ONLINE APPENDIX OF "FISHER-SCHULTZ LECTURE: 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. This Appendix contains additional results and proofs, together with empirical results omitted from the main text.

APPENDIX A. DEFERRED DISCUSSION AND PROOFS FOR SECTION 3

Comment A.1 (Motonicity Restrictions on GATES)**.** Suppose we observe $s_0(Z)$. In this case we can define the ideal GATES as:

$$\gamma_{0k} := \mathrm{E}\left[s_0(Z) \mid G_{0k}\right], \quad k = 1, \ldots, K,$$

where $G_{0k} := \{s_0(Z) \in I_{0k}\}$, with $I_{0k} = \left[\ell_{0,k-1}, \ell_{0,k}\right)$ and $-\infty = \ell_0 < \ell_1 < \ldots < \ell_K = +\infty$. By construction the ideal GATES obey the monotonicity restriction:

$$\gamma_{01} \leqslant \ldots \leqslant \gamma_{0K}.$$

If $S(Z)$ provides a good approximation to $s_0(Z)$, it is reasonable to expect that the GATES also obey the monotonicity restriction: $\gamma_1 \leqslant \ldots \leqslant \gamma_K$, but there is no guarantee. However, we can always replace $\gamma = \{\gamma_k\}_{k=1}^K$ by the non-decreasing rearrangement (sorted vector) $\gamma^* = \left\{\gamma_k^*\right\}_{k=1}^K$, such that γ^* obeys the monotonicity condition $\gamma_1^* \leqslant \ldots \leqslant \gamma_K^*$. The benefit is that γ^* is always closer to $\gamma_0 = \{\gamma_{0k}\}_{k=1}^K$ than γ in the sense that

$$\|\gamma^* - \gamma_0\|_\infty \leqslant \|\gamma - \gamma_0\|_\infty,$$

where $\|\cdot\|_\infty$ is the sup-norm. This follows from the contraction property of the rearrangement (e.g., Chernozhukov et al., 2009). Therefore, we can always use sorting to target the ideal GATES better. Similarly, when performing estimation, we can replace $\widehat{\gamma} = \{\widehat{\gamma}_k\}_{k=1}^K$ by their non-decreasing rearrangement (sorted vector) $\widehat{\gamma}^* = \left\{\widehat{\gamma}_k^*\right\}_{k=1}^K$, which results in an estimator with lower estimation error in the sense that surely:

$$\|\widehat{\gamma}^* - \gamma_0\|_\infty \leqslant \|\widehat{\gamma} - \gamma_0\|_\infty.$$

Date: October 10, 2023.

Proof of Theorem 3.1. The subset of the normal equations, which correspond to $\alpha := (\alpha_1, \alpha_2)'$, are $\mathrm{E}[w(Z)(Y - \alpha_0' X_1 - \alpha' X_2)X_2] = 0$. 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:

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

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

$$\begin{aligned}
&\mathrm{E}[w(Z)\{s_0(Z)D - \alpha_1(D - p(Z))\}(D - p(Z))] = 0,\\
&\mathrm{E}[w(Z)\{s_0(Z)D - \alpha_2(D - p(Z))(S - \mathrm{E}S)\}(D - p(Z))(S - \mathrm{E}S)] = 0.
\end{aligned}$$

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

$$\begin{aligned}
\alpha_1 &= \frac{\mathrm{E}[w(Z)\{s_0(Z)D(D - p(Z))]\}}{\mathrm{E}[w(Z)(D - p(Z))^2]} = \frac{\mathrm{E}[w(Z)s_0(Z)w^{-1}(Z)]}{\mathrm{E}[w(Z)w^{-1}(Z)]} = \mathrm{E}s_0(Z),\\
\alpha_2 &= \frac{\mathrm{E}[w(Z)\{s_0(Z)D(D - p(Z))(S - \mathrm{E}S)\}]}{\mathrm{E}[w(Z)(D - p(Z))^2(S - \mathrm{E}S)^2]}\\
&= \frac{\mathrm{E}[w(Z)s_0(Z)w^{-1}(Z)(S - \mathrm{E}S)]}{\mathrm{E}[w(Z)w^{-1}(Z)(S - \mathrm{E}S)^2]} = \frac{\mathrm{Cov}(s_0(Z), S)}{\mathrm{Var}(S)}.
\end{aligned}$$

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

$$\mathrm{E}\{[s_0(Z) - \alpha_1 - \alpha_2(S - \mathrm{E}S)][1, (S - \mathrm{E}S)]'\} = 0,$$

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

Proof of Theorem 3.2. The subset of the normal equations, which correspond to $\mu := (\mu_1, \mu_2)'$, are $\mathrm{E}[(YH - \mu_0' X_1 H - \mu' \tilde{X}_2)\tilde{X}_2] = 0$. 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:

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

Hence the normal equations simplify to: $\mathrm{E}[(s_0(Z)DH - \mu'\tilde{X}_2)\tilde{X}_2] = 0$. Since 1 and $S - \mathrm{E}S$ are orthogonal, the normal equations above further simplify to

$$\mathrm{E}\{s_0(Z)DH - \mu_1\} = 0, \;\; \mathrm{E}[\{s_0(Z)DH - \mu_2(S - \mathrm{E}S)\}(S - \mathrm{E}S)] = 0.$$

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

$$\mathrm{E}\{s_0(Z) - \mu_1\} = 0, \;\; \mathrm{E}[\{s_0(Z) - \mu_2(S - \mathrm{E}S)\}(S - \mathrm{E}S)] = 0.$$

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 Definition 3.1. ∎

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 $\mu := (\mu_k)_{k=1}^K$, are given by $\mathrm{E}[(YH - \mu_0'X_1H - \mu'\tilde{W}_2)\tilde{W}_2] = 0$. Substituting $Y = b_0(Z) + s_0(Z)D + U$, and using the definition $\tilde{W}_2 = \tilde{W}_2(Z) = [1(G_k)_{k=1}^K]'$, $X_1 = X_1(Z)$, and the law of iterated expectations, we notice that:

$$\begin{aligned}
&\mathrm{E}[b_0(Z)H\tilde{W}_2(Z)] = \mathrm{E}[b_0(Z)\underbrace{\mathrm{E}[H(D,Z) \mid Z]}_{=0}\tilde{W}_2(Z)] = 0,\\
&\mathrm{E}[UH\tilde{W}_2(Z)] = \mathrm{E}[\underbrace{\mathrm{E}[U \mid Z,D]}_{0}H(D,Z)\tilde{W}_2(Z)] = 0,\\
&\mathrm{E}[X_1H\tilde{W}_2(Z)] = \mathrm{E}[X_1(Z)\underbrace{\mathrm{E}[H(D,Z) \mid Z]}_{=0}\tilde{W}_2(Z)] = 0.
\end{aligned}$$

Hence the normal equations simplify to: $\mathrm{E}[\{s_0(Z)DH - \mu'\tilde{W}_2\}\tilde{W}_2] = 0$. 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 $\mathrm{E}[\{s_0(Z)DH - \mu_k 1(G_k)\}1(G_k)] = 0$. Using that $\mathrm{E}[DH \mid Z] = 1$, $S = S(Z)$, and the law of iterated expectations, the equations simplify to

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

The asserted result follows. ∎

Proof of Comment 3.2. We assume that all variables are square integrable. We consider a pure RCT where $p(Z) = p$. Define

$$\gamma := \mathrm{Cov}(B,S)/\mathrm{Var}(S) = p^{-1}\mathrm{Cov}(B,DS)/\mathrm{Var}(S),$$
$$R := D(S-\mathrm{E}S) - p\gamma(B-\mathrm{E}B).$$

We can re-parameterize

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

where $\tilde{X} = (1, B(Z), D, D(S-\mathrm{E}S))$, as follows:

$$Y = \bar{\alpha}_1 + \tilde{\alpha}_2(B-\mathrm{E}B) + \tilde{\beta}_1(D-p) + \tilde{\beta}_2 D(S-\mathrm{E}S) + \varepsilon,$$

for $\bar{\alpha}_1 = \tilde{\alpha}_1 + \tilde{\alpha}_2\mathrm{E}B + \tilde{\beta}_1 p$. Also,

$$Y = b_0(Z) + s_0(Z)D + U, \quad \mathrm{E}[U \mid Z, D] = 0,$$

where U is independent of any function of Z and D.

Note that $B-\mathrm{E}B$, $D-p$, 1, and U are mutually orthogonal. Also $D(S-\mathrm{E}S)$ is orthogonal to $1, D-p$, and U. Using these notes and Frisch-Waugh-Lovell theorem, it is standard to verify that $\tilde{\beta}_1 = \beta_1$. Using these notes and the Frisch-Waugh-Lovell theorem, we obtain

$$\begin{aligned}
\tilde{\beta}_2 &= \mathrm{E}YR/\mathrm{E}R^2 \\
&= \mathrm{E}(b_0(Z) + s_0(Z)D + U)R]/\mathrm{E}R^2 \\
&= \mathrm{E}[s_0(Z)D + b_0(Z)]R/\mathrm{E}R^2 \\
&= \mathrm{E}[s_0(Z)DR]/\mathrm{E}R^2 + \mathrm{E}[b_0(Z)R]/\mathrm{E}R^2 \\
&= \mathrm{Cov}(s_0(Z), DR)/\mathrm{Var}(R) + \mathrm{Cov}(b_0(Z), R)/\mathrm{Var}(R).
\end{aligned}$$

This expression does not simplify to $\beta_2 = \mathrm{Cov}(s_0(Z), S(Z))/\mathrm{Var}(S(Z))$ in general.

Here are some sufficient conditions for the simplification: Suppose $B-\mathrm{E}B$ spans $S-\mathrm{E}S$, so we can set $B = S$ without loss of generality, then $\gamma = 1$, and

$$R = (D-p)(S-\mathrm{E}S),$$

in which case

$$\mathrm{Cov}(s_0(Z), DR)/\mathrm{Var}(R) = \beta_2, \;\; \mathrm{E}[b_0(Z)R]/\mathrm{Var}(R) = 0.$$

Another sufficient condition is when $b_0(Z)$ and $s_0(Z)$ are both uncorrelated to B and S, in which case

$$\tilde{\beta}_2 = \beta_2 = 0.$$

Finally, consider the case where S and B are uncorrelated. In this case $\gamma = 0$ so that

$$R = D(S-\mathrm{E}S),$$

which gives

$$\tilde{\beta}_2 = \beta_2 + p\mathrm{E}b_0(Z)(S-ES)/\mathrm{E}R^2,$$

which simplifies to β_2 if $b_0(Z)$ is also uncorrelated to S. ∎

Comparison of the Second Stage Estimation Strategies for BLP of CATE. We focus on the estimation of the BLP. The analysis can be extended to the GATES using analogous arguments.

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

$$Y_i = X_{1i}'\widehat{\alpha}_0 + (D_i - p(Z_i))X_{2i}'\widehat{\alpha} + \widehat{\varepsilon}_i, \quad i \in M, \quad \mathbb{E}_{N,M}[w(Z_i)\widehat{\varepsilon}_i X_i] = 0,$$

where $w(Z) = \{p(Z)(1-p(Z))\}^{-1}$, $X_i = [X_{1i}', (D_i - p(Z_i))X_{2i}']'$, and $\widehat{\alpha} := (\widehat{\alpha}_1, \widehat{\alpha}_2)'$. Let $\widehat{\theta} := (\widehat{\alpha}_0', \widehat{\alpha}')'$. Then,

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

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

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

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,

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

using that $\mathrm{E}[w(Z)XX']$ is block-diagonal between $\widehat{\alpha}_0$ and $\widehat{\alpha}$.

