AEA Papers and Proceedings 2019, 109: 71–76
https://doi.org/10.1257/pandp.20191070

Augmenting Pre-Analysis Plans with Machine Learning[†]

By Jens Ludwig, Sendhil Mullainathan, and Jann Spiess*

I. Pre-Analysis Plans Pose a Specificity Conundrum

Randomized controlled trials (RCTs) are costly in terms of both time and dollars: just the New Jersey site alone of the Negative Income Tax experiments cost $60 million (Kershaw 1972), while Congress set aside $100 million for the Moving to Opportunity experiment, which has taken 25 years (and counting).[1] Analyzing the data that result from these RCTs is not straightforward. Investigators must specify which outcome variables (or combinations or transformations of outcomes) are explored, how to test for baseline balance, which subgroups are considered, and which control variables are used (and with what functional form).

The many degrees of freedom available to researchers raise fears of post-hoc analyses ("*p*-hacking"). Without properly sized tests we do not know whether to believe the findings, rendering these data—so costly to produce—nearly useless. One solution is to restrict these freedoms and have researchers specify their analyses in detail ahead of time. And indeed such pre-analysis plans (PAPs) have become common:[2] over 150 PAPs were entered into the American Economic Association registry in 2017, versus 20 in 2013.[3]

While pre-analysis plans reduce ex post fudging, they typically require a level of specificity that is usually not directly implied by the hypothesis being tested. Imagine for example writing the PAP for the $300 million RAND Health Insurance Experiment before it happened. A key question is whether health insurance affects health. This initially seems straightforward. But then you sit down to set pen to PAP and realize it requires answering the very specific question: What is "health"? Is it self-reported health status? Or number of physically unhealthy days? Or specific conditions like heart disease, diabetes, or cancer (or others)? What cut-points should we use—just for example body mass index (BMI) over 30 (obesity), or also over 25 (overweight), or 40 (morbid obesity)? How should we weight outcomes together into a single "health" index? There are many defensible specifications. None is self-evidently best for testing whether health insurance affects health. The other specific choices that PAPs require to prevent unhelpful human data mining, like subgroups, control variables, or functional forms, are also often made arbitrarily given how little theory has to say about them. We call this the *specificity conundrum.*

Recent developments in machine learning (ML) provide a different, more productive way for ex post analysis of experimental data. Supervised ML algorithms focus on finding prediction functions that are as accurate as possible out of sample, using the data to select the right variables and functional forms rather than relying on the investigator to specify these choices (for reviews see Hastie, Tibshirani, and Friedman 2001; Mullainathan and Spiess 2017). While designed for prediction, recent work shows how they can also help with many

*Ludwig: University of Chicago, 1155 East 60th Street, Chicago, IL 60637, and NBER (email: jludwig@uchicago.edu); Mullainathan: Chicago Booth School of Business, 5807 South Woodlawn Avenue, Chicago, IL 60637, and NBER (email: sendhil.mullainathan@chicagobooth.edu); Spiess: Microsoft Research New England, 1 Memorial Drive, Cambridge, MA 02142 (email: jspiess@stanford.edu). Paper prepared for the 2019 meetings of the American Economic Association. We thank Johann Gagnon-Bartsch, Hunt Allcott, and Steve Pischke for valuable comments. All opinions and any errors are of course our own.

[†] Go to https://doi.org/10.1257/pandp.20191070 to visit the article page for additional materials and author disclosure statement(s).

[1] All dollar amounts are in 2018 US dollars.

[2] Alongside the growing use of PAPs has been increased attention by economists to their strengths and limitations. See, for example, Casey, Glennerster, and Miguel (2012); Coffman and Niederle (2015); Olken (2015); Heckman and Singer (2017); Christensen and Miguel (2018).

[3] www.povertyactionlab.org/blog/2-15-18/addressing-challenges-publication-bias-rct-registration.

of the key analysis tasks associated with causal inference in RCTs (Table 1). However, generic off-the-shelf ML algorithms are data-intensive. Because our social science RCTs often do not reach the scale of data typically used for ML, these ML methods are not a perfect substitute for the pre-specified analyses of current PAPs.

In this paper we show one way to augment current PAPs with ML in a way that results in a "cheap-lunch": how to benefit from the flexibility ML tools provide, but at limited worst-case cost to power. Our goal is not to develop new ML methods. Rather, we show how existing ML tools can be productively deployed in the PAP framework.

