

## Pure-AMC

### Causal blind spots when using prediction models for treatment decisions

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

Published: 27/02/2024

#### *Citation for published version (APA):*

Geloven, N. V., Keogh, R. H., Amsterdam, W. V., Cinà, G., Krijthe, J. H., Peek, N., Luijken, K., Magliacane, S., Morzywołek, P., Ommen, T. V., Putter, H., Sperrin, M., Wang, J., Weir, D. L., & Didelez, V. (2024). *Causal blind spots when using prediction models for treatment decisions*.

#### **General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

#### **Take down policy**

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

# Causal blind spots when using prediction models for treatment decisions

Nan van Geloven (1\*), Ruth H Keogh (2), Wouter van Amsterdam (3), Giovanni Cinà (4,5,6), Jesse H. Krijthe (7), Niels Peek (8,9), Kim Luijken (3), Sara Magliacane (5), Paweł Morzywołek (10,11), Thijs van Ommen (12), Hein Putter (1), Matthew Sperrin (8), Junfeng Wang (12), Daniala L. Weir (12), Vanessa Didelez (13)

- (1) Leiden University Medical Center, Leiden, The Netherlands
- (2) London School of Hygiene and Tropical Medicine, London, United Kingdom
- (3) University Medical Center Utrecht, Utrecht, The Netherlands
- (4) Amsterdam University Medical Center, Amsterdam, The Netherlands
- (5) University of Amsterdam, Amsterdam, The Netherlands.
- (6) Pacmed, Amsterdam, The Netherlands
- (7) Delft University of Technology, Delft, The Netherlands
- (8) University of Manchester, Manchester, United Kingdom
- (9) THIS Institute, University of Cambridge, United Kingdom
- (10) Ghent University, Ghent, Belgium
- (11) University of Washington, Seattle, United States
- (12) Utrecht University, Utrecht, The Netherlands
- (13) Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany

\*corresponding author: n.van\_geloven@lumc.nl

Keywords: clinical prediction models, risk-based decision making, causal inference, prediction under interventions

## Abstract

Prediction models are increasingly proposed for guiding treatment decisions, but most fail to address the special role of treatments, leading to inappropriate use. This paper highlights the limitations of using standard prediction models for treatment decision support. We identify "causal blind spots" in three common approaches to handling treatments in prediction modelling and illustrate potential harmful consequences in several medical applications. We advocate for an extension of guidelines for development, reporting, clinical evaluation and monitoring of prediction models to ensure that the intended use of the model is matched to an appropriate risk estimand. For decision support this requires a shift towards developing predictions under the specific treatment options under consideration ('predictions under interventions'). We argue that this will improve the efficacy of prediction models in guiding treatment decisions and prevent potential negative effects on patient outcomes.

### Key messages

- When prediction models are used to inform treatment decisions, confounders, colliders and mediators, as well as changes in treatment protocols over time may lead to misinformed decision-making.
- We illustrate how this can occur in real examples of implemented prediction models.
- To avoid patient harm, prediction models should specify upfront the treatment decisions they aim to support and target a prediction estimand in line with that goal.
- When prediction models are intended to inform treatment decisions, they should be developed and validated using causal reasoning and inference techniques.

## Introduction

Prediction models are increasingly used to support decisions about medical treatments or lifestyle changes that aim to lower a patient's individual risk of experiencing an adverse outcome (e.g., [1]). The claim is that we could make better decisions if we knew an individual's risk of experiencing the adverse outcome of interest. For example, clinicians may choose to treat patients with a high risk more aggressively, or may reserve expensive or scarce treatments for such patients.

However, the exact way in which predictions are intended to support treatment decisions is seldom specified in papers that present newly developed prediction models and is not part of current reporting guidelines.[2] A recent review of prediction models developed for hospitalized covid-19 patients found that 64% of papers suggested that their model could support treatment decisions, but none specified how they would do so. The majority of papers did not even specify what treatments were applied to the patients that were used for development of the model.[3]

Likewise, medical guidelines currently recommend use of prediction models for decision support, but do not provide guidance on how to do so. The American Joint Committee on Cancer recently published criteria for using prediction models as decision support in cancer. According to their criteria a prediction model may or may not include treatments as predictors.[4] However, if models do not explicitly take treatments into account, they implicitly assume that treatment decisions will continue to be made in the same way as they had been historically.[5] As a result, doctors may try to optimize treatment choices based on a prediction model that actually assumes the treatment would be given based on past treatment guidelines - a contradictory situation that may potentially bring harm to patients.[6]

Recent methodological works highlight that decision support requires predictions of outcomes under the specific treatment options that patients and doctors are considering, for instance, 'the risk if we do not initiate treatment'.[7,5,8,9] This has been referred to as 'predictions under interventions' and development and performance assessment of such models requires causal inference techniques.[10,11,12,13] The need for involving causal reasoning into prediction research may seem counterintuitive, as traditionally the tasks of prediction and causal inference have been seen as separate.[14] However, when estimated risks from prediction models are used to inform treatment decisions, this is no longer a purely predictive task.