In the second strategy, we run the linear regression

$$H_iY_i = H_iX_{1i}'\widehat{\mu}_0 + X_{2i}'\widehat{\mu} + \widehat{\varepsilon}_i, \quad \mathbb{E}_{N,M}\widehat{\varepsilon}_i\tilde{X}_i = 0,$$

where $H_i = (D_i - p(Z_i))w(Z_i)$, $\tilde{X}_i = [H_iX_{1i}', X_{2i}]'$ and $\widehat{\mu} := (\widehat{\mu}_1, \widehat{\mu}_2)'$. Let $\tilde{\theta} := (\widehat{\mu}_0', \widehat{\mu}')'$. Then,

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

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,

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

where we use that $\mathrm{E}[\tilde{X}\tilde{X}']$ is block-diagonal between $\widehat{\mu}_0$ and $\widehat{\mu}$, 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 $\widehat{\alpha}$ and $\widehat{\mu}$ have the same asymptotic distribution because they have the same first order representation.

APPENDIX B. DEFERRED DISCUSSION, RESULTS AND PROOFS FOR SECTION 4

Comment B.1 (Robustness of the Coverage Property)**.** In our numerical results, the coverage property in Theorem 4.3 is satisfied even if only (R1) holds. This suggests that it may be possible to establish the coverage property under much weaker conditions. In particular, the coverage property holds without the concentration conditions (R2) and (R3) if

$$\mathrm{P}\left(|\mathrm{M}(\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta^*) \mid \mathrm{Data})| > z\right) \leqslant \mathrm{P}\left(|\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta_A)| > z\right) + \gamma_N''. \tag{B.1}$$

Perhaps surprisingly, this property does hold in numerical experiments even when θ_A does not concentrate around θ^*; see, e.g. Figure 1. However, formally demonstrating this property proved difficult and remains an unresolved problem for future research.

Other Issues: Stratified Splitting, Small Variation of Proxies. The idea of stratified sample splitting is to balance the proportions of treated and untreated units in both a and m samples so that the proportion of treated units is equal to the experiment's propensity scores across strata. This balance potentially improves the performance of the inferential algorithms. Stratified sampling formally requires us to replace the i.i.d. assumption with an i.n.i.d. assumption (independent but not identically distributed observations). The inference results continue to apply as long as the conditions (R1), (R2), (R3) hold. We conjecture that these conditions continue to be plausible under stratified splitting.

Another issue is that the analysis may generate proxy predictors S that have little variation, so we can think of them as "weak". This causes some target parameters to be weakly identified, e.g., the BLP parameter, leading to the potential breakdown of the basic normal approximation (4.5), which our inferential results rely on. To avoid this issue, we can add small noise to the proxies (jittering) so that inference results go through.

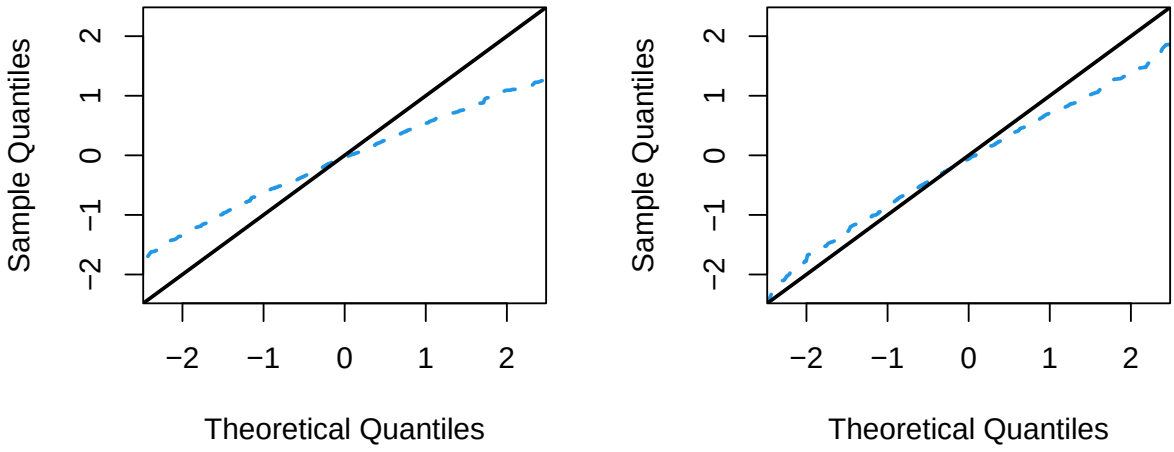

FIGURE 1. A simple Monte-Carlo experiment illustrating inferential robustness with and without concentration conditions.

NOTES. This example shows that the actual quantiles of the statistic $\mathrm{M}(\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta^*) \mid \text{Data})$ are conservatively bounded by those of $N(0,1)$ with concentration and without concentration. The estimand θ_A is generated from $U(0,K)$, with $K = 1/\sqrt{N}$ in the left panel (almost homogeneous case) or with $K = 10$ in the right panel (strong heterogeneous case). The estimator $\widehat{\theta}_A$ is generated as $\theta_A + 1/|M| \sum_{i \in M} \varepsilon_i$, where ε_i's are i.i.d. exponential random variables centered to have mean zero. The main sample indices M are randomly drawn from $\{1, ..., N\}$ without replacement, with $N = 600$ and the subsample size $n/N = 1/3$. $\widehat{\sigma}_A$ is the classical standard error for the sample mean. In the left figure we get 99.5% coverage, and in the right 98.2% coverage for the nominal level of 95%. The results are based on 100 splits, and 1,000 replications.

Proof of Lemma 4.1. To show (4.2) note that,

$$\mathrm{E}|\widehat{\theta} - \theta'| = \mathrm{EE}[|\widehat{\theta} - \theta'| \mid \text{Data}] \leqslant \mathrm{EE}[|\widehat{\theta}_A - \theta'| \mid \text{Data}] \leqslant \mathrm{E}|\widehat{\theta}_A - \theta'|,$$

where the inequality follows from (any) median minizing average absolute loss and its equivariance property. The equalities hold by the law of iterated expectation. The claim (4.3) follows in the same way.

To show (4.4), let $U^* = \{U_a^*\}_{a \in \mathscr{A}}$ and $L^* = \{L_a^*\}_{a \in \mathscr{A}}$ denoted non-decreasing monotone rearrangements of $\{U_a\}_{a \in \mathscr{A}}$ and $L = \{L_a\}_{a \in \mathscr{A}}$. Then

$$|U - L| \leqslant \|U^* - L^*\|_\infty \leqslant \|U - L\|_\infty,$$

where the second inequality follows from the rearrangement having contractive property in the max distance. ∎

Proof of Theorem 4.1. We demonstrate the result for p^+. The proofs for other p-values follow similarly. We use $M_{\mathscr{A}}[\cdot]$ as short hand for M$[\cdot|$Data$]$, with overlined and underlined versions defined similarly.

To show claim (ii) we note that for $z = \Phi^{-1}(1-\alpha)$ and using that $\Phi(z) = 1-\alpha$:

$$\begin{aligned}
\mathrm{P}\Big(M_{\mathscr{A}}[1-\Phi(\widehat{\sigma}_A^{-1}(\widehat{\theta}_A-\theta_0))] < \alpha\Big) &= \mathrm{P}\Big(M_{\mathscr{A}}[\Phi(-\widehat{\sigma}_A^{-1}(\widehat{\theta}_A-\theta_0))] < \alpha-1\Big) \\
&= \mathrm{P}\Big(\, M_{\mathscr{A}}[\widehat{\sigma}_A^{-1}(\theta_A-\widehat{\theta}_A)] < -z\Big) \\
&\leqslant \mathrm{P}\Big(\widehat{\sigma}_A^{-1}(\theta_A-\widehat{\theta}_A) < -z\Big) + \gamma_N'' \\
&\leqslant \Phi(-z) + \gamma_N' + \gamma_N'' = \alpha + \gamma_N' + \gamma_N'',
\end{aligned}$$

where the first inequality uses the concentration of median assumption, and the last inequality follows from the approximate normality assumption (4.5).

To show claim (i), we note that

$$\begin{aligned}
\mathrm{P}\Big(M_{\mathscr{A}}[1-\Phi(\widehat{\sigma}_A^{-1}(\widehat{\theta}_A-\theta_0))] < \alpha\Big) &= \mathrm{P}\Big(\, M_{\mathscr{A}}[\widehat{\sigma}_A^{-1}(\theta_A-\widehat{\theta}_A)] < -z\Big) \\
&\leqslant \mathrm{P}\left(\frac{1}{\mathscr{A}}\sum_{a\in\mathscr{A}} 1\{\, \widehat{\sigma}_A^{-1}(\theta_A-\widehat{\theta}_A)] < -z\} \geqslant 1/2\right) \\
&\leqslant 2\mathrm{E}\left[\frac{1}{\mathscr{A}}\sum_{a\in\mathscr{A}} 1\{\, \widehat{\sigma}_A^{-1}(\theta_A-\widehat{\theta}_A)] < -z\}\right] \\
&= 2\mathrm{P}\{\, \widehat{\sigma}_A^{-1}(\theta_A-\widehat{\theta}_A)] < -z\} \\
&\leqslant 2\Phi(-z) + 2\gamma_N' = 2\alpha + 2\gamma_N',
\end{aligned}$$

where the first equality reused the previous calculation, the first inequality holds by definition of the numerical median, the second inequality holds by Markov inequality, and the equality that follows holds by

$$\begin{aligned}
2\mathrm{E}\left[\frac{1}{\mathscr{A}}\sum_{a\in\mathscr{A}} 1\{\, \widehat{\sigma}_a^{-1}(\theta_a-\widehat{\theta}_a)] < -z\}\right] &= 2\mathrm{EP}\Big(\, \widehat{\sigma}_A^{-1}(\theta_A-\widehat{\theta}_A) < -z \,|\, \text{Data}\Big) \\
&= 2\mathrm{P}\{\, \widehat{\sigma}_A^{-1}(\theta_A-\widehat{\theta}_A)] < -z\},
\end{aligned}$$

and the last inequality follows from the approximate normality assumption (4.5).

Proof of Theorem 4.2. Define $\mathscr{D} = \{\theta_a, [L_a, U_a] : a \in \mathscr{A}\}$, and let $A \sim U(\mathscr{A})$ given $\mathscr{D}$. Then,

$$\begin{aligned}
\mathrm{P}(\theta_A < L) &= \mathrm{EP}(\theta_A < L \mid \mathscr{D}) = \mathrm{E}\left[\frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}} 1\{\theta_a < L\}\right] \\
&= \mathrm{E}\left[\frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}} 1\{\theta_a < L,\, L_a < L\} + \frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}} 1\{\theta_a < L,\, L_a \geqslant L\}\right] \\
&\leqslant \mathrm{E}\left[\frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}} 1\{L_a < L\}\right] + \mathrm{E}\left[\frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}} 1\{\theta_a < L_a\}\right] \\
&\leqslant \mathrm{E}(\beta) + \mathrm{E}\left[\frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}} 1\{\theta_a < L_a\}\right] \\
&\leqslant \beta + \mathrm{EP}(\theta_A < L_A \mid \mathscr{D}) \\
&\leqslant \beta + \mathrm{P}\{\theta_A < L_A\} \leqslant \beta + \alpha/2 + o(1),
\end{aligned}$$

where the first equality holds by the law of iterated expectations; the second by the fact that, given $\mathscr{D}$, L is fixed but $A \sim U(\mathscr{A})$; the second inequality holds by definition of L:

$$\frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}} 1\{L_a < L\} \leqslant \frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}} 1\{L_a < \overline{Q}_\beta[L_A \mid \mathrm{Data}]\} \leqslant \beta,$$

by the definition of the upper quantile; and the third by the same argument as the second equality, the penultimate inequality holds by the law of iterated expectations, and the last inequality holds by assumption (4.8). We conclude similarly

$$\mathrm{P}(\theta_A > U) \leqslant \beta + \mathrm{P}\{\theta_A > U_A\} \leqslant \beta + \alpha/2 + o(1).$$

The asserted result follows. ∎

Proof of Theorem 4.3. In the proof let $z = \Phi^{-1}(1-\alpha/2)$, and use $M_{\mathscr{A}}[\cdot]$ and $Q_{\mathscr{A}}[\cdot]$ as short hand for $\mathrm{M}[\cdot|\mathrm{Data}]$ and $\mathrm{Q}[\cdot|\mathrm{Data}]$, respectively, with overlined and underlined versions defined similarly.