Concretely, we augment a standard linear regression that we would usually pre-specify in a PAP with new regressors that come from ML. Properly sized tests require additional precise detail on *how* and *what* ML procedures will be used. The level of specificity required is often greater than with traditional PAPs, although for answering these questions we now have some guidance from data science. When done properly, ML augmentation produces limited costs in power in the worst case, relative to the original unaugmented PAPs, while providing power gains when the PAP was missing signal (from unspecified remainders). The careful integration of ML thus provides two gains over existing PAPs, by (i) limiting the need of researchers to make arbitrary analysis choices that are not implied by the initial conceptual hypothesis being tested, and (ii) integrating ex post analysis without fear of p-hacking.

II. Machine-Augmented Pre-Analysis Plans Provide a Cheap Lunch

To illustrate our approach we return to the example of whether there is any effect of health insurance on "health." A standard PAP would fully specify how different health variables are aggregated into a single test, while a pure ML approach could start with a set of variables and aggregate them into a single index by solving a prediction problem (Ludwig, Mullainathan, and Spiess 2017). In one case we must be completely specific about what outcomes to examine, while the other case allows the investigator to be unspecific, or at least only partially specific. Rather than picking one of those two extremes, we combine both approaches into a single test:

(i) We fit a machine-learning prediction function $\hat{f}$ of treatment assignment T from the group of outcomes $Y = (Y_1, Y_2, \ldots)$ (where subscripts denote different variables) that minimizes out-of-sample mean-squared error (MSE). We do this in a K-fold cross-validation form, where we first split the data into K parts and repeatedly fit a prediction function $\hat{f}$ on all but the current fold and then obtain fitted values $\hat{f}(Y)$ for all units in the left-out fold. By cycling through all folds, we obtain an "outcome index" $\hat{f}(Y)$ for each unit in our sample.

(ii) Given specific outcomes $Y^* = (Y_1^*, Y_2^*, \ldots, Y_J^*)$, which are a small number of (transformations of) variables from Y, and the outcome index $\hat{f}(Y)$, we run a joint (Wald) test on whether there is an average effect on any of these variables in our "health" group.

The idea is to not lose much power relative to testing the specific outcomes that we *can* pre-specify, Y^*, while avoiding the need to specify any ambiguous remainder of our test and still unlock an upside from the ML part: if there is indeed signal in the index $\hat{f}(Y)$, then we are able to detect it.[4] Indeed, the following proposition provides conditions that imply for this setting that, asymptotically:[5]

(i) The resulting joint test is valid for the null of "no effect on this group of outcomes";

(ii) If the effect is captured by the specific part, then the power loss of adding ML is at most that of adding one unaffected outcome; and

(iii) If $\hat{f}(Y)$ predicts T better than trivial (with respect to MSE), then power goes to 100 percent as the sample size grows.

[4] The ML part does not directly answer which health variables are affected, and how. However, augmentation has the advantage that we retain some interpretability from the specific part and can test whether the unspecific part adds signal.

[5] The results below are for the case of standard linear regression with an ML regressor on the right, but they carry over to this specific setting of multivariate regression with an ML outcome.

To provide some general results, we consider a researcher who fits an ML predictor $\hat{f}(Z)$ using K-fold cross-validation, and then runs a regression of some univariate outcome Y on regressors $Z^0, Z^*, \hat{f}(Z)$ to obtain OLS estimates $\left(\hat{\alpha}, \hat{\beta}, \hat{\gamma}\right)$, where Z^0, Z^* can be multivariate, and Z^0, Z^*, Z can include arbitrary covariates and treatment assignment.[6] Here, we think of Z^0 as the part of the regression we include but do not directly care about (such as a constant, in which case $\hat{\alpha}$ would be an intercept), and Z^* as the specific regressors related to the quantity we do care about (such as a treatment dummy, in which case $\hat{\beta}$ would be a treatment-effect estimate). We are interested in what $\left(\hat{\beta}, \hat{\gamma}\right)$ implies for the coefficients (β, γ) in the population linear regression

$$Y = \alpha' Z^0 + \beta' Z^* + \gamma f(Z) + \varepsilon,$$

where f is the limit of $\hat{f}$. In the following proposition we compare the output of this procedure to a standard linear regression of Y on Z^0 and Z^* to make inference on β. The online Appendix provides a full formal statement and a proof.

