Carriero *et al. Diagnostic and Prognostic Research* (2025) 9:29
https://doi.org/10.1186/s41512-025-00213-8

Diagnostic and Prognostic Research

COMMENT Open Access

Explainable AI in healthcare: to explain, to predict, or to describe?

Alex Carriero[1*], Anne de Hond[1], Bram Cappers[2], Fernando Paulovich[2], Sanne Abeln[3], Karel GM Moons[1] and Maarten van Smeden[1]

Abstract

Explainable Artificial Intelligence (AI) methods are designed to provide information about how AI-based models make predictions. In healthcare, there is a widespread expectation that these methods will provide relevant and accurate information about a model's inner-workings to different stakeholders (ranging from patients and healthcare providers to AI and medical guideline developers). This is a challenging endeavor since what qualifies as relevant information may differ greatly depending on the stakeholder. For many stakeholders, relevant explanations are causal in nature, yet, explainable AI methods are often not able to deliver this information. Using the Describe-Predict-Explain framework, we argue that Explainable AI methods are good descriptive tools, as they may help to describe *how* a model works but are limited in their ability to explain *why* a model works in terms of true underlying biological mechanisms and cause-and-effect relations. This limits the suitability of explainable AI methods to provide actionable advice to patients or to judge the face validity of AI-based models.

Introduction

In healthcare, there is a long tradition of developing prediction models to estimate the probability that a given health outcome is present (diagnosis) or will occur in the future (prognosis) on an individual patient basis [1]. This information can be helpful to both healthcare providers and patients. For example, a cardiologist might use SCORE-2 to estimate the risk that a patient will develop cardiovascular disease in the next ten years [2]. Proponents claim that the recent developments in AI will make such risk estimation become even more accurate and tailored to the specific patient as more diverse data sources are used (e.g., images, text) and as more complex models are considered [3]. Meanwhile criticism of these more complex models is often related to their black-box nature, especially when used in sensitive settings like healthcare [4–8].

A variety of explainable AI techniques have been suggested to help circumvent the black-box nature of AI, making the inner-workings of AI-based prediction models more transparent. Examples of explainable AI methods include the commonly reported Local Interpretable Model-Agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP) and Gradient Class Activation Mapping (GradCAM) [5, 6, 9–11]. The aim of these methods is to provide understandable and relevant information about an AI-based prediction model's input and functionality. However, the degree to which the information from more transparent AI is useful and relevant may depend on the stakeholder receiving the information, and acting upon it[12]. Moreover, as we will argue in this

*Correspondence:
Alex Carriero
a.j.carriero@umcutrecht.nl
[1]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands
[2]Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands
[3]Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands

paper, the level of transparency that can be reached by explainable AI methods may not align with what important stakeholders, such as healthcare professionals and patients, want or expect from these methods.

In particular, stakeholders may seek information that is causal in nature, since evidence-based (causal) information is typically necessary for medical decision-making. Yet, current explainable AI methods are ill-equipped to deliver such causal guarantees when applied to explain the predictions of correlation-based prediction models [13]. For example, if a cardiologist communicates to a patient that their risk of cardiovascular disease is high based on a prediction from the SCORE-2 model, it is highly desirable for a patient to understand *why* they are at high risk, and what they might be able to *do* to improve their prognosis; these are inherently causal questions. With explainable AI methods, it is technically feasible to determine what patient characteristics led to the prediction and what input values could change in order for the model to give a lower risk estimate. However, guidance based on explainable AI output typically does not come with causal guarantees, meaning that it may not align with guidance that is based on true cause-and-effect relationships. In other words, the guidance may be ineffective or provide potentially harmful recommendations. Thus, information derived from explainable AI methods typically does not meet the standards necessary for use in medical decision-making.

In this paper, we use the Describe-Predict-Explain framework introduced by Shmueli [14] to highlight the type of information that explainable AI generally provides. We argue that within the framework, explainable AI generally provides descriptions, not explanations. Using an illustrative example we demonstrate the risks of misinterpreting the results of explainable AI methods for health(care) decision making.

Framework: describe, predict, or explain

The distinction between description, prediction and explanation is important in the fields of statistics and data science [14–16], representing three distinct motivations for why one might develop a statistical model (e.g., a regression or machine-learning based model).

