Advance Access publication September 25, 2016

Political Analysis (2016) 24:434–456
doi:10.1093/pan/mpw019

Retrospective Causal Inference with Machine Learning Ensembles: An Application to Anti-recidivism Policies in Colombia

Cyrus Samii
Department of Politics, New York University, 19 West 14th Street, New York, NY 10012
e-mail: cds2083@nyu.edu (corresponding author)

Laura Paler
Department of Political Science, University of Pittsburgh, 4600 Wesley W. Posvar Hall, Pittsburgh, PA 15260
e-mail: lpaler@pitt.edu

Sarah Zukerman Daly
Department of Political Science, University of Notre Dame, 217 O'Shaughnessy Hall, Notre Dame, IN 46556
e-mail: sarahdaly@nd.edu

Edited by R. Michael Alvarez

We present new methods to estimate causal effects retrospectively from micro data with the assistance of a machine learning ensemble. This approach overcomes two important limitations in conventional methods like regression modeling or matching: (i) ambiguity about the pertinent retrospective counterfactuals and (ii) potential misspecification, overfitting, and otherwise bias-prone or inefficient use of a large identifying covariate set in the estimation of causal effects. Our method targets the analysis toward a well-defined "retrospective intervention effect" based on hypothetical population interventions and applies a machine learning ensemble that allows data to guide us, in a controlled fashion, on how to use a large identifying covariate set. We illustrate with an analysis of policy options for reducing ex-combatant recidivism in Colombia.

1 Introduction

Retrospective causal studies are essential in the social sciences, but they present acute challenges. They are essential insofar as for some important causal questions there are often no feasible alternatives to a retrospective analysis. Such situations include studies of rare outcomes or outcomes that take many years to come about, such as violence or institutional changes. Adequately powered prospective studies, whether in the form of a randomized experiment or not, may take too long and be too logistically difficult to be practical or may prove unethical.

Authors' note: Authors are listed in reverse alphabetical order and are equal contributors to the project. All replication materials are available at the *Political Analysis* Dataverse (article url: http://dx.doi.org/10.7910/DVN/QXCFO2). We thank Carolina Serrano for excellent research assistance in Colombia and the team at *Fundación Ideas para la Paz* for their collaboration in the data collection. We also thank the Organization of American States, *Misión de Apoyo al Proceso de Paz* and the *Agencia Colombiana para la Reintegración* for their collaboration. Daly acknowledges funding from the Swedish Foreign Ministry, the Smith Richardson Foundation, and the Carroll L. Wilson Award. For helpful discussions, the authors thank Michael Alvarez, two anonymous *Political Analysis* reviewers, Deniz Aksoy, Peter Aronow, Neal Beck, Matthew Blackwell, Drew Dimmery, Ryan Jablonski, Michael Peress, Fredrik Savje, Maya Sen, Teppei Yamamoto, Rodrigo Zarazaga, and seminar participants at the American Political Science Association annual meetings, European Political Science Association annual meetings, Empirical Studies of Conflict working group, Massachussetts Institute of Technology, Midwest Political Science association annual meetings, New York University, and the University of Rochester. Supplementary materials for this article are available on the *Political Analysis* website.

Retrospective studies present acute challenges because they try to make causal inferences about the effects of policies, exposures, or processes that are beyond the control of analysts. This introduces problems of endogeneity and confounding. Moreover, generating results that can inform policy requires estimates that are relevant for one's target population, but sources of quasi-random variation (e.g., instrumental variables or discontinuities) may be too specific in the subpopulations to which they apply to meet these needs. The relevant counterfactual comparisons may not be obvious either.

We draw on new methods from epidemiology and apply a machine learning approach to overcome these challenges (Van der Laan and Rose 2011). Our approach makes use of familiar "conditional independence" assumptions; however, we do so in a way that circumvents problems that arise in simpler uses of regression, matching, or propensity scores (Angrist and Pischke 2009), 58-94.[1] Specifically, we use a very large number of covariate control variables and a machine learning ensemble. Using a very large number of covariates allows us to make conditional independence more believable, which in principle also moves us safely past concerns about "bias amplification" (Myers et al. 2011).[2] But having such a rich covariate set raises questions about how to properly employ the covariates. We face the daunting task of having to choose from among the vast possibilities for terms (e.g., squared, cubed) or interactions to include in a model. We use a machine learning ensemble that lets the data guide us, in a controlled fashion, in using an identifying covariate set. We use a simulation experiment to show how a machine learning ensemble is more robust than conventional methods in extracting identifying variation from irregular functional relationships in a noisy covariate space.

To obtain causal estimates that properly inform realistic policy options, we define our counterfactuals in terms of substantively motivated "retrospective intervention effects" (RIEs) for the target population. The RIE establishes a compelling counterfactual comparison that incorporates different types of information than alternative estimands such as the average treatment effect (ATE), average effect of the treatment on the treated (ATT), or average effect of the treatment on the controls (ATC). (We provide a formal characterization of the differences below.) Consider an analysis of the effects of employment on criminality. The RIE compares what actually occurred in the population to a counterfactual where everyone in the population is ensured to be employed. In contrast, the ATE would estimate how criminality differs when everyone is employed versus when no one is employed, an unrealistic population counterfactual. The ATT and ATC are less unrealistic than the ATE in that they compare how things would change were we to intervene on the employment status among those with and without jobs, respectively. But they cannot speak to the importance of such interventions in the population because they do incorporate pre-intervention levels of employment. Taking pre-existing rates of employment into account is especially important if one wanted to compare an employment intervention to, say, cognitive behavioral therapy for reducing overall crime rates. That said, in some cases estimands other than the RIE may be preferable—it would depend on the goals of the analysis. The ensemble methods that we apply here could be used for other estimands.

This paper contributes to the political methodology literature on causal inference in two ways. First, we offer a didactic presentation of how one can apply the power of machine learning ensembles to causal inference and policy analysis problems. In doing so, we demonstrate how causal inference problems are extensions of ensemble prediction problems, something with which political scientists are already somewhat familiar (Montgomery, Hollanbach, and Ward 2012). Second, we demonstrate the use of hypothetical interventions as a way to target the analysis toward a substantively meaningful counterfactual comparison that yields the RIE. Our application to

[1] We define conditional independence formally below. The idea is that we can identify the set of confounding factors and "condition" them, thereby removing the confounding covariation.

[2] Bias amplification can occur when omitted variables confound estimates of a causal effect and one incorporates additional covariates that purge substantial variation from the treatment variables but fail to purge variation from the outcome variables (Pearl 2010). Risk of bias amplification depends on the specificities of a given data set. Myers et al. (2011) find empirically that such biases tend not to be a major concern in epidemiological applications with reasonable sets of control variables.

https://doi.org/10.1093/pan/mpw019 Published online by Cambridge University Press

retrospective studies extends the existing literature on machine learning for causal inference, which includes work on characterizing heterogenous treatment effects (Imai and Strauss 2011; Green and Kern 2012; Imai and Ratkovic 2013; Grimmer, Messing, and Westwood 2014; Athey and Imbens 2015), locating subpopulations within which conditional ignorability holds (Ratkovic 2014), and non parametrically estimating counterfactual response surfaces (Hill 2011). Third, the high-dimensional propensity score and reweighting methods that we use are readily applicable to other types of reweighting methods, such as for dynamic treatment regimes (Blackwell 2013).

We begin by establishing the inferential setting, and then we discuss potential perils in standard practice for retrospective studies. Next, we develop an approach to identification of causal effects based on hypothetical interventions. Following that, we discuss estimation, practical implementation, and inference. We apply the methods to an illustrative case study that evaluates policy options for reducing recidivism among ex-combatants in Colombia. A conclusion draws out implications and ideas for further research.

2 Setting

Our approach in this paper is based on the innovations of Hubbard and Van der Laan (2008), Van der Laan and Rose (2011), and Young et al. (2009), and we adopt their notation so as to allow readers to refer back to these reference works easily. We start with a target population and then obtain from it a random sample of observations.[3] The observations consist of treatment variables denoted as the vector of random variables $A = (A_1, ..., A_j, ..., A_J)'$, covariates denoted as the vector of random variable $W = (W_1, ..., W_p, ..., W_P)'$, and an outcome variable Y. These observations are defined collectively by the random vector $O = (W, A, Y)'$ that is governed in the target population by some probability distribution, P_0. The task is to estimate the average causal effects of components of A for our target population. An arbitrary component of the treatment vector A is labeled as A_j, the complement of elements in A is labeled as A_{-j}, and the support for A_j is denoted as $\mathcal{A}_j$.

The causal structure is assumed to follow the graph depicted in Fig. 1 (Pearl 2009). We have circled the elements of A to highlight our interest in estimating causal effects for the components of that vector. The causal graph indicates two sources of confounding, originating in W and U, with the variable U standing in to characterize any unobserved determinants of the elements of A. The assumptions embedded in this graph indicate that for estimating the effect of A_j confounding originating in W can be blocked by conditioning on W, where as confounding originating in U can be blocked by conditioning on A_{-j}. An important assumption that this graph encodes is that, aside from the dependencies due to U and W, there are no direct causal relationships between the elements of A. These are substantive assumptions about the causal structure.[4]

Using the "potential outcomes" notation to define causal effects (Holland 1986; Rubin 1978; Sekhon 2009), we can write the outcome that would be observed if treatments $(A_1, ..., A_J)$ were set to $(a_1, ..., a_J)$ as follows:

$$Y(a) = Y(a_1, ..., a_J), \tag{1}$$

where $a \in \prod_{j=1}^{J} \mathcal{A}_j \equiv \mathcal{A}$. Thus, potential outcomes depend on the combinations of treatments a unit receives, with these combinations denoted by the vector a. For an arbitrary unit i in our target population, the causal effect of fixing $A_{ji} = a$ versus $A_{ji} = a'$ is defined as

$$\tau_{ji}(a, a') = Y_i(a, A_{-j}) - Y_i(a', A_{-j}), \tag{2}$$

[3] A subsequent section deals with questions associated with unequal probability sampling or cluster sampling.

[4] If the assumptions are wrong, the analysis will not generally yield unbiased or consistent estimates of causal effects. In an applied setting, one would want to check robustness of one's estimates to a variety of assumptions about the causal graph. For example, one would want to check to see whether estimates change if one assumes that some elements of A are causally dependent on others. Under such alternative assumptions, one would set up the analysis in ways that avoid post-treatment bias by including in the set of covariate controls only the elements of A_{-j} that are not causally dependent on A_j (Rosenbaum 1984; King and Zeng 2006). Once that is done, the analysis would proceed as we describe below. Our primary interest in this paper is to elaborate methods given a causal graph, and so to save space we do not conduct such robustness checks here.

https://doi.org/10.1093/pan/mpw019 Published online by Cambridge University Press

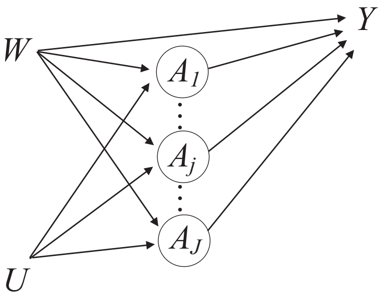

Fig. 1 Assumed causal graph, showing that confounding in W for the effect of A_j can be blocked by conditioning on (W, A_{-j}), and then confounding originating in U can be blocked by conditioning on A_{-j}.

where the introduction of the i subscripts highlights our focus on possible heterogeneity in these effects across units. $\tau_j(a, a')$ is defined as $\mathrm{E}[\tau_{ji}(a, a')]$, the average causal effect with the average taken over the units indexed by i. This target quantity, $\tau_j(a, a')$, is non parametrically identified under the so-called conditional independence assumption (Imai and van Dyk 2004; Imbens 2004; Angrist and Pischke 2009, 52–59; Imbens and Wooldridge 2009):

$$A_j \perp\!\!\!\perp (Y_i(a, A_{-ji}), Y_i(a', A_{-ji}))' | (A_{-ji}, W)'. \tag{3}$$

Figure 1 implies this assumption (although other graphs could also be drawn under this assumption too). Here, A_{-j} and W form a conditioning vector that blocks sources of confounding variation (or "back door paths," Pearl 2009, 16–18, 78–81) in the relationship between A_j and our potential outcomes, $Y_i(a, A_{-ji})$ and $Y_i(a', A_{-ji})$.

