Statistical Science
2020, Vol. 35, No. 3, 430–433
https://doi.org/10.1214/20-STS772
Main article: https://doi.org/10.1214/20-STS721

Comment: Invariance and Causal Inference

Stefan Wager

The problem of distinguishing causal effects from non-causal correlations is one of the oldest and most challenging questions in statistics. In recent years, Professor Bühlmann and co-authors have outlined new methodology for estimating causal effects that starts from an invariance postulate: A set of variables X is causally relevant to an outcome Y if the distribution of Y conditionally on X, $\mathcal{L}(Y \mid X)$, is invariant across all relevant environments. This hypothesis then leads to statistical methodologies that seek causal effects by fitting models that are robust across numerous environments (Peters, Bühlmann and Meinshausen, 2016, Rothenhäusler et al., 2019). The present paper, generously prepared by Professor Bühlmann, is an enlightening summary of this groundbreaking line of work and a valuable addition to the literature.

This invariance hypothesis presents a marked and thought-provoking departure from the currently dominant paradigm for understanding causal effects in epidemiology and econometrics, which defines causal effects in terms of potential outcomes and emphasizes the role of experimental design in identifying causal effects (Neyman, 1923, Holland, 1986, Robins and Richardson, 2010, Rubin, 1974, Rubin, 2005). In general, the potential outcomes based approach allows treatment effects to vary arbitrarily with both observed and unobserved features and is focused on defining, identifying and estimating various (weighted) treatment effect functionals under minimal assumptions. Characterizing how the invariance hypothesis fits into the potential outcomes framework is important to understanding how the results of Peters, Bühlmann and Meinshausen (2016) and Rothenhäusler et al. (2019) connect to more classical approaches.

Potential outcomes and weighted treatment effects. The earliest application of the potential outcomes framework was Neyman's analysis of the randomized controlled trial. In this setting, we are interested in measuring the effect of a binary treatment W_i on a real-valued outcome Y_i. We posit the existence potential outcomes $\{Y_i(0), Y_i(1)\}$ corresponding to the outcome the ith observation would have experienced had they received treatment assignment 0 or 1, respectively, such that $Y_i = Y_i(W_i)$, and then define the sample average treatment effect[1]

$$\tau_{\text{SATE}} = \frac{1}{n}\sum_{i=1}^{n}\big(Y_i(1) - Y_i(0)\big). \tag{1}$$

The seminal result of Neyman (1923) is that, if the treatment assignment W_i is randomized, that is, the treatment assignment is exchangeable and $\{W_i\}_{i=1}^n \perp\!\!\!\perp \{Y_i(0), Y_i(1)\}_{i=1}^n$, then we can construct an unbiased estimate of τ_{SATE} without assumptions: No modeling assumptions are made on the potential outcomes $Y_i(w)$, and in fact the potential outcomes may even be taken as deterministic such that only W_i is random.[2] In particular, it is not necessary to assume that the causal effect is the same for each unit, for example, that $Y_i(1) - Y_i(0) = \tau$ for some shared (or invariant) causal parameter τ.

Starting with Rubin (1974), there has been considerable interest in generalizing the ideas of Neyman (1923) beyond the randomized controlled trial, and in developing appropriate treatment effect estimators that remain justified without making structural assumptions on the per-unit treatment effects $Y_i(1) - Y_i(0)$. One setting that has received considerable attention is that of Rosenbaum and Rubin (1983), where treatment assignment W_i is not randomized, but we observe covariates X_i such that W_i is as good as random after we condition on them, $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \mid X_i$. Under an IID sampling model, the semiparametric efficient variance V for estimating the average treatment effect $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$ can be written in terms of the propensity score $e(x) = \mathbb{P}[W_i = 1 \mid X_i = x]$ (Hahn, 1998, Robins and Rotnitzky, 1995),

$$V = \operatorname{Var}\big[\mathbb{E}\big[Y_i(1) - Y_i(0) \mid X_i\big]\big] + \mathbb{E}\left[\frac{\operatorname{Var}[Y_i(0) \mid X_i]}{1 - e(X_i)} + \frac{\operatorname{Var}[Y_i(1) \mid X_i]}{e(X_i)}\right],$$

and efficient estimators satisfy $\sqrt{n}(\hat{\tau} - \tau) \Rightarrow \mathcal{N}(0, V)$.

One complaint about this result, however, is that V scales with the inverse of the propensity score, and can get quite large if we have poor overlap (i.e., $e(X_i)$ can get

Stefan Wager is Assistant Professor of Operations, Information and Technology, and Assistant Professor of Statistics (by courtesy), Graduate School of Business, Stanford University, Stanford, California 94305, USA (e-mail: swager@stanford.edu).