We first note

$$\begin{aligned}
&\mathrm{P}(\theta^* \notin [L,U]) = \mathrm{P}(\theta^* > M_{\mathscr{A}}(\widehat{\theta}_A + z\widehat{\sigma}_A)) + \mathrm{P}(\theta^* < M_{\mathscr{A}}(\widehat{\theta}_A - z\widehat{\sigma}_A)) \\
&= \mathrm{P}(0 > M_{\mathscr{A}}(\widehat{\theta}_A - \theta^* + z\widehat{\sigma}_A)) + \mathrm{P}(0 < M_{\mathscr{A}}(\widehat{\theta} - \theta^* - z\widehat{\sigma}_A)).
\end{aligned}$$

To show part (i) with $\beta = 1/2$,

$$\begin{aligned}
\mathrm{P}(0 < M_{\mathscr{A}}(\widehat{\theta} - \theta^* - z\widehat{\sigma}_A)) &\leqslant \mathrm{P}\left(\frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}} 1\left(\widehat{\sigma}_a^{-1}(\widehat{\theta}_a - \theta^*) > z\right) \geqslant 1/2\right) \\
&\leqslant 2\mathrm{E}\left[\frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}} 1\left(\widehat{\sigma}_a^{-1}(\widehat{\theta}_a - \theta^*) > z\right)\right] \\
&= 2\mathrm{E}\mathrm{P}\left(\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta^*) > z \mid \text{Data}\right) \\
&= 2\mathrm{P}\left(\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta^*) > z\right) \\
&\leqslant 2\mathrm{P}\left(\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta_A) > z - r_N\right) + 2\gamma_N''' \\
&\leqslant 2(1 - \Phi(z - r_N)) + 2\gamma_N' + 2\gamma_N''' \\
&\leqslant 2\alpha/2 + 2r_N/\sqrt{2\pi} + 2\gamma_N' + 2\gamma_N''',
\end{aligned}$$

where the first inequality follows from the definition of the numerical median, the second from the Markov inequality; the first equality holds by $A \sim U(\mathscr{A})$ given Data, the second by the law of iterated expectations; the third inequality holds by the concentration condition (R3) and the union bound, the penultimate inequality holds by the approximation normality conditions (R1), and the last from the properties of Φ.

We derive similarly that

$$\mathrm{P}(0 > M_{\mathscr{A}}(\widehat{\theta}_A - \theta^* + z\widehat{\sigma}_A)) \leqslant 2\alpha/2 + 2r_N/\sqrt{2\pi} + 2\gamma_N' + 2\gamma_N'''.$$

Therefore the part (i) holds for the term

$$o(1) := 4r_N/\sqrt{2\pi} + 4(\gamma_N' + \gamma_N''').$$

To show part (ii), from the analysis of part (i),

$$\mathrm{P}(0 < M_{\mathscr{A}}(\widehat{\theta} - \theta^* - z\widehat{\sigma}_A)) \leqslant \mathrm{P}\left(\frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}} 1\left(\widehat{\sigma}_a^{-1}(\widehat{\theta}_a - \theta^*) > z\right) \geqslant 1/2\right).$$

Then we bound

$$T = \frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}} 1\left(\widehat{\sigma}_a^{-1}(\widehat{\theta}_a - \theta^*) > z\right) \leqslant T_1 + T_2$$

$$T_1 := \frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}} 1\left(\widehat{\sigma}_a^{-1}(\widehat{\theta}_a - \theta_a) > z - r_N\right), \quad T_2 := \frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}} 1\left(\widehat{\sigma}_a^{-1}(\theta_a - \theta^*) > r_N\right).$$

By the union bound

$$\mathrm{P}(T \geqslant 1/2) \leqslant \mathrm{P}\left(T_1 > 1/2 - 2\sqrt{\gamma_N'''}\right) + \mathrm{P}\left(T_2 \geqslant \sqrt{\gamma_N'''}\right).$$

Then for $\beta_N = 1/2 - 2\sqrt{\gamma_N'''}$,

$$\begin{aligned}
\mathrm{P}(T_1 > \beta_N) &\leqslant \mathrm{P}\left(\overline{\mathrm{Q}}_{\beta_N,\mathscr{A}}[\widehat{\sigma}_A^{-1}(\theta_A - \widehat{\theta}_A)] < -z + r_N\right) \\
&\leqslant \mathrm{P}\left(\mathrm{Q}_{\beta_N,\mathscr{A}}[\widehat{\sigma}_A^{-1}(\theta_A - \widehat{\theta}_A)] < -z + r_N\right) \\
&\leqslant \mathrm{P}\left(\widehat{\sigma}_A^{-1}(\theta_A - \widehat{\theta}_A) < -z + r_N\right) + \gamma_N'' \\
&\leqslant \Phi(-z + r_N) + \gamma_N' + \gamma_N'' \\
&\leqslant \alpha/2 + r_N/\sqrt{2\pi} + \gamma_N' + \gamma_N'',
\end{aligned}$$

where first inequality holds by the definition of the numerical quantile, the third by the concentration of medians assumption (R2), and the fourth by the approximate normality (R1).

Also, by Markov inequality

$$\mathrm{P}\left(T_2 \geqslant \sqrt{\gamma_N'''}\right) \leqslant \mathrm{E}T_2/\sqrt{\gamma_N'''} = \mathrm{P}\left(\sigma_A^{-1}|\theta_A - \theta^*| > r_N\right)/\sqrt{\gamma_N'''} \leqslant \gamma_N'''/\sqrt{\gamma_N'''},$$

where we are using (R3) and the relation

$$\begin{aligned}
\mathrm{E}T_2 &= \mathrm{E}\left[\frac{1}{\mathscr{A}}\sum_{a\in\mathscr{A}} 1\left(\sigma_a^{-1}|\theta_a - \theta^*| > r_N\right)\right] \\
&= \mathrm{EP}(\sigma_A^{-1}|\theta_A - \theta^*| > r_N \mid \mathrm{Data}) = \mathrm{P}\left(\sigma_A^{-1}|\theta_A - \theta^*| > r_N\right),
\end{aligned}$$

using our formalism that $A \sim U(\mathscr{A})$ independently of Data.

Collecting terms conclude

$$\mathrm{P}(0 < M_{\mathscr{A}}(\widehat{\theta} - \theta^* - z\widehat{\sigma}_A)) \leqslant \alpha/2 + r_N/\sqrt{2\pi} + \gamma_N' + \gamma_N'' + \sqrt{\gamma_N'''}.$$

We derive similarly that

$$\mathrm{P}(0 > M_{\mathscr{A}}(\widehat{\theta}_A - \theta^* + z\widehat{\sigma}_A)) \leqslant \alpha/2 + r_N/\sqrt{2\pi} + \gamma_N' + \gamma_N'' + \sqrt{\gamma_N'''}.$$

Therefore the part (ii) holds for the term

$$o(1) = 2\left(r_N/\sqrt{2\pi} + \gamma_N' + \gamma_N'' + \sqrt{\gamma_N'''}\right).$$

To show claim (iii) note that by construction $L \leqslant \widehat{\theta} \leqslant U$ and the coverage event $L \leqslant \theta^* \leqslant U$ implies that $|\widehat{\theta} - \theta^*| \leqslant U - L$. ∎

Concentration of Estimands Around Their Median. The purpose of this section is to demonstrate that the concentration assumptions made in the inference section are plausible. To show this we focus on the BLP parameter

$$\theta_A = \frac{\mathrm{Cov}_Z(s_0(Z), S_A(Z)}{\mathrm{Var}_Z S_A(Z))},$$

where A is a uniform variable on $\mathscr{A}$, and the variance and covariance are taken with respect to the marginal distribution of Z. We want to show the concentration of this parameter around

$$\theta^* = \text{Med}[\theta_A \mid \text{Data}].$$

We show the difference can be bounded using measures of estimation and algorithmic stabilities; we derive inspiration from Chernozhukov et al. (2021) and Chen et al. (2022)).

In what follows, we assume the same set-up as in the main text, in particular the exchangeability.

Estimation Stability or Pseudo-Consistency. Statistical learning theory, for example, results in Section 5, provides bounds on estimation errors of the form:

$$\mathrm{E}\mathrm{E}_Z(S_A(Z) - s_\bullet(Z))^2 = \mathrm{E}(S_A(Z) - s_\bullet(Z))^2 \leqslant R^2_{|A|},$$

where $s_\bullet$ is a fixed "pseudo-true" function that does not depend on A, and this function does not have to be the CATE s_0 in the misspecified case. Here E_Z denotes the expectation taken with respect to the marginal distribution of Z. For example, in Section 5, $s_\bullet$ minimizes the mean square approximation error

$$\min_{s \in \mathscr{S}} \mathrm{E}[s_0(Z) - s(Z)]^2 = \mathrm{E}[s_0(Z) - s_\bullet(Z)]^2,$$

but $s_\bullet$ above does not to be defined in this way.

Define the BLP parameter corresponding to $s_\bullet$ as:

$$\theta_\bullet = \frac{\mathrm{Cov}_Z(s_0(Z), s_\bullet(Z))}{\mathrm{Var}_Z\, s_\bullet(Z))}.$$

This is a fixed estimand.

If $R_{|A|} \to 0$ as $|A| \to \infty$, then S_A converges to the pseudo-true value $s_\bullet$. We call this property "pseudo"-consistency. The lemma shows that in this case, the random estimand θ_A approaches $\theta_\bullet$ at the rate $R_{|A|}$.

Lemma B.1 (Concentration from "Pseudo"-Consistency)**.** *Assume that $S_a \in \mathscr{S}$ for all $a \in \mathscr{A}$ and $s_\bullet \in \mathscr{S}$, that the elements of $\mathscr{S}$ and s_0 are all bounded above by a finite constant K, and that $\mathrm{Var}_Z S(Z)$ is bounded below by a positive constant $k > 0$ for all $S \in \mathscr{S}$. Then*

$$\mathrm{E}|\theta_A - \theta_\bullet| \leqslant C_{K,k}[R_{|A|} \wedge 1]$$

where $C_{K,k}$ is a numeric constant that only depends on K and k.