If the objective is *description* [14, 16], the aim is to describe patterns in available patient data. For example, medical researchers may compute summary statistics, correlations, or develop a (multi-variable) model which relates patient characteristics to a health outcome (e.g., disease, side effect, treatment effect, etc.). In the latter case, the interest lies in studying how the model combines the features in order to compute the predicted outcome. The expression the model uses to relate patient characteristics to a health outcome is seen as a description of the data since it captures relationships between the features and the outcome observed in the model development data. For example, given data from multiple prospective cohorts following breast cancer diagnosis, descriptive information may include a summary of what characteristics are more common in women with breast cancer vs. those without breast cancer [17].

If the objective is *prediction* [14, 16], the aim in healthcare is often to develop a prediction model that gives accurate risk predictions for a diagnostic or prognostic outcome in individual patients [1, 18]. In this context, it is helpful to think about clinical prediction models as tools for use by healthcare professionals and patients. These tools should be easy to use (e.g., involve readily available information) and function effectively in the setting where they will be applied (e.g., provide accurate, reliable, and clinically useful predictions) [16]. The quality of the predictions for future patients is the primary concern. Examples of such prediction models can be found in almost every field of medicine, from disease prevention, to care, and to cure. For example, a prediction model might be used to estimate the risk of femur fracture from a patient's radiograph (diagnostic model) [19] or to estimate a patient's 10-year risk of fatal cardiovascular disease based on their age, smoking status, systolic blood pressure, total- and HDL-cholesterol (prognostic model) [2].

If the objective is *explanation* [14, 16], the aim is to gain an understanding of the world around us[18]. In healthcare this often equates to understanding the direct, causal effect of a patient characteristic, exposure or an administered treatment on patient health outcomes. Intervention studies (i.e., randomized experiments) are generally perceived as the gold standard for causal evidence in medicine. When no randomization is possible or ethical, causal effects may be studied from non-randomized studies using causal inference methodologies. In this domain, there is a specific focus on answering *"What if?"* questions with respect to a single actionable intervention [20]. For example, when estimating a patient's 10-year risk of fatal cardiovascular disease, one may be interested in studying the consequence of an actionable change (e.g., an intervention) on the patient's outcome: *what if* a given patient stops smoking?

To describe or explain AI-based prediction models

Diagnostic and prognostic prediction models, unsurprisingly, align most clearly with Shmueli's motivational domain of prediction. However, stakeholders (i.e., model developers and model users) may require information that is not clearly aligned with the domain of prediction. Instead, model developers and users often seek information about *how* a given prediction model functions (i.e., how does the input information influence a patient's risk estimate). To which domain does this information

belong? I.e., do explainable AI methods provide descriptions or explanations? We argue that the results of explainable AI methods often align closely with Shmueli's domain of description.

The domain of explanation is distinct from description as the aim is to understand cause-and-effect relationships. A true (causal) explanation in Shmueli's framework would provide an understanding of the effect a change in X (input variable) would have on Y (modeled outcome). However, prediction models (regression or machine-learning based) are often not designed to capture causal information about the effects of their input variables on the outcome [16, 21]. Even though AI-based prediction models that function effectively at their prescribed task may provide the illusion that they have an internal understanding of the world, these models have no ability to distinguish causes and effects [22]. A key realization is that prediction models are often trained to learn statistical associations apparent in *observational data*. Identifying causal mechanisms from observational data requires more than just a statistical model (regression or machine learning based), it requires theory, domain knowledge, and often mechanistic studies [16, 20]. Thus, when prediction models are trained to learn statistical associations in observational data, any corresponding model explanations likely align well with Shmueli's domain of description, not explanation.

Consider the earlier example from the domain of description: what characteristics are more common in women with breast cancer vs. without breast cancer given data from many prospective cohorts following a breast cancer diagnosis? For a prediction model developed using a database of observations from these cohorts, the same questions can be reformulated as: what patient characteristics contribute to the prediction of breast cancer? In this case, the prediction model makes predictions based on patterns apparent in the model's training data. Information about how the model works is then, by extension, a description of the model training data. As we illustrate next, these descriptions or model "explanations" may or may not align with true underlying cause-and-effect relationships, depending on the nature of the training data, for instance, what variables were measured and included in the model.

