., 2019). Finally, 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 the treatment effect depends on covariates, such as gender, age, etc. This heterogeneity is essential to assess if the impact of the program would generalize to a 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 (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 might be a large number of potential ways to form subgroups. Often researchers collect rich baseline surveys, which give them access to a large number of covariates: 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 the 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 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

[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.

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 available 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, i.e. inference that remains valid under a large class of data generating processes. In fact, in high dimensional settings, absent strong assumptions, generic ML tools may not even produce consistent estimators of the *conditional average treatment effect* (CATE), the difference in the expected potential outcomes between treated and control states conditional on covariates.

Previous attempts to solve this problem focused 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 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 target features of the CATE based on this proxy predictor. In particular, we consider three objects, which are likely to be of interest to applied researchers and policy makers: (1) **Best Linear Predictor** (BLP) of the CATE on the ML proxy predictor; (2) **Sorted Group Average Treatment Effects** (GATES) or average treatment effect by heterogeneity groups induced by the ML proxy predictor; and (3) **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 are associated 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. For example, Crepon et al. (2019) used the same technique to construct and estimate a specific form of heterogeneity at the post-processing

[2]In the recent past, several new empirical papers in economics used ML methods to estimate heterogenous effects. E.g. Rigol et al. (2016) showed that villagers outperform the machine learning tools when they predict heterogeneity in returns to capital. Davis and Heller (2017) predicted who benefits the most from a summer internship projects. Deryugina et al. (Forthcoming) used the methods developed in the present paper to evaluate the heterogeneity in the effect of air pollution on mortality. Crepon et al. (2019) also built 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.

stage. Even then, as we will see, getting robust and conservative standard error for heterogeneity requires a larger sample size than just estimating average treatment effects. This reflects a different trade-off: if we don't assume that we can predict *ex ante* where the heterogeneity might be (in which case we can write it down in a pre-analysis plan), power will be lower, and detecting heterogeneity will require a larger sample. This is a consideration that applied researchers will need to keep in mind when designing and powering their experiments.

We apply our method to a large-scale RCT of nudges to encourage immunization in the state of Haryana, Northern India, designed and discussed in Banerjee et al. (2019), an important practical application in its own right. Immunization is generally recognized as one of the most effective and cost effective ways to prevent illness, disability, and diseases. Yet, worldwide, close to 20 million children every year do not receive critical immunizations (Unicef, 2019). While early policy efforts have focused mainly on improving the infrastructure for immunization services, a more recent literature suggests that "nudges" (such as small incentives, leveraging the social network, SMS reminders, social signalling, etc.) may have large effect on the use of those services.[3] This project was a collaboration with the government of Haryana, which was willing to experiment with a combination of nudges, with the goal of choosing the most effective policy and implement it at scale. It built a custom vaccination platform, and ran a large scale experiment covering seven districts, 140 Primary health centers, 2,360 villages involved in the experiment (including 915 at risk for all the treatments), and 295,038 children in the resulting data base. Immunization was very low at baseline: in every single village of the district, the fraction of children whose parents had reported they received the measles vaccine (the last in the sequence) was 39%, and only 19.4% had received the vaccine before the age of 15 months, whereas the full sequence is supposed to be completed in one year. The experiment was a cross randomized design of three main nudges: providing incentives, sending SMS reminders, and seeding ambassadors. It included several variants for each policy: the level and schedule of the incentives, the number of people receiving reminders, and mode of selection of the ambassadors, leading to a large number (75) of finely differentiated bundles.

Banerjee et al. (2019) developed a methodology to identify the most effective and cost effective bundle of policies, based on an application of LASSO to a "smart pooling" specification that imposes some structure on the bundles, and in particular, the idea that policy variants (level of incentives, or level of coverage of SMS reminders) may be indistinguishable in practice. They found that the most cost-effective policy is to combine "information hubs" (people identified by others

[3]See, for example, Banerjee, Duflo, Glennerster, and Kothari (2010); Bassani, Arora, Wazny, Gaffey, Lenters, and Bhutta (2013); Wakadha, Chandir, Were, Rubin, Obor, Levine, Gibson, Odhiambo, Laserson, and Feikin (2013); Johri, Pérez, Arsenault, Sharma, Pai, Pahwa, and Sylvestre (2015); Oyo-Ita, Wiysonge, Oringanje, Nwachukwu, Oduwole, and Meremikwu (2016); Gibson, Ochieng, Kagucia, Were, Hayford, Moulton, Levine, Odhiambo, O'Brien, and Feikin (2017); Karing (2018); Domek, Contreras-Roldan, OâĂŹLeary, Bull, Furniss, Kempe, and Asturias (2016); Uddin, Shamsuzzaman, Horng, Labrique, Vasudevan, Zeller, Chowdhury, Larson, Bishai, and Alam (2016); Regan, Bloomfield, Peters, and Effler (2017); Alatas, Chandrasekhar, Mobius, Olken, and Paladines (2019); Banerjee, Chandrasekhar, Duflo, Dalpath, Floretta, Jackson, Francine, Kannan, and Schrimpf (2020).

as good at diffusing information) and SMS reminders. This is cheap and can be done everywhere. In fact, they showed that this policy is the only one among those tested that would actually save money to the government for each measles shot, while increasing immunization. But the most *effective* policy, i.e. the policy that increases immunization the most, is the combination of incentives, immunization ambassadors, and SMS reminders, which is much more expensive. Yet, while this policy increase the cost per immunization, the effects are important: in our sample, the number of monthly measles shot (the last vaccine in the schedule, and thus a marker for full immunization) delivered increases by 2.77, a 38% of the a control group that got neither SMS nor increasing incentives nor information hubs. The government was therefore interested in finding out where the program would be most effective, to implement it only in those places even at the higher cost per immunization.

The pre-analysis plan specified to look for heterogeneity by gender and by "Village-level baseline/national census variables: including assets, beliefs, knowledge, and attitudes towards immunization" but did not identify one or two specific baseline variables to look at: this reflected genuine uncertainty (as is often the case). Many factors can influence policy impact, from attitudes to implementation capabilities, to baseline levels, and we did not have a specific theory of where to look for. It is precisely the type of context that requires a principled approach to avoid overfitting, and provide a policy-relevant recommendation.[4]

The rest of the paper is organized as follows. Section 2 formalizes the framework, describes our approach and compares it with the existing literature. Section 3 presents identification and estimation strategies for the key features of the CATE of interest. Section 4 introduces our variational inference method that accounts for uncertainty coming from parameter estimation and sample splitting. Section 5 discusses practical considerations such as the choice of ML method, stratified sample splitting and dealing with ML proxies with little variation, together with extensions to other prediction and causal inference problems. Section 6 reports the results of the empirical application and provides detailed implementation algorithms. Section 7 concludes with some remarks. The Appendix gathers the proofs of the main theoretical results, power calculations based on numerical simulations, and additional technical results.

2. Our Agnostic Approach

2.1. **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; Rubin, 1974). Let Z be a possibly high-dimensional vector of covariates that characterize the observational units. The main causal functions are the baseline conditional average (BCA):

$$b_0(Z) := \mathrm{E}[Y(0) \mid Z], \tag{2.1}$$

[4]This approach of finding the best treatment, and then looking at where it works the best, gets closer to the idea of "personalized medicine". Using the same data, Agarwal et al. (2020) go one step further and use a "synthetic intervention" approach to look for the policy that works the best for each kind of village.

and the conditional average treatment effect (CATE):

$$s_0(Z) := \mathrm{E}[Y(1) \mid Z] - \mathrm{E}[Y(0) \mid Z]. \tag{2.2}$$

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

$$D \perp\!\!\!\perp (Y(1), Y(0)) \mid Z, \tag{2.3}$$

and the propensity score is known and is given by

$$p(Z) := \mathrm{P}[D = 1 \mid Z] = \mathrm{P}[D = 1 \mid Z_1], \tag{2.4}$$

which we assume is bounded away from zero or one:

$$p(Z) \in [p_0, p_1] \subset (0, 1). \tag{2.5}$$

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,

$$b_0(Z) = \mathrm{E}[Y \mid D = 0, Z], \tag{2.6}$$

and

$$s_0(Z) = \mathrm{E}[Y \mid D = 1, Z] - \mathrm{E}[Y \mid D = 0, Z]. \tag{2.7}$$

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 $\mathbb{E} = \mathbb{E}_P$.

