

## Causal Machine Learning Approaches for Identifying Factors Influencing Predictive Outcomes

**Dr. Zofia Malinowska**

Associate Professor, Baltic Centre for Applied Machine Intelligence, Poland

### Abstract

Machine learning has become an important analytical approach for identifying relationships among explanatory variables and predicting outcomes across complex datasets. Conventional predictive models frequently emphasize statistical associations that improve forecasting accuracy without establishing whether the identified relationships represent causal influences. Consequently, variables receiving substantial predictive importance may not necessarily produce changes in the outcome when modified through an intervention. Causal machine learning combines principles of causal inference with flexible computational modeling to investigate treatment effects, confounding structures, and differences in intervention responses across observational units. The present paper develops a hypothetical methodological framework for examining causal machine learning approaches used to identify factors influencing predictive outcomes. The proposed investigation considers propensity-score adjustment, inverse probability weighting, doubly robust estimation, Bayesian outcome modeling, and honest causal tree methods. A fictional residential energy-efficiency dataset containing 6,000 buildings is introduced to illustrate differences between statistical prediction and causal effect estimation. Building age, baseline electricity consumption, floor area, occupancy, and weather conditions are considered potential explanatory characteristics.

The proposed methodology incorporates causal question formulation, treatment definition, directed acyclic graph specification, confounder assessment, propensity modeling, treatment-effect estimation, and sensitivity analysis. Two illustrative analytical tables demonstrate the influence of unequal treatment-group composition on observed electricity consumption comparisons. The constructed example distinguishes unadjusted associations from age-standardized contrasts while emphasizing that neither automatically establishes causality. The methodological discussion examines unmeasured confounding, treatment-selection bias, overlap violations, model misspecification, heterogeneous treatment effects, and uncertainty in causal interpretation. Particular attention is given to the differences among predictive feature importance, conditional association, and intervention-related effects. All numerical values are hypothetical and do not represent completed computational experiments or verified causal findings.

**Keywords:** Causal Machine Learning, Causal Inference, Predictive Outcomes, Treatment Effects, Propensity Score, Confounding Adjustment, Doubly Robust Estimation, Causal Trees, Heterogeneous Treatment Effects, Counterfactual Analysis.

## 1. Introduction

Machine learning has become a significant component of contemporary computational research because it enables automated systems to examine complex datasets and identify relationships associated with measurable outcomes. Predictive algorithms are frequently used to estimate future values, classify observations, and identify variables associated with particular results. These methods can achieve substantial predictive accuracy when trained using representative observations and appropriately evaluated on independent data. However, successful prediction does not necessarily explain why an outcome occurs. A variable may contribute useful predictive information without representing a factor whose modification would produce a corresponding change in the outcome.

The distinction between prediction and causation is particularly important when analytical findings are intended to inform decisions about interventions. A predictive model may identify a strong relationship between building characteristics and electricity consumption, but this relationship does not independently establish whether changing a particular characteristic would reduce energy use. The observed association may reflect differences in building age, occupancy, environmental conditions, or other variables that influence both the explanatory factor and the measured outcome. Consequently, decision-making applications require analytical procedures capable of distinguishing statistical relationships from defensible causal effects.

Causal inference provides a theoretical foundation for examining how outcomes might differ under alternative interventions. The potential-outcomes framework represents these alternatives through the outcomes that would occur under different treatment conditions. However, observational data generally reveal only the outcome associated with the treatment actually received by each unit. The unobserved alternative creates a fundamental identification challenge. Researchers therefore employ assumptions concerning treatment assignment, measured confounding variables, and comparability between treated and untreated observations. These assumptions are central to the interpretation of estimated causal effects and cannot be established through prediction accuracy alone.