Even though causal knowledge is not required to develop a prediction model, it is possible, and often very helpful, for model developers to incorporate causal knowledge into a prediction model. This might include using domain knowledge to determine what information should be included in the model (i.e., what predictors should be measured) or to determine how best to incorporate existing treatment strategies in the model [18, 23]. Prediction models may also be used to directly answer "What if?" questions with respect to a specific intervention (e.g., prediction under intervention models) [21, 24]. Yet, including causal information during model development or using a prediction model to estimate a *single* causal effect does not necessarily improve our ability to interpret corresponding model explanations for *all* included variables as causal. This is because the assumptions necessary to estimate a causal effect are typically met for at most one variable (e.g., treatment), not all variables included in the model.

Illustration

The phrase "correlation does not equal causation" is widely remembered, yet the consequences of this statement are not necessarily intuitive. Using an illustrative example, we highlight how the correlations present in a model training data set may not match the underlying causal effects (e.g., correlation with opposite sign to causal effect), yet still result in a model with adequate predictive performance. In this illustration, we create a simple data-generating mechanism that captures three causal mechanisms (Fig. 1). We generate training and validation data accordingly, develop a black-box prediction model and validate the model's performance. Subsequently, we study the inner-workings of the prediction model using a common post-hoc explainable AI method. The code for this illustration is freely available on GitHub: https://github.com/alexcarriero/illustrative_example.

In our data-generating mechanism (Fig. 1) the Outcome has three direct causes X, B, and Z, represented by arrows terminating at the Outcome. First, consider the path highlighted in blue: variable A causes variable X which in turn causes the Outcome. This might represent a dynamic such as: smoking (A) causing increased hypertension (X) which in turn causes increased risk of kidney disease (Outcome). Next, consider the path highlighted in yellow. In this case, both variable B and the Outcome cause variable Y. This might represent a dynamic such as: age (B) and kidney disease (Outcome) both causing increased risk of hospitalization (Y). Finally, consider the pathway highlighted in purple: variable Z is a common cause of both variable C and the Outcome. This might represent a dynamic such as: diabetes (Z) causing kidney disease (Outcome) and also causing a patient to be prescribed insulin (C). In summary, the data-generating process includes four causal paths to the Outcome: an intermediate path (blue), collider path (yellow), confounder path (purple), and a direct path (grey).

Consider the following hypothetical setting: for a group of consecutive patients admitted to a hospital, we wish to develop a model to predict a diagnosis of kidney disease. We are able to measure the candidate predictors (input features): patient smoking status (A), hypertension (X), age (B), and if the patient has been prescribed insulin (C). Given we are studying hospitalized patients (a subgroup