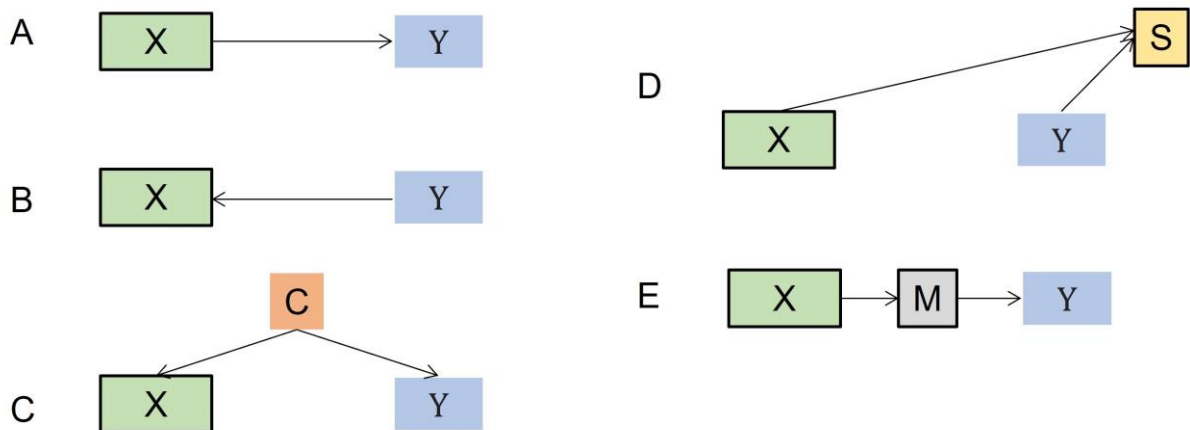
In this paper we aim to raise awareness about the potential harm of inappropriately using prediction models to support treatment decisions when the models were not designed for this purpose. We use causal reasoning aided by directed acyclic graphs (DAGs) to highlight ways in which prediction models could fail in supporting treatment decisions in real examples. This is discussed by considering 'causal blind spots' in three common ways of handling treatments during model development: 1) including treatment as a predictor, 2) restricting the development data based on treatment status, and 3) ignoring treatments in the development data. We conclude by proposing a way forward in which we can make use of causally grounded predictions under interventions for decision support.

## Setting

Our focus is on the common setting in which end users obtain estimates of the risk of an outcome from a prediction model and then use this to inform decisions about the best out of multiple treatment options. For ease of our arguments we will focus on a binary treatment and on decisions to either start or cease using it. We use the word ‘treatment’ here but it could be any intervention, lifestyle exposure or other action about which patients and doctors want to make a decision informed by the prediction model.

Figure 1 depicts a series of simple DAGs to illustrate different causal structures under which a predictor  $X$  (possibly a treatment) and an outcome  $Y$  in a prediction model can be associated.[15] A predictive algorithm detects associations, but is not designed to deduce the mechanisms through which these associations came about. We argue below that considering the underlying causal mechanism related to treatment use in the development data for a prediction model is crucial to determining whether the outputs from that model can be used to support future treatment decisions. In subsequent sections we address real examples which contain combinations of the basic structures illustrated in Figure 1.

The population of individuals on which the prediction model is developed may include prevalent users of the treatment, as well as individuals who are non-users or incident users at baseline. Later, we also discuss the case where individuals start treatment during follow up.



**Figure 1 Prediction models through the lens of causal DAGs.** Different causal structures that may lead to an association between a predictor X (green shaded) and outcome Y (blue shaded) in a prediction model. Rectangles indicate variables, arrows indicate a causal effect, black outline of the rectangle indicates that the variable is conditioned on in the prediction model. Orange shaded variable is a confounder, yellow shaded variable a selection variable (collider) and grey shaded variable a mediator. Firstly, X and Y may be associated because X is indeed a cause of Y (Figure 1A). However, also Y being a cause of X may have induced the association (Figure 1B). In clinical prediction models the chronology of first measuring predictors and later on measuring the outcome will typically prevent associations induced by such *reverse causation*. However, some settings have been described where the disease Y is already latently present at baseline (when X is measured). For instance, decreased food intake (X) may predict a cancer diagnosis (Y) because the cancer was not yet detected at baseline, while it was already causing appetite loss. One may also see the latent cancer as an instance of the third option, which is the setting where a third variable 'C' is a cause of both the predictor and the outcome, but this third variable is not included in the prediction model (Figure 1C). This introduces a non-causal association between X and Y, referred to as *confounding*. Figure 1D depicts a setting where a non-causal association is introduced between X and Y by conditioning on a third variable that is a causal consequence of both X and Y (called a *collider*).<sup>[16]</sup> Conditioning on a third variable may occur due to the third variable being included as one of the additional predictors in the model or by basing the selection of the sample used to develop the prediction model on the value of this third variable. As the former would not be expected due to chronology mentioned above, we focus below on selection into the sample. Figure 1E depicts a mediating variable 'M' on the path from X to Y. If the mediating variable is included as an additional predictor in the model, (part of) the causal effect of X on Y will be 'blocked' (or masked) by M.

