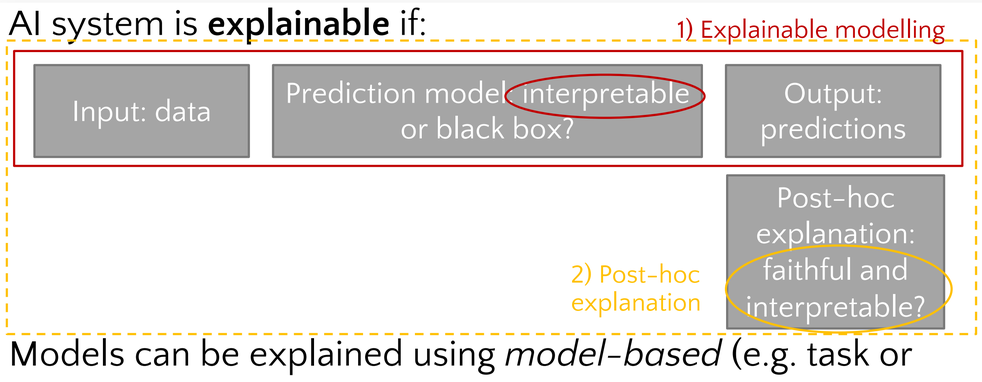
“The human body is a black box”, do we need Explainable AI (XAI)?

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

Background: Artificial intelligence (AI) has the potential to improve patient care, 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. This work explores different types of explanations to overcome the transparency problem of AI in health care.

AI system is **explainable** if:



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).

Explanations often asked for when:

1) cost of misclassification is high or,

2) performance of model is not sufficiently proven in practice.

XAI Benefits	XAI Costs
To assist in verifying (or improving) other model desiderata	Potentially lower model (or human-machine task) performance
To manage social interaction	Time to design and use explanations
To discover new insights	

Empirical research using OMOP data

Using the PLP framework, we applied different types of XAI techniques to real-world data, which led to the following findings:

- **Interpretable models** with a limited number of covariates and good predictive performance **can be developed** for various prediction tasks (e.g. using clinical expertise, feature selection or rule-based methods).
- Model are **unstable** both in terms of the variables included in the model and in the sign of their coefficients. Similarly, different feature importance methods result in different generated explanations.
- There is some **trade-off** between model performance and interpretability, but it **varies** across prediction tasks and seems to be **stronger** for high levels of model complexity.

Identified risks of XAI

1. Often **multiple explanations** possible (e.g. model unstability, feature importance disagreement).
2. Explanations can be **overinterpreted** (e.g. as causal relation, to identify risk factors).
3. Requiring (certain types of) explanations might come **at cost of** predictive performance.
4. Explanations can have **unintended (adverse) effects** (e.g. decreasing human-machine task performance)

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), explanations are not sufficient by itself and not the ultimate goal. It is important to link the need an explanation strives to fulfil with design choice.

Aniek Markus, Jan Kors, Katia Verhamme, Peter Rijnbeek

Erasmus MC University Medical Center Rotterdam

Health Data Science