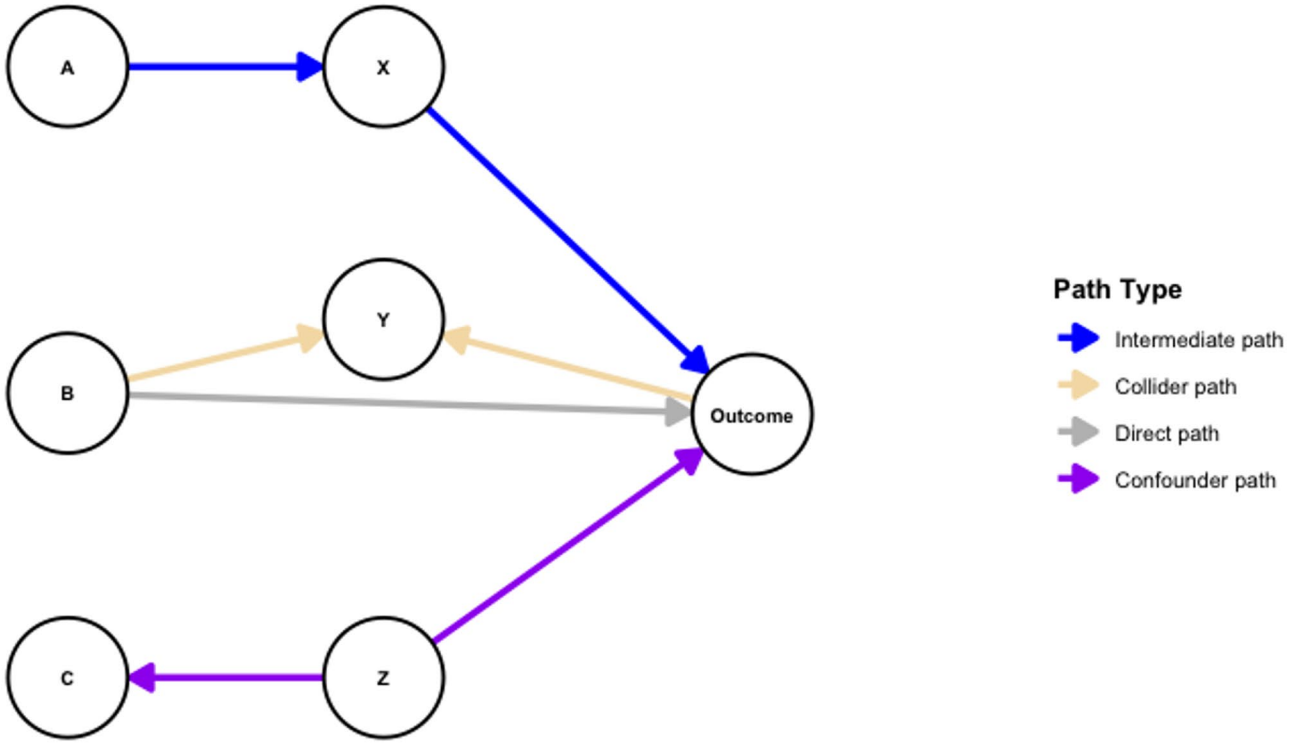

Fig. 1 Data-generating mechanism. All arrows indicate a positive causal effect

of our original population), information about hospitalization (Y) is now implicitly included in the prediction model we develop (i.e., since we have restricted our analysis to patients in-hospital).

A prediction model for the diagnosis of kidney disease with the candidate predictors was trained using a large, simulated data set of 20,000 observations. We used XGBoost to develop the prediction model [25] and assessed model performance using a large (n = 100,000) independent validation set generated using the same data-generating mechanism; this model had an AUROC of 0.80 and was well-calibrated. Shapley value explanations were subsequently generated for all individuals in the model training data [9] (Fig. 2). We used interventional-TreeSHAP [26] to compute the SHAP values and the reference population was a random sample of 5000 individuals from the training data. For a brief introduction to SHAP see TextBox 2. To see the complete model training, validation, and post-hoc explanation code and results please see our codebook.

In Fig. 2, we present the output of the chosen explainable AI method (a SHAP bee-swarm plot). This plot provides insight into the model's overall functionality. Since we know the true causal data-generating mechanism for our data-set, we can compare these insights with our expectations based on the true data-generating mechanism.

Intermediate in the causal chain

Based on the data-generating mechanism, we know that smoking status (A) has a causal relationship with the occurrence of kidney disease (Outcome). Yet, smoking status is identified as the predictor with the least effect on the predictions. Hypertension (X) contributes much more to the model's predictions than smoking (A), which is the original cause and hypertension is the intermediate variable (see Fig. 1). Of course, it is misleading to conclude that smoking is not a cause of kidney disease based on the model explanation. The correct interpretation here is that, once information about hypertension is present, the information about smoking is less or even no longer helpful in the prediction model, since the information contained by hypertension and smoking is largely interchangeable for making the prediction of kidney disease.