3 Perils of Standard Practice

Conditional independence of the treatments offers the promise of being able to identify causal effects. But one still faces the challenge of operationalizing conditional independence. Imbens (2004) reviews general approaches rooted in either (i) propensity scores and a focus on the "assignment mechanism" that determines the relationship between covariates, $(A_{-j}, W)'$, and the causal factor of interest, A_j, or (ii) response surface modeling and a focus on outcome data generating processes that relate covariates, $(A_{-j}, W)'$, to outcomes, $(Y(a, A_{-j}), Y(a', A_{-j}))'$. As Imbens shows, accounting for either assignment or response is sufficient for identifying a causal effect under the conditional independence assumption. Analysts have put forward various arguments for whether it is preferable to emphasize assignment (Rosenbaum and Rubin 1983; Rubin 2008), response surfaces (Pearl 2010; Hill 2011), or a combination of the two in the construction of "doubly robust" estimators (Robins and Rotnitzky 1995; Bang and Robins 2005).

Regression modeling, the workhorse method in the social sciences, can be variously conceptualized as following either approach. Following Angrist and Pischke (2009, 52–59), suppose that effects are homogenous such that $\tau_{ji}(a, a') = \tau_j(a, a')$ for all units and that one defines the conditioning vector as $X_i \equiv (A_{-ji}, W_i)'$ in a regression model of the form

$$Y_i = \alpha + \beta A_{ji} + X_i\gamma + \epsilon_i. \tag{4}$$

We suppose that the error term, ϵ_i, equals the ordinary least squares (OLS) residual from the regression of $Y_i - \alpha - \beta A_{ji}$ on X_i when this regression is carried out on the full population for which one wants to make inference. Then, so long as the control vector specification in X_i is adequate to ensure that the linearity assumption holds—that is, $\mathrm{E}[Y_i - \alpha - \beta A_{ji} | X_i] = X_i\gamma$ holds—the OLS estimate of β is consistent for the homogenous effect, $\tau_j(a, a')$ (Angrist and Pischke 2009, 57–59). This is in essence a response modeling approach. In contrast, Angrist and Krueger (1999) and Aronow and Samii (2016) develop the case where the control function, $X_i\gamma$, models the assignment process. In this case, the homogenous effects assumption again implies that the OLS estimator for β is consistent for $\tau_j(a, a')$.

https://doi.org/10.1093/pan/mpw019 Published online by Cambridge University Press

These two assumptions—homogenous effects and correct specification for the control vector, X—are unrealistic in many applied settings, making the naïve use of linear regression a problematic tool for exploiting conditional independence of the treatment. Furthermore, it would be heroic to presume that all relevant heterogeneity could be modeled. The linearity assumption is especially vexing when conditional independence of the treatment requires a large covariate set as this introduces a bewildering array of possible higher-order terms and interactions that one must decide on including or excluding. If either homogenous effects or correct linear specification fails to hold, causal effects estimated with linear regression may fail to characterize the average causal effects for the target population. First, even if linearity in X holds but effects are *heterogeneous*, then the OLS estimator recovers a distorted estimate of the average causal effect. The distortions are based on an implicit weighting that linear regression produces based on the conditional variance of A_j (Angrist and Krueger 1999; Angrist and Pischke 2009, 75; Aronow and Samii 2016).[5] Second, when the specification based on X is wrong, residual confounding may remain and bias the results. Beyond these risks of getting it wrong, there is also the question of researcher discretion through which terms in X may be manipulated to produce "desirable" results (King and Zeng 2006).

Direct covariate matching is an alternative to regression and it relieves the analyst from some of the modeling burdens necessary with regression (Ho et al. 2007). Nevertheless, direct covariate matching becomes difficult when the covariate space is large. When that is the case, one is forced to apply some method of characterizing distance in the covariate space in order either to identify "nearest neighbors" or, in kernel matching, generate kernel-weighted approximations of counterfactual outcomes (Imbens and Wooldridge 2009). Generally speaking, distance metrics for direct covariate matching convey no optimality criteria with respect to bias minimization. Matching on propensity scores (Rosenbaum and Rubin 1983) or prognostic scores (Hansen 2008) can resolve such dimensionality problems and in a manner that is targeted toward bias minimization, but in practice one is left with the task of determining a specification for the propensity or prognostic scores. When the covariate space is very large, similar challenges make it difficult to use other "direct balancing" methods such as entropy balancing (Hainmueller 2011).

The idea we pursue is that a machine learning approach might allow us to sift through the information content in a large covariate set to target bias minimization in an efficient manner. Machine learning methods are distinguished from other statistical methods in their emphasis on "regularization," which is the use of penalties for model complexity (Bickel and Li 2006; Hastie, Tibshirani, and Friedman 2009, 34), as well as processes of tuning models so as to minimize cross-validated prediction error. Our machine learning ensemble targets prediction error for propensity scores. By combining regularization and cross-validation, the ensemble is built to wade through the noisy variation in a large covariate set and extract meaningful predictive covariate variation. Because we are predicting propensity scores, this predictive variation is also variation that provides the basis for causal identification. As Van der Laan and Rose (2011) show, one could also use machine learning in a response-surface modeling approach. However, using propensity scores allows for one round of machine learning that can then be used to estimate effects on a variety of outcomes, whereas a response modeling approach would require a separate machine learning step for each outcome. Busso, DiNardo, and McCrary (2014) show that when covariate distributions have good overlap over the treatment values, estimation using inverse propensity score weights (IPWs) exhibits favorable efficiency properties. Below, we use a simulation study to illustrate these points.

4 Defining RIEs

The first step of our approach is to define coherent causal quantities given that effects are possibly heterogeneous and nonlinear. We do so through the definition of the RIE. Following Hubbard and Van der Laan (2008), we consider hypothetical population interventions on the components of A. Such hypothetical interventions are conceptualized as taking a treatment, say A_j, and imagining a manipulation that changes $A_j = a_j$ to $A_j = a_j'$. Defining hypothetical interventions has two

[5] Although the key results in these papers are developed with respect to OLS regression, as Aronow and Samii (2016) showed, the very same results apply in the first order to estimates for generalized linear models such as logit, probit, and so on.

https://doi.org/10.1093/pan/mpw019 Published online by Cambridge University Press

methodological benefits. First, it allows us to define clear causal estimands under effects that vary not only from unit to unit, but also over different values of the underlying causal factors (e.g., non linear or threshold effects). Second, we can define potential interventions in a manner that takes into account real-world options and therefore establish estimands that are directly relevant for policy analysis (Manski 1995, 54–58). Different hypothetical interventions can be compared with each other in terms of their costs and estimated effects so as to come up with a ranking of the kinds of manipulations that are most promising from a practical perspective.

Our goal is to estimate, retrospectively, the effects of hypothetical interventions associated with each component of A on the outcome distribution for the population. That is, we seek to estimate the difference between what has *actually happened* against a counterfactual of *what would have happened* had there been an intervention on variable A_j. The way that one defines hypothetical interventions depends on the types of practical questions that one wants to answer. Consider an intervention on A_j defined as fixing $A_j = \underline{a}_j$ for all members of the population. If $\underline{a}_j$ were the minimum value of A_j, for example, then the RIE would be equivalent to what epidemiologists refer to as the "attributable risk" (Rothman, Greenland, and Lash 2008, 63), which measures the average consequence of the observed level of A_j relative to a counterfactual of A_j being kept to its minimum throughout the population.

Another type of hypothetical intervention is one that manipulates values of a continuous treatment, but does so in a manner that varies depending on individuals' realized values of the treatment variable. For example, suppose the causal factor of interest is income. We could define an intervention that ensures that all individuals have some minimum level of income, $\underline{c}$. Then, we apply this intervention to all individuals, in which case we would be changing the incomes for all individuals with incomes of less than $\underline{c}$ to be, counterfactually, $\underline{c}$. For individuals with incomes higher than $\underline{c}$, the intervention would have no effect, and so their incomes would remain as observed.

For outcome Y, define the RIE for A_j and intervention value $\underline{a}_j$ as

$$\psi_j = \underbrace{\mathrm{E}[Y(\underline{a}_j, A_{-j})]}_{\text{counterfactual mean}} - \underbrace{\mathrm{E}[Y],}_{\text{observed mean}} \tag{5}$$

where A_{-j} refers to elements of A other than A_j. The RIE has a direct relationship to the ATT or ATC depending on the nature of the intervention that one wants to study. To see this, suppose that there is a binary intervention variable, $A_j = 0, 1$ and that the intervention of interest is one that sets A_j at 0 (e.g., it is an intervention that protects individuals from a harmful exposure). Then,

$$\begin{aligned} \psi_j &= \mathrm{E}[Y(0, A_{-j})] - E[Y] \\ &= \{\mathrm{E}[Y(0, A_{-j})|A_j = 0]Pr[A_j = 0] + \mathrm{E}[Y(0, A_{-j})|A_j = 1]Pr[A_j = 1]\} \\ &\quad -\{\mathrm{E}[Y(0, A_{-j})|A_j = 0]Pr[A_j = 0] + \mathrm{E}[Y(1, A_{-j})|A_j = 1]Pr[A_j = 1]\} \\ &= \{\mathrm{E}[Y(0, A_{-j})|A_j = 1] - \mathrm{E}[Y(1, A_{-j})|A_j = 1]\}Pr[A_j = 1]. \end{aligned} \tag{6}$$

Now note that the ATT for A_j is defined as

$$\mathrm{ATT} \equiv \mathrm{E}[Y(1, A_{-j})|A_j = 1] - \mathrm{E}[Y(0, A_{-j})|A_j = 1] = -\frac{\psi_j}{Pr[A_j = 1]}. \tag{7}$$

For this intervention, the RIE has a close relationship to the ATT. A similar decomposition would follow for the ATC if we defined the intervention of interest as one that sets A_j to 0. What is important to note here is how the RIE depends on the nature of the intervention that is being considered and how it incorporates information on the proportion of units that would be affected by the intervention.

We set the RIE as our target for a few reasons. First, it compares a policy-relevant counterfactual to what has actually happened. It allows us to answer the question of whether it would have been "worth it" to have pursued various interventions, using observed reality as a benchmark. We

https://doi.org/10.1093/pan/mpw019 Published online by Cambridge University Press

feel that this provides a very coherent way to assess the policy relevance of different causal factors. It takes as a starting place considerations of whether a causal factor could be manipulated, to what extent and at what cost, and then quantifies the effects. Second, the nature of the comparison limits the number of "unknowns" that we need to address in the analysis while still allowing us to address policy-relevant questions clearly. Given our sampling design, the observed outcome mean ($\mathrm{E}[Y]$) is identifiable from our data with no special assumptions. Our analytical task is merely to characterize the counterfactual mean ($\mathrm{E}[Y(\underline{a}_j, A_{-j})]$). This makes for a more tractable analysis than would be the case, say, of comparing two counterfactual means when estimating an ATE (e.g., comparing two hypothetical interventions against each other). Our approach is consistent with the recommendations of Manski (1995, chap. 3), who proposes that one should target causal estimands depending on the data at hand, the policy questions one wants to answer, and the treatment regimes that different policies might imply.