Causal machine learning extends conventional causal inference by incorporating flexible statistical algorithms into treatment assignment modeling, outcome prediction, and heterogeneous effect estimation. Methods involving propensity scores, doubly robust estimators, Bayesian regression trees, and causal partitioning provide different procedures for examining relationships between interventions and outcomes. These methods can represent nonlinear associations and interactions among explanatory variables. Nevertheless, algorithmic flexibility cannot automatically remove bias arising from unmeasured confounding or inadequate study design. Reliable causal investigation therefore requires appropriate identification assumptions, transparent variable selection, and careful examination of the treatment assignment mechanism.

The present paper develops a hypothetical methodological framework for investigating causal machine learning approaches through an illustrative residential building energy-efficiency application. A fictional dataset of 6,000 buildings is considered, with energy-efficiency retrofitting defined as the intervention and subsequent electricity consumption defined as the outcome. The proposed analysis compares descriptive associations with covariate-adjusted treatment-effect procedures and considers how treatment responses may differ among building categories. Two hypothetical tables demonstrate the influence of treatment-group composition on observed comparisons. No real building measurements were analyzed, and no causal machine learning models were executed.

## 2. Objectives of the Study

The primary objective of this study is to examine how causal machine learning methods can distinguish factors that predict outcomes from factors that may exert causal influence under a clearly defined intervention. The investigation focuses on treatment assignment, confounding adjustment, heterogeneous responses, and methodological reliability. Three specific objectives guide the proposed research framework.

### 2.1 Identification of Predictive Associations and Potential Causal Factors

The first objective is to examine the distinction between predictive relationships and potential causal influences within observational datasets. Conventional machine learning methods identify characteristics that contribute to accurate outcome predictions, but these relationships may reflect confounding, reverse causation, or indirect associations. The proposed study considers how building age, prior electricity consumption, floor area, and occupancy could influence the observed relationship between energy-efficiency retrofitting and subsequent electricity use. Particular attention is given to identifying variables relevant to causal adjustment rather than selecting covariates solely through predictive importance.

The investigation further aims to establish a structured approach for defining causal questions before selecting analytical methods. The treatment, outcome, target population, and relevant observation period must be specified to support meaningful causal interpretation. A factor may be strongly associated with electricity consumption yet remain unsuitable as a treatment variable because it cannot be modified through the proposed intervention. Consequently, the framework distinguishes intervention-related variables from contextual predictors and emphasizes the importance of specifying scientifically interpretable comparisons.

### 2.2 Comparative Assessment of Causal Estimation Strategies

The second objective focuses on examining alternative strategies for estimating treatment effects using observational information. Propensity-score adjustment, inverse probability weighting, doubly robust estimation, and flexible outcome modeling are selected to represent complementary approaches. The intended comparison considers how each procedure accounts for measured differences between treatment groups. Particular attention is given to the assumptions supporting causal interpretation, the treatment populations represented by different estimators, and the potential sensitivity of results to model specification.

A further objective involves examining how machine learning methods may support the estimation of heterogeneous treatment effects. The influence of an intervention may differ according to building characteristics, baseline electricity use, or environmental conditions. Honest causal partitioning and flexible regression methods provide potential approaches for investigating these differences. However, apparent heterogeneity can also arise from random variation or model overfitting. The proposed study therefore emphasizes independent assessment and uncertainty estimation before interpreting subgroup differences as meaningful causal effects.

### 2.3 Evaluation of Causal Reliability and Identification Assumptions

The third objective is to identify methodological conditions necessary for reliable causal analysis. Observational treatment assignment may depend on characteristics that also influence the outcome, creating differences between treated and untreated units before the intervention. Appropriate causal evaluation therefore requires consideration of exchangeability, treatment overlap, outcome consistency,

and the timing of measured covariates. The proposed framework examines how these assumptions can be supported through study design, diagnostic assessment, and transparent documentation.

The investigation additionally aims to identify limitations that remain after computational adjustment. Propensity weighting and flexible predictive modeling can address selected measured differences but cannot guarantee the removal of unobserved confounding. Statistical estimation may also become unstable when treatment groups have limited covariate overlap. Consequently, this objective emphasizes sensitivity analysis, covariate balance assessment, and careful interpretation. Reliable causal conclusions must reflect both the observed information and the assumptions required to interpret adjusted comparisons.