2.2. **Properties of Machine Learning Estimators of $s_0(Z)$ Motivating the Agnostic Approach.** In modern high-dimensional settings, estimation and inference is challenging because the target function $z \mapsto s_0(z)$ lives in a very complex class. ML methods effectively explore various forms of sparsity to yield "good" approximations to $s_0(z)$. In its simplest form, sparsity reduces the complexity of $z \mapsto s_0(z)$ by assuming that $s_0(z)$ can be well-approximated by a function that only depends on a low-dimensional subset of z, making estimation and inference possible. 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. Sparsity, however, is an untestable assumption that must be used with caution as often cannot be rooted in economic arguments.

Without some form of sparsity, 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 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 (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 assumptions are made. Examples of such assumptions include structured forms of linear and non-linear sparsity and super-smoothness.[5] 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).[6] 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 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.

In this paper we take an agnostic view. We neither rely on any sparsity or super-smoothness assumption nor impose conditions that make the ML estimators consistent. We simply treat ML as providing proxy predictors for the objects of interest.

2.3. **Our Approach.** We propose strategies for estimation and inference on

$$\textit{key features of} \quad s_0(Z) \text{ rather than } s_0(Z) \text{ itself.}$$

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

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 $\mathrm{Data} = (Y_i, D_i, Z_i)_{i=1}^N$ into a main sample, denoted by $\mathrm{Data}_M = (Y_i, D_i, Z_i)_{i \in M}$, and an auxiliary sample, denoted by $\mathrm{Data}_A = (Y_i, D_i, Z_i)_{i \in A}$. We will sometimes

[5]The function $z \mapsto s_0(z)$ is super-smooth if it has continuous and bounded derivatives of all orders.

[6]Let $z \mapsto s_0(z)$ be a target function that lives in an infinite-dimensional class with unknown regularity s (e.g., smoothness or degree of sparsity). Adaptive consistent estimation (resp. inference) for $z \mapsto s_0(z)$ with respect to s is possible, if there exists a consistent estimator (resp. valid confidence set) with a rate of covergence (resp. diameter) that changes with s in a rate-optimal way.

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. After splitting the sample, we carry out two stages:

Stage 1: from the auxiliary sample A, we obtain ML estimators of the baseline and treatment effects, which we call the proxy predictors,

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

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 estimators of $b_0(z)$ and $s_0(z)$.

Stage 2: We post-process the proxies from Stage 1 to estimate and make inference on features of the CATE $z \mapsto s_0(z)$ in the main sample M. We condition on the auxiliary sample Data_A, so we consider these maps as frozen, when working with the main sample.

The *key features of* $s_0(Z)$ that we target include:

(1) **Best Linear Predictor** (BLP) of the CATE $s_0(Z)$ 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 estimators. 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 estimation, we report the median of the estimated key features over different random splits of the data. For interval estimation, 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 them 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.[7]

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

This work is related to the literature that uses linear and semiparametric regression methods for estimation and inference on heterogeneous effects. Crump et al. (2008) developed tests of treatment effect homogeneity for low-dimensional settings based on traditional series estimators of the CATE. 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 of 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.

Another approach is to use 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 to inform the process of building causal trees, with the main goal of predicting CATE. They also provided 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) proposed a subsampling-based construction of a causal random forest, providing valid pointwise inference for CATE (see also the review in Wager and

[7]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 also avoids taking this risk.

Athey (2017) on prior uses of random forests in causal settings) for the case when covariates are very low-dimensional (and essentially uniformly distributed).[8] Unfortunately, this condition rules out the typical high-dimensional settings that arise in many empirical problems, especially in current RCT where the number of baseline covariates is potentially very large.

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 the 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.

The idea of using a "hold out" sample to validate the result of a ML procedure to discover heterogeneity was suggested in Davis and Heller (2017), who used the method proposed in Wager and Athey (2017) and compared their results to the heterogeneity in a hold out sample. Our inference approach is different because it calls for multiple splits. This procedure itself 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) and Meinshausen et al. (2009), where the ideas are related but the details are quite different, as we shall explain below. The premise is the same, however, as in Meinshausen et al. (2009) and 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) and 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).

[8]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.

3. Main Identification Results and Estimation Strategies

3.1. **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 $L^2(P)$.

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

$$Y = \alpha' X_1 + \beta_1 (D - p(Z)) + \beta_2 (D - p(Z))(S - \mathrm{E}S) + \epsilon, \;\; \mathrm{E}[w(Z)\epsilon X] = 0, \tag{3.1}$$

where $S := S(Z)$, $w(Z) = \{p(Z)(1 - p(Z))\}^{-1}$, $X := (X_1, X_2)$, $X_1 := X_1(Z)$, e.g., $X_1 = [1, B(Z)]$, and $X_2 := [D - p(Z), (D - p(Z))(S - \mathrm{E}S)]$. 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.[9]

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 3.1 (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* (3.1) *also solves the best linear predictor/approximation problem for the target $s_0(Z)$:*

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

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:

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

$$\mathbb{E}_{N,M}[w(Z_i)\widehat{\epsilon}_i X_i] = 0,$$

[9]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.

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

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

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 3.1 (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 in $s_0(Z)$ and $S(Z)$ is a 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 1,000 observations between the auxiliary and main samples, Z is uniformly distributed in $(-1, 1)$, $b_0(Z) = 3Z$, U is normally distributed with variance 0.5 independently of Z, 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 3.2 (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 $s_0(Z)$. ■

Second Strategy. The second strategy to identify the BLP of CATE 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:

$$YH = \mu' X_1 H + \beta_1 + \beta_2(S - \mathrm{E}S) + \tilde{\epsilon}, \quad \mathrm{E}\tilde{\epsilon}\tilde{X} = 0, \tag{3.2}$$

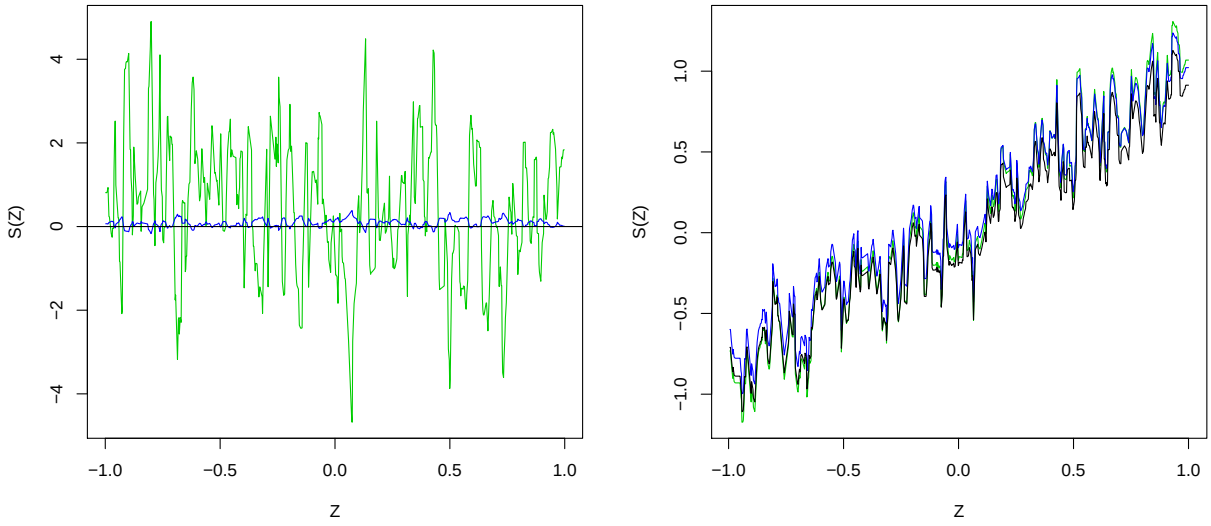

FIGURE 1. Example. In the left panel we have a homogeneous CATE $s_0(Z) = 0$; in the right panel we have heterogeneous CATE $s_0(Z) = Z$. The proxy predictor $S(Z)$ is produced by the Random Forest, shown by green line, the true BLP of CATE is shown by black line, and the estimated BLP of CATE is shown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

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*.

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 3.2 (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* (3.2) *solves the best linear predictor/approximation problem for the target $s_0(Z)$:*

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

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 3.3 (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.

3.2. **GATES.** The target parameters are

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

where G is an indicator of group membership.

Comment 3.4. 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,

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

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

$$Y = \alpha' X_1 + \sum_{k=1}^{K} \gamma_k \cdot (D - p(Z)) \cdot 1(G_k) + \nu, \quad \mathrm{E}[w(Z)\nu W] = 0, \tag{3.3}$$

for $B := B(Z)$, $S := S(Z)$, $W = (X_1', W_2')'$, and $W_2 = (\{(D - p(Z))1(G_k)\}_{k=1}^K)'$.