5 Identification and Estimation

The identification of the RIE, ψ_j, requires the following assumptions.

Assumption 1. *$A = a$ implies $Y = Y(a)$.*

Van der Laan and Rose (2011) and VanderWeele (2009) call this the "consistency" assumption, and it also forms the basis of what Rubin (1990) calls the "stable unit treatment value assumption," or SUTVA. It means that when we observe $A = a$ for a unit, we are sure to observe the corresponding potential outcome $Y = Y(a)$ for that unit, and this is true regardless of what we observe in other units.[6] This assumption would be violated in situations of "interference," where units' outcomes are affected by the treatment status of other units (Cox 1958). In such cases, one could try to redefine units of analysis to some higher level of aggregation such that Assumption 1 is plausible.

Assumption 2. *For any $\underline{a}_j$ considered in the analysis, $Y(\underline{a}_j, A_{-j}) \perp\!\!\!\perp A_j | (W, A_{-j})$.*

This conditional independence assumption requires that conditioning on W and A_{-j} breaks any dependence between the realized value of the particular exposure, A_j, and potential outcomes when $A_j = \underline{a}_j$. The causal graph in Fig. 1 establishes that this assumption allows for causal identification. This assumption would be violated if the true data-generating process departed from Fig. 1 in particular ways, including causal relations between the elements of A, or the existence of other unmeasured confounders that causally determined Y and elements of A. In such cases, one would either have to limit the analysis to elements of A for which Fig. 1 is valid or collect additional data to restore the causal dependence and independence assumptions encoded by Fig. 1.

Assumption 3. *For all $\underline{a}_j$ considered in the analysis, $Pr[A_j = \underline{a}_j | W, A_{-j}] > b$ for some $b > 0$.*

This "positivity" or "covariate overlap" assumption allows us to construct the counterfactual distribution of potential outcomes under the intervention, $A_j = \underline{a}_j$, using the set of observations for which $A_j = \underline{a}_j$ in the sample (Petersen et al. 2011). This assumption is necessary to identify the population-level counterfactual and therefore to obtain the population-level RIE. If it does not hold, then identification would be restricted to the subpopulation with values of W and A_{-j}, for which Assumption 3 does hold.

These assumptions above identify the population-level counterfactual mean, $\mathrm{E}[Y(\underline{a}_j, A_{-j})]$, as follows:

[6] This usage of the word "consistency" should not be confused with its other meaning with reference to the asymptotic convergence of an estimator to a target parameter.

https://doi.org/10.1093/pan/mpw019 Published online by Cambridge University Press

$$\begin{aligned}\mathrm{E}[Y(\underline{a}_j, A_{-j})] &= \mathrm{E}[\mathrm{E}[Y(\underline{a}_j, A_{-j})|W, A_{-j}]] \\ &= \mathrm{E}[\mathrm{E}[Y(\underline{a}_j, A_{-j})|W, A_{-j}, A_j = \underline{a}_j]] \\ &= \mathrm{E}[\mathrm{E}[Y|W, A_{-j}, A_j = \underline{a}_j]],\end{aligned} \tag{8}$$

where the last term can be estimated using the observed Y outcomes for units with $A_j = \underline{a}_j$. The outer expectation is what is key: in constructing this counterfactual population average, one needs to weight the contributions of the (W, A_{-j})-specific Y means in a manner that corresponds to the distribution of (W, A_{-j}) in the population. The IPW approach that we explain below reweights the subpopulation of units with $A_j = \underline{a}_j$ such that it resembles the target population.

We use this identification result to construct an IPW estimator of the RIE:

$$\hat{\psi}_j^{\mathrm{IPW}} = \frac{1}{N}\sum_{i=1}^{n}\left(\frac{I(A_{ji} = \underline{a}_j)}{\hat{g}_j(\underline{a}_j|W_i, A_{-ji})} Y_i\right) - \overline{Y}, \tag{9}$$

where N is the sample size and $\hat{g}_j(\underline{a}_j|W_i, A_{-ji})$ is a consistent estimator for $Pr[A_j = \underline{a}_j|W_i, A_{-ji}]$. In essence, we take a weighted average of the outcomes of those units for which $A_j = \underline{a}_j$ without an intervention, where the weighting essentially expands each of these units' outcome contributions so that it proxies for the appropriate share of the population with $A_j \neq \underline{a}_j$. For example, if the intervention is the establishment of the income floor, $\underline{c}$, then the share of the population for which $A_j \neq \underline{a}_j$ is the share with incomes less than $\underline{c}$. To construct the counterfactual mean under the income floor intervention, we expansion-weight certain individuals with incomes higher than $\underline{c}$ to approximate contributions from those with incomes below $\underline{c}$. The way that we identify individuals to expansion-weight is through their covariate profiles, (W, A_{-j}). In the Supplementary Materials, we show that under mild conditions on the data, $\hat{\psi}_j^{IPW}$ is consistent for ψ_j, and we can construct conservative confidence intervals. In our application below, we also account for unequal probability cluster sampling.

6 Ensemble Methods for Propensity Scores

We do not typically know the functional form for the propensity score, $g_j(\underline{a}_j|W_i, A_{-ji})$, and so we use a machine learning ensemble method known as "super learning" to approximate such knowledge (Van der Laan, Polley, and Hubbard 2007; Polley, Rose, and Van der Laan 2011). The super learner methodology is very similar to ensemble Bayesian model averaging (EBMA) discussed by Montgomery, Hollanbach, and Ward (2012). Both super learning and EBMA compute a weighted average of the output of an ensemble of models, where each model is weighted on the basis of some loss criterion and loss scores for the members of the ensemble are generated using cross-validation. Ensemble methods relieve the analyst from having to make arbitrary choices about what estimation method to use and what specifications to fix for a given estimation method. Rather, the analyst is free to consider a variety of estimation methods (linear regression methods, tree-based methods, etc.). Then, the analyst uses cross-validation to determine the loss (e.g., the mean square prediction error) associated with each method. Finally, the loss value associated with each method is used to determine the weight given to predictions from each method in the analysis. Using cross-validated loss helps to minimize risks associated with over fitting.

To obtain our super learner ensemble estimate of the propensity score, we first obtain propensity score estimates from a set of candidate estimation algorithms. Then, to construct the ensemble estimate, we take a weighted average of estimates from the candidate algorithms. The weighting is done in a way that minimizes the expected mean squared error (MSE).

https://doi.org/10.1093/pan/mpw019 Published online by Cambridge University Press

Formally, we have a set of candidate estimation algorithms indexed by $c = 1, ..., C$. For each candidate algorithm we have an estimator, $\hat{g}_j^c(\cdot)$, that we fit to the data from each of the cross-validation splits, which are indexed by $v = 1, ..., V$. The cross-validation splits are constructed by randomly partitioning the data into V subsets; then each split consists of an estimation subsample of size $N - (N/V)$ and a hold-out sample of size $N_v = N/V$. For each candidate algorithm, we fit the model on the estimation subsample to obtain $\hat{g}_j^{c,v}(\cdot)$, and then we generate predictions to the units in the hold-out sample. From that, the average MSE over the cross-validation splits for candidate algorithm c is

$$\begin{aligned}\ell_j^c &= \frac{1}{V}\sum_{v=1}^{V}\frac{1}{N_v}\sum_{i=1}^{N_v}[I(A_{ji} = \underline{a}_j) - \hat{g}_j^{c,v}(\underline{a}_j|W_i, A_{-ji})]^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}[I(A_{ji} = \underline{a}_j) - \hat{g}_j^{c,v(i)}(\underline{a}_j|W_i, A_{-ji})]^2,\end{aligned} \qquad (10)$$

where $v(i)$ indexes the cross-validation split that contains unit i in the hold-out sample. The last line shows that each unit receives a set of predicted values generated by each algorithm from when the unit was in a hold-out sample. Moving from a single-candidate algorithm to the ensemble, we seek the minimum MSE weighted average of candidate algorithm estimates, which we obtain by solving for the ensemble weights as

$$(w_j^{1*}, ..., w_j^{C*}) = \arg\min_{(w_j^1,...,w_j^C)}\frac{1}{N}\sum_{i=1}^{N}\left[I(A_{ji} = \underline{a}_j) - \sum_{c=1}^{C} w_j^c \hat{g}_j^{c,v(i)}(\underline{a}_j|W_i, A_{-ji})\right]^2, \qquad (11)$$

$$\text{subject to} \sum_{c=1}^{C} w_j^c = 1 \text{ and } w_j^c \geq 0 \text{ for all } c.$$

One can obtain the $(w_j^{1*}, ..., w_j^{C*})$ weights vector by fitting a constrained non-negative least squares regression of the observed $I(A_{ji} = \underline{a}_j)$ values on the estimated $(\hat{g}_j^{c,v(i)}(\cdot), ..., \hat{g}_j^{C,v(i)}(\cdot))$ values (Van der Laan, Polley, and Hubbard 2007). Given these weights, we fit the candidate algorithms on the complete data, and the ensemble prediction for the propensity score is given as

$$\hat{g}_j(\underline{a}_j|W_i, A_{-ji}) = \sum_{c=1}^{C} w_j^{c*}\hat{g}_j^c(\underline{a}_j|W_i, A_{-ji}). \qquad (12)$$

(Van der Laan, Polley, and Hubbard 2007), Thm. 1 showed that under mild regularity conditions, the mean square error of prediction for $\hat{g}_j(\cdot)$ converges in N_v to the mean square error of the best candidate algorithm. Therefore, the consistency properties of $\hat{g}_j(\cdot)$ are inherited from the best candidate algorithm.

The candidate algorithms in our ensemble include the following: (i) logistic regression, (ii) t-regularized logistic regression (Gelman et al. 2008), (iii) kernel regularized least squares (KRLS) (Hainmueller and Hazlett 2014), (iv) Bayesian additive regression trees (BART) (Chipman, George, and McCulloch 2010), and (v) ν-support vector machine (SVM) classification (Chen, Lin, and Schoelkopf 2005; Hastie, Tibshirani, and Friedman 2009, chap. 12). This ensemble includes methods that are demonstrably effective in hunting out nonlinearities (e.g., KRLS and support vector classification) and interactions (e.g., BART).[7] We use ten cross-validation splits ($V = 10$ in our ensemble). Polley, Rose, and Van der Laan (2011) demonstrated that a ten-fold cross-validation super learner using some of these algorithms (they did not include KRLS) performs well in a wide range of data settings, including in estimating highly irregular and non monotonic conditional mean functions.

[7] This ensemble represents the full set of algorithms for which the authors know of research demonstrating effectiveness in relevant applied settings. In using the approach developed in this paper, researchers are free to consider other, potentially superior algorithms in their ensemble.

https://doi.org/10.1093/pan/mpw019 Published online by Cambridge University Press

In our illustration below, we use a rich covariate set, and so our ensemble relies primarily on regularized methods that reward sparsity (i.e., they shrink partial effects of covariates to zero) in order to further control over fitting (Bickel and Li 2006). Such regularization is likely to be important when the covariate set contains large amounts of noise that obscure identifying variation. The only non regularized method is logistic regression, which does not reward sparsity but is a method that we include because it remains the workhorse approach to propensity score estimation in political science. This provides a useful benchmark to evaluate gains from the much more computationally complicated algorithms and the ensemble routine overall since we can view the weight given by the super learner to logistic regression relative to the other methods.