## Including treatment as a predictor

Suppose that the prediction model included baseline treatment status as one of the predictors. If end users apply the prediction model by specifying 'no' for the treatment variable and interpret the result as the risk that may be expected when a new patient chooses not to start (or to cease) treatment, this is a 'causal interpretation' of risk. The causal interpretation being made is that the model gives the expected risk under the decision not to give (or to cease) the treatment. In general, prediction models do not allow this interpretation. They provide predictions based on the subset of patients who happened to be untreated. In an observational data setting, the untreated subset may not be representative for the whole population. Predictions only have a causal interpretation in the special case where the other predictors in the model form a so-called 'valid adjustment set'. This means that the other predictors must resolve any confounding, that the sample selection should not have introduced collider bias and the other predictors should not be consequences of the treatment.[15] Typically, researchers select the set of predictors in order to maximise predictive performance of the model given the available variables, rather than with the listed causal requirements in mind. In general, the set of predictors selected in this way will not form a valid adjustment set for a treatment variable. Below we describe three situations illustrated by real examples of prediction models in which causal interpretation of risks obtained by including treatment as a predictor clearly fails.

### *Unmeasured confounding – PREDICT equations for cardiovascular risk*

The PREDICT study used routinely collected primary care data from over 400,000 individuals in New Zealand to develop a model for predicting 5-year risk of cardiovascular disease.[17] One of the predictors in the model was the use of blood pressure lowering medication at baseline (representing both prevalent and incident use). Use of blood pressure lowering medication according to the model is associated with a higher hazard of developing cardiovascular disease compared to patients not taking this medication (HR for men 1.40 95% CI (1.31-1.50), HR for women 1.34 95% CI (1.27-1.42)). Applying the prediction model for a woman who is on blood pressure lowering medication, smokes, has systolic BP 120 and TC/HDL ratio 1.1, results in an estimated risk of 11% for cardiovascular disease in the next five years. Using the same inputs, but changing it so that this woman does not use blood pressure lowering medication results in an estimated risk of only 8%. This should surely not be interpreted as support for stopping the blood pressure lowering medication for this woman. Instead, the higher estimated risk of the outcome with medication use is most likely due to confounding. For example, previous episodes of hypertension (not included in the model) may have formed an indication to start the medication and at the same time led to increased cardiovascular risk (Figure 2A).

### *Collider bias due to sample selection – covid-19 app and UK Biobank*

Interpretations of predictions may also be affected by how the development data was selected, in particular when the data were restricted based on consequences or causes of both predictor and outcome.[18]

During the covid-19 pandemic several prediction models for covid-19 infection were developed using data collected with mobile phone apps, to which participants could contribute daily information on their activities and on covid-19 infection status (e.g., [19]). It can be expected that app users are not a random sample from the population.[20] Suppose that a model for covid-19

infection included exercise level as a predictor and was used to inform lifestyle recommendations about whether modifying exercise level could change one's infection risk. Also suppose that people who don't exercise much are more likely to be app users, and that people with symptoms of underlying covid-19 infection or who have recently been in contact with infected individuals were more inclined to contribute daily data to the app. This self-selection would induce a non-causal association between exercising and risk of infection among app users, leading to higher estimated infection risks for individuals who exercise more (Figure 2B). These risks are therefore not appropriate for informing lifestyle changes in new individuals. This is an example of collider bias due to selecting the data in a way that depends on treatment and outcome.