Collider

From Fig. 2, we observe another surprising result: age (B) has a negative relation with the risk of kidney disease (i.e., higher age means lower predicted risk of kidney disease). In our causal structure (Fig. 1), we specified that this relation is positive (i.e., higher age means higher risk of kidney disease). This illustrates a phenomenon called "collider bias". In this case, by selectively including individuals in the training data based on hospitalization (the collider variable) we introduced a spurious

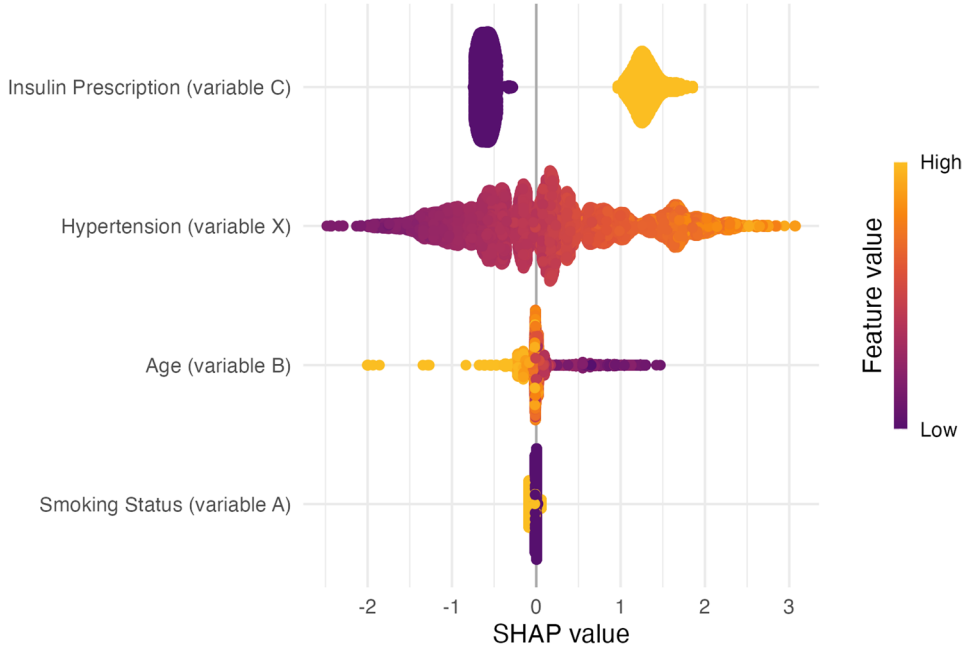

Fig. 2 SHAP bee-swarm plot for our illustrative example. Each dot represents a SHAP value, there is one dot for each feature's input value for every individual in the model training set. On the horizontal axis, SHAP values that deviate from 0 indicate positive or negative associations with the model's predictions. On the vertical axis, the features are ranked from top to bottom based on their importance to the model's predictions. Purple dots correspond to low feature values and yellow dots to high feature values. For example, individuals who were prescribed insulin (yellow dots in the insulin prescription row) have positive SHAP values, indicating that they had higher predicted risks than those who were not prescribed insulin

correlation between age and kidney disease. This correlation is spurious in the sense that it is not seen in the general population (where the opposite is actually true). Yet, this spurious correlation is completely representative of the model development population (e.g., younger people who find themselves in the hospital might have a higher incidence of kidney disease than older people who may be in the hospital for many other reasons). Thus, the negative relation is an accurate description of the training data, but not of the causal mechanism of age and kidney disease. Therefore, while the correlation for this variable likely does not generalize outside of hospitalized patients, it could be worth including in the model under the premise that the model will be deployed in similar hospital settings. In other words, correlations that show a direction contraindicative to the expected direction are not necessarily a clue that the prediction model is not functioning well, they may be the result of collider bias.

Confounder

Finally, consider the predictor insulin prescription. We know from the data-generating mechanism that insulin prescription (C) has no causal relationship with kidney disease (Outcome), yet, it has the strongest association with the outcome among the candidate predictors. The correlation learned by the model is the result of the (unmeasured) confounder diabetes, which was not considered in this example as a candidate predictor. In other words, a strong relationship between insulin and the Outcome is introduced as a result of not recording diabetes diagnosis in the model development population and subsequently, not including it as a predictor in the prediction model. If we had measured and included information regarding a patient's diabetes diagnosis, the prescription of insulin would likely cease to be an important predictor of kidney disease. This illustrates that using model explanations as the basis for actionable medical decisions can also be dangerous (e.g., ceasing to prescribe insulin would certainly not prevent people from getting kidney disease and recommending that patients stop taking insulin would likely worsen their health).