The KRLS, BART, and ν-support vector classification and regression algorithms are based on models that grow in complexity with the data,[8] although such growth is constrained by regularization parameters. In a manner similar to Taylor approximation, allowing for more complexity helps to ensure improved approximations and consistency for the predicted mean conditional on the covariates included in the analysis (Greenshtein and Ritov 2004).

In our ensemble, we economize on computational costs by using the default rule-of-thumb settings for the regularization parameters that approximate MSE minimization.[9] In principle, one could incorporate into the ensemble multiple versions of each algorithm, with each version applying a different regularization parameter, and then construct the cross-validated error-minimizing combination, although this could entail relatively high computational costs.

7 Simulation Study

We provide evidence on finite sample performance of the ensemble method using a simulation study that illustrates the challenge of extracting meaningful variation in covariate sets as the noise-to-signal ratio increases.[10] We consider a situation in which we have observational data on an outcome Y, a single binary treatment variable $A = 0, 1$, and then a vector of covariates, W. Our estimand is the RIE for a hypothetical intervention that removes exposure to the treatment—that is, it sets A at 0 for everyone. This corresponds to the case that we explored above in the decomposition that relates the RIE to the ATT. The outcome Y depends on the value of A and underlying potential outcomes, $(Y(1), Y(0))$—that is, $Y = AY(1) + (1 - A)Y(0)$. We set up the simulation so that outcomes and treatment assignment probabilities are a function of only one covariate, W_1:

$$Y(0) = W_1 + .5(W_1 - \min(W_1))^2 + \epsilon_0 \tag{13}$$

$$Y(1) = W_1 + .75(W_1 - \min(W_1))^2 + .75(W_1 - \min(W_1))^3 + \epsilon_1$$

$$Pr[A = 1|W_1] = \text{logit}^{-1}\big(-.5 + .75W_1 - .5[W_1 - \text{mean}(W_1)]^2\big),$$

where $\epsilon_0 \sim \text{N}(0, 5^2)$, $\epsilon_0 \sim \text{N}(0, 10^2)$, $W_1 \sim \text{N}(0, 1)$, and $\min(W_1)$ and $\text{mean}(W_1)$ take the minimum and mean, respectively, of the sample draws of W_1 prior to producing the $(A, Y(0), Y(1))$ values.[11] Figure 2 displays data from an example simulation run.

One goal of the simulation is to show how our machine learning ensemble handles non linear and non monotonic functions such as the ones displayed in Fig. 2. Another goal is to study the challenge of working with a high-dimensional covariate set in which the identifying variation in W_1 is obscured by the existence of other covariates with little identifying power. Therefore, in addition to working with just W_1, we add first five and then ten dimensions of pure white noise to the covariate set—that is, five and then ten additional covariates, each drawn independently as $\text{N}(0, 1)$,

[8] Estimators that grow in complexity like this are known as "sieve" estimators (Geman and Hwang 1982).

[9] The rule-of-thumb methods are specific for each algorithm. See Gelman et al. (2008, 1364–65) for t-regularized logistic regression; Hainmueller and Hazlett (2014, 6–7) for KRLS; and Chipman, George, and McCulloch (2010, 269–73) for BART; and Chalimourda, Schoelkopf, and Smola (2004, 129) for ν-support vector classification.

[10] For replication materials, see Samii (2016).

[11] Using the minimum and mean in this way are simple ways to control how the non linearity appears in the sample.

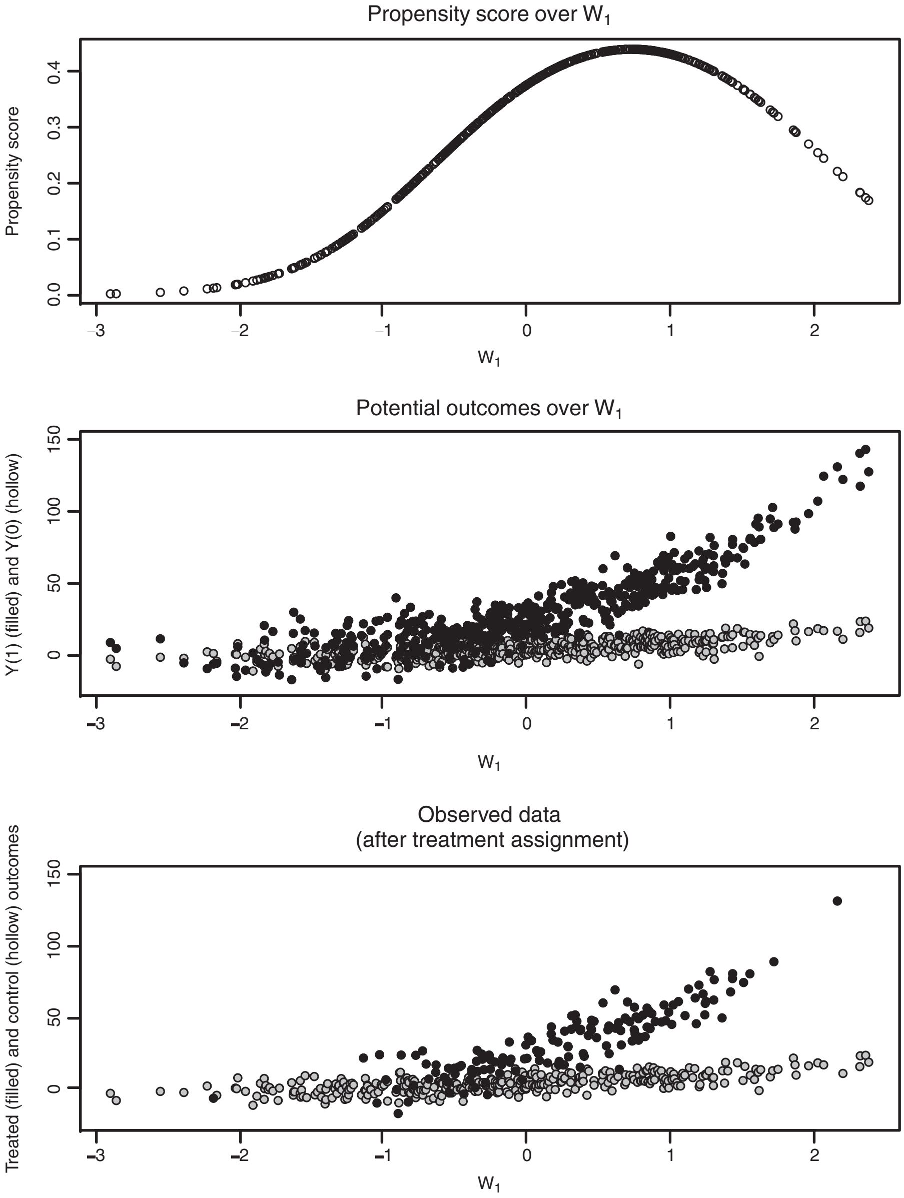

Fig. 2 Plots from an example simulation run. The top plot shows the expected value of the propensity score over the confounding covariate, W_1. The middle plot shows potential outcomes under treatment (filled) and control (hollow) for the full sample. The bottom plot shows observed outcomes for those assigned to treatment (solid) and control (hollow).

and thus unrelated to either Y or A. We want to see how well various methods perform in sorting through all of this noise to extract the variation that is meaningful for causal identification.

In our study, we compare four methods to estimate the RIE:

(i) OLS regression where we regress Y on W_1 and then the other covariates, with no interactions or higher-order terms, where the coefficient on A serves as our estimate;

(ii) Naïve IPW where we first estimate the propensity score using a logistic regression of A on W_1 and then the other covariates, with no interactions or higher-order terms; then, we use the estimated propensity score to construct the RIE estimate;

https://doi.org/10.1093/pan/mpw019 Published online by Cambridge University Press

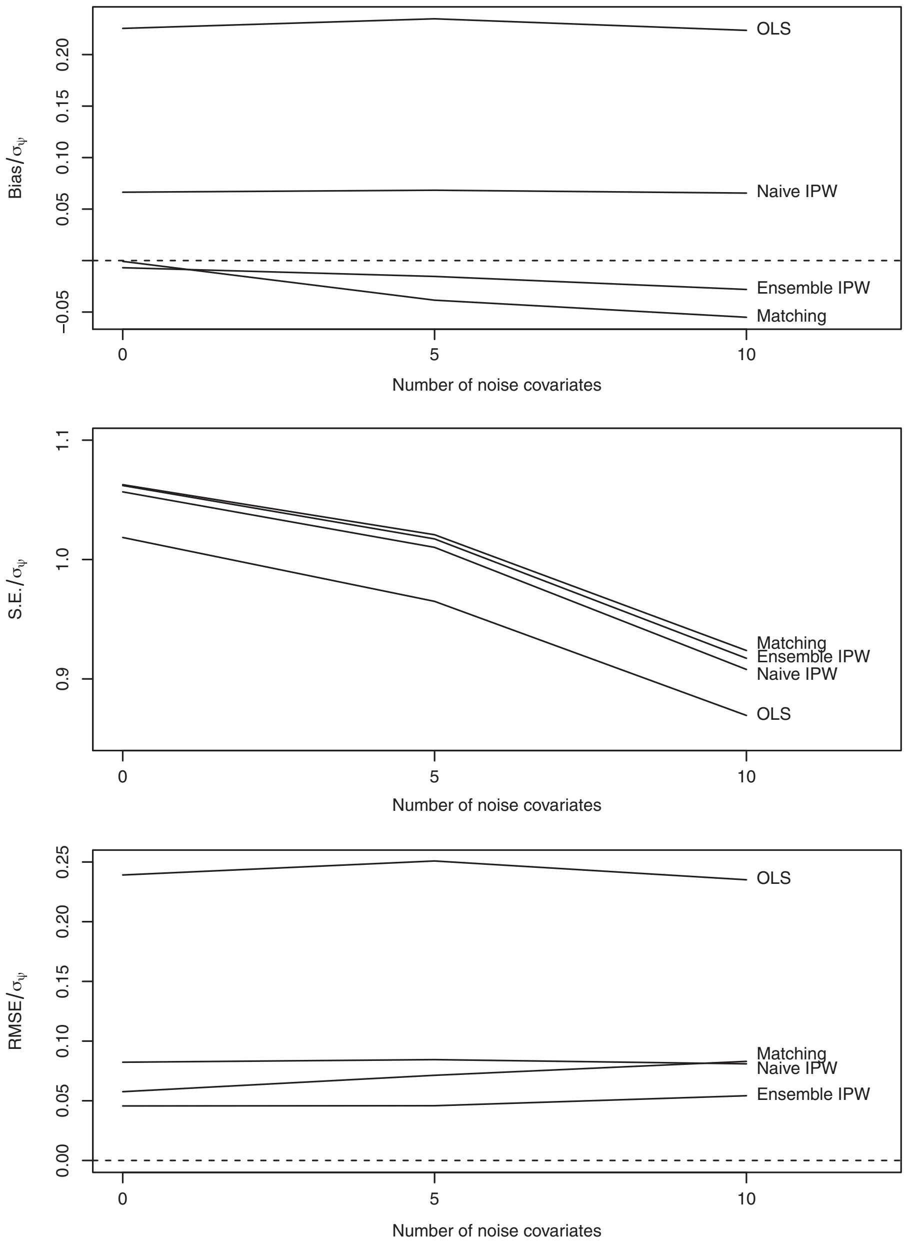

Fig. 3 Simulation results. From top to bottom, the graphs show bias, standard error (SE), and root mean square error (RMSE) for the different estimators of the RIE from 250 simulation runs as the number of noise covariates increases from 0 to 10. All results are standardized relative to the standard deviation of the true sample RIE across the simulation runs.