Concentration under Algorithmic Stability. On the other hand, algorithmic or statistical influence analysis often implies that

$$\mathrm{E}\mathrm{E}_Z(S_A(Z) - S_{A'}(Z))^2 \leqslant R'^2_{|A|},$$

where A and A' are independent uniform variables on $\mathscr{A}$. To explain the notion, let M and M' be the complements of A and A' relative to $\{1, ..., N\}$. The symmetric difference between A and A' is

$M \cap M'$. If the latter set is small in cardinality relative to the cardinality of A, then we would expect S_A and $S_{A'}$ not to differ if the machine producing S's is a smooth function of data. The definition above provides one way to measure this stability. We provide further discussion below.

By triangle inequality, the algorithmic stability can be bounded by estimation stability:

$$\sqrt{\mathrm{EE}_Z(S_A(Z) - S_{A'}(Z))^2} \leqslant 2\sqrt{\mathrm{EE}_Z(S_A(Z) - S_\bullet(Z))^2}$$

Therefore algorithmic stability is more general.

Lemma B.2 (Concentration from Algorithmic Stability)**.** *Suppose the assumptions of the previous lemma hold. Then if $R'_{|A|} \to 0$ as $|A| \to \infty$, then*

$$\mathrm{E}|\theta_A - \theta_{A'}| \leqslant C_{K,k}[R'_{|A|} \wedge 1],$$

where $C_{K,k}$ is a numeric constant that only depends on K and k.

Putting it Together: Concentration Around Median. The following result shows that the desired concentration condition holds if either estimation stability or algorithmic stability is strong enough.

Lemma B.3 (Stability of Median Target from Estimation or Algorithmic Stability)**.** *Suppose the assumptions of the previous lemma hold. Then*

$$\mathrm{E}|\theta_A - \theta^*| \leqslant \mathrm{E}|\theta_A - \theta_{A'}| \wedge \mathrm{E}|\theta_A - \theta_\bullet|.$$

Therefore, if

$$\sqrt{n}C_{K,k}[R'_{|A|} \wedge R_{|A|}] \leqslant \delta_N \tag{B.2}$$

for $\delta_N \to 0$ as $N \to \infty$, then

$$\mathrm{P}(\sqrt{n}|\widehat{\theta}_A - \theta^*| > \sqrt{\delta_N}) \leqslant \sqrt{\delta_N}.$$

The latter implies the condition we want provided $\widehat{\sigma}_A/\sqrt{n} + \sqrt{n}/\widehat{\sigma}_A = O_P(1)$.

We conclude here with some comparisons of the two notions of stability. Estimation stability readily follows from the available statistical learning theory. In particular $R^2_{|A|}$ scales like $d/|A|$ where d is the intrinsic dimension of the function class $\mathscr{S}$, as we discussed in Section 5. Therefore, n needs to be much smaller than $d(N-n)$ to satisfy the last condition of the last lemma.

Algorithmic stability does not require estimation stability, even though the latter property seems quite mild. On the other hand, its characterizations are not well-studied and are much less available. See Chernozhukov et al. (2021) for analysis of constrained Lasso and Ridge that is applicable here; see also Chen et al. (2022) for leave-one-out stability analysis for bagged estimators over the subsamples (this analysis requires extension to the present framework).

It is useful to give a simple example to compare the two measures of stability. If $S_A(Z)$'s are generated by linear least squares $Z'\widehat{\beta}_A$ with $d = \dim(Z)$, then we have a crude upper bound on the

algorithmic stability bound $R_A'^2$ scaling like $nd/(N-n)^2$. This is generally smaller than R_A scaling like $d/(N-n)$. It implies a weaker same qualitative requirement on n: n needs to be smaller than $\sqrt{d}(N-n)$ to satisfy the condition (B.2) of Lemma B.3.

Proof of Lemmas B.1- B.3. To show Lemma B.1, it is convenient to define $f^o(Z) = f(Z) - \mathrm{E}_Z f(Z)$. Then, using the boundedness assumption we have

$$\begin{aligned} |\mathrm{Cov}_Z(s_0(Z), S_A(Z)) - \mathrm{Cov}_Z(s_0(Z), s_\bullet(Z))| &= |\mathrm{E}_Z[s_0^o(Z)S_A(Z)] - \mathrm{E}[s_0^o(Z)s_\bullet(Z)]| \\ &\leqslant K\mathrm{E}_Z|S_A(Z) - s_\bullet(Z))|, \end{aligned}$$

$$\begin{aligned} |\mathrm{Var}_Z(S_A(Z)) - \mathrm{Var}_Z(s_\bullet(Z))| &= |\mathrm{E}_Z(S_A^o(Z))^2 - \mathrm{E}_Z(s_\bullet^o(Z))^2| \\ &\leqslant \mathrm{E}_Z|S_A^o(Z) + s_\bullet^o(Z))||S_A^o(Z) - s_\bullet^o(Z))| \\ &\leqslant 2K\mathrm{E}_Z|S_A^o(Z) - s_\bullet^o(Z))|. \end{aligned}$$

Then using elementary inequalities and boundedness assumptions conclude

$$|\theta_A - \theta_\bullet| \leqslant (k^{-1}K + 2k^{-2}K^2)\mathrm{E}_Z|S_A(Z) - s_\bullet(Z)|$$

Taking expectation over A,

$$\mathrm{E}|\theta_A - \theta_\bullet| \leqslant (k^{-1}K + 2k^{-2}K^2)\mathrm{E}\mathrm{E}_Z|S_A(Z) - s_\bullet(Z))| \leqslant C_{K,k}R_{|A|},$$

where the last inequality follows from the norm inequality.

Lemma B.2 follows analogously, replacing $s_\bullet$ with $S_{A'}$ to obtain

$$|\theta_A - \theta_{A'}| \leqslant (k^{-1}K + 2k^{-2}K^2)\mathrm{E}_Z|S_A(Z) - S_{A'}(Z)|$$

Taking expectation over (A, A'), we obtain:

$$\mathrm{E}|\theta_A - \theta_{A'}| \leqslant (k^{-1}K + 2k^{-2}K^2)\mathrm{E}\mathrm{E}_Z|S_A(Z) - S_{A'}(Z))| \leqslant C_{K,k}R'_{|A|},$$

where the last inequality follows from the norm inequality.

To show Lemma B.3, we note that

$$\begin{aligned} \mathrm{E}|\theta_A - \theta^*| &= \mathrm{E}\mathrm{E}\left[\frac{1}{\mathscr{A}}\sum_{a\in\mathscr{A}}|\theta_a - \theta^*| \mid \mathrm{Data}\right] \\ &\leqslant \mathrm{E}\mathrm{E}\left[\frac{1}{\mathscr{A}}\sum_{a\in\mathscr{A}}|\theta_a - \theta_\bullet| \mid \mathrm{Data}\right] = \mathrm{E}|\theta_A - \theta_\bullet|, \end{aligned}$$

where the first property holds by the law of iterated expectation and by $A \sim U(\mathscr{A})$ independently of the Data, the inequality holds by definition of θ^* as the median of the sample $\{\theta_a : a \in \mathscr{A}\}$, and the last equality holds by iterating expectations again.

Similarly, we note

$$\begin{aligned}
\mathrm{E}|\theta_A - \theta^*| &= \mathrm{EE}\left[\frac{1}{\mathscr{A}}\sum_{a\in\mathscr{A}}|\theta_a - \theta^*| \mid \mathrm{Data}\right] \\
&= \mathrm{E}\frac{1}{|\mathscr{A}|}\sum_{a'\in\mathscr{A}}\mathrm{E}\left[\frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}}|\theta_a - \theta^*| \mid \mathrm{Data}\right] \\
&\leqslant \mathrm{E}\frac{1}{|\mathscr{A}|}\sum_{a'\in\mathscr{A}}\mathrm{E}\left[\frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}}|\theta_a - \theta_{a'}| \mid \mathrm{Data}\right] \\
&= \mathrm{E}\left[\frac{1}{|\mathscr{A}|}\sum_{a'\in\mathscr{A}}\frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}}|\theta_a - \theta_{a'}| \mid \mathrm{Data}\right] = \mathrm{E}|\theta_A - \theta_{A'}|,
\end{aligned}$$

where the first property holds by the law of iterated expectation and by $A \sim U(\mathscr{A})$ independently of the Data, the inequality holds by definition of θ^* as the median of $\{\theta_A : a \in \mathscr{A}\}$, the last equality holds by iterating expectations and independence of A and A'.

Finally, the second claim of the Lemma follows by the Markov inequality. ∎

Appendix C. Deferred Discussion and Proofs for Section 5

Comment C.1 (Extensions of Theorem 5.2)**.** The result of Theorem 5.2 follows from combining Theorem 3 of Liang et al. (2015) with Theorem 5.1. We assumed boundedness conditions to make the statement as simple as possible. Bounds on excess risk without the boundedness conditions follow from Theorem 4 in Liang, Rakhlin, and Sridharan (2015). If the class $\mathscr{S}$ is not convex, similar performance bound is attained by Audibert (2007)'s "star" modification of the optimizer S, by Theorems 3 and 4 by Liang et al. (2015). We refer to Liang et al. (2015) and Vijaykumar (2021) for detailed general discussion.

Using Losses (A) **and** (B) **for Choosing the Best ML Method.** The loss functions (A) and (B) can also be used to aggregate several learning methods using separate auxiliary subsets.[1] To fix ideas, suppose we have a set of methods giving scores S_k and B_k, $k = 1, ..., K$, where K is small, obtained using a subset $A_1 \subset A$. Then, we can combine these scores into

$$S(Z) = \sum_{k=1}^{K} \lambda_k^S S_k(Z); \;\; B(Z) = \sum_{k=1}^{K} \lambda_k^B B_k(Z),$$

and then we can learn the weights λ^S and λ^B by optimizing the loss functions (A) or (B) evaluated on subset A_2, such that A_2 does not overlap with A_1, e.g. $A_2 = A \setminus A_1$.

[1] This is in contrast to our main proposal, where we choose the best ML method based on goodness-of-fit measures in the second stage.

Comment C.2 (Learning Guarantee for Aggregation)**.** Let $\widehat{\lambda}^S$ and $\widehat{\lambda}^B$ denote the weights learned in this way. Then, another application of the results of Liang et al. (2015) for linear regression, under the assumption that $|Y|$, $|B|$, $|S|$, $|w(Z)|$ are all bounded by R, gives the excess risk bound:

$$\mathrm{E}\Big[s_0(Z) - \sum_{k=1}^K \widehat{\lambda}_k^S S_k(Z)\Big]^2 - \overbrace{\min_{\{\lambda_k^S\}_{k=1}^K} \mathrm{E}\Big[s_0(Z) - \sum_{k=1}^K \lambda_k^S S_k(Z)\Big]^2}^{\text{oracle risk}} \leqslant C_R K/|A_2|,$$

where C_R is some constant that depends on R and $|A_2|$ is the sample size used to perform the aggregation. Thus, if the right-hand side is small, the excess risk of this estimator relative to the oracle aggregation method is negligible. Since the oracle aggregation risk here is weakly smaller than the oracle risk of choosing the best prediction rule $\min_k \mathrm{E}[s_0(Z) - S_k(Z)]^2$, convex aggregation here is approximately better than choosing the best ML method.

Comment C.3 (Large K)**.** The method above gives a small excess risk when $K/|A_2|$ is small; otherwise, the excess risk can be large. In the latter case, we can apply Lasso to select a sparse linear combination of rules, and the sharp bounds on excess risk follow from Example 4 in Koltchinskii et al. (2011). Finally, we may choose the "best" machine learning algorithm using objective functions (A) and (B) evaluated on the data subset A_2. Results of Wegkamp et al. (2003) imply certain good guarantees for the "best" approach, but sharp bounds on the excess risk that scale like $\log K/|A_2|$ only hold for the "star" modification of the "best" method (Audibert, 2007).

