TU Delft

Delft University of Technology

Document Version
Final published version

Citation (APA)
van Geloven, N., Keogh, R. H., van Amsterdam, W., Cinà, G., Krijthe, J. H., Peek, N., Luijken, K., Magliacane, S., Morzywołek, P., & More Authors (2025). The Risks of Risk Assessment: Causal Blind Spots When Using Prediction Models for Treatment Decisions. *Annals of internal medicine*, *178*(9), 1326-1333. https://doi.org/10.7326/ANNALS-24-00279

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

Green Open Access added to TU Delft Institutional Repository
as part of the Taverne amendment.

More information about this copyright law amendment
can be found at https://www.openaccess.nl.

Otherwise as indicated in the copyright section:
the publisher is the copyright holder of this work and the
author uses the Dutch legislation to make this work public.

Annals of Internal Medicine RESEARCH AND REPORTING METHODS

The Risks of Risk Assessment: Causal Blind Spots When Using Prediction Models for Treatment Decisions

Nan van Geloven, PhD; Ruth H. Keogh, PhD; Wouter van Amsterdam, MD, PhD; Giovanni Cinà, PhD; Jesse H. Krijthe, PhD; Niels Peek, PhD; Kim Luijken, PhD; Sara Magliacane, PhD; Paweł Morzywołek, PhD; Thijs van Ommen, PhD; Hein Putter, PhD; Matthew Sperrin, PhD; Junfeng Wang, PhD; Daniala L. Weir, PhD; and Vanessa Didelez, PhD

Clinicians increasingly rely on prediction models to guide treatment choices. Most prediction models, however, are developed using observational data that include some patients who have already received the treatment the prediction model is meant to inform. Special attention to the causal role of those earlier treatments is required when interpreting the resulting predictions.

"Causal blind spots" were identified in 3 common approaches to handling treatment when developing a prediction model: including treatment as a predictor, restricting to persons taking a certain treatment, and ignoring treatment. Through several real examples, this article illustrates how the risks obtained from models developed using such approaches may be misinterpreted and can lead to misinformed decision making. The discussion covers issues attributable to confounding, selection, mediation, and changes in treatment protocols over time.

An extension of guidelines for the development, reporting, and evaluation of prediction models is advocated to avoid such misinterpretations. Developers must ensure that the intended target population for the model, and the treatment conditions under which predictions hold, are clearly communicated. When prediction models are intended to inform treatment decisions, they need to provide estimates of risk under the specific treatment (or intervention) options being considered, known as "prediction under interventions." Next to suitable data, this requires causal reasoning and causal inference techniques during model development and evaluation. Being clear about what a given prediction model can and cannot be used for prevents misinformed treatment decisions and thereby prevents potential harm to patients.

Ann Intern Med. doi:10.7326/ANNALS-24-00279
For author, article, and disclosure information, see end of text.
This article was published at Annals.org on 29 July 2025.

Clinical prediction (or prognostic) models estimate the risk for future outcomes based on observable patient characteristics. They do so by learning from a historical, often observational, data set (the development data set) in which outcomes are already known. Partly fueled by the recent surge in artificial intelligence, clinicians increasingly use these models, expecting that knowing a person's risk will help them make better decisions on medical treatments or lifestyle advice that may lower that risk. For example, clinicians may choose to treat high-risk patients more aggressively, while managing low-risk patients more conservatively. In this article, we illustrate how using clinical prediction models derived from observational data in this way may inadvertently lead to inappropriate or even harmful treatment decisions.

We focus on the use of predictions to inform a decision on a particular treatment (the "target treatment"), but the problems we point out apply equally to, say, lifestyle advice. The key issue is that most prediction models are derived from observational data in which some persons have already received the target treatment (1). Predictions will be influenced by the mix of patients who did or did not receive these target treatments in the development data as well as the reasons (indications) for those treatment choices. This makes predictions challenging to interpret: for instance, a new patient may have a low estimated risk because similar patients in the development data received aggressive treatments that were successful, hence conservative management is likely inappropriate.

A patient and physician deciding on a target treatment would ideally know the patient's risk "if they were to take the treatment" and their risk "if they were not to take it." We will refer to such risks that carry a "what-if" interpretation as potential risks. A difference between the potential risks with and without treatment indicates a causal treatment effect. Although treatment decisions may be based on the size of the causal treatment effect, the 2 separate potential risks under treatment and under no treatment provide more complete information. In certain cases, knowing the patient's potential risk under a single treatment option could suffice for the treatment decision. For instance, if a patient's risk is low if they were to remain untreated, treatment may be deemed unnecessary regardless of the size of the causal treatment effect. Alternatively, if a patient's prognosis is poor even if they were to be treated, this may prompt the decision to refrain from treatment.

As we will illustrate in this article, many prediction models derived from observational data using standard

See also:

Web-Only
Supplement

Downloaded from https://annals.org by 145.94.227.50 on 07/30/2025.

methods do not provide valid estimates of causal treatment effects, nor do they estimate the potential risk under a given treatment option properly. The method for developing models that can do this properly has been called "counterfactual prediction" or "prediction under interventions" (2–4). This links causal inference approaches to prediction methods, which traditionally have been treated separately.