(iii) Mahalanobis distance nearest-neighbor matching with replacement on W_1 and the other covariates to construct the counterfactual quantities in the RIE expression and then combining them to compute the RIE; note that the Mahalanobis distance metric corresponds precisely to the joint normality of the covariates;

(iv) Ensemble IPW that first uses the machine learning ensemble that we described above to estimate the propensity score with W_1 and the rest of the covariates and then uses the estimated propensity score to construct the RIE estimate.

The data-generating process exhibits a combination of issues that complicate causal effect estimation in the real world: (i) effect heterogeneity, (ii) non linearities in the relationship between covariates and potential outcomes, (iii) non linearity in the relationship between covariates and propensity scores, and (iv) covariates of differing value for determining assignment and outcomes. The methods described above handle these issues differently, with consequences for expected bias. The OLS estimator ignores all four of the issues. The naïve IPW estimator ignores non linearity in the propensity score (issue 3) and the differing importance of covariates (issue 4). The matching estimator ignores the differing importance of covariates (issue 4). The ensemble IPW estimator attends, in principle, to all four issues.

Results from 250 simulation runs with a sample size of 500 are displayed in Fig. 3.[12] The graphs display bias, the standard error (i.e., standard deviation of estimates across the simulation runs), and then root mean square error (RMSE) for zero noise covariates, five noise covariates, and then ten noise covariates. These results are all standardized relative to the standard deviation of the true RIE over simulation runs ($\sigma_\psi = 3.60$). In terms of bias, the OLS and naïve IPW estimates are clearly poorest, owing to misspecification which for OLS fails to characterize the dramatically increasing effects in W_1 and for naïve IPW fails to capture the peak in the propensity score. The increase in noise covariates does not appreciably affect their biases. With no noise covariates, matching and ensemble IPW are similarly unbiased. Matching, however, is very sensitive to the increase in noise covariates. The problem is that, as we introduce more covariates, the meaningful differences (in terms of bias minimization) in W_1 are overwhelmed by meaningless differences in the other dimensions. As a result, matches tend to become more random relative to W_1, and because of the way the data are distributed in the covariate space, we get negative bias. The ensemble IPW estimator is much less sensitive to these problems—bias is half the magnitude when we get to ten covariates. All methods perform similarly in terms of their SEs, with matching performing slightly worse than the rest. RMSE combines these effects, showing that the ensemble IPW estimator is barely affected by higher dimensions of covariate noise. By the time we get to ten noise covariates, matching is performing as poorly (in an RMSE sense) as the misspecified naïve IPW estimator. The misspecified OLS estimator performs by far and away the worst.

The simulation captures the two reasons that we turn to machine learning ensembles. First, the ensemble is effective in the presence of irregular functional forms, and, unlike OLS or naïve IPW, we do not have to pre-specify these functional forms. Second, the ensemble is not overwhelmed by noise in the covariate space the way that matching is. Both estimators are consistent in terms of sample size for the RIE, but they differ in their finite sample performance depending on the amount of covariate noise. Matching's performance degrades substantially even with five or ten noise covariates. In the application below, the number of covariates is much higher.

8 Application to Anti-recidivism Policies in Colombia

Our application is to a study of policy alternatives to reduce recidivism among demobilized paramilitary and guerrilla fighters in Colombia. "Recidivism" refers to the committing of crimes such as murder, assault, extortion, or robbery after demobilization. Such recidivism among former combatants is at the heart of the troubling emergence of "*bandas criminales*" that have taken charge of narcotics trafficking and threatened social order across Colombia (International Crisis Group 2012). The analysis was meant to shed light on the kinds of interventions that might be most promising for the government to undertake to battle recidivism and increase former militants' reintegration into civilian life. Of particular interest was how funds might best be allocated across potential interventions targeting economic welfare, security, relations with authorities, psychological health, and relations among ex-combatants.

[12] The ensemble method is fairly slow to run because it employs ten-fold cross-validation, meaning that the simulations also run quite slowly. The results become quite stable after about 150 simulation runs; letting it run for 250 provides some extra security on convergence.

https://doi.org/10.1093/pan/mpw019 Published online by Cambridge University Press

Our data are from a representative multistage sample of 1158 ex-combatants fielded in forty-seven Colombian municipalities between November 2012 and May 2013 in collaboration with a Colombian think tank, *Fundación Ideas para la Paz*; the Colombian governmental department charged with the reintegration of former combatants, the *Agencia Colombiana para la Reintegración*; and the Organization of American States' *Misión de Apoyo al Proceso de Paz*. The survey sought to achieve representativeness for the population of demobilized combatants in crime-affected areas of Colombia and included prisoners and "hard to locate" ex-combatants, as well as ex-combatants in good standing with the authorities.[13] In addition to the survey responses for the individuals in the sample, we obtained a rich set of variables from administrative records of the Colombian attorney general's office (*Fiscalia General de la Nación*) and government agencies in charge of ex-combatant reintegration programs.

The first step of the analysis required that we define a set of risk factors and associated hypothetical interventions. We defined these in consultation with relevant government authorities, establishing a list of six risk factors and associated hypothetical interventions. These risk factors, associated variables, and hypothetical interventions are shown in Table 1. In some cases, the nature of the intervention has a clear programmatic interpretation, such as ensuring that the ex-combatant is employed. In other cases, the nature of the interventions is, admittedly, a bit vague. For example, ensuring that ex-combatants have confidence in government at a level that is above five in a ten-point scale does not have an immediately actionable interpretation. What we imagine is that there could be an intervention that generates such a change in attitudes.

Having established the risk factors and interventions, the next step was to establish a covariate set that would allow for credible causal identification. Our covariate set includes data extracted from the administrative files, measures obtained through the surveys, and then municipality fixed effects, for a total of 114 covariates. The covariates account for individuals' household, personal and various contextual circumstances prior to joining their respective armed groups, various facets of their experiences during their time in the armed groups, and the nature of their demobilization and reintegration experiences. To reduce measurement error, we performed a preliminary stage of dimension reduction using a one-factor latent trait analysis that reduced the dimensionality of our covariate set to a set of twenty-three indices constructed by taking inverse covariance–weighted averages of variables that can reasonably be assumed to capture common traits (O'Brien 1984). This preliminary step of dimension reduction was pre specified prior to data collection, which established ex ante the sets of items that were meant to capture common traits. The covariate set for our final analysis uses these twenty-three indices along with a vector of nine demographic traits and dummy variables for the forty-seven municipalities in which the subjects had demobilized, and so there was a total of seventy-nine covariates.

Having defined treatments and covariates, the last step in the data preparation was in defining and measuring outcomes. Given the sensitive nature of recidivism outcomes, we constructed a "recidivism vulnerability index." The index takes its highest value of three for known recidivists, and values ranging between zero and two on the basis of the number of clues that our data show suggest that the respondent is vulnerable to being recidivist. The index is based on information from attorney general records (history of arrest, charges, and imprisonment), responses to survey questions on crimes committed, responses to survey questions on the extent to which illegal behavior might be condoned, and responses to survey questions on exposure to opportunities in which crimes might be committed. The latter three were obtained via a self-administered questionnaire answered in private, following best practice in the survey literature for sensitive questions (Tourangeau and Yan 2005). Proven recidivists were those identified as such through the attorney general data or who, in our survey, admitted to being recidivist.

Table 2 displays the distribution of the recidivism index in the population and for subpopulations defined on the basis of the intervention variables. We estimate that the population is fairly evenly distributed over the recidivism index levels. For the intervention variables, however, we see that in some cases the population is not divided into two equally sized groups. For example,

[13] Details on the methods that we used to construct the sample are given in Daly, Paler, and Samii (2016).

https://doi.org/10.1093/pan/mpw019 Published online by Cambridge University Press

Table 1 Risk factors and hypothetical interventions

Risk factor	*Target variable description*	*Hypothetical intervention*
Economic welfare	Employed 1 year after demobilization	Unemployed are made employed.
Sense of security	Felt secure 1 year after demobilization	Insecure are made to feel secure.
Confidence in government	Confident 1 year after demobilization that government would keep promises	Not confident are made to feel confident.
Emotional well-being	Scale constructed from variables measuring how psychologically upbeat 1 year after demobilization	Psychologically depressed are made to feel upbeat.
Horizontal network relations with ex-combatants	Of 5 closest acquaintances, how many were ex-combatants 1 year after demobilization	Those with more than half ex-combatant peers are made to have less than half.
Vertical network relations with commanders	How regularly respondent spoke to commander 1 year after demobilization	Those who spoke to commander are made to rarely speak to commander.

Table 2 Recidivism vulnerability index outcome and intervention variables ($N = 1158$)

	Recidivism index value[a] =	*0*	*1*	*2*	*3*	*Mean*	*(SE)*
			(% in each category)				
i	Unweighted full sample	27	26	15	33	1.53	(0.04)
	Weighted full sample[b]	28	31	16	23	1.38	(0.06)
ii	Has employment = 0 (18%)	25	37	14	23	1.35	(0.09)
	Has employment = 1 (82%)	29	29	16	26	1.39	(0.07)
iii	Has security = 0 (39%)	23	25	22	20	1.60	(0.08)
	Has security = 1 (61%)	32	35	12	22	1.24	(0.07)
iv	Confidence in govt. = 0 (42%)	16	29	23	32	1.70	(0.07)
	Confidence in govt. = 1 (58%)	37	33	11	20	1.15	(0.07)
v	Not depressed = 0 (23%)	18	24	24	34	1.74	(0.14)
	Not depressed = 1 (77%)	31	33	14	22	1.27	(0.06)
vi	Few ex-com. peers = 0 (19%)	21	26	18	35	1.67	(0.12)
	Few ex-com. peers = 1 (81%)	30	32	15	23	1.31	(0.06)
vii	Doesn't speak to commander = 0 (15%)	22	23	17	38	1.71	(0.13)
	Doesn't speak to commander = 1 (85%)	29	32	16	24	1.32	(0.06)

Notes: i–vii, multiple imputation estimates of sample proportions; ii–vii, estimates use survey weights.
[a] Recidivism index values range from 0 = "non-recidivist" to 3 = "proven recidivist."
[b] Incorporates survey weights to account for unequal sampling probabilities across sample strata.

only 18% of the population reports that they were without employment one year after demobilization, and so it is only for this 18% that the hypothetical employment intervention would apply. Similar circumstances hold for the individuals who are depressed, have a large fraction of ex-combatants in their social networks, or continued to speak to their commanders. That being the case, the potential for interventions on these variables to make a major impact is limited to some extent. Only if the effects were very pronounced would the RIE be of substantial magnitude. We stress that this is a feature, not a bug, of the RIE approach: it tells us what kinds of policies might have the largest return, all things considered. This takes into account the possibility that the share of the population for which there is a particular "problem" may be quite small. Table 2 also shows differences in the recidivism index values over the intervention variables. We see pronounced differences for all but the employment variable. Of course, these comparisons could be biased by confounding. Our propensity score approach addresses this possibility.