PROPOSITION 1 (Cheap Lunch): *Assume regularity conditions* (*see the online Appendix*) *and*

$$E\left[\left(\hat{f}(Z) - f(Z)\right)^2\right] \to 0$$

as $n \to \infty$. *Then, asymptotically for* $n \to \infty$:

(i) $\left(\hat{\beta}, \hat{\gamma}\right) \xrightarrow{P} (\beta, \gamma)$;

(ii) *If* $\gamma = 0$ *and* $E\left[\varepsilon \hat{f}(Z)\right] = 0$, *then* $\sqrt{n}\left(\left(\hat{\beta}, \hat{\gamma}\right) - (\beta, \gamma)\right)$ *has the same weak Normal limit as in an OLS regression on* $Z^0, Z^*, f(Z)$; *in particular*:

(a) *A Wald test of a null hypothesis with* $\gamma = 0, E\left[\varepsilon \hat{f}(Z)\right] = 0$ *is still valid*;

(b) *For alternatives with* $\beta \neq \mathbf{0}$, *the power loss is at most that of adding one irrelevant regressor in the OLS regression on* Z^0, Z^*;

(c) *If* $f(Z)$ *predicts the residual of the linear regression on* Z^0, Z^* *better than trivial* (*with respect to MSE*), *then power goes to* 100 *percent as* $n \to \infty$.

(iii) *If* $E\left[Z^* \hat{f}(Z)\right] = \mathbf{0}$ *and* $E\left[Z^* \left(Z^0\right)'\right] = \mathbf{00}'$ *then* $\sqrt{n}\left(\hat{\beta} - \beta\right)$ *has the same weak Normal limit as in an OLS regression on* $Z^0, Z^*, f(Z)$; *in particular:*

(a) *OLS inference on* β *remains valid;*

(b) *If* $\gamma = 0$, *then the variance of* $\hat{\beta}$ *is that in an OLS regression on* Z^0, Z^*;

(c) *If* $f(Z)$ *predicts the residual of the linear regression on* Z^0, Z^* *with* $E[Z^* \varepsilon^2 (Z^*)'] \prec E[Z^* (\gamma f(Z) + \varepsilon)^2 (Z^*)']$, *then the variance of* $\hat{\beta}$ *goes down.*

The first part of this proposition shows that the OLS estimates are still consistent for an appropriate population counterpart, assuming that the ML variance vanishes asymptotically. However, the fact that $\hat{f}(Z)$ is itself estimated from the data will in general affect standard errors and render standard OLS inference invalid. But our results also yield two specific cases for which OLS tests and standard errors remain valid. They encapsulate the idea of the example: under assumptions, adding ML does not invalidate the test, comes at a limited worst-case cost, and offers an upside.

With these results in hand, we show their relevance by applying them to the tasks associated with analyzing experimental data with randomly assigned binary treatment T.

When thinking about *which control variables to include* in estimating an average treatment effect, we can run the regression

$$Y = \alpha + \tau T + \beta' X^* + \gamma \hat{f}(X) + \varepsilon,$$

which combines specific control variables X^* with an ML index $\hat{f}(X)$ that predicts Y from the full vector X of controls (similar to Wager et al. 2016; Bloniarz et al. 2016; Wu and Gagnon-Bartsch 2018). By (iii), inference on β remains valid. For balanced treatment ($E[T] = 0.5$), the variance of the resulting estimator can only improve asymptotically, and will if $\hat{f}(X)$ predicts the OLS residual.

[6] By the Frisch-Waugh-Lovell theorem, we effectively fit machine learning on residuals of the linear regression on Z^0 and Z^* to absorb additional variation or uncover additional signal.

For testing *baseline balance*, the reverse regression

$$T = \alpha + \beta' X^* + \gamma \hat{f}(X) + \varepsilon$$

includes specific baseline controls X^* and an ML index $\hat{f}(X)$ that predicts treatment assignment T from the full vector X of controls (building upon Gagnon-Bartsch and Shem-Tov 2016). By (ii), a test of $\beta = \mathbf{0}, \gamma = 0$ is still valid for the null of "no difference between treatment and control distributions at baseline"; its loss in power relative to the specific test is at most that of adding one unaffected control to the original test; and if $\hat{f}(X)$ predicts treatment assignment better than random, power will approach 1.