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:

$$Y_i = \widehat{\alpha}' X_{1i} + \widehat{\gamma}' W_{2i} + \widehat{\nu}_i, \quad i \in M, \quad \mathbb{E}_{N,M}[w(Z_i)\widehat{\nu}_i W_i] = 0. \tag{3.4}$$

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.

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

$$YH = \mu' X_1 H + \sum_{k=1}^{K} \gamma_k \cdot 1(G_k) + \nu, \quad \mathrm{E}[\nu \tilde{W}] = 0, \tag{3.5}$$

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:

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

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

$$Y_i H_i = \widehat{\mu}' X_{1i} H_i + \widehat{\gamma}' \tilde{W}_{2i} + \widehat{\nu}_i, \quad i \in M, \quad \mathbb{E}_{N,M}[\widehat{\nu}_i \tilde{W}_i] = 0. \tag{3.6}$$

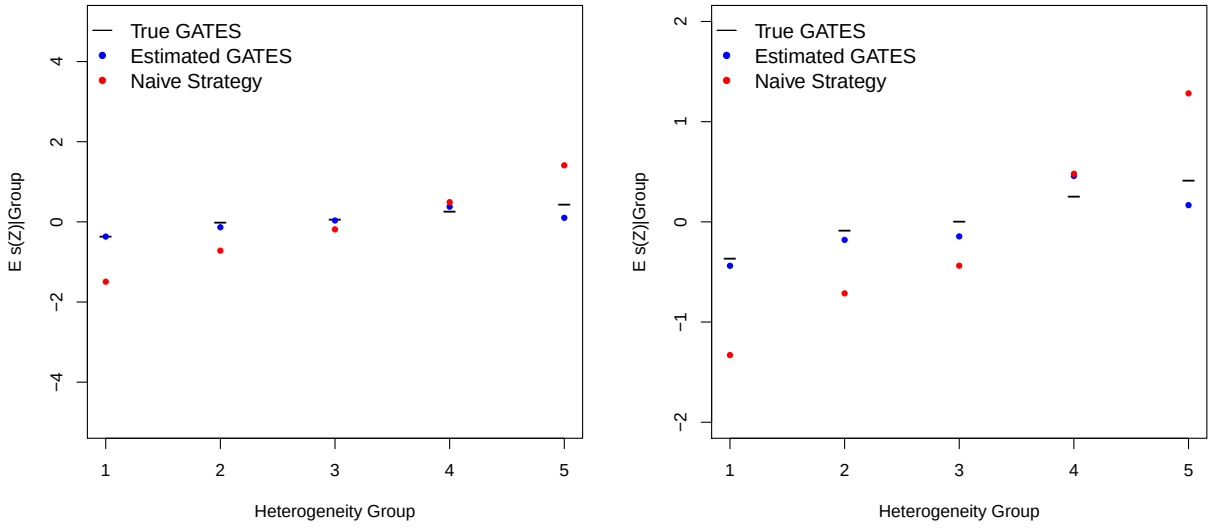

FIGURE 2. In the left panel we have the homogeneous CATE $s_0(Z) = 0$; in the right panel we have heterogeneous CATE $s_0(Z) = Z$. The proxy predictor $S(Z)$ for CATE is produced by the random forest, whose sorted averages by groups are shown as red dots, exhibiting large biases. These are the naive 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.

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 3.3 (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 (3.3) or by the regression equation (3.5). These parameters defined in two different ways are equivalent and are equal to the expectation of $s_0(Z)$ conditional on the proxy group $G_k = \{S \in I_k\}$:*

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

3.3. **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 affected 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 and single out the covariates that are associated with the heterogeneity in the CATE. We call this type of comparisons as classification analysis or CLAN.

4. "Variational" Estimation and Inference Methods

4.1. **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.

4.2. **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 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 4.1. 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):

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

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 $1 - 2\alpha$:

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

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 4.2. For a random variable X with law P_X we define:

$$\begin{aligned} \underline{\text{Med}}(X) &:= \inf\{x \in \mathbb{R} : \text{P}_X(X \leqslant x) \geqslant 1/2\}, \ \ \overline{\text{Med}}(X) := \sup\{x \in \mathbb{R} : \text{P}_X(X \geqslant x) \geqslant 1/2\}, \\ \text{Med}(X) &:= (\underline{\text{Med}}(X) + \overline{\text{Med}}(X))/2. \end{aligned}$$

Note that the lower median $\underline{\text{Med}}(X)$ is the usual definition of the median. The upper median $\overline{\text{Med}}(X)$ 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 $\underline{\text{Med}}(X) = 2$ and $\overline{\text{Med}}(X) = 3$; and if X is uniform on $\{1, 2, 3\}$, then $\overline{\text{Med}}(X) = \underline{\text{Med}}(X) = 2$. 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 4.3. Our construction of p-values builds upon the false-discovery-rate type adjustment ideas in Benjamini and Hochberg (1995) and 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 4.1 (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.[10] ■

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:

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

[10]The inequality is tight in the sense that it can hold with equality. For example, if $J = \{1,2\}$ and $U_2 = 1 - U_1$, then $M = U_1 \wedge (1 - U_1)$ and $\mathbb{P}[M \leqslant \alpha/2] = \alpha$.

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 4.1 (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:

$$\text{CI} := \{\theta \in \mathbb{R} : p_u(\theta) > \alpha/2,\ p_l(\theta) > \alpha/2\}, \tag{4.1}$$

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

$$p_l(\theta) := \underline{\text{Med}}(1 - \Phi[\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta)] \mid \text{Data}), \tag{4.2}$$

$$p_u(\theta) := \underline{\text{Med}}(\Phi[\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta)] \mid \text{Data}). \tag{4.3}$$

The confidence interval CI has the following representation in terms of the medians of t-statistics implied by the proof Theorem 4.2 stated below:

$$\text{CI} = \left\{ \theta \in \mathbb{R} : \begin{array}{c} \overline{\text{Med}}\left[\frac{\theta-\widehat{\theta}_A}{\widehat{\sigma}_A} - \Phi^{-1}(1-\alpha/2) \mid \text{Data}\right] < 0 \\ \underline{\text{Med}}\left[\frac{\theta-\widehat{\theta}_A}{\widehat{\sigma}_A} + \Phi^{-1}(1-\alpha/2) \mid \text{Data}\right] > 0 \end{array} \right\}. \tag{4.4}$$

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 4.2 (Uniform Validity of Variational Confidence Intervals)**.** CI *can be represented as (4.4) 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 $P \in \mathcal{P}$.

5. Other Considerations and Extensions

5.1. **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 $z \mapsto S(z)$.

(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. This approach works well when there is clearly a ML method that dominates the others – so that we choose the best method with probability approaching 1.[11]

5.2. **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

$$\Lambda := |\beta_2|^2 \mathrm{Var}(S(Z)) = \mathrm{Corr}^2(s_0(Z), S(Z)) \mathrm{Var}(s_0(Z)). \tag{5.1}$$

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 $S(Z)$.

5.3. **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

$$\bar{\Lambda} = \mathrm{E} \left(\sum_{k=1}^{K} \gamma_k 1(S \in I_k) \right)^2 = \sum_{k=1}^{K} \gamma_k^2 \mathrm{P}(S \in I_k). \tag{5.2}$$

This is the part of variation of $s_0(z)$, $\mathrm{E} s_0^2(Z)$, 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

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

[11] When this is not the case, we could still formal inference using the conservative Bonferroni approach.

5.4. **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 an 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.

5.5. **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.

5.6. **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 $s_0(Z)$ on $S(Z)$.
(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 can post-process $S(Z)$ in the main sample, by estimating the BLP and GATs.
(5) We can perform 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 constructions 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 the 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 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 individual characteristics. 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, which we assume is known. That is, $\mathrm{E}[\tilde{Y} \mid Z] = s_0(Z)$ under mild conditions on the density using the integration by parts formula.

6. Application: Where are nudges for immunization the most effective?

We apply our methods to a RCT in India that was conducted to improve immunization, and provide a detailed implementation algorithm.

