9TH STATISTICS & BIOPHARMACY CONFERENCE
19 | 21 SEPT 2022

Key notions in health technology assessment of artificial intelligence-based medical devices: what healthcare stakeholders need to know

Line Farah*[1,2], Juliette Murris*[3,4,5], Isabelle Borget[1,6,7], Nicolas Martelli[1,8], Sandrine Katsahian[3,4,9,10]

[1] Groupe de Recherche et d'accueil en Droit et Economie de la Santé (GRADES) Department, Université Paris Paris-Saclay, Châtenay-Malabry, France; [2] Centre d'innovation des Dispositifs Médicaux (CiDM), Délégation à la Recherche Clinique et à l'Innovation, Hôpital Foch, Suresnes, France; [3] Inserm, Centre de Recherche des Cordeliers, Sorbonne Université, Université de Paris Cité, Paris, France; [4] Inria, HeKA, PariSantéCampus, Paris, France; [5] RWE & Data, Pierre Fabre, Boulogne-Billancourt, France; [6] Department of Biostatistics and Epidemiology, Gustave Roussy, University Paris-Saclay, 94805 Villejuif, France; [7] Oncostat U1018, Inserm, Université Paris-Saclay, Équipe Labellisée Ligue Contre le Cancer, Villejuif, France; [8] Hôpital Européen Georges Pompidou, Service Pharmacie, Paris, France; [9] Inserm, Centre d'Investigation Clinique 1418 (CIC1418) Épidémiologie Clinique, Paris, France; [10] Hôpital Européen Georges Pompidou, Service d'informatique médicale, biostatistiques et santé publique, AP-HP, Paris, France.

CONTEXT & OBJECTIVES

- **Understanding of algorithms** in artificial intelligence (AI) in healthcare has become an **essential criterion** following the new regulation processes for AI, data and medical devices;
- **AI based-medical devices** (AI-based MD) Softwares as a Medical Device when the algorithms are intended to prevent, diagnose, treat, mitigate, or cure diseases;[1]
- To assess these technologies, **specific methodological frameworks** are required by health technology assessment (HTA) agencies;[2]
- The inability to understand such algorithms, even if their performance has been prioritized, raises **serious concerns.**

Study objectives

1. To specify the different concepts which are essential to the development of AI-based MD and to ensure the health technology assessment process by evaluators and HTA agencies
2. To analyze how performance, interpretability, and explainability are key to help actors developing their health technologies models accordingly

STATE OF ART: HEALTH TECHNOLOGY ASSESSMENT OF AI-BASED MD

I To assess AI-based MD, HTA agencies aim to evaluate them with a **standardized method** through **multiple domains** such as safety, clinical effectiveness, costs and economic evaluation, organizational aspects, patients, social and legal aspects. A need for specific criteria was highlighted to assess these solutions

II The European guidelines for trustworthy AI include the notions of **"explicability" and "interpretability" as principle of trustworthy AI** in addition to prevention of harm and fairness. In the case that "explicability" is not well defined or not possible with **'black box' algorithms**, other explicability measures such as traceability, auditability and transparent communication on system capabilities could be needed..

III According to the HAS (Haute Autorité de Santé, French HTA agency), these notions are **essential and need to be defined in the reimbursement dossier** of AI-based MDs which can be submitted by companies.

PREFORMANCE, INTERPRETABILITY, EXPLAINABILITY

Measuring AI-based MD's performance

Performance consists in evaluating the error between predictions and observed data, through goodness of fit (mainly used for explanatory models) and/or of prediction (applied for predictive models). Rigorous performance evaluation lies in the fine use of available data:

Data preprocessing → Training and tuning the algorithm → Testing the predictive model → Computing adequate performance metrics

Evaluating interpretability & explainability in AI health technologies

HTA agencies distinguish explainability (*Why?*) and interpretability (*How?*) during the evaluation process.[5-7] Three levels have been identified across ML designs (Figure 1):

- High level of interpretability, provided by models that are intrinsically interpretable;
- Medium level, characterized by random forests, graphical models, causal inference, and Bayesian networks;
- Low level with the most complex models such as SVM, ensemble methods and (deep) neural networks.