### 3. Review of Literature

#### 3.1 Foundations of Causal Identification and Observational Study Design

Robins et al. (2000) developed marginal structural modeling approaches for causal inference when treatment assignment and covariate history create complex observational relationships. Rosenbaum (2002) examined the design and analysis of observational studies, emphasizing strategies for addressing differences between treatment groups. Shadish et al. (2002) provided a comprehensive treatment of experimental and quasi-experimental designs. These contributions established methodological foundations for evaluating interventions and distinguishing observed associations from causal effects supported by appropriate research designs.

Imbens (2004) reviewed nonparametric treatment-effect estimation under assumptions concerning observed confounding variables. Imai and van Dyk (2004) examined extensions of propensity-score methods for general treatment regimes. Pearl (2009) developed a systematic treatment of causal models and graphical representations, while Imbens and Rubin (2015) presented a comprehensive framework based on potential outcomes. These contributions demonstrate the importance of specifying causal assumptions and defining the target treatment effect before selecting statistical or machine learning procedures.

#### 3.2 Propensity Modeling and Doubly Robust Estimation

Lunceford and Davidian (2004) investigated stratification and weighting procedures for estimating treatment effects using propensity scores. Bang and Robins (2005) examined doubly robust estimation and demonstrated how combining treatment assignment and outcome models can provide protection against selected forms of model misspecification. Caliendo and Kopeinig (2008) provided practical guidance concerning propensity-score matching, common support, and matching quality. Together, these studies established important procedures for improving comparability between treated and untreated observational units.

Austin (2009) examined covariate balance diagnostics for propensity-score-matched samples and emphasized the importance of evaluating whether measured group differences were adequately reduced. Stuart (2010) reviewed matching methods for causal inference, while Austin (2011) explained propensity-score approaches for reducing measured confounding in observational studies. These contributions demonstrate that propensity scores are tools for implementing an identification strategy rather than substitutes for that strategy. Their use depends on appropriate variable selection, overlap, and transparent assessment of post-adjustment balance.

### 3.3 Machine Learning, Flexible Outcome Models, and Effect Heterogeneity

Breiman (2001) introduced Random Forest as a flexible ensemble approach for prediction. Strobl et al. (2007) examined bias in Random Forest variable importance measures, demonstrating the need for caution when interpreting predictive importance. Chipman et al. (2010) developed Bayesian Additive Regression Trees, providing a flexible statistical approach for modeling nonlinear outcome relationships. Hill (2011) investigated Bayesian nonparametric modeling for causal inference and demonstrated its relevance to treatment-effect estimation in observational settings.

Athey and Imbens (2016) introduced recursive partitioning procedures for examining heterogeneous causal effects and developed methods designed to support valid inference for subgroup treatment differences. Hernán and Robins (2016) discussed the use of observational data to emulate clearly specified target trials. VanderWeele (2015) examined causal explanation, mediation, and interaction. These contributions highlight the distinction between estimating average intervention effects and understanding how those effects may vary across observational units. They also emphasize the need to establish appropriate research design assumptions before interpreting complex computational models causally.

## 4. Research Methodology

### 4.1 Research Design and Causal Identification Strategy

The present paper adopts a hypothetical observational research design intended to examine the relationship between residential energy-efficiency retrofitting and subsequent electricity consumption. The proposed treatment variable indicates whether a building underwent a specified energy-efficiency retrofit before the follow-up period. The outcome is average monthly electricity consumption during the subsequent six months, measured in kilowatt-hours. The principal causal estimand is the average treatment effect across the eligible building population. The study is illustrative and does not represent a completed observational investigation.