Immunization is widely believed to be one of the most cost-effective ways to save children lifes. Much progress has been made in increasing immunization coverage since the 1990s. For example, according to the World Health Organization (WHO), global measles deaths have decreased by 73% from 536,000 estimated deaths in 2000 to 142,000 in 2018. In the last few years, however, global vaccination coverage has remained stuck, around 85% (until the COVID-19 epidemics, when they plummeted). In 2018, 19.7 million children under the age of one year did not receive basic vaccines. Around 60% of these children lived in ten countries: Angola, Brazil, Democratic Republic of the Congo, Ethiopia, India, Indonesia, Nigeria, Pakistan, the Philippines and Vietnam. The WHO estimates that immunization saves 2-3 million deaths every year, and that an additional 1.5 million deaths could be averted every year if global vaccination coverage improves (this is comparable to 689,000 deaths from COVID-19 between January and August 2020).[12]

While most of the early efforts have been devoted to building an immunization infrastructure and making sure that immunization is available close to people's homes, there is a growing recognition that it is important to also address the demand for immunization. Part of the low demand reflects deep seated mistrust, but in many cases, parents seem to be perfectly willing to immunize their children. For example, in our data for Haryana, India, among the sample's older siblings who should all have completed their immunization course, 99% had received polio drops and about 90% had an immunization card. 90% of the parents claimed to believe immunization is beneficial and 3% claimed to believe that it is harmful. However, only 37% of the older children had completed the course and received the measles vaccine, according to their parents (which is likely to be an overestimate), and only 19.4% had done so before the fifteen month of life, when it is supposed to be done between the 10th and the 12th month. It seems that parents lose steam over the course of the immunization sequence, and nudges could be helpful to boost demand. Indeed, a recent literature cited in the introduction suggests that "nudges," such as small incentives, leveraging the social network, SMS, etc., may have large effect on the use of those services.

In 2017, Esther Duflo, one of the authors of this paper, led a team that conducted a large scale experiment with the government of Haryana, in North India, to test various strategies to increase the take up of immunization services. The government health system rolled out an e-health platform designed by a research team and programmed by a MIT group (SANA health), in which nurses

[12]See WHO "10 facts on immunization", `https://www.who.int/features/factfiles/immunization/facts/en/index1.html`

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.[13] From the individual data, we constructed the monthly sum of the number of children eligible for the program (i.e. age 12 months or younger at their first vaccines) who received each particular immunization at a program location. These children were aged between 0 and 15 months. In this paper, we focus on the number of children who received the measles shot, as it is the last vaccine in the sequence, and thus a reliable marker for full immunization.

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 at the village level:

(1) Small incentives for immunization: parents/caregivers receive mobile phone credit upon bringing children for vaccinations.
(2) Immunization ambassador 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 SMS 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 immunization ambassadors were either randomly selected, or chosen to be information hubs, using the "gossip" methodology developed by Banerjee et al. (2020), 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 75 possible treatment combinations.

Banerjee et al. (2020) developed and implemented a two-step methodology to identify the most cost effective and the most effective policy to increase the number of children completing the full course of immunization at the village level, and estimate its effects (correcting for bias due to the fact that the policy is found to be the best). First, they used a specific version of LASSO to determine which policies are irrelevant, and which policy variants can be pooled together. Second, they obtained consistent estimates of these restricted set of pooled policies using post-LASSO (Chernozhukov et al., 2015). They found that the most cost effective policy (and the only one to reduce the cost of each immunization compared to the status quo) is to combine information hub ambassadors and SMS reminders. But the policy that increases immunization the most is the combination of information-hub ambassador, the presence of reminders, and increasing incentives (regardless

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

of levels). This is also the most expensive package, so the government was interested in prioritizing villages: where should they scale up the full package? This is an excellent application of this methodology, because there was no strong prior.

We compare 25 villages where this particular policy bundle 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 15 months of younger in a given month in a given village that receive the measles shot. 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 such as religion, caste, financial status, marriage and family status, education, and baseline immunization. The propensity score is constant.

Table 1 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). The combined treatment was very effective on average. During the course of the intervention, on average 7.30 children per month aged 15 months or less got the measles shot that completes the immunization sequence in control villages, and 10.08 did so in treatment villages. This is a raw difference of 2.77, or 38%of the control mean. Note that while these effects are not insignificant, we are far from reaching full immunization: The baseline survey suggest that about 38% of children aged 1-3 had received the measles shot at baseline, and 19.4% had received it before they turned 15 months. These estimate imply that the fraction getting their measles shot before 15 months would only go up to $26.7\%.(19.4 + 0.38 * 19.4)$

The implementation details for the heterogeneity analysis follows the algorithm described below, with three characteristics 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 2 compares the four ML methods for producing proxy predictors $S(Z_i)$ using the criteria in (5.1) and (5.2). We find that Elastic Net and Neural Network outperform the other methods, with Elastic Net beating Neural Network for the GATES by a small margin and viceversa for the BLP. Accordingly, we shall focus on these two methods for the rest of the analysis.

Table 3 presents the results of the BLP of CATE on the ML proxies. 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. The ATE estimates in column 1 and 3 indicate that the package treatment increases the number of immunized children by 2.81 based on elastic net and by 2.53 based on neural network. Reassuringly these estimates are on either side of the raw difference in means (2.77). Focusing on 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 3 present the estimated GATES coefficients $\gamma_1 - \gamma_5$ along with joint confidence bands and the ATE estimates. In Table 4 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 16.34 and 16.80 based on elastic net and neural network methods, respectively, and statistically significant. Given that the ATE estimates in the whole population are about 2.5, these results suggest a large and potentially policy-relevant heterogeneity.

The analysis so far reveals very large heterogeneity, with two striking results. First, the results are very large for the most affected villages. In these villages, on average 10.36 extra of the children who were eligible for incentives at baseline get the measles vaccines every month (starting from a mean of 2.19 in the elastic net estimation). Second, the impact is *negative* and significant in the least affected villages (an average decline of 6.13 immunization per month, starting from 12.24 in the elastic net estimation). It looks like in some contexts, the combined package of small incentives, reminder, and persuasion by members of the social network actually put people off immunization.

Given these large differences, it is important to find out whether this heterogeneity seems to be associated with pre-existing characteristics. To answer this question, we explore what variables are associated with the heterogeneity detected in BLP and GATES via CLAN. Table 5 reports the CLAN estimates for a selected set of covariates and Tables 9–10 in the appendix for the rest of covariates. Regardless of the method used, the estimated differences in means between most and least affected groups for the number of vaccines to pregnant mother, number of vaccines to child since birth, number of polio drops to child, fraction of children receiving immunization card, and fraction of children receiving measles vaccines, are negative and statistically significant. Those are various measures of pre-treatment immunization levels, all survey based, that have nothing to do with our measure of impact. 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 from the CLAN. Thus, in this instance, the policy that is preferred ex-ante by the government (since it is equality enhancing) 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 light on the negative effect we find for the least affected group. Note that this effect is *not* mechanical. Even in the least affected villages, there was a good number of children who did not receive the measles shot, and they were not close to reaching full immunization, where they could not have experience an increase. It may be that it would have been difficult to vaccinate 10.36 extra children every month, but there was plenty of scope to experience an increase in immunization. We had no prior on that the effect would be larger in the villages with the lowest rate of immunization. On the contrary, immunization rates could have been low precisely because parents were more doubtful about immunization. Immunization are particularly low in muslim-majority villages, for example, and this

is believed to reflect their lack of trust in the health system. There were therefore reasons to be genuinely uncertain about where immunization would have had the largest effect.

One possible interpretation of the negative impact in some villages is that villagers were intrinsically motivated to get immunized. The nudging with small incentives and mild social pressure may have backfired, crowding out intrinsic motivation without providing a strong enough extrinsic motivation to act as in Gneezy and Rustichini (2000). A point estimate of 10.36 extra immunization per months in the most affected group might seem high: this represent a multiplication by 4.7 of the baseline level (based on the elastic net specification). This increase in immunizations is not out of line with the literature, however: in a set of villages with very low immunization rate in Rajasthan, Banerjee et al. (2010) find that small incentives increase immunization from 18% to 39% (relative to a treatment that just improves infrastructures, and 6% relative to the control group) in a low immunization region (in the entire sample, not in the places where it is most effective), which was also a very large increase. Given the restrictions imposed to the data set (only children 1 year or less at their first immunization were included), the data cover children who were 15 months or younger when getting the measles shot. Among the older cohort in the most affected group, only 12.7% of children were vaccinated before 15 months. Taking this as a benchmark for the control group, the estimate would still imply that only 60% of the treatment group was immunized before 15 month: a big improvement, but no implausible.