Collider bias may also occur if, reversely, the factors that influence sample selection influence the treatment and outcome. This issue has for instance been identified in studies examining the association between lifestyle factors and health outcomes using data from the UK Biobank.[21,22,23]. The UK Biobank enrolled over 500,000 individuals in the United Kingdom, but this was only 5.5% of all invited individuals. Participation was voluntary, resulting in a study group that was not representative of the UK population (e.g., the proportion of current smokers was lower). UK Biobank participants also have lower rates of mortality and cancer incidence, which could be related to differences in observed lifestyle factors, but could also be due to differences in other biological factors not recorded.[24] McCarthy et al (2000) used the UK Biobank data to develop a prediction model for head and neck cancer.[25] Several lifestyle factors were associated with higher cancer risk in their model. In their discussion they recommend that their personalized risk estimates 'could support discussions between dental professionals and patients regarding risk behaviours, such as smoking and alcohol consumption. This would also provide the opportunity to discuss health promoting behaviours, such as eating fresh fruit and vegetables and taking regular exercise.' This assumes a causal interpretation of the model in which continuation of poor lifestyle behaviour leads to increased risk, and that this can be used to motivate lifestyle changes. However, lower estimated cancer risks for patients with healthier lifestyle could be in part due to the selection mechanisms. It is for instance known that individuals with higher socio-economic status (SES) were more inclined to participate and it is possible that the same holds for persons with biological risk factors.[24] Since SES is related to lifestyle habits, and biological risk factors relate to cancer incidence, selection based on both will introduce a non-causal association between lifestyle habits and cancer incidence (Figure 2C). This type of collider bias is also referred to as M-bias due to the shape of the DAG.[26] The McCarthy model did include a predictor representing SES of the area that participants lived in (Townsend deprivation index), but this may not fully capture individual participant's SES.

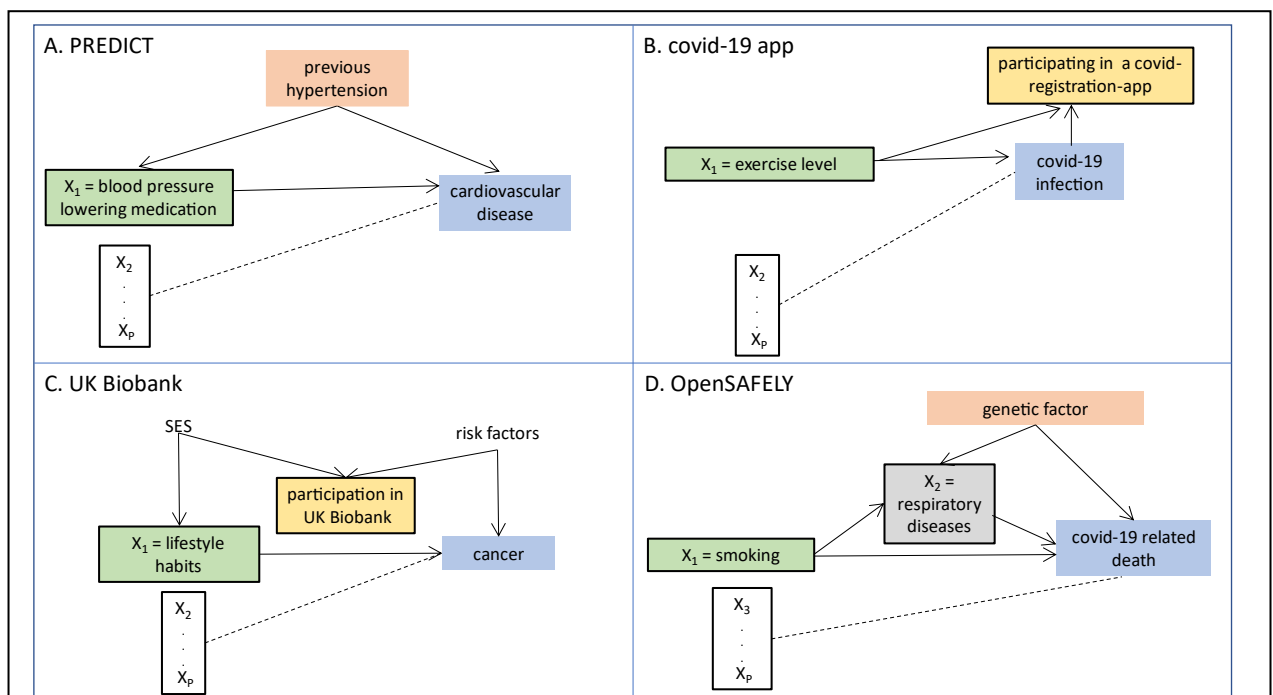
#### *Mediation and collider bias due to including a consequence of the treatment in the model – OpenSAFELY model for risk of covid-19 death*

Williamson et al. evaluated risk factors associated with covid-19-related death using a multivariable model fitted using primary care records of >17 million adults in the United Kingdom from the OpenSAFELY data platform.[27] In the multivariable model, the coefficient of 'current smoking' was less than 1 (HR 0.89 95% CI 0.82-0.97), representing lower hazard of covid-19-related death for smokers than for non-smokers. Though the authors warned against causal interpretation of their model, policy-makers initially interpreted this result to mean that smoking was protective for covid-19 related death, which almost led to exclusion of smokers from shielding policies in France.[28,29]



A likely explanation for the counterintuitive finding for smoking is that the model additionally included factors that are causal consequences of smoking, for example respiratory diseases. If these variables are also causes of the outcome, they may have acted as mediators, masking part of the true causal effect of smoking (Figure 2D). One may think that mediators are not to be expected in a prediction model given that typically all predictors are measured at the same baseline moment, and there is no time to let the treatment variable impact the mediator. However, often the treatment variable represents an existing (prevalent) exposure to that treatment. For example, current smoking is very strongly related to smoking in the past year. And the past exposure to smoking may have led to respiratory diseases.

