American Economic Review: Papers & Proceedings 2015, 105(5): 491–495
http://dx.doi.org/10.1257/aer.p20151023

Prediction Policy Problems[†]

By Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer*

Empirical policy research often focuses on causal inference. Since policy choices seem to depend on understanding the counterfactual—what happens with and without a policy—this tight link of causality and policy seems natural. While this link holds in many cases, we argue that there are also many policy applications where causal inference is not central, or even necessary.

Consider two toy examples. One policymaker facing a drought must decide whether to invest in a rain dance to increase the chance of rain. Another seeing clouds must decide whether to take an umbrella to work to avoid getting wet on the way home. Both decisions could benefit from an empirical study of rain. But each has different requirements of the estimator. One requires causality: Do rain dances cause rain? The other does not, needing only prediction: Is the chance of rain high enough to merit an umbrella? We often focus on rain dance–like policy problems. But there are also many umbrella-like policy problems. Not only are these prediction problems neglected, machine learning can help us solve them more effectively.

In this paper, we (i) provide a simple framework that clarifies the distinction between causation and prediction; (ii) explain how machine learning adds value over traditional regression approaches in solving prediction problems; (iii) provide an empirical example from health policy to illustrate how improved predictions can generate large social impact; (iv) illustrate how "umbrella" problems are common and important in many important policy domains; and (v) argue that solving these problems produces not just policy impact but also theoretical and economic insights.[1]

I. Prediction and Causation

Let Y be an outcome variable (such as rain) which depends in an unknown way on a set of variables X_0 and X. A policymaker must decide on X_0 (e.g., an umbrella or rain dance) in order to maximize a (known) payoff function $\pi(X_0, Y)$. Our decision of X_0 depends on the derivative

$$\frac{d\pi(X_0, Y)}{dX_0} = \frac{\partial \pi}{\partial X_0} \underbrace{(Y)}_{\text{prediction}} + \frac{\partial \pi}{\partial Y} \underbrace{\frac{\partial Y}{\partial X_0}}_{\text{causation}}.$$

Empirical work can help estimate the two unknowns in this equation: $\frac{\partial Y}{\partial X_0}$ and $\frac{\partial \pi}{\partial X_0}$. Estimating $\frac{\partial Y}{\partial X_0}$ requires causal inference: answering how much does X_0 affect Y?

The other term, $\frac{\partial \pi}{\partial X_0}$, is unknown for a different reason. We know the payoff function, but since its value must be evaluated at Y, knowing the exact value of $\frac{\partial \pi}{\partial X_0}$ requires a prediction Y. We know how much utility umbrellas provide only once we know the level of rain.

Choosing X_0 therefore requires solving both causation and prediction problems. Assume

*Kleinberg: Cornell University, Ithaca, NY 14853 (e-mail: kleinber@cs.cornell.edu); Ludwig: University of Chicago, 1155 East 60th Street, Chicago, IL 60637 and NBER (e-mail: jludwig@uchicago.edu); Mullainathan: Harvard University, 1805 Cambridge Street, Cambridge, MA 02138 and NBER (e-mail: mullain@fas.harvard.edu); Obermeyer: Harvard Medical School, Boston, MA 02115 and Brigham and Women's Hospital (e-mail: zobermeyer@partners.org). For financial support we thank operating grants to the University of Chicago Crime Lab by the MacArthur and McCormick foundations and the National Institutes of Health Common Fund grant DP5 OD012161 to Ziad Obermeyer. Thanks to Marianne Bertrand, Kerwin Charles, and Elizabeth Santorella for helpful comments and Brent Cohn, Maggie Makar, and Matt Repka for extremely helpful research assistance. Any errors and all opinions are of course ours alone.

[†] Go to http://dx.doi.org/10.1257/aer.p20151023 to visit the article page for additional materials and author disclosure statement(s).

[1] A longer version of this paper (Kleinberg, Ludwig, Mullainathan, and Obermeyer 2015) fleshes out each of these points, providing greater detail on the model, the empirical work and a more thorough summary of machine learning.

away one of these terms—place an exclusion restriction—and only one problem remains. Rain dances are a pure causal inference problem because rain dances have no direct effect on payoffs $\frac{\partial \pi}{\partial X_0} = 0$. Umbrellas are a pure prediction problem because umbrellas have no direct effect on rain $\frac{\partial Y}{\partial X_0} = 0$.