In this article, we discuss the risks for falsely assigning a what-if interpretation to prediction outputs. We attribute these misinterpretations to "causal blind spots" and explain how they may arise under 3 common ways of handling treatments when developing (or training) a prediction model: 1) including the target treatment as a predictor, 2) restricting the development data to those who did (or those who did not) receive the target treatment, and 3) ignoring the target treatment. We then outline recommendations for deriving and evaluating algorithms that do allow for a what-if interpretation.

Causal Blind Spots When Including Treatment as a Predictor

Suppose that a prediction model was developed on observational data and included the baseline target treatment status as one of the predictors. Patients and physicians who want to know the patient's potential risk if they were to take and if they were not to take the target treatment might be inclined to specify "yes," respectively, "no" for the treatment variable and interpret the resulting predictions as such. Below, we describe 3 situations in which this interpretation is not valid, illustrated by real examples. The assumed causal structures linking the relevant variables underlying these examples are depicted using causal diagrams (directed acyclic graphs) in the **Figure**.

Confounding: Blood Pressure–Lowering Medication and Cardiovascular Risk

It is well known that associations found in prediction models developed using observational data cannot be interpreted causally due to confounding, that is, predictors and outcome may have common causes, not included in the model, that induce their apparent relationship (5). What is less appreciated is that, for the same reason, risks obtained from these models cannot be interpreted as representing potential risks under specified treatment options. Consider the situation where a person wants to know their potential cardiovascular risk if they were or were not to use blood pressure–lowering medication. The PREDICT model, which estimates 5-year risk for cardiovascular disease and was derived from primary care data from more than 400 000 persons in New Zealand, includes as a predictor an indicator of use of blood pressure–lowering medication at baseline (prevalent and incident use). Use of this medication is associated with a *higher* risk for developing cardiovascular disease in the model (6). Applying the prediction model for a 75-year-old woman who is on blood pressure–lowering medication, smokes, has systolic blood pressure of 120 mm Hg and a total cholesterol–to–high-density-lipoprotein ratio of 3.5, results in an estimated 5-year risk for cardiovascular disease of 15%. Using the same inputs, except switching to no use of blood pressure–lowering medication, gives an estimated risk of only 11%. This should certainly not be interpreted as evidence in support of discontinuing the blood pressure–lowering medication for this woman. Instead, the higher risk could be explained by preexisting health conditions in people receiving antihypertensive medication, which are also related to the outcome, but are not included as predictors. For example, a prior diagnosis of hypertension, which is not included in the model, may have prompted the initiation of medication while simultaneously increasing cardiovascular risk. This is sometimes called "confounding by indication" (**Figure**, *A*).