The causal identification strategy would require consistency of treatment definitions, conditional exchangeability given measured pretreatment covariates, and sufficient overlap in treatment probabilities. Relevant characteristics would include building age, baseline electricity consumption, floor area, occupancy, and environmental conditions measured before the intervention. A directed acyclic graph would be developed to specify assumed relationships and identify an appropriate adjustment set. Post-treatment variables would not be controlled indiscriminately because they could mediate the retrofit effect or introduce bias under certain causal structures.

### 4.2 Hypothetical Dataset and Causal Feature Specification

For illustrative purposes, the study assumes a fictional dataset containing 6,000 residential buildings. Of these, 2,400 buildings receive an energy-efficiency retrofit, while 3,600 remain untreated during the defined observation period. The dataset contains 3,000 older buildings and 3,000 newer buildings. Treatment assignment is assumed to be nonrandom, with older buildings more frequently represented in the retrofit group. These quantities are manually constructed and do not describe actual buildings, verified retrofit programs, or measured electricity consumption.

The proposed dataset would include baseline monthly electricity consumption, building age, floor area, occupancy, and relevant environmental variables. The retrofit intervention would be defined using a consistent set of energy-efficiency improvements completed before follow-up. Historical

electricity consumption would be measured before treatment assignment to avoid temporal ambiguity. Data preparation would include verification of treatment dates, measurement units, missing records, and observation eligibility. Machine learning models would estimate treatment probabilities and outcome relationships using pretreatment information, with diagnostic assessment of covariate overlap and balance.

### 4.3 Causal Estimators and Analytical Evaluation

The proposed analytical framework would begin with unadjusted consumption comparisons and covariate-standardized descriptive analysis. Propensity-score models would estimate the probability of receiving the retrofit based on measured baseline characteristics. Inverse probability weighting would then construct a weighted comparison under the assumed identification conditions. Doubly robust estimation would combine treatment assignment models with outcome regression. Bayesian Additive Regression Trees and honest causal partitioning would be considered for flexible outcome modeling and the exploration of heterogeneous responses.

The principal effect measure would be the difference in average monthly electricity consumption between the potential outcomes under retrofit and no retrofit. Negative effects would correspond to reduced electricity consumption under the intervention. Model evaluation would examine propensity overlap, standardized mean differences, weight instability, and sensitivity to estimation choices. Sample splitting would be used where appropriate to reduce overfitting during effect-heterogeneity assessment. Confidence intervals would require suitable uncertainty procedures, and sensitivity analyses would assess the vulnerability of conclusions to unmeasured confounding. No causal estimators were executed for this manuscript.

## 5. Analysis

### 5.1 Illustrative Relationship Between Building Characteristics and Observed Consumption

An unadjusted comparison of treated and untreated buildings may be misleading when the groups differ in characteristics associated with electricity consumption. Table 1 presents hypothetical follow-up consumption values for older and newer buildings. Older buildings constitute a larger proportion of the retrofit group, while newer buildings are more common among untreated units. This constructed distribution demonstrates how unequal group composition can affect the overall comparison, even when within-category consumption differences follow a different pattern.

The illustrative values show lower observed electricity consumption among treated buildings within both age categories. However, the retrofit group's larger proportion of older buildings increases its pooled average consumption relative to what would be expected from an age-balanced comparison. Consequently, the pooled association differs from the within-category contrasts. These numerical relationships illustrate confounding by measured characteristics, but they do not establish a causal retrofit effect. Actual causal interpretation would require additional adjustment and defensible identification assumptions.

**Table 1: Hypothetical Building Distribution and Observed Electricity Consumption**

| Building Category | Retrofit Status | Number of Buildings | Mean Follow-Up Consumption (kWh/month) |
|-------------------|-----------------|---------------------|----------------------------------------|
| Older Buildings   | Retrofitted     | 1,500               | 370                                    |
| Older Buildings   | Not Retrofitted | 1,500               | 430                                    |
| Newer Buildings   | Retrofitted     | 900                 | 250                                    |
| Newer Buildings   | Not Retrofitted | 2,100               | 300                                    |

The constructed data demonstrate how pretreatment characteristics can influence comparisons between observational groups. The reported averages are descriptive quantities and should not be treated as measurements of intervention effects.