Proof of Theorem 5.1. For the objective (B), write $Y = b_0(Z) + Ds_0(Z) + \varepsilon$, where $\mathrm{E}[\varepsilon \mid D, Z] = 0$. Then

$$YH = \{Hb_0(Z) + (HD-1)s_0(Z)\} + s_0(Z) + \varepsilon H,$$

where the first term can be expressed as:

$$\{Hb_0(Z) + (HD-1)s_0(Z)\} = H(b_0(Z) + (1-p(Z))s_0(Z)) = H\bar{b}_0(Z).$$

So that we can decompose:

$$YH - b(Z)H - s(Z) = \{H(\bar{b}_0(Z) - b(Z))\} + \{s_0(Z) - s(Z)\} + \varepsilon H.$$

Then the result follows taking the square and expectation, by using (i) orthogonality of the three terms in the decomposition above:

$$\mathrm{E}[\varepsilon H^2(\bar{b}_0(Z) - b(Z))] = 0, \ \ \mathrm{E}[\varepsilon H(s_0(Z) - s(Z))] = 0, \ \ \mathrm{E}[H(\bar{b}_0(Z) - b(Z))(s_0(Z) - s(Z)] = 0,$$

where the last relation follows from $\mathrm{E}[H \mid Z] = 0$, and (ii) also noting that $\mathrm{E}[H^2|Z] = w(Z)$.

For the objective (A), write similarly,

$$Y - b(Z) - (D - p(Z))s(Z) = [\tilde{b}_0(Z) - b(Z)] + [D - p(z)](s_0(Z) - s(Z)) + \varepsilon,$$

and then conclude that the three terms are orthogonal to each other. The result follows by completing the square and taking expectation, where we also observe that

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

since $\mathrm{E}[w(Z)(D-p(Z))^2 \mid Z] = 1$. ∎

Proof of Theorem 5.2. We demonstrate the result for type B loss; the demonstration for type A follows similarly. Application of Theorem 3 of Liang et al. (2015) gives the following bound on the excess risk $\mathscr{R}$ of the estimator (B,S):

$$0 \leqslant \mathscr{R} := \mathrm{E}[YH - B(Z)H - S(Z)]^2 - \mathrm{E}[YH - b_\bullet(Z)H - s_\bullet(Z)]^2 \leqslant C_K \mathscr{R}^o(A, \mathscr{H}, c_K),$$

where $(b_\bullet, s_\bullet)$ minimize $\mathrm{E}[YH - b_\bullet(Z)H - s_\bullet(Z)]^2$ over $b \in \mathscr{B}$ and $s \in \mathscr{S}$, and C_K and c_K are positive constants that only depend on K, and $\mathscr{H} := 4(H\mathscr{B} + \mathscr{S})$. Theorem 5.1 then implies that

$$\begin{aligned} \mathscr{R} &= \mathrm{E}[s_0(Z) - S(Z)]^2 - \mathrm{E}[s_0(Z) - s_\bullet(Z)]^2 \\ &+ \mathrm{E}[w(Z)(\bar{b}_0(Z) - B(Z))]^2 - \mathrm{E}[w(Z)(\bar{b}_0(Z) - b_\bullet(Z))]^2, \end{aligned}$$

where the second term is non-negative. Therefore,

$$\mathrm{E}[s_0(Z) - S(Z)]^2 - \mathrm{E}[s_0(Z) - s_\bullet(Z)]^2 \leqslant C_K \mathscr{R}^o(A, \mathscr{H}, c_K).$$

The lower bound

$$\mathrm{E}[s_0(Z) - S(Z)]^2 - \mathrm{E}[s_0(Z) - s_\bullet(Z)]^2 \geqslant \mathrm{E}[S(Z) - s_\bullet(Z)]^2$$

follows from Pythagorian inequality for obtuse triangles and the fact that $s_\bullet$ minimizes $\mathrm{E}[s_0(Z) - S(Z)]^2$ over the convex set $\mathscr{S}$. ∎

APPENDIX D. GAUSSIAN APPROXIMATION FOR SPLIT-SAMPLE LEAST SQUARES UNIFORMLY OVER CONVEX SETS AND IN P.

We present a set-up that covers not only the split-sample least square estimators of the main text, but also other potential cases of interest. Let W denote a generic data vector. All the linear regressions or mean estimators used on the main sample M could be viewed as ordinary least squares with a suitable definition of W.

Throughout we assume that $\{(W_i)\}_{i=1}^N$ are i.i.d. copies of vector W that has law P. We abbreviate $(\mathscr{D}_A, \mathscr{D}_M) := (\mathrm{Data}_A, \mathrm{Data}_M)$. There is a learning algorithm that inputs $\mathscr{D}_A$ and outputs a map $f(\cdot; \mathscr{D}_A)$, which maps the support of W to $\mathbb{R}^{d+1}$ for a fixed d. This map defines the split-specific outcome and regressors:

$$(Y_{A,i}, X_{A,i}) = f(W_i; \mathscr{D}_A), \quad i \in M.$$

Let $\widehat{\beta}_A$ be a solution to $\mathbb{E}_{N,M}[X_{A,i}\widehat{\varepsilon}_{A,i}] = 0$ for $\widehat{\varepsilon}_{A,i} = Y_{A,i} - X'_{A,i}\widehat{\beta}_A$. Let $\widehat{V}_A$ denote the Eicker-Huber-White sandwich

$$\widehat{V}_A := (\mathbb{E}_{N,M}X_{A,i}X'_{A,i})^{-1}\mathbb{E}_{N,M}\widehat{\varepsilon}^2_{A,i}X_{A,i}X'_{A,i}(\mathbb{E}_{N,M}X_{A,i}X'_{A,i})^{-1},$$

whenever it exists.

Fix some positive finite constants c and C. Let β_A denote a solution to $\mathrm{E}[X_A\varepsilon_A] = 0$, for $\varepsilon_A = Y_A - X'_A\beta_A$, if it exists. And let

$$V_A := (\mathrm{E}_P[X_AX'_A \mid \mathscr{D}_A])^{-1}\mathrm{E}_P[\varepsilon^2_AX_AX'_A \mid \mathscr{D}_A](\mathrm{E}_P[X_AX'_A \mid \mathscr{D}_A])^{-1},$$

if it exists. Let $\mathscr{E}_{A,N}$ be the event that

$$\mathrm{E}_P|Y_A|^{4+\delta} + \mathrm{E}_P[\|X_A\|^{4+\delta} \mid \mathscr{D}_A] \leqslant C, \quad \min_{\|a\|=1}\mathrm{E}_P[(a'X_A)^2 \mid \mathscr{D}_A] > c.$$

On this event β_A and ε_A are well defined. Let $\mathscr{E}'_{A,N} \subset \mathscr{E}_{A,N}$ be the event such that

$$\min_{\|a\|=1}\mathrm{E}_P[(\varepsilon_Aa'X_A)^2 \mid \mathscr{D}_A] > c.$$

On this event V_N is well-defined. Let $CS(\mathbb{R}^d)$ denote the collection of the convex sets in $\mathbb{R}^d$.

We observe that, by the i.i.d. sampling and $A \sim U(\mathscr{A})$ independently of Data, $(\mathscr{D}_A, \mathscr{D}_M)$ has the same distribution as $(\mathscr{D}_a, \mathscr{D}_m)$, for a fixed partition (a,m). This is an exchangeability property. Therefore, we can fix (A,M) to be a fixed partition $\{a,m\}$ in what follows. Moreover $(X_{a,i}, Y_{a,i})_{i=1}^{N-m}$ are i.i.d. conditional on $\mathscr{D}_a$. These observations simplify the verification of the following result.

Lemma D.1 (Gaussian Approximation)**.** *Using the setup above, let γ_N be a sequence of positive constants tending to zero. Suppose that for all $P \in \mathscr{P}$, we have $\mathrm{P}_P(\mathscr{E}'_{N,A}) \geqslant 1 - \gamma_N$. Then, uniformly in $P \in \mathscr{P}$, as $(n,N) \to \infty$:*

$$\sup_{R\in CS(\mathbb{R}^d)}\left|\mathrm{P}_P[\widehat{V}_A^{-1/2}(\widehat{\beta}_A - \beta_A) \in R \mid \mathscr{D}_A] - \mathrm{P}(N(0,I_d) \in R)\right| \xrightarrow{\mathrm{P}_P} 0,$$

$$\sup_{R\in CS(\mathbb{R}^d)}\left|\mathrm{P}_P[\widehat{V}_A^{-1/2}(\widehat{\beta}_A - \beta_A) \in R] - \mathrm{P}(N(0,I_d) \in R)\right| \longrightarrow 0,$$

and the same results hold with $\widehat{V}_A$ replaced by V_A; moreover, $\widehat{V}_NV_N^{-1} \to_{\mathrm{P}_P} I$ both conditional on $\mathscr{D}_N$ and unconditionally.

Proof of Lemma D.1. It suffices to demonstrate the argument for an arbitrary sequence $\{P_N\}$ in $\mathscr{P}$. Let

$$\widehat{t}_A := \widehat{V}_A^{-1/2}(\widehat{\beta}_A - \beta_A),\ t_A := V_A^{-1/2}(\widehat{\beta}^o_A - \beta_A),\ \widehat{\beta}^o_A := [\mathrm{E}X_AX'_A]^{-1}\mathbb{E}_{N,M}X_AY_A.$$

Consider the event $\mathscr{E}''_{N,A} \subseteq \mathscr{E}'_{N,A}$ such that:

$$\mathscr{E}''_{N,A} = \left\{(\widehat{t}_A, t_A, \widehat{V}_N) \text{ exist and } \|\widehat{t}_A - t_A\| + \|\widehat{V}_N - V_N\| \leqslant r_N\right\}.$$

It follows from the standard arguments for asymptotic theory for least squares under i.i.d. sampling of data arrays $(Y_{A,i}, X_{A,i})_{i=1}^N$, e.g. Gallant and White (1988), that there exists a sequence of positive constants $\{r_N, \delta_N\} \searrow 0$ such that $\mathrm{P}\left(\mathscr{E}''_{N,A} \mid \mathscr{D}_A\right) \geqslant 1 - \delta_N$ on the event $\mathscr{E}'_{N,A}$. Therefore by the union bound

$$\mathrm{P}\left(\mathscr{E}''_{N,A}\right) \geqslant 1 - \delta_N - \gamma_N, \tag{D.1}$$

for γ_N defined in the statement of the lemma. For $r > 0$ let $R^r = \{x \in \mathbb{R}^d : d(x,R) \geqslant r\}$ and $R^{-r} = \{x \in R : d(x, \mathbb{R}^d \setminus R) \geqslant r\}$, where $d(x,R) := \min_{x' \in R} \|x' - x\|$. Note that R^{-r} can be an empty set. Then, on the event $\mathscr{E}''_{N,A}$,

$$\begin{aligned}
\mathrm{P}\left(\widehat{t}_A \in R \mid \mathscr{D}_A\right) &\geqslant \mathrm{P}\left(t_A \in R^{-r_N} \mid \mathscr{D}_A\right) \\
&\geqslant \mathrm{P}(N(0, I_d) \in R^{-r_N}) - B_N d^{1/4}/\sqrt{n}, \\
&\geqslant \mathrm{P}(N(0, I_d) \in R) - 4d^{1/4} r_N - B_N d^{1/4}/\sqrt{n},
\end{aligned}$$

where $B_N = C' \mathrm{E}[\|V_N^{-1/2} X_A \varepsilon_A\|^3 \mid \mathscr{D}_A]$, where C' is a numerical constant. The second inequality follows by the Bentkus bounds (Bentkus, 2003; Raič, 2019), which extend the Berry-Essen bounds to the multidimensional case, and the last inequality follows from the Ball's reverse isoperimetric inequality of the standard Gaussian vector (Ball, 1991). It follows similarly that

$$\begin{aligned}
\mathrm{P}\left(\widehat{t}_A \in R \mid \mathscr{D}_A\right) &\leqslant \mathrm{P}\left(t_A \in R^{r_N} \mid \mathscr{D}_A\right) \\
&\leqslant \mathrm{P}(N(0, I_d) \in R^{r_N}) + B_N/\sqrt{n}, \\
&\leqslant \mathrm{P}(N(0, I_d) \in R) + 4d^{1/4} r_N + B_N d^{1/4}/\sqrt{n}.
\end{aligned}$$