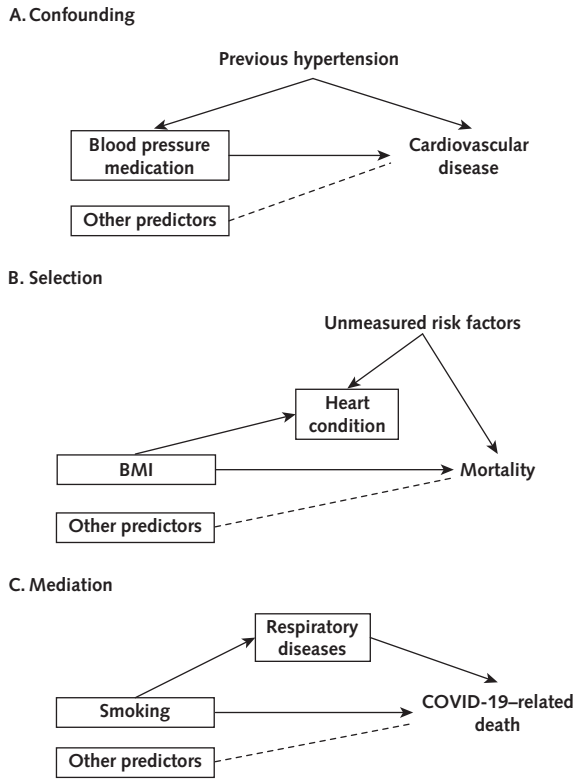
Selection: BMI and Mortality Risk

Interpretations of predictions are also impacted by how patients are selected for inclusion in observational development data, in particular when they are selected based on a particular feature that is causally related to both the target treatment and the outcome (7–9). Consider, for example, the question of whether a patient with a heart condition should or should not gain weight. For such lifestyle advice, physicians would want to know a patient's potential risk if they were to gain weight or were not to do so. Several prediction models developed in data sets selecting exclusively patients with a heart condition have included baseline body mass index (BMI) as a predictor (which as a modifiable lifestyle factor is the target treatment in this example) and reported, somewhat unexpectedly, that patients with obesity (high BMI) had lower mortality risk compared with patients without obesity (for example, Peterson and colleagues [10] and Gruberg and colleagues [11]). This should certainly not lead to advising patients with heart conditions to gain weight to improve their survival. Instead, the underlying reason for the apparently paradoxical protective survival effect of obesity could be that obesity is a major cause of heart problems and patients in the development data *without* obesity were more likely to have other (unmeasured) factors (for example, genes) that led to their heart condition. If these characteristics are also risk factors for mortality, then, in persons with a heart condition, obesity may act as a marker for the *absence* of much more serious risk factors (**Figure**, *B*) (12). Lifestyle interventions to increase BMI will not eliminate these other risk factors. Similar spurious protective effects of obesity have been reported in other settings where patients were selected based on underlying health conditions such as diabetes, and this has been called the "obesity paradox" (12, 13).

In **Supplement Section A** (available at Annals.org), we give 2 additional examples of prediction models where the patient selection likely introduced bias.

Downloaded from https://annals.org by 145.94.227.50 on 07/30/2025.

Figure. Causal structures underlying the examples, depicted in causal-directed acyclic graphs.



A box around a variable indicates that this variable is included in the prediction model or that the data selection was based on the variable. See Supplement Section B for background information on these graphs. BMI = body mass index. A. Confounding: The confounding effect of previous hypertension (not included in the prediction model) induces a noncausal association between "blood pressure medication" and "cardiovascular disease." B. Selection: By selecting only patients with a heart condition, a noncausal association is induced between "BMI" and "unmeasured risk factors" and therefore with "mortality." C. Mediation: By including a consequence of "smoking" in the model (the mediator "respiratory diseases"), part of the effect of "smoking" on "COVID-19–related death" could be masked.

Mediation: Smoking and Risk for COVID-19 Death

A third situation where a prediction model does not provide risks that can inform on alternative treatment options is when the model includes a consequence of the target treatment as predictor. Williamson and colleagues evaluated risk factors associated with COVID-19–related death using a multivariable model fitted using primary care records of more than 17 million adults in the United Kingdom (14). The multivariable model suggested a lower hazard of COVID-19–related death for smokers compared with nonsmokers. Although the authors warned against causal interpretation of their model and did not advise its use for decision making, policymakers initially interpreted this result as evidence that smoking was protective for COVID-19–related death, which nearly resulted in smokers being excluded from shielding policies in France (15, 16).

One likely mechanism contributing to the counterintuitive finding about smoking is that the model included factors that are causal consequences of smoking, such as respiratory diseases. If these factors are also causes of the outcome, they may have acted as mediators, that is, intermediary variables transmitting the influence of smoking to mortality (**Figure**, *C*). Including mediators in the model may mask (or "adjust away") part of the true causal effect of smoking. It might be assumed that mediators cannot be included in a prediction model, given that predictors are typically measured at the same baseline moment, and there is no time for the treatment variable to impact the mediator. However, the "treatment variable" (here, smoking) often represents a prevalent exposure, which may have influenced other factors included as predictors (here, development of respiratory diseases). In **Supplement Section B** (available at Annals.org), we point out a second mechanism that may be at play in this example.

Causal Blind Spots When Restricting the Development Data Based on the Target Treatment

Another approach that is often taken in an attempt to obtain potential risks *if a patient were* (*or were not*) *to take treatment* is to restrict the observational development data to patients who did (or did not) use the target treatment (17). However, except in historical cohorts where treatment was not yet available, persons who are (un)treated in observational data are generally not representative of those for whom that decision is still to be made. In fact, restricting the development data based on treatment status brings the same issues as the earlier examples in which treatment was included as a predictor in the model as well as potentially introducing issues due to selection similar to those described in the above BMI example (in Selection: BMI and Mortality Risk) where selection was not on the basis of treatment.

For instance, in the above PREDICT model, suppose that—in an attempt to estimate patients' potential risks if they were not to use blood pressure–lowering medication—investigators had restricted the development data to persons not using blood pressure–lowering medication at baseline. Applying this model to the example patient considered earlier (75-year-old woman; smokes; systolic blood pressure, 120 mm Hg; total cholesterol–to–high-density-lipoprotein ratio, 3.5) to obtain a 5-year risk for cardiovascular disease under the condition of not using blood pressure–lowering medication would again output a risk of around 11%. This risk is likely based on relatively healthy persons (for example, those without prior diagnosis of hypertension) and, as before, it is not a valid estimate of the potential risk if the example patient were to stop taking the treatment. In addition, immortal time bias can occur if the development data

Downloaded from https://annals.org by 145.94.227.50 on 07/30/2025.

are restricted to untreated persons who do not initiate blood pressure–lowering medication during follow-up because this restriction would be based on postbaseline treatment status (18, 19).

