# Trade-offs in the design of explainable prediction models for health care

**Aniek F. Markus[1], Jan A. Kors[1], Katia M.C. Verhamme[1], Peter R. Rijnbeek[1]**
**[1]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands**

## Background

Artificial intelligence (AI) has the potential to improve patient care by personalizing treatments and can help address challenges in growing expenditures, but implementation of prediction models in clinical practice is still limited. Lack of transparency is – at least in the current state of AI maturity – often seen as one of the main problems. In recent years, eXplainable Artificial Intelligence (XAI) has gained a lot of attention in the machine learning (ML) community. However, many explainable AI techniques have not been applied and tested at scale on real-world data. This work explores different types of explanations to overcome the transparency problem of AI in health care.

## Methods

We formulated five research objectives and designed various studies to achieve those. Methodological details of each study can be found in the respective papers.

I) Review current literature and improve formalization of the field of explainable AI[1].
II) Develop intrinsically interpretable patient-level prediction (PLP) models[2-4].
III) Apply various explainable modelling and post-hoc explanation methods on real-world health care data[1-4].
IV) Evaluate the limitations of different types of explanations for prediction models[1,2,5,6].
V) Provide insight in trade-offs to guide development of explainable AI in the context of health care[3,7].

We summarize the main takeaways in this abstract.

## Results

We define an AI system (i.e. prediction model) to be explainable if the task model is intrinsically interpretable or if the non-interpretable task model is complemented with an interpretable and faithful explanation[7]. Models can be explained using model-based (e.g. task or surrogate model), attribution-based (e.g. feature importance), and example-based explanations (e.g. counterfactual explanation). These are the main findings:

I) We find that evidence of the usefulness of explainability is still lacking in practice and recognize that complementary measures might be needed to create trustworthy AI (e.g. reporting data quality, performing extensive (external) validation, and regulation)[7].
II) Using the PLP framework, we can develop models with a limited number of covariates and good predictive performance for various prediction tasks. Different techniques can be used such as phenotype generation with clinical expertise[8], feature selection[4], or rule-based methods[3].
III) We applied different types of explainable AI techniques to real-world data. However, computation times might be a hurdle in practice as several methods are not scalable to the high dimensionality of health care data[1,6].

IV) We showed predictions model are unstable both in terms of the variables included in the model and in the sign of their coefficients. Hence, it is important to be careful to identify 'risk factors' and not to over-interpret the developed models in general[5]. Similarly, different feature importance methods result in different generated explanations[1,2]. Also for counterfactual explanations we show these often do not consistently depict real-world relations[6].

V) There is some trade-off between model performance and interpretability (as expected), but it varies across prediction tasks and seems to be stronger for high levels of model complexity[3]. We conclude that explainable modelling might be preferred over post-hoc explanations when using explainable AI to create trustworthy AI for health care, as post-hoc explanations might not be faithful (i.e. accurately describe model behavior)[7].

**Conclusion**

Although explanations can be useful to assist implementation in practice by allowing for a human in the loop to detect and correct problems (e.g. existing biases), there are several risks that should be considered. First, there are often multiple explanations possible. Second, the presented explanations can be overinterpreted in various ways (e.g. as causal relations). Third, requiring (certain types of) explanations might come at the cost of predictive performance. Finally, explanations can have unintended (adverse) effects (e.g. decreasing human-machine performance). Hence, it is important to remember explanations are not sufficient by itself and not the ultimate goal.

**References**


1. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. J Biomed Inform. 2021;113:103655.

2. Williams RD, Markus AF, Yang C, Duarte-Salles T, DuVall SL, Falconer T, et al. Seek COVER: using a disease proxy to rapidly develop and validate a personalized risk calculator for COVID-19 outcomes in an international network. BMC Med Res Methodol. 2022;22(1):35.

3. Markus AF, Kors JA, Fridgeirsson EA, Verhamme KM, Rijnbeek PR, editors. EXPLORE: learning interpretable rules for patient-level prediction. Observational Health Data Sciences and Informatics Europe Symposium; 2023; [Conference abstract].

4. Markus AF, Fridgeirsson EA, Williams RD, editors. Creating parsimonious patient-level prediction models using feature selection. Observational Health Data Sciences and Informatics Global Symposium; 2023; [Conference abstract].

5. Markus AF, Fridgeirsson EA, Kors JA, Verhamme K, Rijnbeek PR. Challenges of Estimating Global Feature Importance in Real-World Health Care Data. Caring is Sharing–Exploiting the Value in Data for Health and Innovation: IOS Press; 2023. p. 1057-61.

6. Markus AF, Fridgeirsson EA, Kors JA, Verhamme KM, Reps JM, Rijnbeek PR, editors. Understanding the Size of the Feature Importance Disagreement Problem in Real-World Data. ICML 3rd Workshop on Interpretable Machine Learning in aHealthcare (IMLH); 2023.

7. Markus AF, Rijnbeek PR, Reps JM, editors. Why predicting risk can't identify 'risk factors': empirical assessment of model stability in machine learning across observational health databases. Machine Learning for Healthcare Conference; 2022: PMLR.

8. Höllig J, Markus AF, de Slegte J, Bagave P, editors. Semantic Meaningfulness: Evaluating Counterfactual Approaches for Real-World Plausibility and Feasibility 2023; Cham: Springer Nature Switzerland.