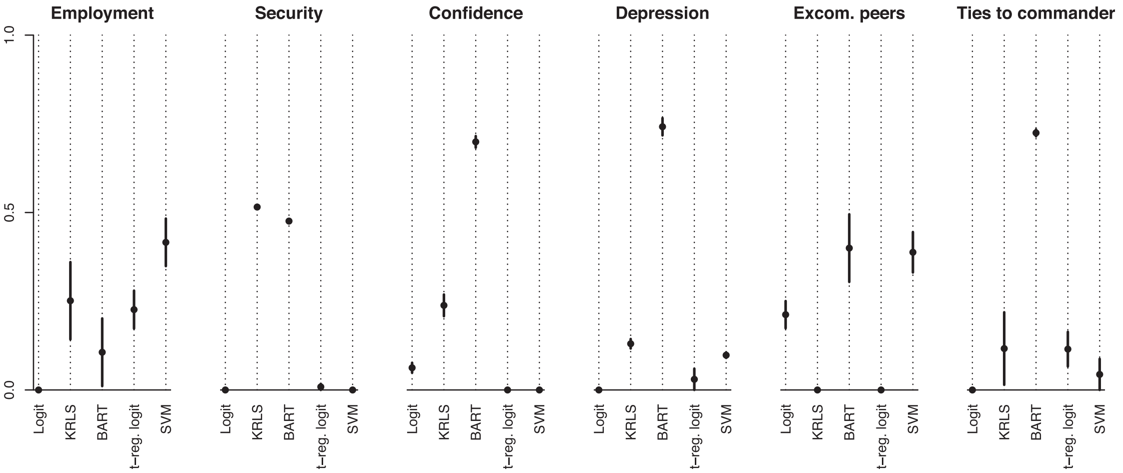

Fig. 4 Weights applied to propensity score predictions from each prediction method. The values of weights run along the y axis, and prediction methods run along the x axis. Results are grouped by intervention. The weights are constrained to be no less than zero and to sum to one for each intervention. The black bars show the range of the weights over the ten imputation runs, and the dots show the means.

The survey data exhibited small amounts of item-level missingness on the various covariates; however, such missingness adds up and would have resulted in dropping a non-negligible amount of data. We used ten-round multiple imputation, with imputations produced via predictive mean matching (Royston 2004). Because of the low item-level missingness, the imputation method is unlikely to make much of a difference in the results, and predictive mean matching is robust to misspecification. Estimates were constructed from the imputation-completed data sets using the usual combination rules, with point estimates computed as the mean of estimates across imputations and SEs computed in a manner accounting for both the within- and between-imputation variances (Little and Rubin 2002, 85–89). (In the Supplementary Materials, we show the workflow.) We fit the components of the ensemble using associated R (v.3.0.3) packages for each of the estimation methods. These were then fed into the *SuperLearner* package for R (Polley and Van der Laan 2012) to perform the cross-validation and MSE-based averaging that produced our propensity score estimates. Then, effects, SEs, and confidence intervals were constructed based on our survey design with the *survey* package in R (Lumley 2010).

Figure 4 shows the weights that the prediction methods received in the ensembles predicting the different intervention propensity scores. Recall that for each intervention, the weights are obtained from a constrained regression of the observed treatment values on the propensity scores from each prediction method, with the constraint being that coefficients cannot be less than zero and that they must sum to one. The figure shows the predictive performance of each method. Logistic regression performs very poorly, receiving zero weight in all ensembles except for the one predicting the propensity score for having few ex-combatant peer relationships. The weight given to the other methods varies over interventions. BART very regularly receives high weight—indeed, it is the only method that receives positive weight in all interventions. But BART's weight is surpassed for the employment and security intervention and essentially ties for first place for the ex-combatant peers intervention. Understanding why one or another method tends to perform well for different prediction problems could be a useful avenue for further research. But the main take away point here is that no single method would have been as reliable as the ensemble for these six prediction problems.

Figures 5 and 6 demonstrate how the IPW adjustment removed confounding for estimating the RIEs. Figure 5 shows the results of a placebo test that estimates pseudo-RIEs using *covariates* as outcome variables. Thicker horizontal bars are 90% confidence intervals, and thinner bars are 95% intervals. This plot allows us to see how the subpopulations that we use to form the counterfactual approximations differ from the overall population in terms of covariate means. The plot shows a high degree of imbalance. If we did not reweight by the inverse of $\hat{g}_j(.)$, these covariate imbalances would confound the RIE estimates. Figure 6 shows that the IPW adjustment removes these mean

https://doi.org/10.1093/pan/mpw019 Published online by Cambridge University Press

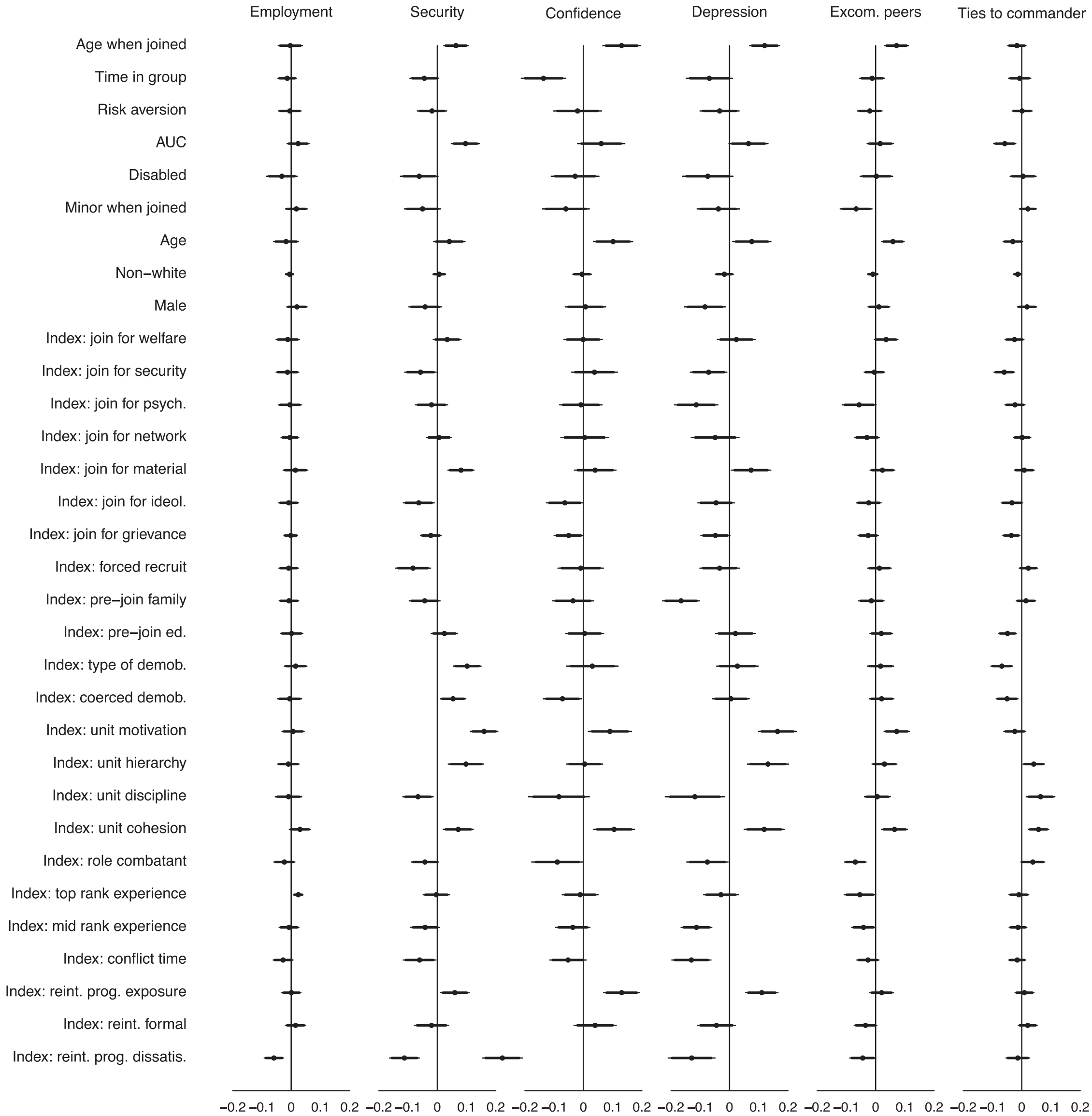

Fig. 5 Tests of mean balance for covariates and covariate indices in the raw data, prior to IPW adjustment. Mean differences are shown in standard deviation units. The horizontal bars passing through the points are the 95% (thin) and 90% (thicker) confidence intervals for the mean differences. "AUC" refers to membership in the *Autodefensas Unidas de Colombia*, the umbrella organization for paramilitary forces.

differences and the potential for confounding. A few covariates remain slightly out of balance in terms of their means, but no more than would be expected by chance (as evident from rates at which the confidence intervals fail to cover zero).

Figure 7 shows the distribution of propensity scores estimated by the ensemble for each intervention. The histograms display the propensity scores of units for which $A_{ji} = \underline{a}_j$. These are the units that are *not* subject to intervention and thus provide the outcome data used to construct the counterfactual mean for units that *are* subject to the interventions (i.e., for which $A_{ji} \neq \underline{a}_j$). The propensity scores are clearly bounded away from zero, which is important for estimator stability. In some cases, propensity scores are very close to the value of one, which is indicative of covariate combinations for which there would be few, if any, units subject to intervention in expectation.

Figure 8 plots RIE estimates and respective 95% confidence intervals. The figure displays the estimates based on the ensemble IPW method (black dots) and then estimates from the following

https://doi.org/10.1093/pan/mpw019 Published online by Cambridge University Press

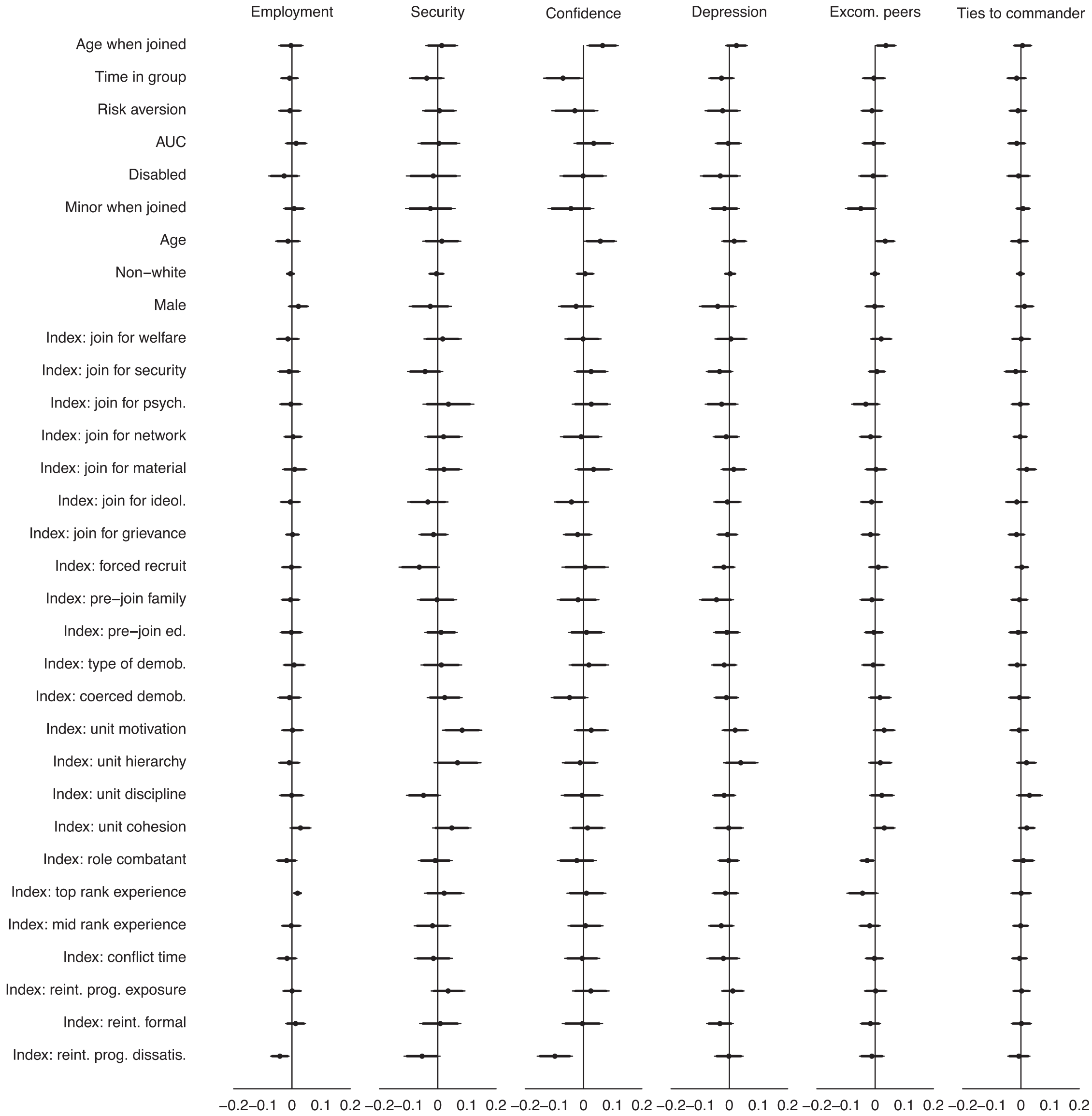

Fig. 6 Tests of mean balance for covariates and covariate indices with the IPW-adjusted data. Mean differences are shown in standard deviation units. The horizontal bars passing through the points are the 95% (thin) and 90% (thicker) confidence intervals for the mean differences.