This derivative also illustrates two key features of prediction problems. First, the need for prediction arises exactly because $\frac{\partial \pi}{\partial X_0}$ depends on Y. Prediction is necessary only because the benefit of an umbrella depends on rain. As we illustrate in the final section, this kind of dependency is common for many important policy problems. Second, because only $\hat{Y}$ enters the decision, prediction problems only require low error in $\hat{Y}$; they do not require the coefficients to be unbiased or causal.

II. Machine Learning

Standard empirical techniques are not optimized for prediction problems because they focus on unbiasedness. Ordinary least squares (OLS), for example, is only the best linear *unbiased* estimator. To see how it can lead to poor predictions, consider a two variable example where OLS estimation produced $\hat{\beta}_1 = 1 \pm 0.001$ and $\hat{\beta}_2 = 4 \pm 10$, suggesting a predictor of $x_1 + 4x_2$. But given the noise in $\hat{\beta}_2$, for prediction purposes one would be tempted to place a smaller (possibly 0) coefficient on x_2. Introducing this bias could improve prediction by removing noise.

This intuition holds more generally. Suppose we are given a dataset D of n points $(y_i, x_i) \sim G$. We must use this data to pick a function $\hat{f} \in \mathcal{F}$ so as to predict the y value of a new data point $(y, x) \sim G$. The goal is to minimize a loss function, which for simplicity we take to be $(y - \hat{f}(x))^2$.

OLS minimizes in-sample error, choosing from $\mathcal{F}_{lin}$, the set of linear estimators:

$$\hat{f}_{OLS} = \underset{f_\beta \in \mathcal{F}_{lin}}{\arg\min} \sum_{i=1}^{n} (y_i - f(x_i))^2,$$

but for prediction we are not interested in doing well *in sample*: we would like to do well *out of sample*. Ensuring zero bias in-sample creates problems out of sample. To see this, consider the mean squared error at the new point x, $MSE(x) = E_D[(\hat{f}(x) - y)^2]$. This can be decomposed as

$$\underbrace{E_D[(\hat{f}(x) - E_D[\hat{y}_0])^2]}_{\text{Variance}} + \underbrace{(E_D[\hat{y}_0] - y)^2}_{\text{Bias}^2}.$$

Because the f varies from sample to sample, it produces variance (the first term). This must be traded off against bias (the second term). By ensuring zero bias, OLS allows no trade-off.

Machine learning techniques were developed specifically to maximize prediction performance by providing an empirical way to make this bias-variance trade-off (Hastie, Tibshirani, and Friedman 2009 provide a useful overview). Instead of minimizing *only* in-sample error, ML techniques minimize:

$$\hat{f}_{ML} = \underset{f \in \mathcal{F}}{\arg\min} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda R(f).$$

Here $R(f)$ is a *regularizer* that penalizes functions that create variance. It is constructed such that the set of functions $\mathcal{F}_c = \{f \,|\, R(f) \leq c\,|\}$ create more variable predictions as c increases. For linear models, larger coefficients allow more variable predictions, so a natural regularizer is $R(f_\beta) = \|\beta\|^d$, which is the lasso and ridge estimators for $d = 1$ and 2 respectively. In effect, this minimization now explicitly includes a bias (in-sample error) and variance term $(R(f))$, where λ can be thought of as the price at which we trade off variance to bias. OLS is a special case where we put an infinite (relative) price on bias $\left(\frac{1}{\lambda} = \infty\right)$.

A key insight of machine learning is that this price λ can be chosen *using the data itself*. Imagine we split the data into f subsets (often called "folds"). For a set of λ, we estimate the algorithm on $f - 1$ of the folds and then see which value of λ produces the best prediction in the fth fold. This cross-validation procedure effectively simulates the bias-variance trade-off by creating a way to see which λ does best "out of sample."

These two insights—regularization and empirical choice of the regularization penalty—together also change the kinds of predictors we can consider. First, they allow for "wide" data, to predict even when we have more variables than data points. For example, researchers using language data often have ten or a hundred times as

many variables as data. Second, this allows for far more flexible functional forms. One can include many higher order interaction terms or use techniques such as decision trees which by construction allow for a high degree of interactivity.

Machine learning techniques are in one sense not new: they are a natural offshoot of non-parametric statistics. But they provide a disciplined way to predict $\hat{y}$ which (i) uses the data itself to decide how to make the bias-variance trade-off and (ii) allows for search over a very rich set of variables and functional forms. But everything comes at a cost: one must always keep in mind that because they are tuned for $\hat{y}$ they do not (without many other assumptions) give very useful guarantees for $\hat{\beta}$.