Post-hoc explanations enable to thoroughly check what is happening for medium and low level: model-agnostic, example-based and model-specific methods (Table 1).[8-10]

Figure 1. Performance towards interpretability / explainability

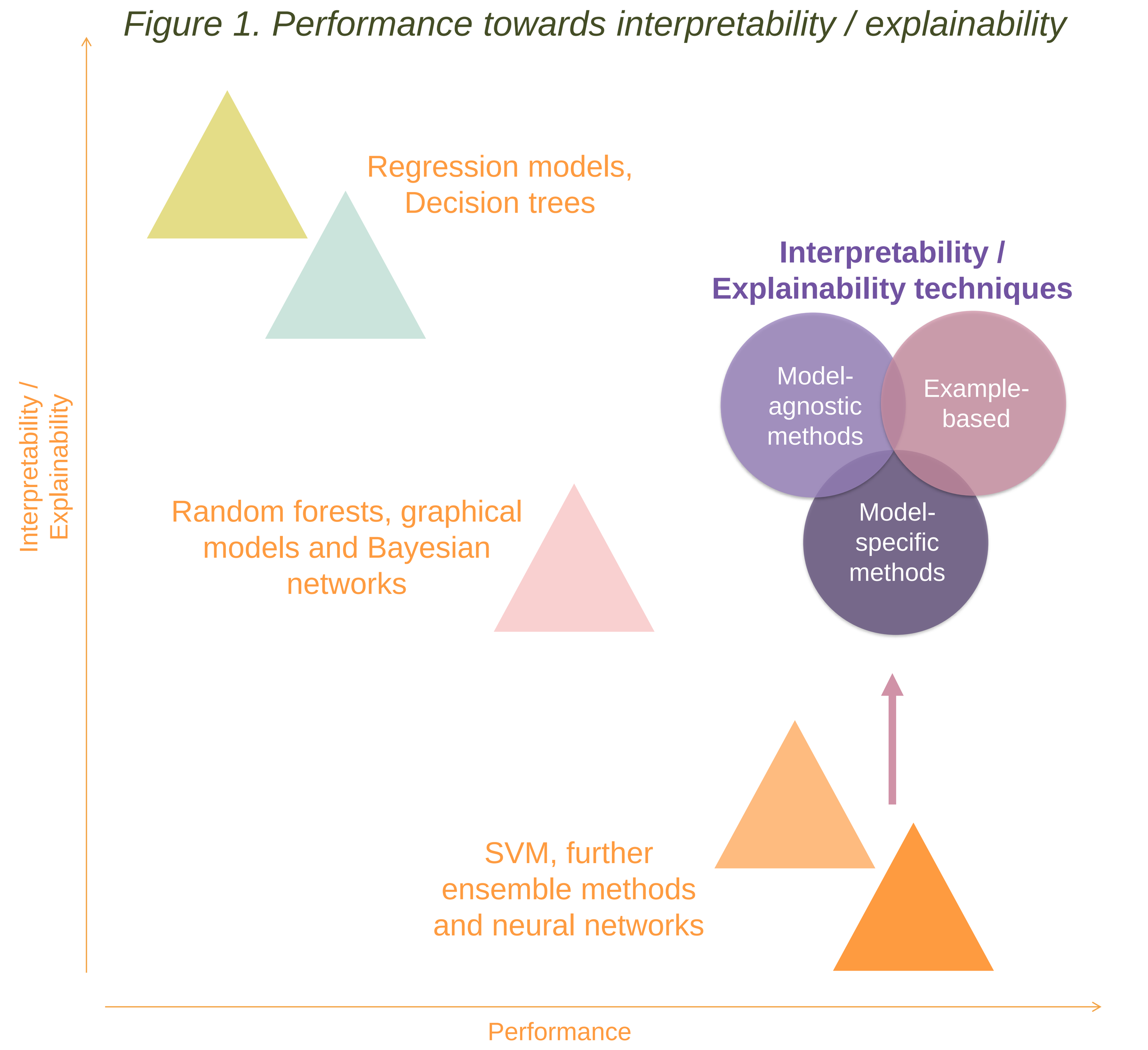

Table 1. Post-hoc explanations serving interpretability and explainability

	Data type	Method type	Main advantage	Some limitations
Feature importance, SHAP, LIME[11-13]	Image, text, tabular	Model-agnostic	Possible application in a post-hoc manner to any kind of algorithm	Feature importance – Sensitive to multicollinearity SHAP – Sensitive to categorical variables and feature interactions LIME – Difficulty to set a distance threshold
Counterfactual explanations[14]	Mainly tabular	Example-based	Easy to understand for the end user	Difficulty for generating feasible and actionable explanations Causal constraints
Gradient-based saliency maps	Mainly image	Model-specific	Easy to understand for the end user	Hardly generalizable

To date, **no consensual approach exists** for the evaluation of interpretability and explainability. Doshi-Velez and Kim[15] have however undertaken rigorous work to answer this need:

- **Involving end users**, and confront the algorithm and reality;
- **Functionally-grounded** evaluations to formalize the algorithm's components as an indicator of the quality of the explanation, favoring **ease of use** and **simplicity**.

DISCUSSION & CONCLUSION

There is a **complex trade-off** between performance and interpretability / explainability

- Predictive performance is a major issue in adopting an AI system;
- There is a need of transparency in medical AI.

Performance, interpretability and explainability are key requirements for a **trustful AI.** Decline in trust in AI may be due to:

- Few randomized clinical trials to assess the performance of AI-based MDs
- Lack of transparency within these technologies
- Risk of inequity introduced by AI biases
- Insufficient regulatory clarity

- The level of confidence in an algorithm relies on **transparency** (interpretability and explicability of outputs) and on **ethics** (in trustworthy and regulatory terms).
- To provide the interpretability, methodologies to 'explainable AI' **need to be associated with ethical and legal analysis**.