comparison estimators: (i) a survey weighted least squares (WLS) regression, where the latter involved regressing the outcome on the hypothetical intervention variables and then on a control vector that included the twenty-three indices, demographic controls, and municipality fixed effects with no higher-order terms of interactions; (ii) a matching estimator that uses one-to-one Mahalanobis distance nearest neighbors matching with replacement to construct the counterfactual mean for those who would be subject to the intervention, with exact matching on municipality indicators; and (iii) a naïve IPW estimator that uses propensity scores from a logistic regression of the relevant treatment on a linear specification for the control variables.

The different estimators yield similar findings in terms of the general direction of the various effects and the way the different interventions are ranked in terms of their beneficial effects (note that negative estimates are beneficial in this context). Where the real differences lie are in the scale of the point estimates. The ensemble IPW estimates are generally closer to zero than the WLS

https://doi.org/10.1093/pan/mpw019 Published online by Cambridge University Press

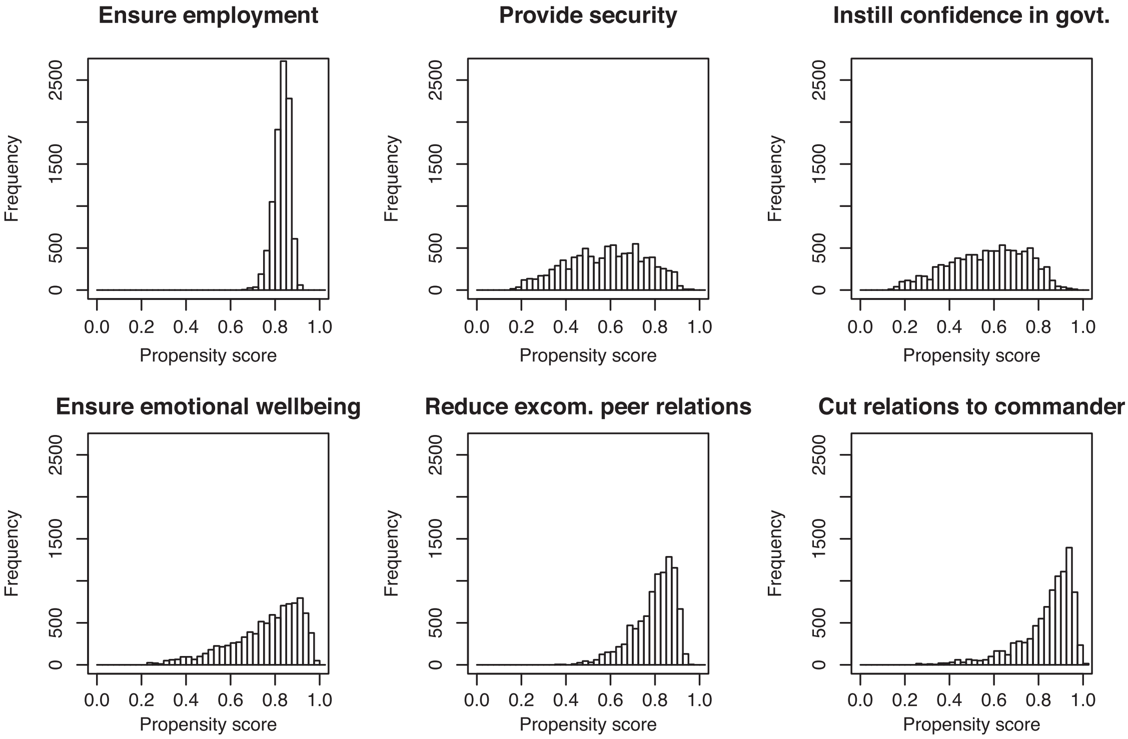

Fig. 7 Histograms of propensity scores estimated by the machine learning ensemble for each of the interventions. The histograms show propensity scores for those not subject to the intervention, as they are the units used to construct the counterfactual outcome distribution for those who are subject to the intervention.

estimates, but generally further away from zero than the matching or naïve IPW estimates. In policy analysis, these scale differences are important because cost-effectiveness analyses depend on the point estimates. The WLS estimates seem to exaggerate the effects of different interventions, where as the matching and naïve IPW estimates seem to heavily understate them.

The RIE estimates are defined in terms of shifts in the population mean. Recall from Table 2 that the population mean in the recidivism index is 1.38 with a standard deviation of 1.14. Thus, the ensemble IPW point estimate for what appears to be the most promising intervention—an intervention that instills confidence in government—is estimated to have reduced the average of recidivism tendencies by about 0.3 on the scale of the index or about a quarter of a standard deviation. That would be a very meaningful effect substantively. Note that the scale of this effect is a product of both the magnitude of the effect and the extent to which such an intervention would require the altering of individuals' treatment values. For this intervention, Table 2 shows that 42% of the ex-combatants had confidence index values below the intervention threshold,[14] and so it is for them that the intervention would induce a counterfactual change. By contrast, the hypothetical employment, emotional well-being, ex-combatant social networks, and relations to commander interventions would introduce counterfactual changes for smaller fractions of the population. For these interventions, the potential for a substantial RIE would be more limited on this basis. Even as such, we still find statistically and substantively significant RIEs for all but the employment intervention. This illustrates how the RIE is a population-level effect estimate, combining average unit-level effects with information on who should be treated. This yields a quantity that is immediately informative for policy.

[14] The percentage is the same with and without the survey weights.

https://doi.org/10.1093/pan/mpw019 Published online by Cambridge University Press

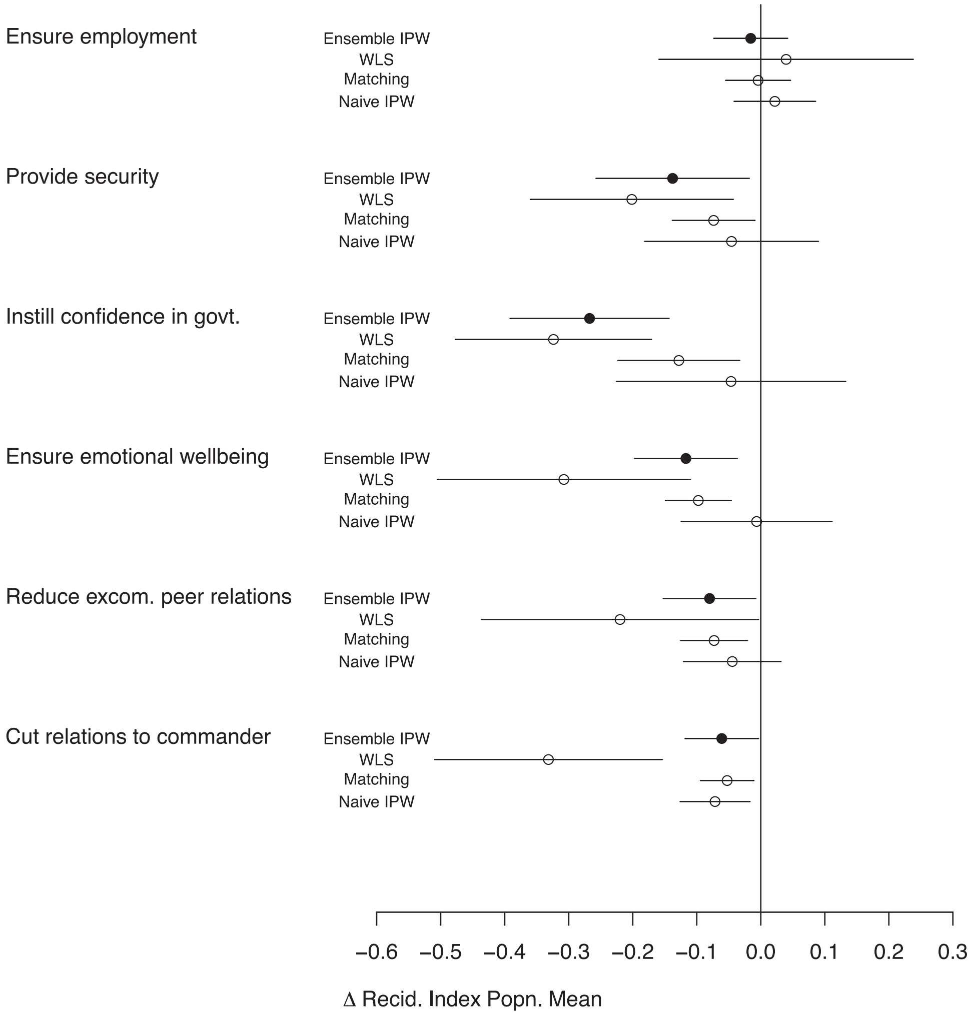

Fig. 8 RIE estimates. The vertical line indicates the location of a null effect. The plot shows point estimates (dots) and 95% confidence intervals (horizontal bars running through the dots). Ensemble IPW = inverse probability weighting RIE estimator, using the ensemble propensity scores; WLS = weighted least squares estimator based on a regression on the intervention variables and a simple linear specification for the covariates; Matching = nearest neighbor Mahalanobis distance matching RIE estimator; Naïve IPW = inverse probability weighting RIE estimator, using propensity scores from a logistic regression with a simple linear specification for the covariates.

9 Conclusion

This paper considers a method for retrospective causal inference that applies machine learning tools to sidestep problems with conventional approaches. Our approach has two core features that each confer benefits. First, we define the RIE. The RIE uses the device of hypothetical interventions to pin down clear population-level counterfactual comparisons. It also allows us to evaluate, in an easy-to-interpret manner, the relative importance of different risk factors and their effects on a population's outcomes. Second, we use a machine learning ensemble to use a large number of control variables for causal identification. A simulation experiment shows the robustness of the ensemble relative to conventional methods in extracting identifying variation from irregular functional relationships in a noisy covariate space. We reweight using predicted propensity scores to approximate the counterfactual defined under hypothetical interventions. This creates a contrast between what actually happened and an estimate of what might have been. An application to anti-

https://doi.org/10.1093/pan/mpw019 Published online by Cambridge University Press

recidivism policies in Colombia led to crisp conclusions about the relative merits of interventions on ex-combatants' confidence in government, social networks, security, and emotions when compared with other risk factors, such as employment.