Similarly, prediction models developed on observational data from patients who have all received the target treatment will only be applicable to patients who would receive the treatment under the treatment assignment strategy used in the development data. Such models may not be valid for a wider group of patients for whom the decision of whether to treat is yet to be made because patients who did not receive treatment under the earlier assignment strategy may differ from those who did in ways not captured by the prediction model. It is not uncommon for prediction models to be developed using data restricted to patients receiving the target treatment. A systematic review of prognostic models for outcomes of extracorporeal membrane oxygenation therapy found that all 58 included observational studies restricted to patients receiving the extracorporeal membrane oxygenation therapy. Meanwhile, 11 of these 58 studies explicitly stated that their primary aim was to inform decisions about extracorporeal membrane oxygenation initiation for new patients (20). Another example is the EuroSCORE (European System for Cardiac Operative Risk Evaluation) model, which predicts mortality after cardiac surgery. It was suggested the model could be used to assist in deciding whether to proceed with the surgery based on the estimated risk, but the model was developed on an observational cohort consisting solely of patients who had undergone the surgery (21).

Causal Blind Spots When Ignoring Treatments Applied in the Development Data

Given the complications outlined in the sections above, one may think it is better to ignore treatment status when developing a prediction model—that is, not to include the target treatment as predictor nor to restrict based on treatment status. When the development data contain a mix of patients—some treated and some untreated—ignoring treatment in the prediction model will result in risks that are only valid under continuation of the implicit treatment policy in the observational development data, which we refer to as risks "under current care."

Now the question arises as to whether risks under current care are suitable for supporting treatment decisions in new patients. One may argue that if the risk under current care is high, the patient should receive more aggressive treatment. But more aggressive treatment will be different from current care and, because we do not have a prediction under this novel treatment strategy, we do not know if and how much it would benefit the patient. Moreover, what constitutes "current care" is often poorly defined, making it difficult to determine what it would mean for a particular person and hence what would constitute more aggressive treatment options. Conversely, one might be tempted to conclude that those with low risk under current care could receive less intense treatment. This advice is problematic because a patient may have low risk precisely because they received appropriate treatment under current care. Withholding treatment from such "low-risk" patients may lead to harm (22). The potential danger of this approach was highlighted in the context of a prediction model estimating mortality risk among patients hospitalized with pneumonia (23). The goal was to identify patients with pneumonia with a low mortality risk who could be safely treated as outpatients whereas those at high risk could be admitted to hospital. This implicitly assumes that a patient with pneumonia treated in the hospital with a low mortality risk would have similarly experienced low risk if treated at home. This assumption has later been called into question because the prediction algorithm estimated that patients with asthma had a low mortality risk. However, this was likely because these patients received effective intensive hospital care in the development data, leading to favorable outcomes (24). If new patients with asthma were instead treated as outpatients based on the low risk under current care, their outcomes would likely be worse than predicted and, crucially, worse than under the current care before introducing the prediction model.

Similar issues have been highlighted for more recently developed prediction models such as QRISK3, which estimates the 10-year risk for cardiovascular disease in the general population and is used in primary care in the United Kingdom for supporting decisions on initiating statins (25). Ideally, those decisions would be supported by estimates of potential risks if patients were not to use statins; if this untreated risk is low, no preventive treatment would be needed. During model development, however, initiation of statin therapy (as well as other treatments) during follow-up was ignored. This means that the risks obtained from the model represent expected outcomes under the statin-initiation policy in the development data. A person might have a low risk according to this model because similar patients initiated treatments at a later stage, and, therefore, they would not necessarily have a low risk without treatment (26, 27).

Both examples illustrate the *prediction paradox*: when predictions are used to support treatment decisions, this will change treatment practice and thus potentially invalidate any prediction that was made under historical practice (28–30). For this reason, ignoring treatments during model development in general does not provide risks that can be used to guide future treatment decisions.

The Way Forward

The notion that "prognosis cannot be divorced from contemplated medical action, nor from action to

Downloaded from https://annals.org by 145.94.227.50 on 07/30/2025.

Table. Recommendations for Prediction Under Interventions

Aspect	What Does It Mean?	Recommendations
Defining the prediction estimand		
Target population	Persons to whom the prediction model will be applied	Describe the patient group for whom all treatment options are possible and for whom a treatment decision is necessary.
Time point of intended use	The moment when the prediction model will be applied	Align time point of intended use with the moment of decision making.
Outcome and prediction horizon	The outcome of interest and the time horizon over which predictions are made	Choose an outcome and time horizon that are relevant to the treatment decision.
Predictors	The clinical and/or demographic patient factors to include in the prediction algorithm	Include predictors that are (readily) available at the time of treatment decision making. Ensuring that predictors are measured pretreatment avoids including mediators as predictors (for example, "smoking and risk of COVID-19 death"). Choose predictors separately from variables needed to adjust for confounding; these do not need to be the same.
Treatment option(s)	The target treatment options under consideration for decision making, in particular the option(s) under which estimates of potential risks are desired	Distinguish between baseline treatments and treatment options that represent longer exposure periods postbaseline.
Assessing the design of the data source		
Sample selection	How persons were included in the data set	Selection bias may arise if inclusion in the data is (either directly or indirectly) related to the treatment and the outcome (for example, "BMI and mortality risk").
Treatment assignment	The policy by which the target treatment was assigned in the data	Specify whether treatment was randomized or chosen by patients/physicians. How was adherence to the treatment? If the indication for treatment assignment is linked to the outcome, then confounding is likely (for example, "blood pressure–lowering medication and cardiovascular risk").
Development and evaluation of the prediction algorithm		
Assumptions	The assumptions required for valid prediction under interventions	Explicitly state and justify the assumptions that allow a "what-if" interpretation of predictions, for example, no unmeasured confounding.
Handling of target treatment	How the treatment was handled in the analysis	Decide whether to use the treatment as a variable in the model, restrict on it (for baseline treatments only), censor or otherwise. Distinguish prevalent from incident use of treatment.
Handling of confounding	How confounding was addressed	Use causal inference methods to adjust for baseline and time-varying confounding when treatment was not randomized.
Time zero	Definition of baseline	Use the time of decision making as time zero and align eligibility, treatment assignment, and start of follow-up at this time point to avoid time-related bias.