A second issue may arise when a variable included in the prediction model is a consequence of treatment. Suppose there is an underlying (e.g. genetic) factor that has a causal effect on risks of both respiratory disease and covid-19 related death (Figure 2D). If this factor is unaccounted for in the prediction model, then a spurious association between smoking and the outcome will be created because respiratory disease is a collider on the non-causal path from smoking to outcome via the unmeasured genetic factor. In other words, someone who does not smoke but does have a respiratory disease is more likely to be genetically predisposed and therefore also at higher risk for covid-19 related death. This problem will arise even if respiratory disease does not directly affect the outcome (i.e., if the arrow from respiratory disease to outcome was omitted in Figure 2D).[30]



**Figure 2** DAGs depicting unobserved confounding (A), collider bias by sample selection (B and C) and mediation and collider bias by conditioning on a consequence of the treatment (D). Rectangles indicate variables, arrows indicate a causal effect, black outline of the rectangle indicates that the variable is conditioned on in the prediction model. Green shaded variable is a treatment (or modifiable exposure), blue shaded variable is the outcome. Orange shaded variable is a confounder, yellow shaded variable a selection variable (collider) and grey shaded variable a mediator.

## Restricting the development data to only untreated or only treated patients

Instead of including the treatment as a predictor in the model, as in the above examples, it is common in prediction modelling to restrict the data based on treatment status, i.e., deliberately using only patients from the development data who did not use treatment or only those who did.[31] It is tempting to think that restricting to untreated patients would result in estimates of the risk if someone were not to get the treatment. However, those who happen to be untreated in the development data are generally not representative of those for whom that decision is still to be made because treatment decisions in the real world are not random. Restricting the development data based on treatment status is just another way of conditioning on treatment status, and therefore this approach faces the same issues as the earlier examples in which treatment was included as a predictor in the model.

For example, in the PREDICT equations described above, suppose that instead of adding 'not using blood pressure lowering medication' as a predictor in the model, the investigators had restricted their analysis to the individuals who were not using blood pressure lowering medication at baseline. They would then have estimated a similar risk for cardiovascular disease in the next five years of around 8% for an untreated individual (versus 11% for a treated individual). And just like before, this estimated risk is not appropriate for informing a decision about stopping blood pressure lowering medication for this individual. Interpreting estimated risks from this model as 'risks if treatment were stopped/withheld' is only valid if all confounding, collider and mediation issues described above could be ruled out. An additional issue would occur if the development data were restricted to untreated individuals who also did not start blood pressure lowering medication during follow up. Such restriction based on future treatment status would be prone to immortal time bias.[32,33]

A systematic review of prognostic models for outcomes of extracorporeal membrane oxygenation (ECMO) therapy via heart-lung by-pass machines found that all 58 studies restricted to patients receiving the ECMO therapy. At the same time 11 of these 58 studies stated that their primary aim was to support the decision of ECMO initiation in individual new patients.[34] A similar example is the EuroSCORE model that predicts mortality after cardiac surgery. It was suggested the model could be used to inform the decision on whether to proceed with the considered surgery after calculating the risk score. But the model was developed on a population of patients who all underwent the surgery.[35] Similarly, prediction models fitted on (solely) transplanted patients have been suggested to be useful for informing decisions on who should be put on the waitlist for organ transplant.[36]

## Ignoring treatments applied in the development data

Given the complications outlined above, one may think it is better to ignore treatment status when developing a prediction model – that is, not include treatment as predictor and not restrict the data based on treatment status. When the development data contains a mix of patients – some treated and some untreated – ignoring treatment in the prediction model will result in estimated risks under the historical treatment policy which can be interpreted 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 since we do not have a prediction under this novel treatment strategy, we don't know if and how much it will benefit the patient. Moreover, what constitutes 'current care' is often poorly defined, making it difficult to determine what it would mean for a particular individual. Therefore, it is in turn unclear 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 is dangerous advice as a patient may be estimated to be at low risk exactly because under current care these patients received adequate treatment. Withholding treatment from this subgroup may lead to harm, again since we do not have an expected outcome under less intense treatment.[37] The danger of this approach is illustrated by a model that estimates the mortality risk among patients hospitalized with pneumonia.[38] The intention was that pneumonia patients estimated to be at low mortality risk could be safely treated as outpatients whereas those at high risk should be admitted to hospital. The assumption was made that if a hospital-treated pneumonia patient had a very low probability of death, then that patient would also have had a very low probability of death if treated at home.[38] It turned out that this assumption was false. In particular, the prediction algorithm assigned patients with asthma low risks. The underlying reason for asthma to be associated with lower risk was that in the past such patients would often get effective intensive hospital care which led to their favourable outcomes.[39] If, based on the apparent low risk assigned by the prediction algorithm, asthma patients would instead be treated as outpatients, their outcomes would likely be worse than predicted by the model.

