European Journal of Epidemiology (2020) 35:615–617
https://doi.org/10.1007/s10654-020-00659-8

COMMENTARY

Counterfactual prediction is not only for causal inference

Barbra A. Dickerman[1] · Miguel A. Hernán[1,2,3]

Published online: 4 July 2020

Clinical researchers generate and analyze health data for three classes of tasks: description, prediction, and counterfactual prediction [1]. Description uses data to provide a quantitative summary of certain features of the world. Prediction uses data to map some features of the world (the inputs) to other features of the world (the outputs). Counterfactual prediction uses data to predict certain features of the world if the world had been different.

Causal inference is a common goal of counterfactual prediction. Indeed, causal inference can be viewed as the prediction of the distribution of an outcome under two (or more) hypothetical interventions followed by a comparison of those outcome distributions. This contrast of outcome distributions may be the basis of subsequent decision-making: we often choose to implement the hypothetical intervention that, according to our analyses, leads to the most favorable outcome.

But counterfactual prediction can also be used to predict the outcome distribution under a single hypothetical intervention. Then the primary goal is not to make decisions between different interventions, but simply to map the inputs to the outputs in hypothetical (counter to the fact) scenarios that differ from the observed world. Like Schulam and Saria [2] and other authors [3] before, van Geloven and colleagues [4] in this issue of *European Journal of Epidemiology* highlight the difference between (factual) prediction and counterfactual prediction in clinical research.

As an example, suppose we are interested in predicting the 5-year risk of death among individuals recently diagnosed with heart failure. To do so, we develop a prediction algorithm that maps the inputs (baseline variables measured at the time of diagnosis) into the output (death). We know that individuals with severe disease at the time of diagnosis are more likely to receive a heart transplant (a post-baseline treatment), and a heart transplant reduces the risk of death. If we develop our prediction algorithm in a setting in which most individuals have access to a heart transplant, our algorithm will determine that having severe disease is associated with a lower risk of death. However, when we deploy our prediction algorithm in a new setting in which few individuals have access to a heart transplant, our algorithm will fail miserably because, in that setting, having severe disease is associated with a higher risk of death.

For an algorithm developed in one setting to yield predictions transportable to other settings with different treatment patterns, we would need to make accurate predictions about what would happen if, somehow, we could intervene on the availability of heart transplants. That is, our algorithm would need to be designed for counterfactual prediction. van Geloven et al. discuss several estimands for prediction and counterfactual prediction, a distinction that they make sharp in their paper even if their chosen umbrella term ("predictimand") tends to obscure it.

All estimands for counterfactual prediction are based on hypothetical interventions or treatment policies, and can therefore be defined in terms of a (hypothetical) target trial with a single arm [5]. For example: Given the baseline predictors, what would be the 5-year risk of death if all individuals had access to a heart transplant? If no individuals had access to a heart transplant? If a random 50% of individuals had access to a heart transplant? Any estimand whose definition includes a "what if" question is a counterfactual estimand for counterfactual prediction. Analogously, any estimand whose definition does not include a "what if" question (say, the 5-year risk of death among individuals receiving the same treatment pattern as those in this population) is a factual estimand for factual prediction. In other words, the estimand for a prediction task is defined in terms of observed variables, whereas the definition of the estimand for a counterfactual prediction

✉ Barbra A. Dickerman
bad788@mail.harvard.edu

[1] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA

[2] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

[3] Harvard-MIT Division of Health Sciences and Technology, Boston, MA, USA

task necessarily involves counterfactual (or potential) outcomes. This simple classification is not always clear in the addendum to the ICH E9 guideline Statistical Principles for Clinical Trials, which van Geloven et al. use to organize their estimands [6, 7].

Counterfactual prediction may be needed when effective treatments exist and differ across settings, particularly when treatment is allocated to individuals based on a risk factor for the outcome, as is often the case in clinical practice. This variation in treatment patterns may substantially weaken and possibly reverse associations between these factors and the outcome. For example, Caruana et al. found that policies that admitted individuals who presented to the hospital with pneumonia and had a history of asthma directly to the intensive care unit resulted in an unexpected inverse association between asthma and death [8]. This may be problematic not only for time-fixed treatments, when treatment is allocated at baseline based on baseline prognostic factors, but also for time-varying treatments, when treatment is allocated over follow-up based on prognostic factors that evolve over time.