## 5.2 Illustrative Comparison of Crude and Covariate-Standardized Contrasts

Causal machine learning investigations must distinguish unadjusted outcome differences from estimates that account for measured pretreatment characteristics. Table 2 presents contrasts calculated directly from the fictional observations in Table 1. The crude comparison uses the pooled means of the treated and untreated groups. Subgroup comparisons examine older and newer buildings separately. An age-standardized comparison gives equal weight to the two age categories, reflecting their equal representation in the overall hypothetical population.

The constructed results demonstrate that the crude difference is smaller in magnitude than the age-standardized contrast. This occurs because the retrofitted group contains a higher proportion of older buildings, which have higher observed electricity consumption in the fictional example. However, standardization by building age alone does not establish that all relevant confounding has been addressed. The proposed propensity-weighted and doubly robust models would need to consider additional baseline characteristics before the adjusted contrast could receive a conditional causal interpretation.

**Table 2: Hypothetical Consumption Contrasts Under Different Comparison Strategies**

| Comparison Strategy     | Mean Retrofitted Consumption (kWh/month) | Mean Untreated Consumption (kWh/month) | Observed Difference (kWh/month) |
|-------------------------|------------------------------------------|----------------------------------------|---------------------------------|
| Crude Pooled Comparison | 325.00                                   | 354.17                                 | −29.17                          |
| Older Buildings Only    | 370.00                                   | 430.00                                 | −60.00                          |

|                                     |        |        |        |
|-------------------------------------|--------|--------|--------|
| Newer Buildings Only                | 250.00 | 300.00 | −50.00 |
| Age-Standardized Comparison (50:50) | 310.00 | 365.00 | −55.00 |

The hypothetical crude contrast is approximately  $-29.17$  kWh per month, compared with an age-standardized contrast of  $-55.00$  kWh per month. The difference illustrates how adjustment for one measured characteristic can alter an observational comparison. It does not prove that retrofitting caused the estimated reduction or that adjustment for building age is sufficient.

## 6. Discussion

### 6.1 Predictive Association and Causal Interpretation

Predictive machine learning models identify statistical relationships that support accurate outcome estimation. These relationships may involve genuine causal mechanisms, shared causes, indirect pathways, or variables serving as proxies for unmeasured conditions. Consequently, strong predictive performance does not establish that changing an important model feature would produce a corresponding outcome change. The proposed framework illustrates this distinction through residential electricity consumption, where building characteristics may predict energy use without necessarily representing independently modifiable intervention effects.

The hypothetical analysis demonstrates how the observed relationship between retrofitting and electricity consumption changes after accounting for building age. This adjustment is informative because it reveals that treatment-group composition influences the pooled comparison. Nevertheless, the age-standardized contrast remains an observational association unless additional identification assumptions are justified. Unmeasured differences in insulation quality, household behavior, or previous energy investments may continue to affect the estimated relationship. Therefore, causal claims require stronger evidence than the statistical comparison alone.

### 6.2 Propensity Scores and Doubly Robust Learning

Propensity-score procedures aim to improve comparability between treated and untreated observations based on measured pretreatment characteristics. Weighting and matching can reduce selected differences in observed covariate distributions, supporting treatment-effect estimation when the necessary identification assumptions are plausible. However, these methods require careful model specification and sufficient overlap. Extreme propensity scores may generate unstable weights, while incomplete adjustment sets can leave substantial confounding. Consequently, propensity modeling should be evaluated through balance diagnostics rather than predictive discrimination alone.

Doubly robust estimation combines treatment probability modeling with outcome regression and offers additional protection against selected forms of model misspecification. Bang and Robins (2005) demonstrated the theoretical importance of such estimators under appropriate assumptions. However, double robustness does not eliminate unmeasured confounding or guarantee correct estimation when both component models are inappropriate. Flexible machine learning may assist in estimating these components, but its use requires suitable regularization and validation. The present manuscript proposes these procedures without reporting their empirical implementation.