Our last exercise is to compute the cost effectiveness of the program in various groups. To do so, we compute in each village the average number of immunizations delivered per dollar spent in a month, in each group. The dollar spent are the fixed cost to run an immunization program per month (nurse salaries, administrative overheads , etc) plus the marginal cost of each vaccine multiplied by the number of vaccines administered (incentives distributed to local health workers, vaccines doses, syringes, etc...) in both treatment and control villages, plus the extra cost of running each particular treatment (the cost of the tablets used for recording in all the treatment villages, the cost of contacting and enrolling the ambassadors, and the cost of the incentives). We then estimate the cost effectiveness in each GATES group as $\mathrm{E}[X(1) - X(0) \mid G = G_k]$, where X is the immunizations per dollar. $\mathrm{E}[X(1) - X(0) \mid G_k] = \mathrm{E}[X \mid D = 1, G = G_k] - \mathrm{E}[X \mid D = 0, G = G_k]$ by the randomization assumption, and we can estimate each of $\mathrm{E}[X \mid D = 1, G = G_k]$ and $\mathrm{E}[X \mid D = 1, G = G_k]$ analogously to CLAN, that is by taking sample averages within treatment groups for each sample split and aggregating over sample splits.

The results are presented in Table 6. They highlight the crucial importance of treatment effect heterogeneity for policy decision in this context. Overall, as shown in Banerjee et al. (2020) the treatment is not cost effective compared to the control (the immunization per dollar spent goes down). This analysis reveals that this is driven (not surprisingly) by negative impacts on cost effectiveness in the groups where it is least effective. However, in the fourth and fifth quintile of cost-effectiveness, we cannot reject that the immunization per dollar spent is the same in the control group and in the treatment group, despite the added marginal cost of the incentives and the

vaccines: this is because the fixed cost of running the program is now spread over a larger number of immunizations.[14]

6.1. **Implementation Algorithm.** We describe an algorithm based on the first identification strategy and provide some specific implementation details for the empirical example.

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 $i \in [N]$.

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 small variance to the proxies, e.g., a $1/20$-th fraction of the sample variance of Y.

b. Estimate the BLP parameters by weighted OLS in M, i.e.,

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

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.,

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

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 $\{S_i\}_{i \in M}$.

[14] Since the main result in Banerjee et al. (2020) is that the most cost effective option on average is the combination of SMS plus Information hubs, an alternative policy question may therefore be whether there are places where it may be more cost effective to add the incentives to this cheaper treatment. We replicated the heterogeneity analysis comparing these two treatment (full package, versus SMS and ambassadors, but no gossips) and look at the cost effectiveness by GATES in these two options. There is also considerable heterogeneity in this comparison (see Figure 4). The results for cost effectiveness are shown in Table 11 in the appendix. There again, we find that in the two quintiles where adding incentives is most effective, it would need be cost effective, even compared to an alternative status quo of just having SMS and information hubs.

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 4.

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.[15] 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.

7. Conclusion

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

[15]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).

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. An empirical application illustrates the practical use of the approach.

Our hope is that applied researchers use the method to discover whether there is heterogeneity in their data in a disciplined way. A researcher might be concerned about the application of our method due to the possible power loss induced by sample splitting. 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.[16] 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, but it is not as high as researchers might fear. 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. Thus, if discovering and exploiting heterogeneity in treatment effect is a key goal of the research, the researcher should indeed plan for larger sample sizes (relative to just testing whether the treatment has an effect), but the required sample size remain within the realm of what is feasible in the field. In many applications we are aware of, there was apparent heterogeneity according to some covariates of interest, but the disciplined ML heterogeneity exercise found no systematic difference. This could be because this heterogeneity was a fluke, or because the method does not have the power to detect it in a smaller sample. In any case, what this experience suggests is that one should not rely on ex-post heterogenous effects in such cases.

The application to immunization in India is of substantive interest. As the world rushes towards a new vaccine against the coronavirus, there are doubts on whether the vaccine take up will be sufficient. We will need to rely on research on other type of vaccines to design effective strategies. Our findings suggest that a combination of small incentives, relay by information hub, and SMS reminders can have very large effect on vaccine take up in some villages where immunization was

[16]More generally, the plan needs to specify a parametric form for the heterogeneity as a low dimensional function of pre-specified covariates (e.g., Chernozhukov et al., 2015). In this case, ML tools can still be used to efficiently estimate the CATEs in the presence of control variables but are not required to detect heterogeneity (Belloni et al., 2017; Chernozhukov et al., 2017).

low at baseline, and even be more cost-effective than the status quo, but can also backfire in other places. This suggests that these type of strategy need to be piloted in the relevant context before being rolled out, and that heterogeneity needs to be taken into account.

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.

Anish Agarwal, Abdullah Alomar, Romain Cosson, Devavrat Shah, and Dennis Shen. Synthetic interventions, 2020.

Vivi Alatas, Arun G Chandrasekhar, Markus Mobius, Benjamin A Olken, and Cindy Paladines. When celebrities speak: A nationwide twitter experiment promoting vaccination in indonesia. Technical report, National Bureau of Economic Research, 2019.

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.

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. 2019.

Abhijit Banerjee, Arun Chandrasekhar, Esther Duflo, Suresh Dalpath, John Floretta, Matthew Jackson, Loza Francine, Harini Kannan, and Anna Schrimpf. Selecting the most effective nudge: Evidence from a large scale experiment on immunization. 2020.

Abhijit Vinayak Banerjee, Esther Duflo, Rachel Glennerster, and Dhruva Kothari. Improving immunisation coverage in rural india: clustered randomised controlled evaluation of immunisation campaigns with and without incentives. *BMJ*, 340, 2010. ISSN 0959-8138. doi: 10.1136/bmj.c2220. URL `https://www.bmj.com/content/340/bmj.c2220`.

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.

Diego G Bassani, Paul Arora, Kerri Wazny, Michelle F Gaffey, Lindsey Lenters, and Zulfiqar A Bhutta. Financial incentives and coverage of child health interventions: a systematic review and meta-analysis. *BMC Public Health*, 13(S3):S30, 2013.

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 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, Christian Hansen, and Martin Spindler. Post-selection and post-regularization inference in linear models with very many controls and instruments. *American Economic Review: Papers and Proceedings*, 105(5):486–490, 2015.

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 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.

Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics*, 90(3):389–405, 2008.

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.

Gretchen J Domek, Ingrid L Contreras-Roldan, Sean T OâĂŹLeary, Sheana Bull, Anna Furniss, Allison Kempe, and Edwin J Asturias. Sms text message reminders to improve infant vaccination coverage in guatemala: a pilot randomized controlled trial. *Vaccine*, 34(21):2437–2443, 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.

Dustin G Gibson, Benard Ochieng, E Wangeci Kagucia, Joyce Were, Kyla Hayford, Lawrence H Moulton, Orin S Levine, Frank Odhiambo, Katherine L O'Brien, and Daniel R Feikin. Mobile phone-delivered reminders and incentives to improve childhood immunisation coverage and timeliness in kenya (m-simu): a cluster randomised controlled trial. *The Lancet Global Health*, 5 (4):e428–e438, 2017.

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.

Mira Johri, Myriam Cielo Pérez, Catherine Arsenault, Jitendar K Sharma, Nitika Pant Pai, Smriti Pahwa, and Marie-Pierre Sylvestre. Strategies to increase the demand for childhood vaccination in low-and middle-income countries: a systematic review and meta-analysis. *Bulletin of the World Health Organization*, 93:339–346, 2015.

Anne Karing. Social signaling and childhood immunization: A field experiment in sierra leone. *University of California, Berkeley Working Paper*, 2018.

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.

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

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.

Angela Oyo-Ita, Charles S Wiysonge, Chioma Oringanje, Chukwuemeka E Nwachukwu, Olabisi Oduwole, and Martin M Meremikwu. Interventions for improving coverage of childhood immunisation in low-and middle-income countries. *Cochrane Database of Systematic Reviews*, (7), 2016.

Annette K Regan, Lauren Bloomfield, Ian Peters, and Paul V Effler. Randomized controlled trial of text message reminders for increasing influenza vaccination. *The Annals of Family Medicine*, 15 (6):507–514, 2017.

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`.

Md Jasim Uddin, Md Shamsuzzaman, Lily Horng, Alain Labrique, Lavanya Vasudevan, Kelsey Zeller, Mridul Chowdhury, Charles P Larson, David Bishai, and Nurul Alam. Use of mobile phones for improving vaccination coverage among children living in rural hard-to-reach areas and urban streets of bangladesh. *Vaccine*, 34(2):276–283, 2016.