BMI = body mass index.

be taken by the patient in response to prognostication" was recognized as early as 1987 (31, 32); however, using established and implemented prediction models as examples, we demonstrated that misinterpretations of prediction models remain widespread.

As a way forward, we argue that developers of clinical prediction models should more carefully describe the target treatment(s) the prediction model is intended to inform, how that treatment was assigned in the development data, and how it was handled during model development (33). An effective way of specifying the role of the target treatment in a prediction model is by using the so-called *prediction estimand* framework (2, 26, 34). In addition to the usual elements required in reporting of prediction models (outcome, target population, moment of intended use, predictors, and so forth), the prediction estimand formally describes the treatment conditions under which predictions hold (for example, risk under current care, and how such care is defined). For decision support, the prediction estimand should specify the level(s) of the target treatment for which the model provides predictions (for example, the potential risk "if the individual were not to start the target treatment") (34). The prediction estimand should be part of the planning stage of model development and evaluation and, importantly, communicated to end users of the model (35, 36). To encourage its use, a description of the prediction estimand could be added to existing reporting guidelines such as TRIPOD + AI and appraisal tools such as PROBAST + AI (33, 37–39).

Recent literature delineates how causal inference techniques for estimating treatment effects when using observational data can be applied (or modified) for valid development and evaluation of potential risks under different treatment options (2, 4, 26, 34, 40–44). Our examples highlight that careful consideration is needed around selection of the data used for model development, and of the variables included as predictors. In addition to collecting predictors of the outcome, additional variables required to control for confounding (which may be time-varying) and to avoid selection bias are needed, plus information on starting and stopping

Downloaded from https://annals.org by 145.94.227.50 on 07/30/2025.

treatment (3). To avoid time-related bias, the data source must allow for alignment of eligibility, treatment assignment, and start of follow-up at the time of decision making (45). The assumptions allowing valid prediction under interventions using observational data need to be carefully stated, justified, and evaluated. For a summary of recommendations for prediction under interventions, see the Table.

Randomized controlled trials are the only setting where purely predictive methods may produce potential risks under different treatment options because randomization ensures that, on average, participant characteristics at baseline are comparable in the treatment arms. Trial data have been used for this purpose, especially in settings with heterogeneity in treatment effects across persons (46, 47). However, trial data may not be suitable for widely applicable individualized prediction due to strict inclusion criteria and small sample sizes. Although trials allow estimation of potential risks under baseline treatment options, they may be insufficient for potential risks under longer periods of treatment exposure (that is, sustained treatments) when there is imperfect treatment adherence or drop out. In that case, again, causal inference methods for prediction under sustained interventions are required (48). Combining trial data with observational data has been proposed to extend trial results to a broader target population (49).

CONCLUSION

We have identified causal blind spots that make prediction models derived from observational data susceptible to misinterpretation and often unsuitable for informing individual treatment decisions.

There is a deluge of prediction models in medical journals, on websites, and in apps. With the growing availability of mostly observational data and the rising popularity of machine-learning techniques, the future will likely see an increase in black-box algorithms proposed as decision support tools. Although flexible methods and large data sets may lead to improved predictive performance, the causal blind spots we have outlined are structural challenges. These problems can only be solved by integrating causal reasoning into predictive modeling.

From Leiden University Medical Center, Leiden, the Netherlands (N.v.G., H.P.); London School of Hygiene and Tropical Medicine, London, United Kingdom (R.H.K.); University Medical Center Utrecht, Utrecht, the Netherlands (W.v.A., K.L.); Amsterdam University Medical Centers, Amsterdam; University of Amsterdam, Amsterdam; and Pacmed, Amsterdam, the Netherlands (G.C.); Delft University of Technology, Delft, the Netherlands (J.H.K.); University of Manchester, Manchester; and THIS Institute, University of Cambridge, Cambridge, United Kingdom (N.P.); University of Amsterdam, Amsterdam, the Netherlands (S.M.); Ghent University, Ghent, Belgium; and University of Washington, Seattle, Washington (P.M.); Utrecht University, Utrecht, the Netherlands (T.v.O., J.W., D.L.W.); University of Manchester, Manchester, United Kingdom (M.S.); and Leibniz Institute for Prevention Research and Epidemiology–BIPS, Bremen; and Department of Mathematics and Computer Science, University of Bremen, Bremen, Germany (V.D.).