Through the above illustration, we hope to have compelled the reader that the patterns apparent in the training data of prediction models are largely affected by which variables were measured and how those variables relate to each other and the outcome in the underlying causal pathways. Prediction models learn to predict an outcome based on patterns in the development population and in turn, model explanations aim to recover the patterns. While the model explanations in this example do not *explain* the data-generating process, they do *describe* the patterns apparent in the development data.

Are all prediction model explanations "just" descriptions?

Within the domain of explainable AI there are two common strategies. The first is to create a prediction model

that is simple enough for a human to understand e.g., prediction based on a simple decision tree or sum score (e.g., simple rules for predicting ICU congestion [27]). Alternatively, if prediction models are more complex (i.e., the modeled relations between the input features and predictions are not clear), then one may apply post-hoc methodologies (e.g., SHAP and LIME [9, 10]) to elucidate information about how the model makes predictions. Note however that the distinction between "simple" and "complex" models is subjective (see TextBox).

The post-hoc explainable AI methods typically prioritize human understanding over a faithful (completely correct) description of how the model computes predictions. Since humans are not typically adept at comprehending multi-dimensional relationships, information about important characteristics of the complex model (e.g., strong interaction effects, non-linearities) can be lost in favor of an explanation which is more easily understood [7, 28–30]. This is disconcerting as the complex patterns that warrant the used of a complex model over a simpler model may remain completely hidden. Furthermore, limitations of the post-hoc methods can result in substantial added uncertainty about the degree of agreement between the information extracted via the post-hoc methods and the true inner-workings of a more complex prediction model [7, 30, 31].

We used a post-hoc explainability method in our illustration, but wish to emphasize that our conclusions hold more generally, even for models that are fully transparent by design. As long as a model relies on learned statistical associations from a data set to make predictions, then information about how the model works, in general, will describe the patterns learned by the model from the training data set. Regardless if the patterns are directly apparent, as they are with simple models, or if they are elucidated by post-hoc techniques (albeit with some added uncertainty), the results do not offer support of causal effects unless accompanied by the necessary assumptions [20]. For regression-based models, this realization is sometimes called the Table 2 Fallacy [32], i.e., it is not advisable to interpret al.l model coefficients in a multivariable regression model as causal effect estimates as typically the assumptions necessary to estimate a causal effect are met for at most one variable (e.g., treatment), not all variables included in the model.

Discussion

In conclusion, "explanation" (within Shmueli's framework) implies causal interpretations whereas "explainable AI" does not guarantee causal interpretations; explainable AI methods do not distill causal information out of a correlation-based prediction model. Descriptions summarize patterns apparent in a data set and such patterns are largely affected by which variables were measured and how those variables relate to each other and the outcome in terms of cause-and-effect relationships. Misinterpreting model explanations as causal explanations can be misleading and lead to misinformed decision-making. Therefore, for stakeholders with causal interests (e.g., model users seeking an explanation for a prediction that is congruent with true underlying biological mechanisms and cause-and-effect relations) descriptive information is not suitable.

In our illustration, it was easy to see that the relationships learned by the prediction model, and subsequently presented to us via the model explanations did not match the data-generating (causal) mechanism. Given our understanding of the data-generating mechanism, we could provide a rationale for why certain correlations were learned (or not learned) by the model, and could examine which correlations matched the causal effects and which did not. In practice, when there may not be 6 but rather, hundreds of predictor variables available, the pursuit of understanding which relationships we can expect to align with the data-generating mechanism (and which we cannot) becomes exponentially more difficult. This is further complicated by an incomplete understanding of the (causal) data-generating mechanisms in many clinical applications. While many recent methods aim to provide causal explanations [33–36], these methods necessarily assume knowledge of a causal graph (a diagram representing the causal relationships between all relevant variables and the outcome) which is rarely available for multi-variable clinical prediction models in practice [37].

In this paper, we sought to highlight that good prediction models can have surprising correlations (opposite from the causal direction) and consequently, model explanations can be poor resources for judging face-validity of a model or for giving actionable advice to patients. However, there are many other reasons to opt for simple transparent models and/or to generate model explanations (e.g., ease of implementation, transparency in general, hypothesis generation, exploratory analyses). Depending on the specific context of a prediction model, we encourage discussion about whether or not explainable AI is necessary, how explainability should be achieved (simple and interpretable models vs. post-hoc methods) and ethical dimensions about the use of machine learning in healthcare.