[1] One major assumption here is that of no interference, that is, that W_i only affects the outcome of the ith unit (Imbens and Rubin, 2015). For a discussion of potential outcomes modeling under interference; see Basse, Feller and Toulis (2019), Hudgens and Halloran (2008) and references therein.

[2] These results can be considerably generalized. For example, Ding, Feller and Miratrix (2019) and Lin (2013) for a discussion of regression adjustments in this setting.

close to 0 or 1). We can avoid this problem by changing estimands. In medical settings, it is common to have some units who are essentially guaranteed to get control and have $e(X_i)$ very close to 0 because the studied treatment is simply not applicable to them. In cases like these, it may be more fruitful to estimate the average treatment effect on the treatment $\tau_{\text{ATT}} = \mathbb{E}[Y_i(1) - Y_i(0) \mid W_i = 1]$, in which case the semiparametric efficiency bound depends inversely on $1 - e(X_i)$ but not $e(X_i)$ (Hahn, 1998). Meanwhile, if we have overlap problems near both 0 and 1, Crump et al. (2008) and Li, Morgan and Zaslavsky (2018) advocate further modifying the estimand to improve the precision we can estimate it with; for example, one could target the overlap-weighted average treatment effect

$$\tau_{\text{ATO}} = \mathbb{E}\big[e(X_i)\big(1 - e(X_i)\big)\big(Y_i(1) - Y_i(0)\big)\big] \Big/ \mathbb{E}\big[e(X_i)\big(1 - e(X_i)\big)\big].$$

The upshot is that, under the unconfoundedness assumption of Rosenbaum and Rubin (1983), the accuracy with which we can estimate treatment effects in different parts of the feature space depends on the propensity score, and it is possible to mitigate excess variance by focusing on those parts of the feature space where $e(\cdot)$ is closest to 0.5. However, in a general sampling design, such re-weighting changes the estimand, and the literature on treatment effect estimation has gone to considerable trouble to understand how.[3]

Questions on how best to weight heterogeneous treatment effects are also of central importance in many settings beyond the above one. Imbens and Angrist (1994) consider treatment effect estimation under noncompliance, and discuss when instrumental variables methods can be used to identify the average treatment effect on the compliers. Regression discontinuity designs exploit sharp treatment assignment rules to identify treatment effects for marginal units close to the treatment boundary (Hahn, Todd and Van der Klaauw, 2001), and again carefully weighting our estimand can lead to gains in precision (Imbens and Wager, 2019). Athey and Wager (2017) and Kitagawa and Tetenov (2018) consider the problem of learning simple treatment assignment rules under arbitrary treatment heterogeneity.

Identification via invariance. The invariance-based approach takes a view that goes in essentially the opposite direction from the potential outcomes approach as discussed above. Instead of starting from a perspective that treatment effects are heterogeneous and then studying how different estimators can recover different weighted averages of the treatment effect function, it starts by positing the targeted causal effects as invariant, and then uses this invariance to design new identification strategies.

The invariance assumption is obviously a powerful idea, and has led to some striking empirical successes. In particular, it is far from clear how one might have approached the impressive gene-knockout study of Meinshausen et al. (2016) starting from a definition of causal effects via potential outcomes. What is not clear to me is the relationship between the kinds of effects that are identified via the invariance-based methods discussed by Bühlmann, and (appropriately weighted) treatment effects that are the focus of epidemiological or econometric studies following the potential outcomes paradigm as outlined in Imbens and Rubin (2015). Understanding this connection further seems like an important topic for further study.

One area where synthesis between the potential outcome- and invariance-based approaches to causal inference may be particularly fruitful is in modeling "external validity", that is, how causal effects measured in once context are relevant in a new context. In a very crude sense, one could argue that the potential outcomes community strives to measure well-defined weighted in-study average treatment effects under minimal assumptions, but has focused less on how these effects transport to different contexts. Conversely, the invariance-based approach takes external validity as given and uses it to highlight causal effects that would remain unidentified in a pure potential outcomes setting. One question that is likely to benefit from insights from both communities is in understanding which representations of causal effects have the most reliable external validity. Such representations would be extremely helpful for data-driven policy making across heterogeneous environments. In any case, effects that are both "invariant" in the sense of Bühlmann's paper and "causal" in the sense of potential outcomes modeling are likely to be of particular scientific interest.

The California GAIN study. To highlight these issues in the context of an application, consider the following stud whose goal was to evaluate the efficacy of a jobs training program. California's Greater Avenues to Independence (GAIN) program is designed to help welfare recipients rejoin the workforce. Starting in 1988, the Manpower Demonstration Research Corporation conducted a randomized evaluation of the GAIN program; see Hotz, Imbens and Klerman (2006) for a detailed discussion of the experiment. The GAIN evaluation was run across multiple counties, including Alameda, Los Angeles, Riverside and San Diego. Each county participating in the experiment was given considerable discretion in how to implement GAIN. Riverside chose to focus on a labor force attachment approach centered around quickly moving people on welfare into jobs—even if they are low-paying

[3] One interpretation of the overlap-weighted estimand τ_{ATO} is the following. Suppose a statistician erroneously believed that all treatment effects were constant, $Y_i(1) - Y_i(0) = \tau$, and sought to estimate the parameter τ using the popular estimator of Robinson (1988). Then, in large samples, the resulting estimate $\hat{\tau}$ would not converge to the average treatment effect, but rather to τ_{ATO}.