Note: Dr. van Geloven is the guarantor for the study, had final responsibility for the decision to submit for publication, and attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Financial Support: Prof. Keogh was funded by UK Research and Innovation (Future Leaders Fellowship MR/S017968/1 and MR/X015017/1). Dr. Wang was funded by the European Union's Horizon 2020 research and innovation program (The HTx project, grant agreement number 825162). Dr. Cinà was a consultant for Pacmed during the writing of this article and owns stock appreciation rights (SARs) of the company. Dr. Sperrin acknowledges support of the UKRI AI program, and the Engineering and Physical Sciences Research Council, for CHAI–Causality in Healthcare AI Hub (grant number EP/Y028856/1).

Disclosures: Disclosure forms are available with the article online.

Corresponding Author: Nan van Geloven, PhD, Leiden University Medical Center, Leiden 2300 RC, the Netherlands; e-mail, n.van_geloven@lumc.nl.

Author contributions are available at Annals.org.

References

1. **Goodacre S.** Using clinical risk models to predict outcomes: what are we predicting and why? Emerg Med J. 2023;40:728-730. [PMID: 37468227] doi:10.1136/emermed-2022-213057
2. **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;35:619-630. [PMID: 32445007] doi:10.1007/s10654-020-00636-1
3. **Dickerman BA, Hernán MA.** Counterfactual prediction is not only for causal inference. Eur J Epidemiol. 2020;35:615-617. [PMID: 32623620] doi:10.1007/s10654-020-00659-8
4. **Lin L, Sperrin M, Jenkins DA, et al.** A scoping review of causal methods enabling predictions under hypothetical interventions. Diagn Progn Res. 2021;5:3. [PMID: 33536082] doi:10.1186/s41512-021-00092-9
5. **Westreich D, Greenland S.** The Table 2 fallacy: presenting and interpreting confounder and modifier coefficients. Am J Epidemiol. 2013;177:292-298. [PMID: 23371353] doi:10.1093/aje/kws412
6. **Pylypchuk R, Wells S, Kerr A, et al.** Cardiovascular disease risk prediction equations in 400000 primary care patients in New Zealand: a derivation and validation study. Lancet. 2018;391:1897-1907. [PMID: 29735391] doi:10.1016/S0140-6736(18)30664-0
7. **Hernán MA, Hernández-Díaz S, Robins JM.** A structural approach to selection bias. Epidemiology. 2004;15:615-625. [PMID: 15308962] doi:10.1097/01.ede.0000135174.63482.43
8. **Didelez V, Kreiner S, Keiding N.** Graphical models for inference under outcome-dependent sampling. Statistical Science. 2010;25:368-387. doi:10.1214/10-STS340
9. **Hernán MA, Monge S.** Selection bias due to conditioning on a collider. BMJ. 2023;381:1135. [PMID: 37286200] doi:10.1136/bmj.p1135

Downloaded from https://annals.org by 145.94.227.50 on 07/30/2025.

10. **Peterson ED, Dai D, DeLong ER, et al; NCDR Registry Participants.** Contemporary mortality risk prediction for percutaneous coronary intervention: results from 588,398 procedures in the National Cardiovascular Data Registry. J Am Coll Cardiol. 2010;55:1923-1932. [PMID: 20430263] doi:10.1016/j.jacc.2010.02.005

11. **Gruberg L, Weissman NJ, Waksman R, et al.** The impact of obesity on the short-term and long-term outcomes after percutaneous coronary intervention: the obesity paradox? J Am Coll Cardiol. 2002;39:578-584. [PMID: 11849854] doi:10.1016/s0735-1097(01)01802-2

12. **Amundson DE, Djurkovic S, Matwiyoff GN.** The obesity paradox. Crit Care Clin. 2010;26:583-596. [PMID: 20970043] doi:10.1016/j.ccc.2010.06.004

13. **Lajous M, Banack HR, Kaufman JS, et al.** Should patients with chronic disease be told to gain weight? The obesity paradox and selection bias. Am J Med. 2015;128:334-336. [PMID: 25460531] doi:10.1016/j.amjmed.2014.10.043

14. **Williamson EJ, Walker AJ, Bhaskaran K, et al.** Factors associated with COVID-19-related death using OpenSAFELY. Nature. 2020;584:430-436. [PMID: 32640463] doi:10.1038/s41586-020-2521-4

15. **Westreich D, Edwards JK, van Smeden M.** Comment on Williamson et al. (OpenSAFELY): the Table 2 fallacy in a study of COVID-19 mortality risk factors. Epidemiology. 2021;32:e1-e2. [PMID: 33065610] doi:10.1097/EDE.0000000000001259

16. **Tennant P.** TABLE 2 FALLACY: or why interpretation needs more than transparency. Presented at the annual general meeting of the Computational Statistics and Machine Learning Section of the Royal Statistical Society, RSS Interpretable Machine Learning & Causal Inference Workshop, 20 December 2015. Accessed at https://rss.org.uk/getattachment/Training-Events/Events/Events-2020/Sections/Interpretable-Machine-Learning-Causal-Inferenc-(1)/PWGT-Table-2-Fallacy-RSS-December-2020-(1).pdf.aspx/?lang=en-GB on 2 January 2024.