In summary, explainable AI techniques have the aim of lessening the risks associated with black-box modeling by providing more or fully transparent AI-based models, yet, when explainable AI is misused or misinterpreted new risks are introduced. Hence, a clear understanding of the boundaries of these methods is necessary for the safe deployment of prediction models in healthcare.

Text Box 1: False Dichotomy

The division of models into "glass box" (or "inherently interpretable") and "black box" models represents a false dichotomy. The line between "glass box" and "black box" models is necessarily subjective. It cannot be drawn on the basis of which algorithm is used to develop the prediction model, as even decision trees and logistic regression models can easily also become too complicated to understand (e.g., multi-feature interactions, hundreds of features), while for instance neural networks may be constrained such that they can be easily understood. Rather, the distinction is dependent on the interpreter, and is subjective as it may vary among stakeholders. While there are certainly models for which no human will have intuition, and likewise, very simple decision trees which all stakeholders may understand, between the extremes there exist models which may be seen as interpretable to some stakeholders while not to others.

Text Box 2: Introduction to SHAP

Shapley value explanations are model explanations that are generated on an individual basis. The explanations are presented as a set of feature contributions i.e., each input value (e.g., patient characteristic) given to the model is given a SHAP value indicating how important it was in the computation of an individual's prediction. Shapley value explanations in our example may be interpreted as an explanation for why a prediction differs from the global average prediction (where the global average prediction is estimated using a reference population e.g., a random sample from the training data set). The SHAP values for each individual have a nice additive property: the difference between an individual's prediction and the global average prediction is distributed perfectly across the feature contributions (SHAP values sum to the difference between an individual's prediction and the global average prediction). When Shapley value explanations are generated for many individuals, they can be visualized together to show information about the model's overall functionality. A beeswarm plot (Figure 2), presents the SHAP values for many individuals organized by each feature in the model.

Acknowledgements

The authors kindly acknowledge Florian van Leeuwen and Lotta Meijerink for helpful feedback and discussions regarding our illustrative example.

Authors' contributions

Concept and Design: AC, MvS, AdH, BC, Drafting Manuscript: AC, MvS, Editing/Critical revision of manuscript for intellectual content: all authors.

Funding

Declaration:
The research for this contribution was made possible by the Artificial Intelligence Programme of the EWUU alliance (https://ai.ewuu.nl/).

Data availability

The code and results for our illustrative example are freely available on GitHub: https://github.com/alexcarriero/illustrative_example.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 12 August 2025 / Accepted: 23 October 2025
Published online: 05 December 2025