Unicef. 20 million children miss out on lifesaving measles, diphtheria and tetanus vaccines in 2018. *https://www.unicef.org/eca/press-releases/20-million-children-miss-out-lifesaving-measles-diphtheria-and-tetanus-vaccines-2018*, 2019.

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

Hotenzia Wakadha, Subhash Chandir, Elijah Victor Were, Alan Rubin, David Obor, Orin S Levine, Dustin G Gibson, Frank Odhiambo, Kayla F Laserson, and Daniel R Feikin. The feasibility of using mobile-phone based sms reminders and conditional cash transfers to improve timely immunization in rural kenya. *Vaccine*, 31(6):987–993, 2013.

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.

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

Figures and tables

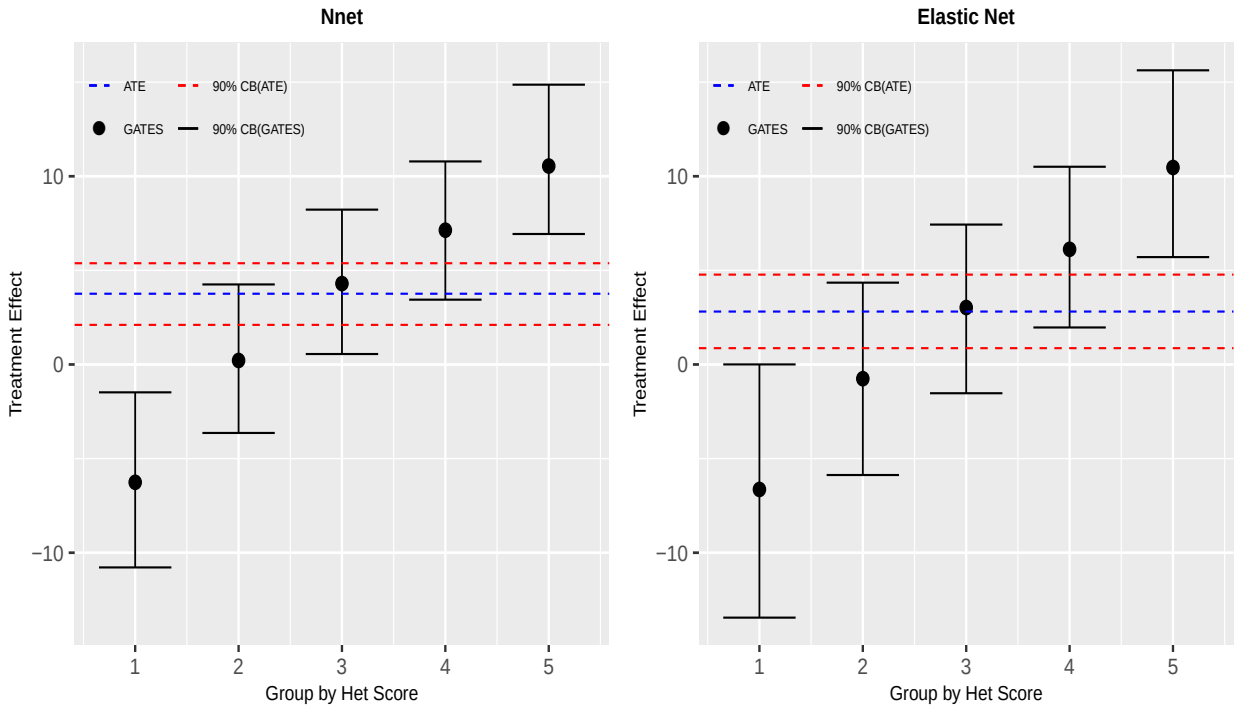

Figure 3. GATES of Immunization Incentives. Point estimates and 90% adjusted confidence intervals uniform across groups based on 100 random splits in half

TABLE 1. Selected Descriptive Statistics of Villages

	All	Treated	Control
Outcome Variables *(Village-Month Level)*			
Number of children who completed the immunization schedule	8.234	10.071	7.304
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
Fraction of kids who received Measles vaccine by 15 months of age	0.194	0.175	0.203
Fraction of kids who received Measles vaccine at credible locations	0.386	0.368	0.395
Number of Observations			
Villages	103	25	78
Village-Months	844	204	640

TABLE 2. Comparison of ML Methods: Immunization Incentives

	Elastic Net	Boosting	Neural Network	Random Forest
Best BLP (Λ)	45.240	27.560	47.940	20.180
Best GATES ($\bar{\Lambda}$)	5.889	4.567	5.443	3.661

Notes: Medians over 100 splits in half.

TABLE 3. BLP of Immunization Incentives

Elastic Net		Neural Network	
ATE (β_1)	HET (β_2)	ATE (β_1)	HET (β_2)
2.812	0.876	2.530	1.059
(0.867,4.774)	(0.656,1.105)	(0.984,4.079)	(0.724,1.401)
[0.008]	[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.

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

	Elastic Net			Nnet		
	20% Most (G_5)	20% Least (G_1)	Difference	20% Most (G_5)	20% Least (G_1)	Difference
GATE $\gamma_k := \widehat{\mathrm{E}}[s_0(Z) \mid G_k]$	10.36	-6.12	16.34	10.39	-6.20	16.80
	(7.42,13.52)	(-9.83,-2.23)	(11.21,21.62)	(6.22,14.60)	(-11.43,-0.73)	(9.50,23.85)
	[0.00]	[0.00]	[0.00]	[0.00]	[0.05]	[0.00]
Control Mean := $\widehat{\mathrm{E}}[b_0(Z) \mid G_k]$	2.19	12.24	-9.87	1.18	10.32	-9.17
	(1.36,2.98)	(11.45,13.10)	(-11.16,-8.73)	(0.44,1.87)	(9.65,11.02)	(-10.17,-8.14)
	[0.00]	[0.00]	[0.00]	[0.00]	[0.00]	[0.00]

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 5. 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	2.161	2.288	-0.128	2.164	2.328	-0.160
to pregnant mother	(2.110,2.212)	(2.237,2.337)	(-0.200,-0.055)	(2.107,2.221)	(2.273,2.385)	(-0.245,-0.082)
	-	-	[0.001]	-	-	[0.000]
Number of vaccines	4.230	4.714	-0.513	3.995	4.670	-0.690
to child since birth	(4.100,4.369)	(4.573,4.860)	(-0.710,-0.311)	(3.816,4.165)	(4.507,4.835)	(-0.937,-0.454)
	-	-	[0.000]	-	-	[0.000]
Fraction of children	1.000	1.000	0.000	0.998	1.000	-0.002
received polio drops	(1.000,1.000)	(1.000,1.000)	(0.000,0.000)	(0.996,1.000)	(0.998,1.002)	(-0.005,0.001)
	-	-	[0.000]	-	-	[0.485]
Number of polio	2.964	2.998	-0.033	2.956	2.994	-0.038
drops to child	(2.954,2.975)	(2.987,3.007)	(-0.047,-0.019)	(2.940,2.971)	(2.980,3.008)	(-0.059,-0.016)
	-	-	[0.000]	-	-	[0.001]
Fraction of children	0.899	0.932	-0.036	0.804	0.930	-0.125
received immunization card	(0.878,0.922)	(0.908,0.956)	(-0.065,-0.004)	(0.765,0.842)	(0.895,0.966)	(-0.178,-0.070)
	-	-	[0.000]	-	-	[0.006]
Fraction of children received	0.127	0.255	-0.131	0.125	0.254	-0.134
Measles vaccine by 15 months of age	(0.100,0.155)	(0.230,0.282)	(-0.167,-0.094)	(0.098,0.152)	(0.229,0.279)	(-0.169,-0.098)
	-	-	[0.052]	-	-	[0.000]
Fraction of children received	0.290	0.435	-0.152	0.275	0.426	-0.151
Measles at credible locations	(0.252,0.327)	(0.400,0.470)	(-0.198,-0.097)	(0.236,0.315)	(0.391,0.461)	(-0.203,-0.100)
	-	-	[0.000]	-	-	[0.000]

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.