Similar issues have been pointed out for more recently-developed prediction models such as QRISK3, which estimates the 10 year risk of cardiovascular disease in the general population and is used in primary care in the UK for supporting decisions on starting the preventive use of statins.[40] During model development 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 period in which the development cohort was followed. A person might appear at low risk in the model because similar patients initiated treatments at a later time point and this does not guarantee they would be at low risk under no treatment.[7,41]

Both examples illustrate the prediction paradox: when predictions are used to support treatment decision, this will change treatment practice and thus potentially invalidate any prediction that was made under historical practice.[42,43,44] For this reason, ignoring treatments during model development in general does not provide actionable risks.

## The way forward

We have pointed out that severe problems may occur when standard prediction models are used to support treatment decisions, and illustrated these with a number of established and implemented prediction models. Now the question arises: how can we avoid these problems?

One option would be to restrict the use of prediction models to purposes other than that of treatment decision support. Standard prediction models can offer important psychological support to patients, helping them to prepare for what to expect. In addition, standard prediction models can be used for care planning.

If prediction models do aim to support decisions about treatments that change the very risk in question, we argue that it is crucial to carefully articulate the risk estimands and discuss the assumptions being made that allow us to estimate these from data. In this case, the risk questions will be inherently causal, e.g. 'what would the risk be for this individual if they started or ceased treatment', or 'what if they were to start a different treatment'? Randomised controlled trials and perhaps historically untreated cohorts are the only sources where pure predictive methods suffice to obtain such interventional predictions from data. Some trials have been used to obtain predictions under intervention options, especially in settings with heterogeneity in treatment effects across individuals.[45,46] But in general, trial data are not ideal for this purpose due to strict inclusion criteria and small sample size. Use of contemporary observational data sources in which at least some individuals already received the treatment of interest requires causal inference techniques both during model development and performance assessment of predictions under interventions.[10,11,12,13] Though causal inference techniques are available for this, but they are currently seldom applied in the predictive context. This also puts extra requirements on data availability: apart from the predictors and potential effect modifiers, additional variables required to control for confounding are needed (which could be and often are time-varying), as well as information on starting and stopping of treatment.[12]

There is a deluge of papers proposing prediction models in medical journals, on websites and in apps. Even more are expected due to increased data availability and popularity of machine learning algorithms. To avoid inappropriate use and potential harm of these models, the intended use of a prediction model needs to be part of the planning stage of model development, and, importantly, communicated to end users at the implementation stage [47,48]. The intended use should be accompanied with a description of an appropriate risk estimand that describes the role of treatment in the model, including the distinction between prevalent, incident and future use of treatment.[7,5,9] The description of such a risk estimand could be added to reporting guidelines such as TRIPOD and appraisal tools such as PROBAST [2,49].

In the end, one would want to know whether patient outcomes improve by using the interventional prediction model for decision making. Careful subgroup analyses of historical trials with sufficient information to examine how outcomes would have been affected if risk-based decision rules had been applied may be possible.[50] Sufficiently rich observational data analyzed with appropriate methods of causal inference could also be used for this purpose and offer the option to compare with outcomes under standard care.[51] Ultimately, the impact of a prediction model should be assessed in a (cluster) randomized trial that compares outcomes of patients in hospitals where patients and doctors used the model to inform their treatment decisions and in hospitals that followed the standard decision making process.[52,53]

## Discussion

In this paper we have identified several blind spots in prediction modelling that make many standard prediction models unsuited for the task of informing decision making.

The problem of causally interpreting an association between a treatment variable included as a predictor in a model and the outcome is related to what has been described as the 'Table 2 fallacy' in causal models: interpreting regression coefficients as causal effects for variables that were not intended to be used in that way when constructing the model.[54] While this issue is becoming more

widely known, the need for causal reasoning in prediction modelling, where the focus is on the risks rather than individual regression coefficients, is largely unrecognized. We have explained how deploying such models to inform treatment decisions could bring serious harm to patients.

The future is likely to bring more (black-box) machine-learning methods proposed as decision support tools. With such models there is less risk of interpreting coefficients as causal, but “explainable” analyses of machine learning algorithms, e.g. with variable importance such as Shapley values, are similarly prone to misinterpretation. Flexible methods and big data can lead to improved predictive accuracy – but the causal blind spots we have addressed are structural problems that cannot be solved by computational power. The predictions resulting from such algorithms remain challenged by the same biases as regression models when used to support treatment decisions.