III. Illustrative Application

Osteoarthritis (joint pain and stiffness) is a common and painful chronic condition among the elderly. Replacement of the affected joints, most commonly hips and knees, provide relief each year to around 500,000 Medicare beneficiaries in the United States. The medical benefits B are well understood: surgery improves quality of life over the patient's remaining life expectancy Y. The costs C are both monetary (roughly \$15,000 calculated using 2010 claims data) and nonmonetary: surgeries are painful and recovery takes time, with significant disability persisting months afterwards. The benefits accrue over time, so surgery only makes sense if someone lives long enough to enjoy them; joint replacement for someone who dies soon afterward is futile—a waste of money and an unnecessary painful imposition on the last few months of life.

The payoff to surgery depends on (eventual) mortality, creating a pure prediction problem. Put differently, the policy challenge is: can we predict which surgeries will be futile using only data available at the time of the surgery? This would allow us save both dollars and disutility for patients.

To study this example we drew a 20 percent sample of 7.4 million Medicare beneficiaries, 98,090 (1.3 percent) of which had a claim for hip or knee replacement surgery in 2010.[2] Of these, 1.4 percent die in the month after surgery, potentially from complications of the surgery itself, and 4.2 percent die in the 1–12 months after surgery. This low rate—roughly the average annual mortality rate for all Medicare recipients—seems to suggest on average surgeries are not futile. But the average is misleading because the policy decision is really about whether surgeries on the *predictably riskiest patients* were futile.

To answer this, we predicted mortality in the 1–12 months after hip or knee replacement using lasso (see Kleinberg, Ludwig, Mullainathan, and Obermeyer 2015 for full details).[3] We used 65,395 observations to fit the models and measured performance on the remaining 32,695 observations. 3,305 independent variables were constructed using Medicare claims dated prior to joint replacement, including patient demographics (age, sex, geography); co-morbidities, symptoms, injuries, acute conditions, and their evolution over time; and health-care utilization.

These predictions give us a way to isolate predictably futile surgeries. In Table 1, we sort beneficiaries by predicted mortality risk, showing risk for the riskiest 1 percent, 2 percent, and so on, *which is highly and predictably concentrated*: for example, the 1 percent riskiest have a 56 percent mortality rate, and account for fully 10 percent of all futile surgeries.[4]

Imagine the dollars from these futile surgeries could instead have been spent on other beneficiaries who would benefit more. To understand the potential savings, we simulated the effect of substituting these riskiest recipients with other

[2] We restricted to fee-for-service beneficiaries with full claims data living in the continental United States, and exclude any with joint replacement in 2009.

[3] This interval reflects two choices. (i) We excluded deaths in the first month after surgery to focus on prediction of Y rather than the short-term causal effect of X_0 on Y (i.e., operative risk, post-surgical complications). (ii) We chose a threshold of 12 months based on studies showing substantial remaining disability six months after surgery, but improved clinical outcomes at the 12-month mark (Hamel et al. 2008). Alternatively, a "break-even" threshold could be derived empirically.

[4] One might wonder whether these riskier patients may also be the ones who also stood to benefit the most from the procedure, potentially justifying surgery. However, variables that should correlate with surgery benefit (number of physician visits for hip or knee pain, physical therapy, and therapeutic joint injections) do not vary significantly by predicted mortality risk. In practice, this exercise is approximate, since some replacements may not have been elective, e.g., for fracture or other acute events. We present alternative specifications in our more detailed paper (footnote 3).

Table 1—Riskiest Joint Replacements

Predicted mortality percentile	Observed mortality rate	Futile procedures averted	Futile spending ($ mill.)
1	0.435 (0.028)	1,984	30
2	0.422 (0.028)	3,844	58
5	0.358 (0.027)	8,061	121
10	0.242 (0.024)	10,512	158
20	0.152 (0.020)	12,317	185
30	0.136 (0.019)	16,151	242