References

1. van Smeden M, Reitsma JB, Riley RD, Collins GS, Moons KG. Clinical prediction models: diagnosis versus prognosis. J Clin Epidemiol. 2021;132:142–5.
2. SCORE2 working group and ESC Cardiovascular risk collaboration. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. Eur Heart J. 2021;42:2439–54.
3. Alowais SA, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ. 2023;23:689.
4. Kundu S. AI in medicine must be explainable. Nat Med. 2021;27:1328–1328.
5. Ali S, et al. The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review. Comput Biol Med. 2023;166:107555.
6. Allgaier J, Mulansky L, Draelos RL, Pryss R. How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare. Artif Intell Med. 2023;143:102616.
7. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1:206–15.
8. Rudin C, Radin J. Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition. Harvard Data Sci Rev. 2019;1(2). https://doi.org/10.1162/99608f92.5a8a3a3d.
9. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Red Hook: Curran Associates Inc.; 2017. p. 4768–77.
10. Ribeiro MT, Singh S, Guestrin C. Why should I trust you? Explaining the predictions of any classifier. (2016) https://doi.org/10.48550/arXiv.1602.04938
11. Selvaraju RR, et al. Grad-CAM: visual explanations from deep networks via Gradient-Based localization. Int J Comput Vision. 2020;128:336–59.
12. Imrie F, Davis R, Van Der Schaar M. Multiple stakeholders drive diverse interpretability requirements for machine learning in healthcare. Nat Mach Intell. 2023;5:824–9.
13. Carloni G, Berti A, Colantonio S. The role of causality in explainable artificial intelligence. (2023) https://doi.org/10.48550/arXiv.2309.09901
14. Shmueli G. To Explain or to Predict?. Stat Sci. 2010;25(3):289–310.
15. Hernán MA, Hsu J, Healy B. A second chance to get causal inference right: A classification of data science tasks. CHANCE. 2019;32:42–9.
16. Shmueli G. To Explain, to Predict, or to describe: figuring out the study goal [Commentary on on the uses and abuses of regression models by Carlin and Moreno-Betancur]. Stat Med. 2025;44:e10307.
17. Hurson AN, et al. Prospective evaluation of a breast-cancer risk model integrating classical risk factors and polygenic risk in 15 cohorts from six countries. Int J Epidemiol. 2021;50:1897–911.
18. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: What, why, and how? *BMJ* 338, b375 (2009).
19. Adams M, et al. Computer vs human: deep learning versus perceptual training for the detection of neck of femur fractures. J Med Imaging Radiat Oncol. 2019;63:27–32.
20. Hernan MA. Causal inference: what if. In: Robins JM, editor. Boca Raton: Taylor & Francis; 2024.
21. van Amsterdam WAC, de Jong PA, Verhoeff JJC, Leiner T, Ranganath R. From algorithms to action: improving patient care requires causality. BMC Med Inf Decis Mak. 2024;24:111.
22. Lipton ZC. The mythos of model interpretability. Commun ACM. 2018;61(10):36–43. https://doi.org/10.1145/3233231.
23. van Geloven N et al. The risks of risk assessment: causal blind spots when using prediction models for treatment decisions. (2024) https://doi.org/10.48550/arXiv.2402.17366
24. Keogh RH, Van Geloven N. Prediction under interventions: evaluation of counterfactual performance using longitudinal observational data. Epidemiology. 2024;35:329.
25. Chen T, Guestrin C, XGBoost:. A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794Association for Computing Machinery, New York, NY, USA, (2016). https://doi.org/10.1145/2939672.2939785

26. Komisarczyk K, Kozminski P, Maksymiuk S, Biecek P. *Treeshap: Compute SHAP Values for Your Tree-Based Models Using the 'TreeSHAP' Algorithm*. (2024).
27. Bravo F, Rudin C, Shaposhnik Y, Yuan Y. Simple Rules for Predicting Congestion Risk in Queueing Systems: Application to ICUs. (2019) https://doi.org/10.2139/ssrn.3384148
28. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3:e745–50.
29. Lipton ZC. The Mythos of Model Interpretability. (2017).
30. Molnar C et al. General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models. in *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers* (eds. Holzinger, A. Springer International Publishing, Cham, 39–68 (2022). https://doi.org/10.1007/978-3-031-04083-2_4
31. Aas K, Jullum M, Løland A. Explaining individual predictions when features are dependent: more accurate approximations to Shapley values. Artif Intell. 2021;298:103502.
32. Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. Am J Epidemiol. 2013;177:292–8.
33. Molnar C, et al. Relating the Partial Dependence Plot and Permutation Feature Importance to the Data Generating Process. In: Longo L, editor. Explainable Artificial Intelligence. xAI 2023. Communications in Computer and Information Science, vol 1901. Cham: Springer; 2023. p. 456–79. https://doi.org/10.1007/978-3-031-44064-9_24.
34. Loftus JR, Bynum LEJ, Hansen S. Causal dependence plots. https://doi.org/10.48550/arXiv.2303.04209
35. Frye C, Rowat C, Feige I. Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. (2021) https://doi.org/10.48550/arXiv.1910.06358
36. Wang Z, Samsten I, Papapetrou P. In: Tucker A, Abreu H, Cardoso P, Pereira Rodrigues J, P., Riaño D, editors. Counterfactual explanations for survival prediction of cardiovascular ICU patients. Cham: Springer International Publishing; 2021. pp. 338–48. https://doi.org/10.1007/978-3-030-77211-6_38.
37. Chen H, Covert IC, Lundberg SM, Lee. S.-I. Algorithms to estimate Shapley value feature attributions. Nat Mach Intell. 2023;5:590–601.

Publisher's Note