For *heterogeneous treatment effects*,

$$Y = \alpha + \tau T + \beta' X^* + (\tau^*)' X^* T + \gamma \hat{\tau}(X)(T - E[T]) + \varepsilon$$

incorporates both interaction (or subgroup) effects with X^*, as well as an ML prediction $\hat{\tau}(X)$ of heterogeneous treatment effects (e.g., Athey and Imbens 2016; Wager and Athey 2018). By (ii), this specification allows tests for whether there are treatment effects; whether treatment effects are heterogeneous; and whether all heterogeneity is captured by the specific covariates X^*. The worst-case cost in power of adding ML is limited by that of the inclusion of an irrelevant interaction term, and if the ML estimate indeed picks up additional heterogeneous treatment effects, these tests will detect this in the limit.

By including ML on the right-hand side of OLS, we do not lose any sample size for the full linear regression, while allowing the data to decide how much weight to put on the ML component (via $\hat{\gamma}$).

III. Machine-Learning Components Need Pre-Specification as Well

In existing machine-learning solutions (Table 1), many of the analysis decisions that are so difficult to specify ex ante with the existing PAP approach are no longer key decisions. ML instead helps use the data themselves to answer these questions rather than have to pre-specify them. Instead, in the machine-augmented PAP (Table 2), our mental energy is now focused elsewhere: on the formulation of the larger conceptual research questions themselves, the pre-specification of any parts of the analysis we particularly care about or have strong priors over, and higher-level questions that affect the trade-offs in ML model selection, but do not force us to write down detailed s pecifications.

To see how involvement of machine learning changes the types of decisions that must be made, consider the data pre-processing stage. We still need to make decisions about what key variable transformations to create—indicators for BMI (itself a transformation of two underlying variables, height and weight) over 30, or also 25, 35, 40? With a standard PAP approach, erring on the side of too few transformations to examine risks missing signal, while pre-specifying too many risks reducing power from multiple-testing adjustments. But the machine-augmented PAP suggests an easier answer: since ML chooses which outcome transformations enter the resulting test, err on the side of over-inclusion in the ML component.

We also need to pre-specify some of the "ML engineering" decisions behind our analyses. For example, in testing for baseline balance, we wish to ensure that our procedure for trying to predict the outcome as a function of baseline control variables has as much statistical power as possible. Unlike the question of "what is health?," about which economic theory can provide limited specific guidance, the question of "what procedure yields the most accurate possible predictor in a given dataset?" is ultimately an empirical one, for which data science can provide some guidance.

While machine augmentation takes away the burden of detailed *substantive* pre-specification of aspects not implied by the question or strong priors ("what is health?"), the ML component conversely requires an even more stringent level of *technical* pre-specification. In addition to the code that the researcher runs—which includes all aspects of the machine-learning algorithm, including function classes, regularizers, and data-splitting schemes—this pre-specification extends to the resolution of any residual randomness. Indeed, since many ML algorithms include stochastic components (such as random splits of the data), researchers should also set seeds in their code that make their analysis fully replicable.

Table 1—Example for Existing Machine-Learning Solutions for Experiments

Task	Source of ambiguity	Example ML solution
Balance check	Combination of baseline variables into a single test	Gagnon-Bartsch and Shem-Tov (2016): Can we predict treatment assignment using baseline covariates?
Average treatment effect	Controls to include in regression	Wager et al. (2016); Bloniarz et al. (2016); Wu and Gagnon-Bartsch (2018): Which controls are most predictive of outcome?
Effect on groups of outcomes	Combination of outcome variables into single index	Ludwig, Mullainathan, and Spiess (2017): Which outcomes are most predictive of treatment assignment?
Subgroup analysis/heterogeneous treatment effects	Subgroups/interactions of interest	Athey and Imbens (2016); Wager and Athey (2018); Chernozhukov, Demirer, Duflo, and Fernandez-Val (2018): Which baseline variables are most predictive of treatment-control difference in outcome?

Table 2—Machine-Augmented Pre-Analysis Plan

Category	What is to be specified for ML part	Specification for machine-augmented plan	Comments
Preprocessing	Transformations of raw variables in data frame	Pick variables and transformations of high value or interest	Leave in all info, including missingness; OK to have multiple versions of same variable
Balance check	ML algorithm that predicts treatment assignment from baseline controls	Joint test of treatment effect on selected baseline controls and ML predictions	Important to use well-powered procedure to avoid suspicion of deliberately underpowered test
Average treatment effect	ML algorithm that predicts outcome from controls	Regression of outcome on selected controls and ML predictions	In observational data second prediction task (Belloni, Chernozhukov, and Hansen, 2014; Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins 2018)
Effects on group of outcomes	Group of outcomes and ML algorithm that predicts treatment assignment from these variables	Joint test of treatment effect on selected outcomes and ML predictions	May care about prediction functions that fulfill monotonicity constraints
Subgroup analysis/heterogeneous treatment effects	ML algorithm that predicts treatment effects from covariates	Regression of outcome on selected treatment-control interactions and interaction of treatment and ML prediction	
Any ML algorithm	Code that includes: function class, regularization procedure, constraints on functions (e.g., monotonicity), search algorithm, cross-validation/hold-out split		Set seeds for full replicability