Since R above is arbitrary convex subset of $\mathbb{R}^d$, we have that on the event $\mathscr{E}''_{N,A}$:

$$\begin{aligned}
&\sup_{R \in CS(\mathbb{R}^d)} \left|\mathrm{P}\left(\widehat{t}_A \in R \mid \mathscr{D}_A\right) - \mathrm{P}(N(0, I_d) \in R)\right| \\
&\leqslant \sup_{R \in CS(\mathbb{R}^d)} \left|\mathrm{P}\left(\widehat{t}_A \in R \mid \mathscr{D}_A\right) - \mathrm{P}(N(0, I_d) \in R)\right| \\
&\leqslant 4d^{1/4} r_N + d^{1/4} B_N/\sqrt{n}.
\end{aligned}$$

Using Holder inequalities, we can check that $B_N \leqslant B$ on the event $\mathscr{E}''_N$, for some constant B that depends only on (c, C, d, δ). The first claim follows combining this inequality with (D.1).

To show that second claim note that

$$\begin{aligned}
&\sup_{R \in CS(\mathbb{R}^d)} \left|\mathrm{EP}\left(\widehat{t}_A \in R \mid \mathscr{D}_A\right) - \mathrm{P}(N(0, I_d) \in R)\right| \\
&\quad \leqslant 4d^{1/4} r_N + d^{1/4} B_N/\sqrt{n} + (1 - \mathrm{P}(\mathscr{E}''_{N,A})) \leqslant 4d^{1/4} B_N/\sqrt{n} + \gamma_N + \delta_N.
\end{aligned}$$

Finally, $\|\widehat{V}_N V_N^{-1} - I\| \leqslant \|V_N^{-1}\| r_N \leqslant c r_N$ conditional on $\mathscr{D}_A$ and on the event $\mathscr{E}''_N$. The conditional convergence claim follows from (D.1). It then follows that $\mathrm{EP}(\|\widehat{V}_N V_N^{-1} - I\| \leqslant c r_N \mid \mathscr{D}_A) \geqslant \mathrm{P}(\mathscr{E}_{N,A}) \geqslant 1 - \delta_N - \gamma_N$. ∎

APPENDIX E. EXTENSION TO UNBIASED SIGNAL FRAMEWORK

Our inference approach generalizes to any problem of the following sort, studied in Semenova and Chernozhukov (2021) using more conventional inference approaches. 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 $s_0(Z)$. In experimental settings the unbiased signals arise from multiplying an outcome with a Riesz representer for the effect of interest, as we explain below.

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 split-sample robust inference on functionals of the BLP and GATs.

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 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 $s_0(z) = \mathrm{E}[g'(D,Z) \mid Z = z]$, where $g'(d,z) = \partial g(d,z)/\partial x$ and $g(d,z) = \mathrm{E}[Y \mid D = d, Z = z]$. In the context of demand analysis, Y is the log of individual demand, D is the log-price of a product, and Z includes prices of other products and individual characteristics. Then, the unbiased signal is given by $\tilde{Y} = -Y[\partial \log p(D \mid Z)/\partial d]$, where $p(\cdot \mid \cdot)$ is the conditional density function of D given Z, which is known if D is generated experimentally conditional on Z. That is, using the integration by parts formula, $\mathrm{E}[\tilde{Y} \mid Z] = s_0(Z)$ under mild conditions on the density.

Example 3 (Other Causal Objects)**.** Chernozhukov et al. (2018) presented a number of other examples where a causal parameter of interest $s_0(Z)$ is expressed as a linear functional of the regression function $g(D,Z) = \mathrm{E}[Y \mid D,Z]$, that is, $s_0(Z) = \mathrm{E}[m(Y,D,Z,g) \mid Z]$, for some moment function m that is linear in g; this includes the examples above for instance. Then, under mild conditions, we can construct an unbiased signal

$$\tilde{Y} = Y\alpha(D,Z), \tag{E.1}$$

where $\alpha(D,Z)$ is the Riesz Representer, such that $\mathrm{E}[Y\alpha(D,Z) \mid Z] = s_0(Z)$. For instance, in CATE, the representer $\alpha(D,Z)$ is the HT transform H; in Example 2, $\alpha(D,Z) = [\partial \log p(X \mid Z)/\partial x]$; and in Example 1, the representer $\alpha(D,Z)$ is just 1. In addition to these examples, other examples that fall in this framework include causal effects from transporting covariates and from distributional shift in covariates induced by policies; see Chernozhukov et al. (2018) for more details. In experimental settings, $\alpha(D,Z)$ will typically be known.

The noise reduction strategies, like the ones we used in the context of H-transformed outcomes, can be useful in these cases as well. For this purpose we can use terms of the form $\{\alpha(D,Z) - \mathrm{E}[\alpha(D,Z) \mid Z]\}B(Z)$ for denoising where $\alpha(D,Z)$ now plays the same role as H before.

APPENDIX F. ADDITIONAL EMPIRICAL RESULTS

TABLE 1. 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	0.122	0.020	0.097	0.070	0.030	0.037
	(0.095,0.146)	(-0.002,0.042)	(0.064,0.130)	(0.046,0.094)	(0.008,0.052)	(0.005,0.068)
	-	-	[0.000]	-	-	[0.051]
Fraction Below Poverty Line (BPL)	0.181	0.194	-0.016	0.183	0.177	0.007
	(0.126,0.233)	(0.143,0.247)	(-0.089,0.057)	(0.132,0.234)	(0.133,0.223)	(-0.061,0.072)
	-	-	[1.000]	-	-	[1.000]
Average Financial Status (1-10 scale)	3.271	3.531	-0.267	3.337	3.741	-0.376
	(3.016,3.534)	(3.284,3.762)	(-0.611,0.072)	(3.095,3.587)	(3.499,3.975)	(-0.708,-0.048)
	-	-	[0.227]	-	-	[0.047]
Fraction Scheduled Caste-Scheduled Tribes (SC/ST)	0.169	0.148	0.022	0.190	0.145	0.046
	(0.125,0.213)	(0.106,0.191)	(-0.039,0.084)	(0.146,0.233)	(0.105,0.186)	(-0.014,0.104)
	-	-	[0.917]	-	-	[0.261]
Fraction Other Backward Caste (OBC)	0.268	0.205	0.060	0.331	0.179	0.154
	(0.215,0.319)	(0.155,0.253)	(-0.012,0.133)	(0.284,0.378)	(0.133,0.224)	(0.089,0.220)
	-	-	[0.204]	-	-	[0.000]
Fraction Minority caste	0.005	0.005	0.000	0.004	0.006	-0.002
	(-0.002,0.013)	(-0.002,0.014)	(-0.011,0.010)	(-0.002,0.010)	(0.000,0.011)	(-0.009,0.006)
	-	-	[1.000]	-	-	[1.000]
Fraction General Caste	0.217	0.464	-0.239	0.230	0.500	-0.273
	(0.154,0.280)	(0.403,0.525)	(-0.331,-0.142)	(0.168,0.293)	(0.439,0.564)	(-0.362,-0.177)
	-	-	[0.000]	-	-	[0.000]
Fraction No Caste	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)
	-	-	[1.000]	-	-	[1.000]
Fraction Other Caste	0.000	0.000	0.000	0.001	0.000	0.001
	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)	(0.000,0.002)	(-0.001,0.001)	(-0.001,0.002)
	-	-	[1.000]	-	-	[0.912]
Fraction Dont Know Caste	0.335	0.173	0.162	0.241	0.162	0.076
	(0.274,0.399)	(0.115,0.228)	(0.077,0.244)	(0.186,0.294)	(0.112,0.214)	(0.004,0.151)
	-	-	[0.000]	-	-	[0.080]
Fraction Hindu	0.806	0.956	-0.142	0.955	0.955	-0.002
	(0.725,0.887)	(0.884,1.026)	(-0.253,-0.035)	(0.921,0.987)	(0.907,0.991)	(-0.040,0.037)
	-	-	[0.021]	-	-	[1.000]
Fraction Muslim	0.169	0.017	0.143	0.026	0.017	0.006
	(0.091,0.247)	(-0.047,0.083)	(0.040,0.246)	(0.006,0.048)	(-0.005,0.052)	(-0.013,0.030)
	-	-	[0.013]	-	-	[1.000]
Fraction Christian	0.000	0.004	-0.004	0.000	0.004	-0.004
	(-0.008,0.008)	(-0.004,0.011)	(-0.015,0.007)	(-0.008,0.008)	(-0.004,0.011)	(-0.014,0.007)
	-	-	[1.000]	-	-	[1.000]
Fraction Buddhist	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)
	-	-	[1.000]	-	-	[1.000]
Fraction Sikh	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)
	-	-	[1.000]	-	-	[1.000]
Fraction Jain	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)
	-	-	[1.000]	-	-	[1.000]
Fraction Other Religion	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)
	-	-	[1.000]	-	-	[1.000]

Notes: Medians over 250 splits. Median confidence interval ($\alpha = .05$) in parenthesis. P-values for the hypothesis that the parameter is equal to zero against the two-sided alternatives in brackets.