The range of problems for which these methods can be applied is constrained by the three identifying assumptions: (i) treatment consistency/SUTVA, (ii) conditional independence, and (iii) positivity. The machine learning element frees us from the specification assumptions that previous methods also require. Treatment consistency and SUTVA can be established, in principle, by properly defining interventions and levels of analysis. For example, if SUTVA is thought to be violated at a low level of aggregation (e.g., individuals), there may be the possibility of satisfying it when we operate at a higher level of aggregation. Conditional independence can be made more believable if we measure a very large set of covariates. For methods requiring specification decisions, this in itself creates enormous complications. We overcome this challenge by incorporating regularized methods into our machine learning ensemble. The positivity assumption requires that there exist, in the real world, units that exhibit the diversity in the treatment variables and covariates needed to construct a counterfactual approximation for a hypothetical intervention (King and Zeng 2006). This assumption is perhaps the most restrictive. In some cases, it may be satisfied by redefining the target population (Crump et al. 2009). But doing so sacrifices the population-level inference that motivated us in the first place. As far as we understand, this is an unavoidable limitation for any observational method (and probably experimental too, given practical and ethical limitations on experimental subject pools).

Retrospective studies are a crucial first step in many research programs. They are essential for understanding causes of outcomes that are rare or that emerge only after many years. This includes outcomes such as violence or institutional change. Oftentimes the goal is to sort through a number of potential causal factors to identify points of intervention that should be prioritized for experimental or prospective studies. The conventional approach for doing so in the social sciences relies on multiple regression, for example, in conventional case-control studies (Korn and Graubard 1999; King and Zeng 2002). However, the validity of multiple regression estimates depends on homogeneity and model specification assumptions that cannot be defended in many instances, and especially so when the set of control variables is large. When the number of necessary control variables is large, other estimation methods such as matching, propensity score, or prognostic score methods either require modeling assumptions or make inefficient use of identifying variation. Under such circumstances, there is reason to be concerned about both bias and the potential for researcher discretion to undermine the validity of the analysis. The methods presented here demonstrate ways toward more objective and reliable retrospective causal inference.

The machine learning ensemble allows the researcher to address the bewildering specification challenges that arise when working with a large number of covariates. Having a large number of covariates at one's disposal allows, in principle, for more plausible causal identification under the conditional independence assumption. At the same time, it raises concerns about researchers selecting from among the vast number of potential specifications to manipulate results. The ensemble method can assuage such concerns in that it targets an objective criterion—the minimum expected error of prediction for the propensity score. This limits researcher degrees of freedom in the specification search, although it does not remove them entirely. The researcher still selects the algorithms, tuning parameters, loss functions, and preprocessing steps. Good faith is still required for credible inference.

10 Funding

Daly acknowledges funding from the Swedish Foreign Ministry, the Smith Richardson Foundation, and the Carroll L. Wilson Award.

Conflict of interest statement. None declared.

https://doi.org/10.1093/pan/mpw019 Published online by Cambridge University Press

References

Angrist, Joshua D., and Alan B. Krueger. 1999. Empirical strategies in labor economics. In *Handbook of labor economics*, eds. Orley C. Ahsenfelter and David Card, Vol. 3:1277–1366. Amsterdam: North Holland.

Angrist, Joshua D., and Jorn-Steffen Pischke. 2009. *Mostly harmless econometrics: an empiricist's companion*. Princeton, NJ: Princeton University Press.

Aronow, Peter M., and Cyrus Samii. 2016. Does regression produce representative estimates of causal effects? *American Journal of Political Science* 60(1):250–67.

Athey, Susan, and Guido W. Imbens. 2015. *Machine learning methods for estimating heterogeneous causal effects*. Working paper.

Bang, Heejung, and James M. Robins. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics* 61:962–72.

Bickel, Peter J., and Bo Li. 2006. Regularization in statistics. *Test* 15(2):271–344.

Blackwell, Matthew. 2013. A framework for dynamic causal inference in political science. *American Journal of Political Science* 57(2):504–19.

Busso, Matias, John DiNardo, and Justin McCrary. 2014. New evidence on the finite sample properties of propensity score reweighting and matching estimators. *The Review of Economics and Statistics* 96(5):885–97.

Chalimourda, Athanassia, Bernhard Schoelkopf, and Alex J. Smola. 2004. Experimentally optimal ν in support vector regression for difference noise models and parameter settings. *Neural Networks* 17:127–41.

Chen, Pai-Hsuen, Chih-Jen Lin, and Bernhard Schoelkopf. 2005. A tutorial on nu-support vector machines. *Applied Stochastic Models in Business and Industry* 21:111–36.

Chipman, Hugh A., Edward I. George., and Robert E. McCulloch. 2010. BART: Bayesian additive regression trees. *The Annals of Applied Statistics* 4(1):266–98.

Cox, David R. 1958. *Planning of experiments*. New York: Wiley.

Crump, Richard K., V. Joseph Hotz., Guido W. Imbens, and Oscar A. Mitnik. 2009. Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 96(1):187–99.

Daly, Sarah Zukerman, Paler Laura, and Samii Cyrus. 2016. *Wartime Networks and the Social Logic of Crime*. Typescript, University of Notre Dame, University of Pittsburgh: New York University.

Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. 2008. A weakly informative default prior for logistic and other regression models. *Annals of Applied Statistics* 2(4):1360–83.

Geman, Stuart, and Chii-Ruey Hwang. 1982. Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics* 10(2):401–14.

Green, Donald P., and Holger L. Kern. 2012. Modeling heterogenous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opinion Quarterly* 76(3):491–511.

Greenshtein, Eitan, and Ritov YaAcov. 2004. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* 10(6):971–88.

Grimmer, Justin, Solomon Messing, and Sean J. Westwood. 2014. Estimating heterogenous treatment effects and the effects of heterogenous treatments with ensemble methods. Unpublished manuscript, Stanford University.

Hainmueller, Jens. 2011. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis* 17(4):400–17.

Hainmueller, Jens, and Chad Hazlett. 2014. Kernel regularized least squares: Reducing misspecification bias with a flexible ad interpretable machine learning approach. *Political Analysis* 22(2):143–68.

Hansen, Ben B. 2008. The prognostic analogue to the propensity score. *Biometrika* 95(2):481–88.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

Hill, Jennifer. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20(1):217–40.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 15(3):199–236.

Holland, Paul W. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81(396):945–60.

Hubbard, Alan E., and Mark J. Van der Laan. 2008. Population intervention models in causal inference. *Biometrika* 95(1):35–47.

Imai, Kosuke, and Marc Ratkovic. 2013. Estimating treatment effect heterogeneity in randomized program evaluation. *Annals of Applied Statistics* 7(1):443–70.

Imai, Kosuke, and Aaron Strauss. 2011. Estimation of heterogenous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Political Analysis* 19(1):1–19.

Imai, Kosuke, and David A. van Dyk. 2004. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association* 99(467):854–66.

Imbens, Guido W. 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics* 86(1):4–29.

Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* 47(1):5–86.

International Crisis Group. 2012. Dismantling Colombia's new illegal armed groups: Lessons from a surrender. *International Crisis Group Latin America Report* 41.

https://doi.org/10.1093/pan/mpw019 Published online by Cambridge University Press

King, Gary, and Langche Zeng. 2002. Estimating risk and rate leveks, ratios, and differences in case–control studies. *Statistics in Medicine* 21(10):1409–27.

———. 2006. The dangers of extreme counterfactuals. *Political Analysis* 14(2):131–59.

Korn, Edward L., and Barry I. Graubard. 1999. *Analysis of health surveys*. New York: Wiley.

Little, Roderick J. A., and Donald B. Rubin. 2002. *Statistical analysis with missing data*, 2nd ed. Hoboken, NJ: Wiley.

Lumley, Thomas. 2010. *Complex surveys: A guide to analysis in R*. Hoboken, NJ: Wiley.

Manski, Charles F. 1995. *Identification problems in the social sciences*. Cambridge, MA: Harvard University Press.

Montgomery, Jacob M., Florian M. Hollanbach., and Michael D. Ward. 2012. Improving predictions using ensemble Bayesian model averaging. *Political Analysis* 20:271–91.

Myers, Jessica A., Jeremy A. Rassen., Jashua J. Gagne., Krista F. Huybrechts., Sebastian Schneeweiss, Kenneth J. Rothman., Marshall M. Joffe., and Robert J. Glynn. 2011. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology* 174(11):1213–22.

O'Brien, Peter C. 1984. Procedures for comparing samples with multiple endpoints. *Biometrics* 40(4):1079–87.

Pearl, Judea. 2009. *Causality: Models, reasoning, and inference*, 2nd ed. New York: Cambridge University Press.

Pearl, Judea. 2010. On a class of bias-amplifying variables that endanger effect estimates. In *Proceedings of UAI*, eds. Peter Grunwald and Peter Spirtes, 417–24. Corvallis, OR: AUAI.

Petersen, Maya L., Kristin E. Porter., Susan Gruber, Yue Wang, and Mark J. Van der Laan. 2011. Positivity. In *Targeted learning: Causal inference for observational and experimental data*, eds. Mark J. Van der Laan and Sherri Rose, chap. 10, 161–86. New York: Springer.

Polley, Eric C., and Mark J. Van der Laan. 2012. *SuperLearner: Super learner prediction*. R package version 2.0-9. http://cran.r-project.org/web/packages/SuperLearner/index.html.

Polley, Eric C., Sherri Rose, and Mark J. Van der Laan. 2011. Super learning. In *Targeted learning: Causal inference for observational and experimental data*, eds. Mark J. Van der Laan and Sherri Rose, chap. 3, 43–66. New York: Springer.

Ratkovic, Marc. 2014. Balancing within the margin: Causal effect estimation with support vector machines. Unpublished manuscript, Princeton University.

Robins, James M., and Andrea Rotnitzky. 1995. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* 90:122–29.

Rosenbaum, Paul R. 1984. The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A* 147(5):656–66.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55.

Rothman, Kenneth J., Sander Greenland, and Timothy L. Lash. 2008. *Modern epidemiology*, 3rd ed. Philadelphia, PA: Lippincott, Williams. and Wilkins.

Royston, Patrick. 2004. Multiple imputation of missing values. *Stata Journal* 4(3):227–41.

Rubin, Donald B. 1978. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics* 6(1):34–58.

———. 2008. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics* 2(3):808–40.

Rubin, Donald D. 1990. Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference* 25:279–92.

Samii, Cyrus. 2016. Replication data for: Retrospective causal inference with machine learning ensembles: An application to anti-recidivism policies in Colombia. http://dx.doi.org/10.7910/DVN/QXCFO2, Harvard Dataverse.

Sekhon, Jasjeet S. 2009. Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science* 12(1):487–508.

Tourangeau, Roger, and Ting Yan. 2005. Sensitive questions in surveys. *Psychological Bulletin* 133(5):859–83.

Van der Laan, Mark J., Eric C. Polley, and Alan E. Hubbard. 2007. Super learner. *Statistical Applications in Genetic and Molecular Biology* 6(1):1–21.

Van der Laan, Mark J., and Sherry Rose. 2011. *Targeted learning: Causal inference for observational and experimental data*. New York: Springer.

VanderWeele, Tyler 2009. Concerning the consistency assumption in causal inference. *Epidemiology* 20(6):880–83.

Young, Jessica G., Alan E. Hubbard., Brenda Eshkenazi, and Nicholas P. Jewell. 2009. *A machine-learning algorithm for estimating and ranking the impact of environmental risk factors in exploratory epidemiological studies*. University of California Berkeley Division of Biostatistics Working Paper Series 250.

https://doi.org/10.1093/pan/mpw019 Published online by Cambridge University Press