## Footnotes

**Contributors:** All authors provided a substantial contribution to the design and interpretation of the paper and revised drafts. NvG, RHK and VD initiated the project. NvG wrote the initial draft and is the guarantor for the study. The corresponding author had final responsibility for the decision to submit for publication. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

### Funding:

RHK was funded by UK Research and Innovation (Future Leaders Fellowship MR/S017968/1 and MR/X015017/1)

JW was funded by the European Union's Horizon 2020 research and innovation program (The HTx project, grant agreement No 825162)

GC was a consultant for Pacmed during the writing of this article and owns stock appreciation rights (SARs) of the company.

## References

1. Aleksandrova K, Reichmann R, Kaaks R, et al. Development and validation of a lifestyle-based model for colorectal cancer risk prediction: the LiFeCRC score. *BMC Medicine* 2021;19:1. doi:10.1186/s12916-020-01826-0
2. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Ann Intern Med* 2015;162:W1–73. doi:10.7326/M14-0698
3. Prosepe I, Groenwold RH, Knevel R, Pajouheshnia R, van Geloven N. The disconnect between development and intended use of clinical prediction models for Covid-19: a systematic review and real-world data illustration. *Front. Epidemiol.* 2:899589. doi: 10.3389/fepid.2022.899589
4. Kattan MW, Hess KR, Amin M, et al. AMERICAN JOINT COMMITTEE ON CANCER ACCEPTANCE CRITERIA FOR INCLUSION OF RISK MODELS FOR INDIVIDUALIZED PROGNOSIS IN THE PRACTICE OF PRECISION MEDICINE. *CA Cancer J Clin* 2016;66:370–4. doi:10.3322/caac.21339
5. 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;35:619–30. doi:10.1007/s10654-020-00636-1
6. van Amsterdam WAC, de Jong PA, Verhoeff JJC, et al. Decision making in cancer: Causal questions require causal answers. 2022. <http://arxiv.org/abs/2209.07397> (accessed 2 Aug 2023).
7. Sperrin M, Martin GP, Pate A, et al. Using marginal structural models to adjust for treatment drop-in when developing clinical prediction models. *Statistics in Medicine*. 2018;37:4142–54.
8. Sperrin M, Diaz-Ordaz K, Pajouheshnia R. Invited Commentary: Treatment Drop-in—Making the Case for Causal Prediction. *American Journal of Epidemiology*. 2021;190:2015–8.
9. Luijken K, Morzywołek P, van Amsterdam W, et al. Risk-based decision making: estimands for sequential prediction under interventions. 2023. <https://doi.org/10.48550/arXiv.2311.17547>

10. 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.
11. Dickerman BA, Dahabreh IJ, Cantos KV, et al. Predicting counterfactual risks under hypothetical treatment strategies: an application to HIV. *Eur J Epidemiol.* 2022;37:367–76.
12. Keogh RH, van Geloven N. Prediction under hypothetical interventions: evaluation of counterfactual performance using longitudinal observational data. 2023. <https://doi.org/10.48550/arXiv.2304.10005>
13. Boyer CB, Dahabreh IJ, Steingrimsson JA. Assessing model performance for counterfactual predictions. 2023. <https://doi.org/10.48550/arXiv.2308.13026>
14. Shmueli G. To Explain or to Predict? *Statistical Science.* 2010;25:289–310.
15. Pearl J. *Causality: models, reasoning, and inference.* Cambridge, U.K. ; New York: Cambridge University Press 2000.
16. Hernán MA, Monge S. Selection bias due to conditioning on a collider. *BMJ.* 2023;381:p1135.
17. Pylypchuk R, Wells S, Kerr A, et al. Cardiovascular disease risk prediction equations in 400 000 primary care patients in New Zealand: a derivation and validation study. *The Lancet* 2018;391:1897–907. doi:10.1016/S0140-6736(18)30664-0
18. Didelez V, Kreiner S, Keiding N. Graphical Models for Inference Under Outcome-Dependent Sampling. *Statistical Science.* 2010;25:368–87.
19. Menni C, Valdes AM, Freidin MB, et al. Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nat Med.* 2020;26:1037–40.
20. Griffith GJ, Morris TT, Tudball MJ, et al. Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nat Commun* 2020;11:5749. doi:10.1038/s41467-020-19478-2
21. Munafò MR, Tilling K, Taylor AE, et al. Collider scope: when selection bias can substantially influence observed associations. *International Journal of Epidemiology.* 2018;47:226–35.
22. Keyes KM, Westreich D. UK Biobank, big data, and the consequences of non-representativeness. *Lancet (London, England).* 2019;393:1297.
23. Stamatakis E, Owen KB, Shepherd L, et al. Is Cohort Representativeness Passé? Poststratified Associations of Lifestyle Risk Factors with Mortality in the UK Biobank. *Epidemiology.* 2021;32:179–88.
24. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *American Journal of Epidemiology.* 2017;186:1026–34.
25. McCarthy CE, Bonnet LJ, Marcus MW, et al. Development and validation of a multivariable risk prediction model for head and neck cancer using the UK Biobank. *International Journal of Oncology.* 2020;57:1192–202.
26. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology.* 1999;10:37–48.
27. Williamson EJ, Walker AJ, Bhaskaran K, et al. Factors associated with COVID-19-related death using OpenSAFELY. *Nature* 2020;584:430–6. doi:10.1038/s41586-020-2521-4
28. 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:e1–2. doi:10.1097/EDE.0000000000001259
29. Tennant P, TABLE 2 FALLACY - or why interpretation needs more than transparency. Presentation at the RSS Interpretable machine learning & causal inference workshop. Dec 20 2015, <https://www.youtube.com/watch?v=OS8LZUxi0eg> accessed Jan 2<sup>nd</sup> 2024.