TABLE 6. Cost Effectiveness in GATE quintiles

	Elastic Net			Nnet		
	Mean in Treatment $(\widehat{\mathrm{E}}[X \mid D=1, G_k])$	Mean in Control $(\widehat{\mathrm{E}}[X \mid D=0, G_k])$	Difference	Mean in Treatment $(\widehat{\mathrm{E}}[X \mid D=1, G_k])$	Mean in Control $(\widehat{\mathrm{E}}[X \mid D=0, G_k])$	Difference
Imm. per dollar (G_1)	0.034	0.047	-0.013	0.033	0.047	-0.014
	(0.030,0.037)	(0.045,0.048)	(-0.017,-0.009)	(0.029,0.036)	(0.045,0.049)	(-0.019,-0.010)
	-	-	[0.000]	-	-	[0.000]
Imm. per dollar (G_2)	0.031	0.044	-0.013	0.035	0.044	-0.009
	(0.027,0.036)	(0.042,0.046)	(-0.018,-0.008)	(0.031,0.039)	(0.042,0.046)	(-0.013,-0.004)
	-	-	[0.000]	-	-	[0.000]
Imm. per dollar (G_3)	0.037	0.043	-0.007	0.037	0.043	-0.005
	(0.033,0.041)	(0.041,0.046)	(-0.011,-0.002)	(0.034,0.041)	(0.041,0.045)	(-0.010,-0.001)
	-	-	[0.015]	-	-	[0.027]
Imm. per dollar (G_4)	0.039	0.039	-0.001	0.037	0.041	-0.004
	(0.036,0.042)	(0.036,0.042)	(-0.005,0.004)	(0.034,0.041)	(0.038,0.044)	(-0.008,0.001)
	-	-	[1.000]	-	-	[0.186]
Imm. per dollar (G_5)	0.036	0.035	0.001	0.035	0.034	0.001
	(0.032,0.041)	(0.030,0.040)	(-0.006,0.008)	(0.031,0.040)	(0.029,0.040)	(-0.006,0.008)
	-	-	[1.000]	-	-	[1.000]

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.

Appendix A. Proofs

Proof of Theorem 3.1. 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 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

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

$$\beta_2 = \frac{\mathrm{E}w(Z)\{s_0(Z)D(D - p(Z))(S - \mathrm{E}S)\}}{\mathrm{E}w(Z)(D - p(Z))^2(S - \mathrm{E}S)^2}$$

$$= \frac{\mathrm{E}w(Z)s_0(Z)w^{-1}(Z)(S - \mathrm{E}S)}{\mathrm{E}w(Z)w^{-1}(Z)(S - \mathrm{E}S)^2} = \mathrm{Cov}(s_0(Z), S)/\mathrm{Var}(S).$$

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 3.2. 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 3.3. The proof is similar to the proof of Theorem 3.1- 3.2. 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 4.1. We have that $p_{.5} \leqslant \alpha/2$ is equivalent to $\mathbb{E}_P[1(p_A \leqslant \alpha/2) \mid \mathrm{Data}] \geqslant 1/2$. So

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

By Markov inequality,

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

Moreover,

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

■

Proof of Theorem 4.2. To show the second claim, we note that

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

where the inequality holds by Theorem 4.1 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

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

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 $P \in \mathcal{P}$*

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

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

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

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 $\mathbb{E}_{N,M}[H_iX_{2i}Y_i] = \mathbb{E}_{N,M}[w(Z_i)(D_i - p(Z_i))X_{2i}Y_i]$.

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:

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

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.[17] 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 7 and 8 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.

[17] 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 7. 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 8. 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.

Appendix E. Additional Empirical Results

Table 9. CLAN of Immunization Incentives: Other Covariates-1

	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	0.066	0.033	0.033	0.098	0.033	0.065
	(0.046,0.085)	(0.015,0.052)	(0.007,0.059)	(0.076,0.119)	(0.013,0.053)	(0.037,0.093)
	-	-	[0.026]	-	-	[0.000]
Fraction Below Poverty Line (BPL)	0.192	0.176	0.004	0.166	0.189	-0.025
	(0.150,0.233)	(0.139,0.214)	(-0.053,0.064)	(0.127,0.205)	(0.151,0.225)	(-0.080,0.033)
	-	-	[1.000]	-	-	[0.765]
Average Financial Status (1-10 scale)	3.357	3.722	-0.346	3.267	3.561	-0.298
	(3.147,3.545)	(3.538,3.918)	(-0.632,-0.041)	(3.052,3.460)	(3.360,3.769)	(-0.595,-0.028)
	-	-	[0.049]	-	-	[0.059]
Fraction Scheduled Caste-Scheduled Tribes (SC/ST)	0.190	0.150	0.040	0.159	0.132	0.025
	(0.154,0.224)	(0.117,0.184)	(-0.008,0.087)	(0.124,0.194)	(0.099,0.166)	(-0.021,0.072)
	-	-	[0.226]	-	-	[0.587]
Fraction Other Backward Caste (OBC)	0.331	0.172	0.158	0.320	0.163	0.156
	(0.293,0.368)	(0.136,0.209)	(0.106,0.211)	(0.280,0.360)	(0.125,0.201)	(0.101,0.210)
	-	-	[0.000]	-	-	[0.000]
Fraction Minority caste	0.003	0.006	-0.003	0.006	0.008	-0.003
	(-0.001,0.009)	(0.002,0.011)	(-0.008,0.004)	(-0.002,0.012)	(0.001,0.016)	(-0.013,0.006)
	-	-	[0.952]	-	-	[0.995]
Fraction General Caste	0.233	0.491	-0.249	0.189	0.505	-0.315
	(0.179,0.285)	(0.441,0.539)	(-0.321,-0.176)	(0.136,0.244)	(0.452,0.557)	(-0.387,-0.244)
	-	-	[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	0.001	0.000	0.001	0.001	0.000	0.001
	(0.000,0.002)	(-0.001,0.001)	(-0.001,0.003)	(0.000,0.002)	(-0.001,0.001)	(-0.001,0.002)
	-	-	[0.552]	-	-	[0.739]
Fraction Dont Know Caste	0.235	0.177	0.062	0.325	0.179	0.136
	(0.192,0.284)	(0.134,0.219)	(-0.002,0.123)	(0.273,0.376)	(0.130,0.225)	(0.070,0.209)
	-	-	[0.116]	-	-	[0.000]
Fraction Hindu	0.959	0.945	-0.004	0.789	0.943	-0.147
	(0.930,0.984)	(0.905,0.984)	(-0.033,0.027)	(0.723,0.853)	(0.884,1.002)	(-0.238,-0.060)
	-	-	[1.000]	-	-	[0.002]
Fraction Muslim	0.024	0.019	0.011	0.185	0.022	0.159
	(0.008,0.046)	(-0.006,0.052)	(-0.008,0.029)	(0.126,0.247)	(-0.031,0.078)	(0.076,0.242)
	-	-	[0.592]	-	-	[0.000]
Fraction Christian	0.000	0.004	-0.004	0.000	0.004	-0.004
	(-0.006,0.006)	(-0.002,0.009)	(-0.012,0.005)	(-0.006,0.006)	(-0.002,0.010)	(-0.013,0.005)
	-	-	[0.870]	-	-	[0.811]
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: Classification Analysis of Imminuzation Incentives for all Covariates .
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 10. CLAN of Immunization Incentives: Other Covariates-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 Dont Know religion	0.015	0.029	-0.015	0.023	0.027	-0.006
	(0.001,0.028)	(0.016,0.041)	(-0.033,0.004)	(0.008,0.038)	(0.014,0.041)	(-0.024,0.013)
	-	-	[0.257]	-	-	[1.000]
Fraction Literate	0.816	0.787	0.026	0.781	0.787	-0.005
	(0.804,0.830)	(0.773,0.801)	(0.009,0.045)	(0.762,0.799)	(0.771,0.804)	(-0.032,0.020)
	-	-	[0.007]	-	-	[1.000]
Fraction Single	0.050	0.044	0.006	0.051	0.049	0.002
	(0.045,0.055)	(0.039,0.049)	(-0.001,0.014)	(0.046,0.057)	(0.044,0.054)	(-0.005,0.010)
	-	-	[0.161]	-	-	[1.000]
Fraction of adults Married (living with spouse)	0.514	0.520	-0.009	0.489	0.518	-0.029
	(0.504,0.522)	(0.509,0.530)	(-0.022,0.005)	(0.477,0.501)	(0.507,0.530)	(-0.045,-0.012)
	-	-	[0.411]	-	-	[0.002]
Fraction of adutls Married (not living with spouse)	0.003	0.003	0.001	0.003	0.003	0.000
	(0.002,0.005)	(0.001,0.004)	(-0.001,0.003)	(0.001,0.005)	(0.001,0.005)	(-0.002,0.003)
	-	-	[0.978]	-	-	[1.000]
Fraction of adults Divorced or Separated	0.005	0.001	0.004	0.006	0.001	0.005
	(0.003,0.006)	(-0.001,0.002)	(0.002,0.006)	(0.004,0.008)	(-0.001,0.002)	(0.003,0.008)
	-	-	[0.001]	-	-	[0.000]
Fraction of adults Widow or Widower	0.036	0.039	-0.003	0.033	0.037	-0.004
	(0.031,0.041)	(0.035,0.043)	(-0.009,0.003)	(0.029,0.037)	(0.033,0.041)	(-0.010,0.002)
	-	-	[0.791]	-	-	[0.443]
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"	0.389	0.392	-0.001	0.416	0.392	0.024
	(0.379,0.401)	(0.381,0.404)	(-0.016,0.014)	(0.401,0.432)	(0.377,0.406)	(0.003,0.045)
	-	-	[1.000]	-	-	[0.054]
Fraction who received Nursery level education or less	0.136	0.165	-0.028	0.154	0.165	-0.012
	(0.127,0.145)	(0.156,0.174)	(-0.040,-0.015)	(0.142,0.166)	(0.154,0.176)	(-0.028,0.006)
	-	-	[0.000]	-	-	[0.348]
Fraction who received Class 4 level education	0.079	0.090	-0.010	0.074	0.092	-0.017
	(0.072,0.087)	(0.084,0.097)	(-0.020,-0.001)	(0.068,0.080)	(0.086,0.097)	(-0.026,-0.009)
	-	-	[0.064]	-	-	[0.000]
Fraction who received Class 8 level education	0.160	0.154	0.004	0.174	0.161	0.012
	(0.151,0.169)	(0.146,0.163)	(-0.008,0.017)	(0.164,0.183)	(0.153,0.170)	(-0.001,0.025)
	-	-	[0.945]	-	-	[0.132]
Fraction who received Class 12 level education	0.248	0.226	0.017	0.204	0.231	-0.026
	(0.236,0.261)	(0.214,0.240)	(-0.001,0.036)	(0.188,0.220)	(0.216,0.247)	(-0.049,-0.005)
	-	-	[0.144]	-	-	[0.034]
Fraction who received Graduate or Other Diploma level education	0.082	0.095	-0.013	0.078	0.085	-0.007
	(0.071,0.093)	(0.084,0.104)	(-0.028,0.003)	(0.067,0.090)	(0.074,0.096)	(-0.023,0.009)
	-	-	[0.209]	-	-	[0.755]
Fraction with education level Other or Dont know	0.293	0.262	0.031	0.311	0.261	0.050
	(0.284,0.303)	(0.254,0.271)	(0.019,0.044)	(0.301,0.320)	(0.252,0.271)	(0.037,0.063)
	-	-	[0.000]	-	-	[0.000]