## 6.3 Heterogeneous Effects and Factor Identification

Intervention effects may differ across observational units according to their baseline characteristics. For example, older buildings may respond differently to an energy-efficiency retrofit than newer buildings because their construction features and existing energy performance differ. Causal partitioning methods can investigate such patterns by identifying subgroups with potentially different treatment effects. However, observed subgroup contrasts are not automatically causal. They may also reflect unequal treatment selection or differences in unmeasured characteristics.

Athey and Imbens (2016) introduced procedures designed to support the investigation of heterogeneous causal effects through recursive partitioning. Such approaches provide a methodological basis for examining variation in intervention responses while addressing aspects of estimation bias. Nevertheless, searching extensively for favorable subgroups can introduce overfitting and unstable conclusions. The proposed framework therefore emphasizes honest sample splitting, uncertainty assessment, and replication. Meaningful causal factor identification requires distinguishing genuine effect modification from patterns created through observational differences or statistical noise.

## 7. Challenges

### 7.1 Unmeasured Confounding and Identification Uncertainty

Unmeasured confounding represents a major challenge in observational causal machine learning because important characteristics may influence both intervention assignment and subsequent outcomes. In the proposed building application, household conservation behavior, undocumented maintenance practices, or prior insulation improvements could affect both retrofit participation and electricity consumption. If these factors are not observed and appropriately addressed, estimated treatment effects may remain biased even after extensive statistical adjustment. More flexible predictive algorithms cannot independently recover causal identification from insufficient observational information.

Identification uncertainty also arises when the proposed intervention or outcome is inadequately defined. Different retrofit combinations may produce different effects, while changes in occupancy during follow-up may complicate the interpretation of electricity consumption. A meaningful causal question therefore requires explicit treatment definitions, consistent measurement periods, and suitable assumptions about potential outcomes. The proposed framework emphasizes documentation of these elements before model development. Sensitivity analysis is necessary for examining how conclusions might change under plausible violations of identification assumptions.

### 7.2 Treatment Overlap and Model Specification Challenges

Treatment overlap requires that observational units with relevant covariate characteristics have a meaningful probability of receiving either treatment condition. When certain building categories almost always receive retrofits or almost never receive them, comparisons involving those characteristics may rely heavily on extrapolation. Propensity-score weighting can become unstable under such conditions because observations with extreme treatment probabilities receive large weights. Consequently, a high-performing treatment assignment model does not necessarily produce a reliable causal estimator.

Model specification introduces additional challenges because both treatment assignment and outcome relationships may contain nonlinearities and interactions. Conventional parametric models can be restrictive when the underlying data structure is complex, while flexible machine learning methods may overfit the available observations. Doubly robust procedures offer selected advantages but remain

dependent on appropriate identification assumptions and estimation quality. The proposed methodology therefore recommends overlap assessment, weight diagnostics, model sensitivity analysis, and independent evaluation rather than reliance on a single computational specification.

### 7.3 Causal Effect Heterogeneity and Analytical Transparency

Heterogeneous treatment-effect estimation can become unreliable when observational subgroups contain few comparable treated and untreated units. Partitioning a dataset into increasingly detailed categories may reduce the amount of information available for each comparison. Apparent differences among estimated effects may then reflect random sampling variation rather than meaningful causal relationships. Consequently, causal machine learning procedures require appropriate constraints, honest evaluation, and uncertainty assessment when identifying potentially influential factors across a population.

Analytical transparency presents another difficulty because complex models may generate treatment-effect estimates without providing easily interpretable explanations of the assumptions supporting them. A ranked list of predictive features cannot substitute for a causal model specifying confounders, mediators, and intervention relationships. Similarly, a flexible estimator may produce precise numerical values that remain misleading when its identification assumptions are violated. The proposed framework therefore emphasizes clear causal diagrams, documented adjustment sets, transparent diagnostics, and cautious presentation of results. Such safeguards are necessary for responsible interpretation.