17. **Pajouheshnia R, Damen JAAG, Groenwold RHH, et al.** Treatment use in prognostic model research: a systematic review of cardiovascular prognostic studies. Diagn Progn Res. 2017;1:15. [PMID: 31093544] doi:10.1186/s41512-017-0015-0

18. **Suissa S, Azoulay L.** Metformin and the risk of cancer: time-related biases in observational studies. Diabetes Care. 2012;35:2665-2673. [PMID: 23173135] doi:10.2337/dc12-0788

19. **Suissa S.** Effectiveness of inhaled corticosteroids in chronic obstructive pulmonary disease: immortal time bias in observational studies. Am J Respir Crit Care Med. 2003;168:49-53. [PMID: 12663327] doi:10.1164/rccm.200210-1231OC

20. **Pladet LCA, Barten JMM, Vernooij LM, et al.** Prognostic models for mortality risk in patients requiring ECMO. Intensive Care Med. 2023;49:131-141. [PMID: 36600027] doi:10.1007/s00134-022-06947-z

21. **Nashef SA, Roques F, Michel P, et al.** European system for cardiac operative risk evaluation (EuroSCORE). Eur J Cardiothorac Surg. 1999;16:9-13. [PMID: 10456395] doi:10.1016/s1010-7940(99)00134-7

22. **van Amsterdam WAC, van Geloven N, Krijthe JH, et al.** When accurate prediction models yield harmful self-fulfilling prophecies. Patterns. 2025;6:101229. [PMID: 40264961] doi:10.1016/j.patter.2025.101229

23. **Cooper GF, Aliferis CF, Ambrosino R, et al.** An evaluation of machine-learning methods for predicting pneumonia mortality. Artif Intell Med. 1997;9:107-138. [PMID: 9040894] doi:10.1016/s0933-3657(96)00367-3

24. **Caruana R, Lou Y, Gehrke J, et al.** Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, 10-13 August 2015. Association for Computing Machinery (ACM):1721-1730. doi:10.1145/2783258.2788613.

25. **Hippisley-Cox J, Coupland C, Brindle P.** Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. BMJ. 2017;357:j2099. [PMID: 28536104] doi:10.1136/bmj.j2099

26. **Sperrin M, Martin GP, Pate A, et al.** Using marginal structural models to adjust for treatment drop-in when developing clinical prediction models. Stat Med. 2018;37:4142-4154. [PMID: 30073700] doi:10.1002/sim.7913

27. **Xu Z, Arnold M, Stevens D, et al.** Prediction of cardiovascular disease risk accounting for future initiation of statin treatment. Am J Epidemiol. 2021;190:2000-2014. [PMID: 33595074] doi:10.1093/aje/kwab031

28. **Peek N, Sperrin M, Mamas M, et al.** Hari Seldon, QRISK3, and the prediction paradox. Rapid response to: Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. BMJ. 2017;357:j2099. Accessed at www.bmj.com/content/357/bmj.j2099/rr-0 on 5 December 2023.

29. **Lenert MC, Matheny ME, Walsh CG.** Prognostic models will be victims of their own success, unless…. J Am Med Inform Assoc. 2019;26:1645-1650. [PMID: 31504588] doi:10.1093/jamia/ocz145

30. **Liley J, Emerson S, Mateen B, et al.** Model updating after interventions paradoxically introduces bias. Proceedings of The 24th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research (PMLR). 2021;130:3916-3924. Accessed at https://proceedings.mlr.press/v130/liley21a.html on 5 December 2023.

31. **Hilden J, Habbema JDF.** Prognosis in medicine: an analysis of its meaning and roles. Theor Med. 1987;8:349-365. [PMID: 3424253] doi:10.1007/BF00489469

32. **Windeler J.** Prognosis - what does the clinician associate with this notion? Stat Med. 2000;19:425-430. [PMID: 10694727] doi:10.1002/(sici)1097-0258(20000229)19:4<425::aid-sim347>3.0.co;2-j

33. **van Amsterdam WAC, Giovanni C, Didelez V, et al.** Prognostic models for decision support need to report their targeted treatments and the expected changes in treatment decisions. Rapid response to: TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. Accessed at www.bmj.com/content/385/bmj-2023-078378/rr-1 on 4 June 2025.