Notes: We predict 1–12 month mortality using an L_1 regularized logistic regression trained on 65,395 Medicare beneficiaries undergoing joint replacement in 2010, using 3,305 claims-based variables and 51 state indicators. λ was tuned using ten-fold cross-validation in the training set. In columns 1 and 2 we sort a hold-out set of 32, 695 by predicted risk into percentiles (column 1) and calculate actual 1–12 month mortality (column 2). Columns 3 and 4 show results of a simulation exercise: we identify a population of eligibles (using published Medicare guidelines: those who had multiple visits to physicians for osteoarthritis and multiple claims for physical therapy or therapeutic joint injections) who did not receive replacement and assign them a predicted risk. We then substitute the high risk surgeries in each row with patients from this eligible distribution for replacement, starting at *median* predicted risk. Column 3 counts the futile procedures averted (i.e., replaced with non-futile procedures) and column 4 quantifies the dollars saved in millions by this substitution.

beneficiaries who might have benefited from joint replacement procedures under Medicare eligibility guidelines, but did not receive them. To be conservative, rather than comparing to the lowest-risk eligibles, we draw from the median predicted risk distribution of these eligibles, and simulate effects of this replacement in columns 3 and 4. Replacing the riskiest 10 percent with lower-risk eligibles would avert 10,512 futile surgeries and reallocate the 158 million per year (if applied to the entire Medicare population) to people who benefit from the surgery, at the cost of postponing joint replacement for 38,533 of the riskiest beneficiaries who would not have died.[5]

[5] The existence of a large pool of low-risk beneficiaries potentially eligible for replacement argues against moral hazard as an explanation for these findings, since physicians who predicted well acting consistent with moral hazard would first exhaust the low-risk pool of patients before operating on higher-risk patients.

IV. Prediction Problems are Common and Important

Our empirical application above highlights how improved prediction using machine learning techniques can have large policy impacts (much like solving causal inference problems has had). There are many other examples as well. In the criminal justice system, for instance, judges have to decide whether to detain or release arrestees as they await adjudication of their case—a decision that depends on a prediction about the arrestee's probability of committing a crime. Kleinberg, Lakkaraju, Leskovec, Ludwig, and Mullainathan (2015) show that machine learning techniques can dramatically improve upon judges' predictions and substantially reduce the amount of crime.

Other illustrative examples include: (i) in education, predicting which teacher will have the greatest value added (Rockoff et al. 2011); (ii) in labor market policy, predicting unemployment spell length to help workers decide on savings rates and job search strategies; (iii) in regulation, targeting health inspections (Kang et al. 2013); (iv) in social policy, predicting highest risk youth for targeting interventions (Chandler, Levitt, and List 2011); and (v) in the finance sector, lenders identifying the underlying credit-worthiness of potential borrowers.

Even this small set of examples are biased by what we *imagine* to be predictable. Some things that seem unpredictable may actually be more predictable than we think using the right empirical tools. As we expand our notion of what is predictable, new applications will arise.

Prediction problems can also generate theoretical insights, for example by changing our understanding of an area. Our empirical application above shows that low-value care is not due just to the standard moral-hazard explanation of health economics but also to mis-prediction. The pattern of discrepancies between human and algorithmic decisions can serve as a behavioral diagnostic about decision-making (Kleinberg, Lakkaraju, Leskovec, Ludwig, and Mullainathan 2015). And prediction can shed light on other theoretical issues. For example, understanding

how people change their behavior as regulators or police change the algorithms they use to target monitoring effort can shed light on the game theory of enforcement.

Prediction policy problems are, in sum, important, common, and interesting, and deserve much more attention from economists than they have received. New advances in machine learning can be adapted by economists to work on these problems, but will require a substantial amount of both theoretical and practical reorientation to yield benefits for those currently engaged in policy studies.

REFERENCES

Chandler, Dana, Steven D. Levitt, and John A. List. 2011. "Predicting and Preventing Shootings among At-Risk Youth." *American Economic Review* 101 (3): 288–92.

Hamel, Mary Beth, Maria Toth, Anna Legedza, and Max Rose. 2008. "Joint Replacement Surgery in Elderly Patients With Severe Osteoarthritis of the Hip or Knee." *Archives of Internal Medicine* 168 (13): 1430–40.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Kang, Jun Seok, Polina Kuznetsova, Michael Luca, and Yejin Choi. 2013. "Where *Not* to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews." In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1443–48. Stroudsburg, PA: Association for Computational Linguistics.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2015. "Human decisions and machine predictions." Unpublished.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Policy prediction problems." Unpublished.

Rockoff, Jonah E., Brian A. Jacob, Thomas J. Kane, and Douglas O. Staiger. 2011. "Can you recognize an effective teacher when you recruit one?" *Education Finance and Policy* 6 (1): 43–74.