Notes: Classification Analysis of Imminuzation Incentives for all Covariates .
Notes: Medians over 100 splits. 90% confidence interval in parenthesis.P-values for the hypothesis that the parameter is equal to zero in brackets.

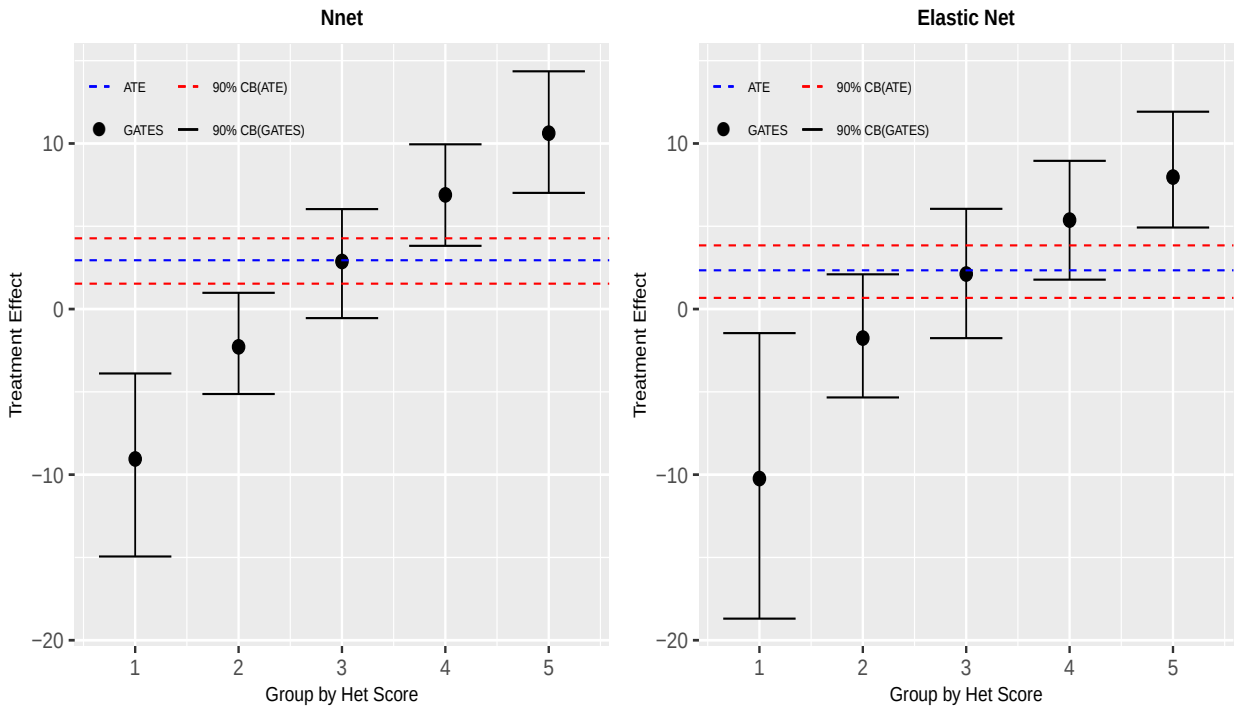

FIGURE 4. GATES of Immunization Full package compared to Ambassadors and SMS only. Point estimates and 90% adjusted confidence intervals uniform across groups based on 100 random splits in half

TABLE 11. Cost effectiveness for GATE Quintiles, Comparing Full treatment to most cost effective treatment

	Elastic Net			Nnet		
	Mean in Treatment ($\widehat{\mathrm{E}}[X \mid D=1, G_k]$)	Mean in Control ($\widehat{\mathrm{E}}[X \mid D=0, G_k]$)	Difference	Mean in Treatment ($\widehat{\mathrm{E}}[X \mid D=1, G_k]$)	Mean in Control ($\widehat{\mathrm{E}}[X \mid D=0, G_k]$)	Difference
Imm. per dollar (All)	0.036	0.043	-0.006	0.036	0.042	-0.006
	(0.034,0.038)	(0.041,0.044)	(-0.008,-0.004)	(0.034,0.038)	(0.041,0.044)	(-0.009,-0.004)
	-	-	[0.000]	-	-	[0.000]
Imm. per dollar (G_1)	0.034	0.047	-0.013	0.033	0.047	-0.014
	(0.030,0.037)	(0.045,0.048)	(-0.017,-0.009)	(0.029,0.036)	(0.045,0.049)	(-0.019,-0.010)
	-	-	[0.000]	-	-	[0.000]
Imm.per dollar (G_2)	0.031	0.044	-0.013	0.035	0.044	-0.009
	(0.027,0.036)	(0.042,0.046)	(-0.018,-0.008)	(0.031,0.039)	(0.042,0.046)	(-0.013,-0.004)
	-	-	[0.000]	-	-	[0.000]
Imm.per dollar (G_3)	0.037	0.043	-0.007	0.037	0.043	-0.005
	(0.033,0.041)	(0.041,0.046)	(-0.011,-0.002)	(0.034,0.041)	(0.041,0.045)	(-0.010,-0.001)
	-	-	[0.015]	-	-	[0.027]
Imm. per dollar (G_4)	0.039	0.039	-0.001	0.037	0.041	-0.004
	(0.036,0.042)	(0.036,0.042)	(-0.005,0.004)	(0.034,0.041)	(0.038,0.044)	(-0.008,0.001)
	-	-	[1.000]	-	-	[0.186]
Imm. per dollar (G_5)	0.036	0.034	0.002	0.036	0.035	0.002
	(0.032,0.040)	(0.030,0.039)	(-0.004,0.007)	(0.033,0.040)	(0.031,0.038)	(-0.003,0.008)
	-	-	[1.000]	-	-	[0.882]

Note: This table compares the cost effectiveness by GATES group when we compare the full package to the most effective package (SMS and information hubs ambassad
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.