34. **Luijken K, Morzywołek P, van Amsterdam W, et al.** Risk-based decision making: estimands for sequential prediction under interventions. Biom J. 2024;66:e70011. [PMID: 39607308] doi:10.1002/bimj.70011

35. **Sendak MP, Gao M, Brajer N, et al.** Presenting machine learning model information to clinical end users with model facts labels. NPJ Digit Med. 2020;3:41. [PMID: 32219182] doi:10.1038/s41746-020-0253-3

36. **van Royen FS, Asselbergs FW, Alfonso F, et al.** Five critical quality criteria for artificial intelligence-based prediction models. Eur Heart J. 2023;44:4831-4834. [PMID: 37897346] doi:10.1093/eurheartj/ehad727

37. **Moons KGM, Damen JAA, Kaul T, et al.** PROBAST + AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505. [PMID: 40127903] doi:10.1136/bmj-2024-082505

38. **Collins GS, Moons KGM, Dhiman P, et al.** TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. [PMID: 38626948] doi:10.1136/bmj-2023-078378

39. **Xu J, Guo Y, Wang F, et al.** Protocol for the development of a reporting guideline for causal and counterfactual prediction models in biomedicine. BMJ Open. 2022;12:e059715. [PMID: 35725267] doi:10.1136/bmjopen-2021-059715

40. **van Amsterdam WAC, de Jong PA, Verhoeff JJC, et al.** From algorithms to action: improving patient care requires causality.

Downloaded from https://annals.org by 145.94.227.50 on 07/30/2025.

BMC Med Inform Decis Mak. 2024;24:111. [PMID: 38664664] doi:10.1186/s12911-024-02513-3
41. **Sperrin M, Diaz-Ordaz K, Pajouheshnia R.** Invited commentary: treatment drop-in-making the case for causal prediction. Am J Epidemiol. 2021;190:2015-2018. [PMID: 33595073] doi:10.1093/aje/kwab030
42. **Dickerman BA, Dahabreh IJ, Cantos KV, et al.** Predicting counterfactual risks under hypothetical treatment strategies: an application to HIV. Eur J Epidemiol. 2022;37:367-376. [PMID: 35190946] doi:10.1007/s10654-022-00855-8
43. **Boyer CB, Dahabreh IJ, Steingrimsson JA**. Estimating and evaluating counterfactual prediction models. arXiv. Preprint posted online on 24 August 2023 (v1), last revised 30 December 2024 (this version, v3). doi:10.48550/arXiv.2308.13026
44. **Keogh RH, Van Geloven N.** Prediction under interventions: evaluation of counterfactual performance using longitudinal observational data. Epidemiology. 2024;35:329-339. [PMID: 38630508] doi:10.1097/EDE.0000000000001713
45. **Hernán MA, Sauer BC, Hernández-Díaz S, et al.** Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. J Clin Epidemiol. 2016;79:70-75. [PMID: 27237061] doi:10.1016/j.jclinepi.2016.04.014
46. **Kent DM, Paulus JK, van Klaveren D, et al.** The Predictive Approaches to Treatment effect Heterogeneity (PATH) statement. Ann Intern Med. 2020;172:35-45. [PMID: 31711134] doi:10.7326/M18-3667
47. **Hingorani AD, van der Windt DA, Riley RD, et al; PROGRESS Group.** Prognosis research strategy (PROGRESS) 4: stratified medicine research. BMJ. 2013;346:e5793. [PMID: 23386361] doi:10.1136/bmj.e5793
48. **Hernán MA, Hernández-Díaz S.** Beyond the intention-to-treat in comparative effectiveness research. Clin Trials. 2012;9:48-55. [PMID: 21948059] doi:10.1177/1740774511420743
49. **Colnet B, Mayer I, Chen G, et al.** Causal inference methods for combining randomized trials and observational studies: a review. Statist Sci. 2024;39:165-191. doi:10.1214/23-STS889

Downloaded from https://annals.org by 145.94.227.50 on 07/30/2025.

Author Contributions: Conception and design: N. van Geloven, R.H. Keogh, W. van Amsterdam, N. Peek, K. Luijken, T. van Ommen, M. Sperrin, D.L. Weir, V. Didelez.
Analysis and interpretation of the data: N. van Geloven, D.L. Weir, V. Didelez.
Drafting of the article: N. van Geloven, R.H. Keogh, G. Cinà, P. Morzywołek.
Critical revision of the article for important intellectual content: N. van Geloven, R.H. Keogh, W. van Amsterdam, G. Cinà, J.H. Krijthe, N. Peek, K. Luijken, P. Morzywołek, H. Putter, M. Sperrin, J. Wang, D.L. Weir, V. Didelez.
Final approval of the article: N. van Geloven, R.H. Keogh, W. van Amsterdam, G. Cinà, J.H. Krijthe, N. Peek, K. Luijken, S. Magliacane, P. Morzywołek, T. van Ommen, H. Putter, M. Sperrin, J. Wang, D.L. Weir, V. Didelez.
Statistical expertise: N. van Geloven, R.H. Keogh, W. van Amsterdam, N. Peek, S. Magliacane, P. Morzywołek, H. Putter, M. Sperrin, J. Wang, V. Didelez.

Downloaded from https://annals.org by 145.94.227.50 on 07/30/2025.