TAKE HOME MESSAGES

- Importance of explainability and interpretability techniques by regulators rises to hold stakeholders more and more accountable for the decisions made by AI-based MDs
- Acceptable standards for explainability are context-dependent depending on the risks of the clinical scenario
- Raising awareness on these concepts is essential for their widespread adoption

BIBLIOGRAPHY, ACKNOWLEDGEMENT

1 Harvey, H. B. & Gowda, V. How the FDA Regulates AI. Academic Radiology 27, 58–61 (2020)
2 Gerke, S., Babic, B., Evgeniou, T. & Cohen, I. G. The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. *npj Digit. Med.* **3**, 1–4 (2020).
3 Dietterich, T. Overfitting and undercomputing in machine learning. *ACM Comput. Surv.* **27**, 326–327 (1995).
4 Hawkins, D. M. The problem of overfitting. *J Chem Inf Comput Sci* **44**, 1–12 (2004).
5 Guidotti, R. *et al.* A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* **51**, 1–42 (2019).
6 Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* **267**, 1–38 (2019).
7 Markus, A. F., Kors, J. A. & Rijnbeek, P. R. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *J. of Biomedical Informatics* **113**, (2021).
8 Molnar, C. *Interpretable Machine Learning*.
9 Linardatos, P., Papastefanopoulos, V. & Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy (Basel)* **23**, E18 (2020).
10 Leiter, C. *et al.* Towards Explainable Evaluation Metrics for Natural Language Generation. *ArXiv* (2022) doi:10.48550/arXiv.2203.11131.
11 Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).
12 Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. in *Proceedings of the 31st International Conference on Neural Information Processing*
13 Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135–1144 (Association for Computing Machinery, 2016). doi:10.1145/2939672.2939778
14. Wachter, S., Mittelstadt, B. & Russell, C. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *ArXiv* (2017) doi:10.2139/ssrn.3063289.
15 Doshi-Velez, F. & Kim, B. Towards A Rigorous Science of Interpretable Machine Learning. (2017).

JM benefits from grant from the Association Nationale de la Recherche et de la Technologie, with Pierre Fabre, Convention industrielle de formation par la recherche number 2020/1701