TABLE 1
Site-specific average treatment effect estimates for the GAIN trial. All confidence intervals were obtained using a Welch two-sample t-test

County:	Alameda	Los Angeles	Riverside	San Diego
95% CI	0.16 +/− 0.19	0.03 +/− 0.09	0.25 +/− 0.10	0.13 +/− 0.11

jobs. In contrast, Alameda, Los Angeles and San Diego adopted more of a focus on human capital development, including education and vocational training.

Table 1 shows estimates of the average treatment effect of GAIN across the four counties. Following Hotz, Imbens and Klerman (2006), the outcome measure is average quarterly income (in $1000s) over the 9-year post-randomization period. We immediately see that the treatment effect estimate in Riverside is much larger than that in any other county. The finding that Riverside achieved a greater treatment effect than the other counties was quickly interpreted by several stakeholders to mean that labor force attachment was a more effective principle for the design of welfare-to-work programs than human capital development. To formalize this claim, one would need to make a kind of invariance argument: If we believed that all counties with a human capital development focused welfare-to-work program should have similar treatment effects, and the Riverside program stands out, then it is natural to attribute the exceptional success of the Riverside program to its use of a labor force attachment approach.

Hotz, Imbens and Klerman (2006), however, question this interpretation. The GAIN dataset includes a number of pre-treatment covariates X_i, and some interesting patterns arise when including them into the analysis. First, the participants in Riverside's GAIN evaluation had a different covariate distribution from GAIN participants.[4] Second, these covariates appear to explain a considerable amount of treatment heterogeneity, that is, the conditional average treatment effect function function $\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x]$ varies in x. Finally, adjusting for these pre-treatment covariates is able to explain at least some of differences between the treatment effects measured in different counties. For example, when training a causal forest for $\tau(x)$ (Athey, Tibshirani and Wager, 2019), the average of the out-of-bag predictions $\hat{\tau}^{(-i)}(X_i)$ in Los Angeles is 0.10 whereas the average of $\hat{\tau}^{(-i)}(X_i)$ in Riverside is 0.16.[5]

Thus, our interpretation of what the GAIN study teaches us depends largely on what we are willing to take as invariant. Initial analyses of the GAIN program implicitly assumed the effect of specific intervention types, such as labor force attachment or human capital development, to be invariant across sites, and used this to argue in favor of the labor force attachment approach. In contrast, once we control for covariates X_i, the data also appears plausibly consistent with an assumption that the conditional average treatment effect function $\tau(x)$ is invariant across sites, regardless of whether the sites framed their interventions in terms of labor force attachment or human capital development. In other words, the first approach assumes invariance across sites to identify differential effects from different variants of the intervention, whereas the latter assumes invariance across types of intervention and instead highlights the effect of covariates $\tau(x)$. Methodological innovations that enable us to synthesize between these two modeling approaches would be of considerable use here.[6]

ACKNOWLEDGMENTS

I am grateful for helpful conversations with Guillaume Basse, Guido Imbens and Dominik Rothenhäusler. This work was supported by National Science Foundation Grant DMS-1916163.

REFERENCES

ATHEY, S., TIBSHIRANI, J. and WAGER, S. (2019). Generalized random forests. *Ann. Statist.* **47** 1148–1178. MR3909963 https://doi.org/10.1214/18-AOS1709

ATHEY, S. and WAGER, S. (2017). Efficient policy learning. Preprint. Available at arXiv:1702.02896.

BAREINBOIM, E. and PEARL, J. (2016). Causal inference and the data-fusion problem. *Proc. Natl. Acad. Sci. USA* **113** 7345–7352.

BASSE, G. W., FELLER, A. and TOULIS, P. (2019). Randomization tests of causal effects under interference. *Biometrika* **106** 487–494. MR3949317 https://doi.org/10.1093/biomet/asy072

CRUMP, R. K., HOTZ, V. J., IMBENS, G. W. and MITNIK, O. A. (2008). Nonparametric tests for treatment effect heterogeneity. *Rev. Econ. Stat.* **90** 389–405.

DING, P., FELLER, A. and MIRATRIX, L. (2019). Decomposing treatment effect variation. *J. Amer. Statist. Assoc.* **114** 304–317. MR3941256 https://doi.org/10.1080/01621459.2017.1407322

[4]For example, the average age of GAIN participants in Riverside was 33.6 years, as opposed to 35.4 years in the other counties. 52% of GAIN participants in Riverside were white, as opposed to 30% in other counties.