There is also a role for counterfactual prediction when we anticipate that treatment policies will change over time. When a useful prediction model is deployed in clinical practice, it will help clinicians identify high-risk individuals, which often prompts action to reduce that risk, thereby disrupting predictor-outcome associations and degrading model performance (prediction models become "victims of their own success") [9, 10]. A model for counterfactual risks provides a natural solution, as its performance may be robust over time even when model deployment influences behaviors that affect risk.

By contrast, factual prediction is only relevant when treatment patterns are effectively constant across settings, either because treatments are similarly distributed or because they are only weakly effective or only a small proportion of individuals is treated, as the resulting impact on model performance might not be clinically meaningful [11]. The question is then, why not always do counterfactual prediction? Because it is riskier than factual prediction.

Valid counterfactual prediction requires the same data, methods, and assumptions as those required for causal inference. Estimating counterfactual risks requires high-quality data on the eligibility criteria, baseline predictors (the inputs), outcome of interest (the output), treatment, and confounders throughout the entire follow-up period. Inverse-probability weighting or the g-formula can be applied to the data to generate predictions had everyone received a certain treatment policy, but the validity of the estimates will depend on unverifiable assumptions such as sequential exchangeability of the treated and the untreated given the measured confounders. By contrast, estimating factual risks only requires high-quality data on the eligibility criteria, baseline predictors, and outcome(s) over follow-up, and no unverifiable assumptions.

A consequence of the above is another important difference between prediction and counterfactual prediction: When doing prediction, the data can be used to evaluate the accuracy of the predictions (for example, by splitting the dataset into training and validation samples). When doing counterfactual prediction, one cannot use the data to evaluate the accuracy of the predictions, precisely because they depend on unverifiable assumptions. Counterfactual prediction allows us to deploy predictive algorithms across settings with different treatment patterns but only if we are willing to make assumptions whose validity cannot be verified.

In summary, we agree with the authors' emphasis on unambiguously specifying the estimand of interest for counterfactual prediction. To do so, we will typically have to describe the hypothetical intervention in terms of the one-arm target trial of interest. Then we will need to use causal inference methods to obtain an estimate whose validity relies on unverifiable conditions. The alternative is pursuing regular (factual) prediction and relying on the potentially unrealistic assumptions required for transportability from one setting to another.

Funding This work was supported by National Institutes of Health Grants K99 CA248335 (B.A.D.) and R37 AI102634.

References

1. Hernán MA, Hsu J, Healy B. A second chance to get causal inference right: a classification of data science tasks. Chance. 2019;32(1):42–9.
2. Schulam P, Saria S. Reliable decision support using counterfactual models. In: Advances in neural information processing systems; 2017.
3. Sperrin M, Martin GP, Pate A, Van Staa T, Peek N, Buchan I. Using marginal structural models to adjust for treatment drop-in when developing clinical prediction models. Stat Med. 2018;37(28):4142–54.
4. van Geloven N, Swanson SA, Ramspek CL, et al. Prediction meets causal inference: the role of treatment in clinical prediction models. Eur J Epidemiol. 2020. https://doi.org/10.1007/s10654-020-00636-1.
5. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. Am J Epidemiol. 2016;183(8):758–64.
6. ICH E9 working group. ICH E9 (R1): addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. EMA/CHMP/ICH/436221/2017. 2020. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-sensitivity-analysis-clinical-trials-guideline-statistical-principles_en.pdf. Accessed 29 June 2020.

7. Hernán MA, Scharfstein D. Cautions as regulators move to end exclusive reliance on intention to treat. Ann Intern Med. 2018;168(7):515–6.
8. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21st ACM SIGKDD international conference on knowledge discovery and data mining; 2015.
9. Lenert MC, Matheny ME, Walsh CG. Prognostic models will be victims of their own success, unless. J Am Med Inform Assoc. 2019;26(12):1645–50.
10. Sperrin M, Jenkins D, Martin GP, Peek N. Explicit causal reasoning is needed to prevent prognostic models being victims of their own success. J Am Med Inform Assoc. 2019;26(12):1675–6.
11. Pajouheshnia R, Peelen LM, Moons KGM, Reitsma JB, Groenwold RHH. Accounting for treatment use when validating a prognostic model: a simulation study. BMC Med Res Methodol. 2017;17(1):103.