30. Hernán MA, Hernández-Díaz S, Robins JM. A Structural Approach to Selection Bias: *Epidemiology*. 2004;15:615–25.
31. 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. doi:10.1186/s41512-017-0015-0
32. Suissa S, Azoulay L. Metformin and the Risk of Cancer. *Diabetes Care*. 2012;35:2665–73.
33. 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.
34. Pladet LCA, Barten JMM, Vernooij LM, et al. Prognostic models for mortality risk in patients requiring ECMO. *Intensive Care Med* 2023;49:131–41. doi:10.1007/s00134-022-06947-z
35. Nashef SA, Roques F, Michel P, et al. European system for cardiac operative risk evaluation (EuroSCORE). *Eur J Cardiothorac Surg* 1999;16:9–13. doi:10.1016/s1010-7940(99)00134-7
36. Almond CS, Gauvreau K, Canter CE, et al. A Risk-Prediction Model for In-Hospital Mortality After Heart Transplantation in US Children. *American Journal of Transplantation*. 2012;12:1240–8.
37. van Amsterdam WAC, van Geloven N, Krijthe JH, et al. When accurate prediction models yield harmful self-fulfilling prophecies. 2023. <https://doi.org/10.48550/arXiv.2312.01210>
38. Cooper GF, Aliferis CF, Ambrosino R, et al. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artificial Intelligence in Medicine*. 1997;9:107–38.
39. Caruana R, Lou Y, Gehrke J, et al. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Sydney NSW Australia: : ACM 2015. 1721–30. doi:10.1145/2783258.2788613
40. 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;j2099.
41. Xu Z, Arnold M, Stevens D, et al. Prediction of Cardiovascular Disease Risk Accounting for Future Initiation of Statin Treatment. *American Journal of Epidemiology*. 2021;190:2000–14.
42. Peek N, Sperrin M, Mamas M, Van Staa T, Buchan I. Hari Seldon, QRISK3, and the prediction paradox. *BMJ*. 2017;357:j2099.
43. Lenert MC, Matheny ME, Walsh CG. Prognostic models will be victims of their own success, unless.... *J Am Med Inform Assoc*. 2019;26:1645–50.
44. 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*. PMLR 2021:3916–24. <https://proceedings.mlr.press/v130/liley21a.html> (accessed 5 December 2023)
45. Hingorani AD, Windt DA van der, Riley RD, et al. Prognosis research strategy (PROGRESS) 4: Stratified medicine research. *BMJ*. 2013;346:e5793.
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.
47. Sendak, M. P., Gao, M., Brajer, N., & Balu, S. (2020). Presenting machine learning model information to clinical end users with model facts labels. *NPJ digital medicine*, 3(1), 1-4.
48. Van Royen FS, Asselbergs FW, Alfonso F, et al. Five critical quality criteria for artificial intelligence-based prediction models. *European Heart Journal*. 2023;44:4831–4.
49. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann Intern Med*. 2019;170:51.
50. Karmali KN, Lloyd-Jones DM, Van Der Leeuw J, et al. Blood pressure-lowering treatment strategies based on cardiovascular risk versus blood pressure: A meta-analysis of individual participant data. *PLoS Med*. 2018;15:e1002538.



51. Sachs MC, Sjölander A, Gabriel EE. Aim for Clinical Utility, Not Just Predictive Accuracy. *Epidemiology*. 2020;31:359–64.
52. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Ann Intern Med*. 2006;144:201–9.
53. Kappen TH, van Klei WA, van Wolfswinkel L, et al. Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. *Diagnostic and Prognostic Research*. 2018;2:11.
54. Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *Am J Epidemiol*. 2013;177:292–8.