TABLE 2. 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	0.031	0.023	0.007	0.015	0.023	-0.008
	(0.013,0.047)	(0.006,0.040)	(-0.017,0.031)	(0.000,0.031)	(0.008,0.039)	(-0.030,0.013)
	-	-	[1.000]	-	-	[0.818]
Fraction Literate	0.779	0.797	-0.017	0.820	0.786	0.032
	(0.756,0.801)	(0.775,0.817)	(-0.048,0.015)	(0.804,0.836)	(0.769,0.803)	(0.010,0.052)
	-	-	[0.626]	-	-	[0.008]
Fraction Single	0.053	0.046	0.006	0.051	0.043	0.007
	(0.046,0.060)	(0.040,0.053)	(-0.004,0.016)	(0.044,0.058)	(0.037,0.049)	(-0.001,0.016)
	-	-	[0.465]	-	-	[0.193]
Fraction of adults Married (living with spouse)	0.490	0.521	-0.032	0.516	0.521	-0.007
	(0.475,0.506)	(0.508,0.536)	(-0.053,-0.011)	(0.504,0.527)	(0.507,0.534)	(-0.024,0.011)
	-	-	[0.006]	-	-	[0.894]
Fraction of adults Married (not living with spouse)	0.002	0.004	-0.001	0.003	0.003	0.000
	(0.000,0.005)	(0.001,0.006)	(-0.005,0.003)	(0.001,0.005)	(0.001,0.005)	(-0.002,0.003)
	-	-	[1.000]	-	-	[1.000]
Fraction of adults Divorced or Separated	0.006	0.002	0.005	0.005	0.001	0.004
	(0.004,0.009)	(-0.001,0.004)	(0.001,0.008)	(0.003,0.007)	(-0.001,0.002)	(0.001,0.007)
	-	-	[0.010]	-	-	[0.008]
Fraction of adults Widow or Widower	0.034	0.036	-0.001	0.036	0.040	-0.004
	(0.029,0.040)	(0.031,0.041)	(-0.009,0.006)	(0.030,0.041)	(0.034,0.045)	(-0.012,0.003)
	-	-	[1.000]	-	-	[0.579]
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"	0.413	0.390	0.025	0.389	0.392	-0.001
	(0.394,0.433)	(0.373,0.406)	(0.000,0.050)	(0.375,0.401)	(0.377,0.406)	(-0.020,0.019)
	-	-	[0.103]	-	-	[1.000]
Fraction who received Nursery level educ. or less	0.156	0.158	-0.003	0.133	0.166	-0.032
	(0.140,0.171)	(0.144,0.172)	(-0.024,0.018)	(0.121,0.144)	(0.155,0.177)	(-0.047,-0.017)
	-	-	[1.000]	-	-	[0.000]
Fraction who received Class 4 level educ.	0.077	0.087	-0.009	0.081	0.090	-0.009
	(0.069,0.086)	(0.079,0.095)	(-0.021,0.002)	(0.073,0.090)	(0.082,0.098)	(-0.021,0.002)
	-	-	[0.218]	-	-	[0.222]
Fraction who received Class 8 level educ.	0.171	0.159	0.013	0.160	0.154	0.008
	(0.159,0.183)	(0.148,0.170)	(-0.003,0.030)	(0.148,0.171)	(0.144,0.165)	(-0.009,0.023)
	-	-	[0.220]	-	-	[0.736]
Fraction who received Class 12 level educ.	0.204	0.232	-0.028	0.246	0.226	0.018
	(0.185,0.224)	(0.213,0.250)	(-0.056,0.001)	(0.230,0.263)	(0.210,0.241)	(-0.005,0.041)
	-	-	[0.119]	-	-	[0.251]
Fraction who received Graduate or Other Diploma	0.076	0.093	-0.017	0.085	0.095	-0.011
	(0.062,0.090)	(0.080,0.106)	(-0.036,0.003)	(0.072,0.098)	(0.082,0.108)	(-0.028,0.007)
	-	-	[0.185]	-	-	[0.492]
Fraction with education level Other or Dont know	0.310	0.264	0.046	0.293	0.262	0.031
	(0.298,0.323)	(0.252,0.276)	(0.029,0.063)	(0.283,0.305)	(0.252,0.272)	(0.016,0.046)
	-	-	[0.000]	-	-	[0.000]

Notes: Medians over 250 splits. Median confidence interval ($\alpha = .05$) in parenthesis. P-values for the hypothesis that the parameter is equal to zero against the two-sided alternatives in brackets.

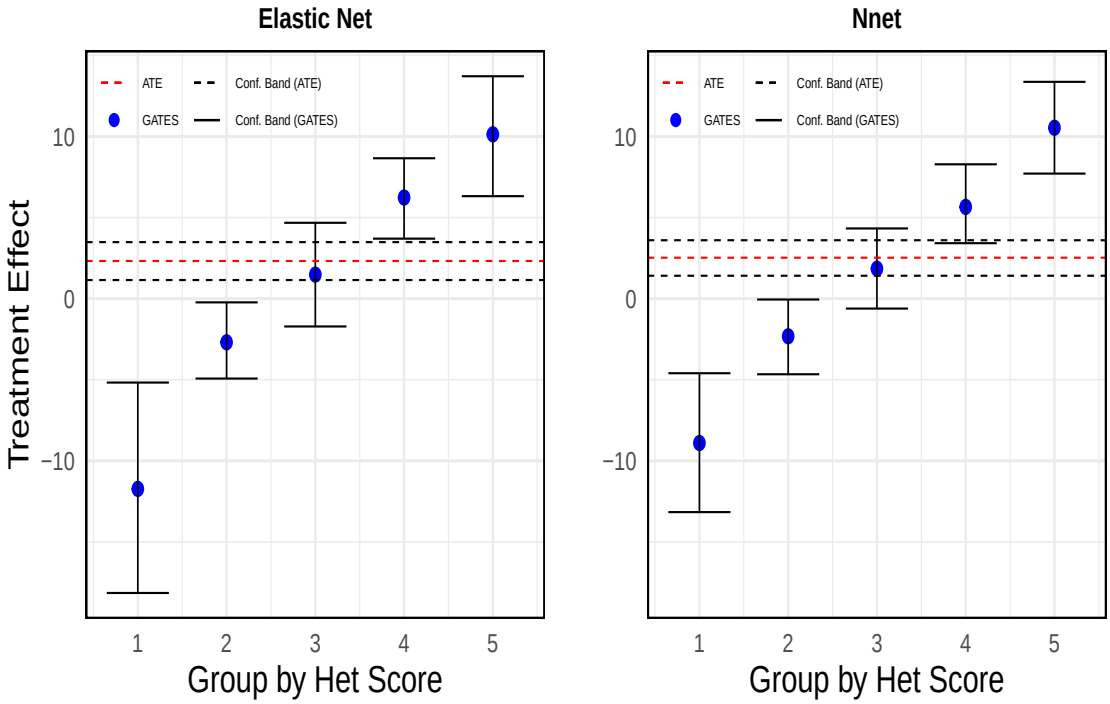

FIGURE 2. GATES of Immunization Full package compared to Ambassadors and SMS only. Point estimates and 90% adjusted confidence intervals uniform across groups based on 250 random splits in half.

TABLE 3. 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)	0.027	0.049	-0.021	0.029	0.049	-0.019
	(0.023,0.032)	(0.048,0.050)	(-0.026,-0.016)	(0.026,0.033)	(0.048,0.050)	(-0.023,-0.016)
	-	-	[0.000]	-	-	[0.000]
Imm.per dollar (G_2)	0.032	0.046	-0.014	0.034	0.046	-0.012
	(0.029,0.036)	(0.045,0.048)	(-0.018,-0.010)	(0.031,0.037)	(0.045,0.048)	(-0.016,-0.009)
	-	-	[0.000]	-	-	[0.000]
Imm.per dollar (G_3)	0.034	0.044	-0.010	0.037	0.044	-0.007
	(0.030,0.037)	(0.042,0.046)	(-0.014,-0.006)	(0.034,0.040)	(0.042,0.046)	(-0.011,-0.003)
	-	-	[0.000]	-	-	[0.001]
Imm. per dollar (G_4)	0.039	0.044	-0.004	0.038	0.044	-0.005
	(0.036,0.042)	(0.042,0.046)	(-0.008,-0.001)	(0.035,0.041)	(0.042,0.046)	(-0.009,-0.002)
	-	-	[0.015]	-	-	[0.008]
Imm. per dollar (G_5)	0.038	0.039	-0.001	0.037	0.040	-0.003
	(0.035,0.041)	(0.036,0.042)	(-0.005,0.003)	(0.033,0.040)	(0.037,0.042)	(-0.007,0.001)
	-	-	[1.000]	-	-	[0.310]

Notes: Medians over 250 splits. Median confidence interval $(\alpha = .05)$ in parenthesis. P-values for the hypothesis that the parameter is equal to zero against the two-sided alternatives in brackets.

APPENDIX G. FIGURES AND TABLES - PREDICTIVE LEARNERS

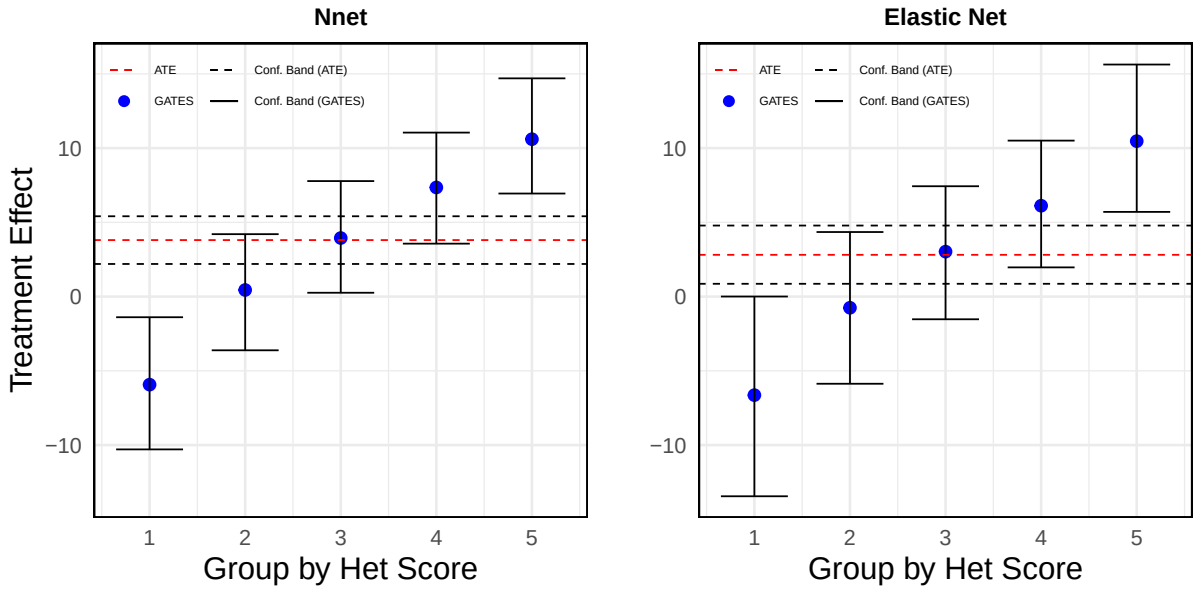

FIGURE 3. GATES of Immunization Incentives, based upon Predictive Learners. Median point estimates and Median confidence interval ($\alpha = .05$) in parenthesis, over 250 splits.

TABLE 4. Comparison of ML Methods: Immunization Incentives

	Elastic Net	Boosting	Neural Network	Random Forest
Best BLP (Λ)	63.670	31.480	51.680	23.400
	[48.879, 77.659]	[23.983, 45.833]	[41.557, 65.97]	[17.538, 31.659]
Best GATES ($\bar{\Lambda}$)	7.950	5.019	5.857	4.185
	[6.803, 8.938]	[4.194, 6.379]	[5.097, 6.459]	[3.142, 5.158]

Notes: Medians over 250 splits in half, based upon Predictive Learners. The brackets report interquartile ranges for goodness-of-fit statistics.

TABLE 5. BLP of Immunization Incentives

Elastic Net		Neural Network	
ATE (β_1)	HET (β_2)	ATE (β_1)	HET (β_2)
2.806	1.070	2.462	0.902
(1.123,4.681)	(0.840,1.312)	(0.879,3.937)	(0.684,1.116)
[0.003]	[0.000]	[0.005]	[0.000]

Notes: Medians over 250 splits, based upon Predictive Learners. Median confidence interval ($\alpha = .05$) in parenthesis. P-values for the hypothesis that the parameter is equal to zero against the two-sided alternatives in brackets.