[5]Reporting the average out-of-bag predictions is a simple but rather crude way of showing that the X_i explain at least part of the difference in treatment effects between Los Angeles and Riverside. The problem of how best to transport potentially heterogeneous effects across sites has received a fair amount of attention, with recent contributions from Bareinboim and Pearl (2016), Hernán and VanderWeele (2011) and Hirshberg, Maleki and Zubizarreta (2019).

[6]It is also possible to have collider-type phenomena, such that $\tau(x)$ is not be invariant across settings, but that there is some coarsening $Z = f(X)$ such that $\tau_Z(z) = \mathbb{E}[Y_i(1) - Y_i(0) \mid f(X_i) = z]$ is invariant. One could attempt to learn the best function for predicting treatment effects in a new environment by using leave-environment-out cross-validation with a loss function that targets the conditional average treatment effect (Nie and Wager, 2017, van der Laan and Dudoit, 2003).

HAHN, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66** 315–331. MR1612242 https://doi.org/10.2307/2998560

HAHN, J., TODD, P. and VAN DER KLAAUW, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* **69** 201–209.

HERNÁN, M. A. and VANDERWEELE, T. J. (2011). Compound treatments and transportability of causal inference. *Epidemiology* **22** 368.

HIRSHBERG, D. A., MALEKI, A. and ZUBIZARRETA, J. (2019). Minimax linear estimation of the retargeted mean. Preprint. Available at arXiv:1901.10296.

HOLLAND, P. W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 945–970. MR0867618

HOTZ, V. J., IMBENS, G. W. and KLERMAN, J. A. (2006). Evaluating the differential effects of alternative welfare-to-work training components: A reanalysis of the California GAIN program. *J. Labor Econ.* **24** 521–566.

HUDGENS, M. G. and HALLORAN, M. E. (2008). Toward causal inference with interference. *J. Amer. Statist. Assoc.* **103** 832–842. MR2435472 https://doi.org/10.1198/016214508000000292

IMBENS, G. W. and ANGRIST, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62** 467–475.

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—For Statistics, Social, and Biomedical Sciences: An introduction*. Cambridge Univ. Press, New York. MR3309951 https://doi.org/10.1017/CBO9781139025751

IMBENS, G. and WAGER, S. (2019). Optimized regression discontinuity designs. *Rev. Econ. Stat.* **101** 264–278.

KITAGAWA, T. and TETENOV, A. (2018). Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica* **86** 591–616. MR3783340 https://doi.org/10.3982/ECTA13288

LI, F., MORGAN, K. L. and ZASLAVSKY, A. M. (2018). Balancing covariates via propensity score weighting. *J. Amer. Statist. Assoc.* **113** 390–400. MR3803473 https://doi.org/10.1080/01621459.2016.1260466

LIN, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Ann. Appl. Stat.* **7** 295–318. MR3086420 https://doi.org/10.1214/12-AOAS583

MEINSHAUSEN, N., HAUSER, A., MOOIJ, J. M., PETERS, J., VERSTEEG, P. and BÜHLMANN, P. (2016). Methods for causal inference from gene perturbation experiments and validation. *Proc. Natl. Acad. Sci. USA* **113** 7361–7368.

NEYMAN, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Rocz. Nauk Rol.* **10** 1–51.

NIE, X. and WAGER, S. (2017). Quasi-oracle estimation of heterogeneous treatment effects. Preprint. Available at arXiv:1712.04912.

PETERS, J., BÜHLMANN, P. and MEINSHAUSEN, N. (2016). Causal inference by using invariant prediction: Identification and confidence intervals. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 947–1012. MR3557186 https://doi.org/10.1111/rssb.12167

ROBINS, J. M. and RICHARDSON, T. S. (2010). Alternative graphical causal models and the identification of direct effects. In *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures* 103–158. Oxford Univ. Press, Oxford.

ROBINS, J. M. and ROTNITZKY, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc.* **90** 122–129. MR1325119

ROBINSON, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica* **56** 931–954. MR0951762 https://doi.org/10.2307/1912705

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974 https://doi.org/10.1093/biomet/70.1.41

ROTHENHÄUSLER, D., MEINSHAUSEN, N., BÜHLMANN, P. and PETERS, J. (2018). Anchor regression: Heterogeneous data meets causality. Preprint. Available at arXiv:1801.06229.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688.

RUBIN, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* **100** 322–331. MR2166071 https://doi.org/10.1198/016214504000001880

VAN DER LAAN, M. J. and DUDOIT, S. (2003). Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. Paper 130, U.C. Berkeley Division of Biostatistics Working Paper Series.