IV. Conclusion

We carry out RCTs because we recognize our understanding of how the world works is so limited. If we knew more, we could just collect data on all the relevant variables and directly estimate structural models.

But these same limits to our understanding complicate the current approach to PAPs, which require a great deal of ex ante specificity in order to avoid post-hoc analyses—the price we pay to limit human data mining. Relegating many of our tests to the category of "secondary analyses" leaves us with a set of invalid *p*-values and results that are hard to interpret. On the other hand, we also do not want to be overly constrained in how we analyze the data.

Machine-augmented PAPs provide us with a way to get a "cheap lunch." Principled ex post analysis lets us realize unexpected discoveries at the cost, in the worst case, of a modest loss of power relative to a standard PAP approach. This approach also capitalizes on the growing availability of large-sample RCTs in economics. Consider that while the RAND Health Insurance Experiment enrolled under 6,000 people, the Oregon Health Insurance Experiment had a sample of nearly 75,000. Given that unlocking the real power of machine learning requires "big data," the time may be increasingly right for a machine-augmented approach to PAPs.

REFERENCES

Athey, Susan, and Guido Imbens. 2016. "Recursive Partitioning for Heterogeneous Causal Effects." *Proceedings of the National Academy of Sciences* 113 (27): 7353–60.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "Inference on Treatment Effects after Selection among High-Dimensional Controls." *Review of Economic Studies* 81 (2): 608–50.

Bloniarz, Adam, Hanzhong Liu, Cun-Hui Zhang, Jasjeet S. Sekhon, and Bin Yu. 2016. "Lasso Adjustments of Treatment Effect Estimates in Randomized Experiments." *Proceedings of the National Academy of Sciences* 113 (27): 7383–90.

Casey, Katherine, Rachel Glennerster, and Edward Miguel. 2012. "Reshaping Institutions: Evidence on Aid Impacts using a Preanalysis Plan." *Quarterly Journal of Economics* 127 (4): 1755–1812.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *Econometrics Journal* 21 (1): C1–C68.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernandez-Val. 2018. "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments." NBER Working Paper 24678.

Christensen, Garret, and Edward Miguel. 2018. "Transparency, Reproducibility, and the Credibility of Economics Research." *Journal of Economic Literature* 56 (3): 920–80.

Coffman, Lucas C., and Muriel Niederle. 2015. "Pre-Analysis Plans Have Limited Upside, Especially Where Replications Are Feasible." *Journal of Economic Perspectives* 29 (3): 81–98.

Gagnon-Bartsch, Johann, and Yotam Shem-Tov. Forthcoming. "The Classification Permutation Test: A Flexible Approach to Testing for Covariate Imbalance." *Annals of Applied Statistics.*

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 1st ed. New York: Springer.

Heckman, James J., and Burton Singer. 2017. "Abducting Economics." *American Economic Review* 107 (5): 298–302.

Kershaw, David N. 1972. "A Negative-Income-Tax Experiment." *Scientific American* 227 (4): 19–25.

Ludwig, Jens, Sendhil Mullainathan, and Jann Spiess. 2017. "Machine Learning Tests for Effects on Multiple Outcomes." arXiv preprint arXiv:1707.01473.

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31(2): 87–106.

Olken, Benjamin A. 2015. "Promises and Perils of Pre-Analysis Plans." *Journal of Economic Perspectives* 29 (3): 61–80.

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association* 113 (523): 1228–42.

Wager, Stefan, Wenfei Du, Jonathan Taylor, and Robert J. Tibshirani. 2016. "High-Dimensional Regression Adjustments in Randomized Experiments." *Proceedings of the National Academy of Sciences* 113 (45): 12673–78.

Wu, Edward, and Johann A. Gagnon-Bartsch. 2018. "The LOOP Estimator: Adjusting for Covariates in Randomized Experiments." *Evaluation Review* 42 (4): 458–88.