TABLE 6. 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 $\gamma_k := \widehat{\mathrm{E}}[s_0(Z) \mid G_k]$	13.300	-7.362	20.88	11.260	-6.063	17.33
	(8.016,18.89)	(-12.63,-2.005)	(12.94,28.43)	(7.866,14.80)	(-9.792,-2.213)	(12.15,22.75)
	[0.000]	[0.016]	[0.000]	[0.000]	[0.006]	[0.000]

Notes: Medians over 250 splits, based upon Predictive Learners.. Median confidence interval ($\alpha = .05$) in parenthesis. P-values for the hypothesis that the parameter is equal to zero against the two-sided alternatives in brackets.

TABLE 7. 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	2.182	2.284	-0.093	2.186	2.282	-0.107
to pregnant mother	(2.110,2.252)	(2.213,2.355)	(-0.193,0.005)	(2.126,2.246)	(2.216,2.343)	(-0.193,-0.019)
	-	-	[0.126]	-	-	[0.033]
Number of vaccines	4.034	4.678	-0.617	4.275	4.722	-0.454
to child since birth	(3.809,4.260)	(4.474,4.891)	(-0.939,-0.327)	(4.103,4.439)	(4.547,4.895)	(-0.690,-0.213)
	-	-	[0.000]	-	-	[0.000]
Fraction of children	0.998	1.000	-0.002	1.000	1.000	0.000
received polio drops	(0.995,1.001)	(0.997,1.003)	(-0.006,0.002)	(1.000,1.000)	(1.000,1.000)	(0.000,0.000)
	-	-	[0.686]	-	-	[0.879]
Number of polio	2.954	2.993	-0.039	2.966	2.998	-0.031
drops to child	(2.933,2.974)	(2.975,3.012)	(-0.066,-0.012)	(2.954,2.978)	(2.985,3.009)	(-0.048,-0.014)
	-	-	[0.009]	-	-	[0.001]
Fraction of children	0.798	0.930	-0.133	0.910	0.929	-0.018
received immunization card	(0.749,0.848)	(0.885,0.975)	(-0.198,-0.062)	(0.888,0.935)	(0.902,0.956)	(-0.052,0.011)
	-	-	[0.001]	-	-	[0.443]
Fraction of children received	0.127	0.250	-0.122	0.125	0.258	-0.130
Measles vaccine by 15 months of age	(0.092,0.163)	(0.216,0.283)	(-0.172,-0.074)	(0.094,0.160)	(0.226,0.289)	(-0.178,-0.085)
	-	-	[0.000]	-	-	[0.000]
Measles at credible locations	0.286	0.406	-0.121	0.288	0.427	-0.142
	(0.237,0.333)	(0.362,0.448)	(-0.187,-0.056)	(0.245,0.331)	(0.388,0.470)	(-0.200,-0.083)
	-	-	[0.001]	-	-	[0.000]

Notes: Medians over 250 splits, based upon Predictive Learners. Median confidence interval ($\alpha = .05$) in parenthesis. P-values for the hypothesis that the parameter is equal to zero against the two-sided alternatives in brackets.

APPENDIX H. FIGURES AND TABLES - PREDICTION INTERVALS

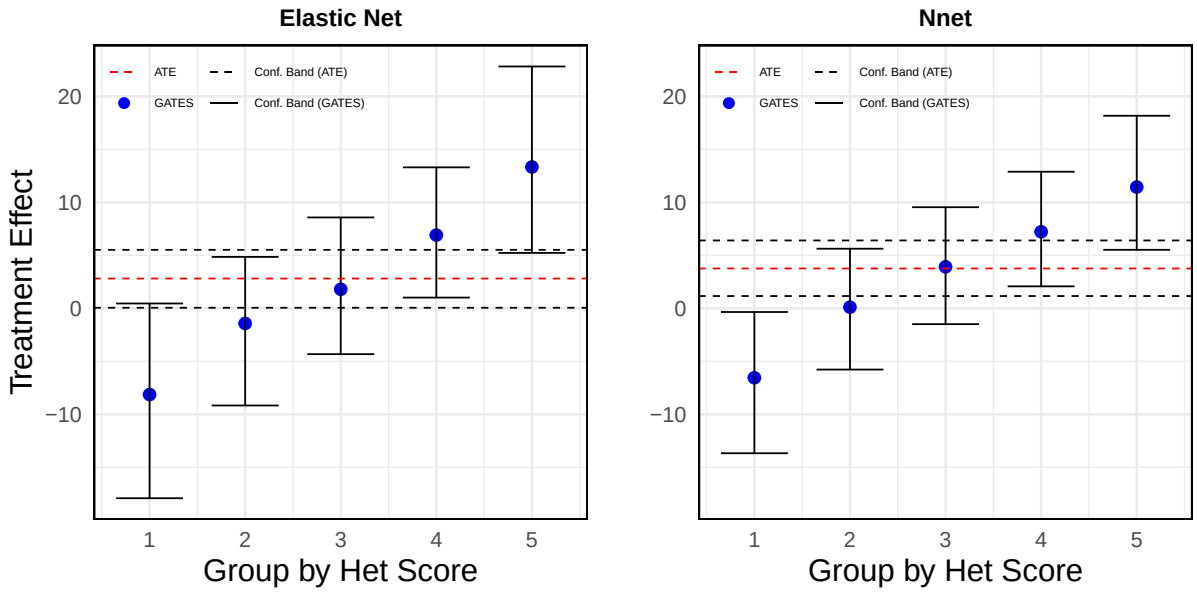

FIGURE 4. GATES of Immunization Incentives, based upon prediction intervals presented in Section 4.3. Median point estimates and Median confidence interval ($\alpha = .05, \beta = .225$), over 250 splits.

TABLE 8. Comparison of ML Methods: Immunization Incentives

	Elastic Net	Boosting	Neural Network	Random Forest
Best BLP (Λ)	67.750	32.900	53.420	25.200
	[51.491, 82.368]	[23.246, 44.665]	[42.516, 67.647]	[18.328, 34.705]
Best GATES ($\bar{\Lambda}$)	8.254	5.104	6.001	4.492
	[7.329, 9.314]	[4.27, 6.079]	[5.087, 6.888]	[3.339, 5.507]

Notes: Medians over 250 splits in half. Note that we used Neural Network Causal Boosting for all methods, using Algorithm 6.2. The brackets report interquartile ranges for goodness-of-fit statistics.

TABLE 9. BLP of Immunization Incentives

Elastic Net		Neural Network	
ATE (β_1)	HET (β_2)	ATE (β_1)	HET (β_2)
2.814	1.047	2.441	0.899
(1.087,4.506)	(0.826,1.262)	(0.846,3.979)	(0.685,1.107)
[0.004]	[0.000]	[0.004]	[0.000]

Notes: Medians over 250 splits, based upon prediction intervals presented in Section 4.3. Confidence Intervals for Median Parameter ($\alpha = .05, \beta = .225$) in parenthesis.

TABLE 10. 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 $\gamma_k := \widehat{\mathrm{E}}[s_0(Z) \mid G_k]$	13.230	-8.000	21.60	11.210	-6.551	18.13
	(6.001,21.43)	(-16.38,-0.192)	(9.310,33.85)	(5.432,17.07)	(-12.85,-0.971)	(9.292,26.56)
	[0.000]	[0.009]	[0.000]	[0.000]	[0.002]	[0.000]

Notes: Medians over 250 splits, based upon prediction intervals presented in Section 4.3. Confidence Intervals for Median Parameter ($\alpha = .05, \beta = .225$) in parenthesis.

TABLE 11. CLAN of Immunization Incentives

	Elastic Net			Nnet		
	20% Most (δ_5)	20% Least (δ_1)	Difference ($\delta_5 - \delta_1$)	20% Most (δ_5)	20% Least (δ_1)	Difference ($\delta_5 - \delta_1$)
Number of vaccines	2.187	2.277	-0.081	2.174	2.285	-0.112
to pregnant mother	(2.115,2.259)	(2.212,2.342)	(-0.180,0.015)	(2.111,2.234)	(2.224,2.345)	(-0.202,-0.028)
	-	-	[0.190]	-	-	[0.019]
Number of vaccines	4.077	4.639	-0.562	4.264	4.734	-0.490
to child since birth	(3.858,4.304)	(4.444,4.859)	(-0.863,-0.260)	(4.091,4.434)	(4.549,4.900)	(-0.739,-0.250)
	-	-	[0.001]	-	-	[0.000]
Fraction of children	0.998	1.000	-0.002	1.000	1.000	0.000
received polio drops	(0.995,1.001)	(0.997,1.003)	(-0.006,0.002)	(1.000,1.000)	(1.000,1.000)	(0.000,0.000)
	-	-	[0.683]	-	-	[0.943]
Number of polio	2.955	2.993	-0.037	2.965	2.998	-0.032
drops to child	(2.935,2.974)	(2.976,3.010)	(-0.063,-0.010)	(2.953,2.977)	(2.985,3.010)	(-0.049,-0.016)
	-	-	[0.013]	-	-	[0.000]
Fraction of children YN	0.803	0.926	-0.121	0.908	0.927	-0.027
received immunization card	(0.754,0.851)	(0.882,0.969)	(-0.187,-0.054)	(0.881,0.932)	(0.898,0.959)	(-0.059,0.007)
	-	-	[0.001]	-	-	[0.217]
Fraction of children received	0.133	0.243	-0.106	0.126	0.260	-0.131
Measles vaccine by 15 months of age	(0.097,0.169)	(0.209,0.276)	(-0.153,-0.056)	(0.095,0.159)	(0.228,0.291)	(-0.176,-0.085)
	-	-	[0.000]	-	-	[0.000]
Measles at credible locations	0.293	0.399	-0.110	0.289	0.433	-0.142
	(0.246,0.338)	(0.358,0.444)	(-0.174,-0.045)	(0.246,0.331)	(0.391,0.475)	(-0.206,-0.084)
	-	-	[0.002]	-	-	[0.000]

Notes: Medians over 250 splits, based upon prediction intervals presented in Section 4.3. Confidence Intervals for Median Parameter $(\alpha = .05, \beta = .225)$ in parenthesis.

REFERENCES

Jean-Yves Audibert. Progressive mixture rules are deviation suboptimal. *Advances in Neural Information Processing Systems*, 20, 2007.

Keith Ball. Volume ratios and a reverse isoperimetric inequality. *Journal of the London Mathematical Society*, 2(2):351–359, 1991.

Vidmantas Bentkus. On the dependence of the berry–esseen bound on dimension. *Journal of Statistical Planning and Inference*, 113(2):385–402, 2003.

Qizhao Chen, Vasilis Syrgkanis, and Morgane Austern. Debiased machine learning without sample-splitting for stable estimators. *arXiv preprint arXiv:2206.01825*, 2022.

Victor Chernozhukov, Whitney Newey, and Rahul Singh. Debiased machine learning of global and local parameters using regularized riesz representers. *arXiv preprint arXiv:1802.08667*, 2018.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association*, 116(536):1849–1864, 2021.

A Ronald Gallant and Halbert White. *A unified theory of estimation and inference for nonlinear dynamic models*. Blackwell, 1988.

Vladimir Koltchinskii, Karim Lounici, and Alexandre B Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Conference on Learning Theory*, pages 1260–1285. PMLR, 2015.

Martin Raič. A multivariate berry–esseen theorem with explicit constants. *Bernoulli*, 25(4A): 2824–2853, 2019.

Vira Semenova and Victor Chernozhukov. Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289, 08 2021.

Suhas Vijaykumar. Localization, convexity, and star aggregation. *Advances in Neural Information Processing Systems*, 34:4570–4581, 2021.

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