## 8. Conclusion

Causal machine learning provides a methodological approach for distinguishing factors that predict outcomes from factors that may influence outcomes through defined interventions. The present paper developed a hypothetical framework incorporating potential-outcomes reasoning, propensity-score adjustment, inverse probability weighting, doubly robust estimation, Bayesian outcome modeling, and honest causal partitioning. Its principal focus was the difference between observational association and intervention-related effect estimation. The proposed framework demonstrates the importance of establishing a defensible causal question before applying flexible computational models.

The illustrative tables demonstrate how differences in treatment-group composition can influence observed outcome comparisons. The constructed building dataset produced a crude electricity consumption difference of approximately  $-29.17$  kWh per month and an age-standardized difference of  $-55.00$  kWh per month. These values were calculated from hypothetical observations and do not represent measured intervention effects. They illustrate how confounding can alter descriptive comparisons, but causal interpretation would require additional assumptions, appropriate adjustment, and verified observational evidence.

Overall, the study identifies unmeasured confounding, treatment overlap, model specification, effect heterogeneity, and analytical transparency as important considerations in causal machine learning. Future empirical investigations should examine these factors using authentic datasets, clearly defined interventions, validated estimation procedures, and sensitivity analysis. Causal machine learning may support intervention evaluation when its statistical flexibility is combined with defensible research design. Nevertheless, reliable conclusions must remain grounded in explicit identification assumptions and independently verified evidence rather than predictive accuracy or algorithmic complexity alone.

## References

1. Athey, S., & Imbens, G. W. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27), 7353–7360. <https://doi.org/10.1073/pnas.1510489113>
2. Bareinboim, E., & Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27), 7345–7352. <https://doi.org/10.1073/pnas.1510507113>
3. Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3), 199–231. <https://doi.org/10.1214/ss/1009213726>
4. Dawid, A. P. (2021). Decision-theoretic foundations for statistical causality. *Journal of Causal Inference*, 9(1), 39–77. <https://doi.org/10.1515/jci-2020-0008>
5. Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if*. Chapman & Hall/CRC.
6. Huntington-Klein, N. (2021). *The effect: An introduction to research design and causality*. Chapman & Hall/CRC. <https://doi.org/10.1201/9781003226055>
7. Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139025751>
8. Janzing, D., Peters, J., Mooij, J. M., & Schölkopf, B. (2012). Identifying confounders using additive noise models. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence* (pp. 349–356). AUAI Press.
9. Johansson, F. D., Shalit, U., & Sontag, D. (2016). Learning representations for counterfactual inference. In *Proceedings of the 33rd International Conference on Machine Learning* (pp. 3020–3029). PMLR.
10. Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10), 4156–4165. <https://doi.org/10.1073/pnas.1804597116>
11. Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., & Welling, M. (2017). Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems* 30 (pp. 6446–6456). Curran Associates.
12. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* 30 (pp. 4765–4774). Curran Associates.
13. Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9780511803161>
14. Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, 3, 96–146. <https://doi.org/10.1214/09-SS057>
15. Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. MIT Press.
16. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?” Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939778>
17. Robins, J. M., Hernán, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5), 550–560.

18. Schölkopf, B. (2022). Causality for machine learning. In H. Geffner, R. Dechter, & J. Y. Halpern (Eds.), Probabilistic and causal inference: The works of Judea Pearl (pp. 765–804). Association for Computing Machinery. <https://doi.org/10.1145/3501714.3501755>
19. Shalit, U., Johansson, F. D., & Sontag, D. (2017). Estimating individual treatment effect: Generalization bounds and algorithms. In Proceedings of the 34th International Conference on Machine Learning (pp. 3076–3085). PMLR.
20. Spirtes, P., Glymour, C. N., & Scheines, R. (2000). Causation, prediction, and search (2nd ed.). MIT Press.