European Journal of Nuclear Medicine and Molecular Imaging (2026) 53:2648–2651
https://doi.org/10.1007/s00259-025-07675-4

SHORT COMMUNICATION

# Explainable AI in nuclear medicine

**Sune Holm[1] · Daria Ferrara[2] · Miriam Pepponi[3] · Elisabetta Abenavoli[3] · Armin Frille[4] · Shaul Duke[1] · Stefan Grünert[5] · Marcus Hacker[5] · Bengt Hennig[5] · Swen Hesse[6] · Lukas Hofmann[4] · Thomas B. Lund[1] · Osama Sabri[6] · Peter Sandøe[1,7] · Roberto Sciagrà[3] · Lalith Kumar Shiyam Sundar[2] · Josef Yu[2,5] · Thomas Beyer[2]**

Received: 19 September 2025 / Accepted: 5 November 2025 / Published online: 25 November 2025
© The Author(s) 2025

## Abstract

**Purpose** In this short communication, we consider the need for explainable AI from the perspective of a large multi-disciplinary research project for predicting cachexia in cancer patients.

**Materials and methods** In a series of meetings, comprising expertise from medicine, data science, sociology, and philosophy, project participants discussed the need for explainability.

**Results** We distinguish between contexts in which a black box AI tool undertakes tasks that users can perform or validate themselves and contexts in which this is not the case.

**Conclusion** We conclude that explanations are likely required when a black box AI tool undertakes tasks that users cannot perform or validate themselves. If the user can verify outputs manually, documented reliability and accuracy may suffice, but explainability can still add value when outputs are uncertain or errors occur. More generally, close collaboration among physicians, AI developers, and other stakeholders is crucial to ensure that AI tools are trustworthy and useful in clinical practice.

**Keywords** Explainable AI · Clinical decision-making · Trustworthy AI · Cachexia · Lung cancer · Medical imaging

## Introduction

The use of AI in clinical settings, particularly for medical imaging and diagnosis, has grown significantly in recent years [1]. Still, it is widely claimed that the black box nature of powerful AI models is an obstacle to their wider adoption in medical practice, because users may not trust the outputs of these AI models [2]. Furthermore, a high predictive accuracy of AI model does not warrant that users are willing to deploy a given AI tool [3, 4]. A common response to the challenge of clinical adoption of black box AI tools is to argue that users should be provided with *explanations* [5] that provide insight into "the reasoning behind decisions" [6]. This has led to a burgeoning explainable AI (xAI) literature discussing the needs and methods for making AI explainable [6–10].

According to another line of argument [11], justified clinical use of black box models should focus on "thorough and rigorous validation" in line with the standards of evaluating the safety and reliability of medical drugs and devices is *sufficient* for justified use of black box AI tools in clinical decision-making.

✉ Sune Holm
suneh@ifro.ku.dk

1 Department of Food and Resource Economics, University of Copenhagen, Copenhagen, Denmark

2 Quantitative Imaging and Medical Physics (QIMP) Team, Medical University of Vienna, Vienna, Austria

3 Nuclear Medicine, Azienda Ospedaliero-Universitaria Careggi, Florence, Italy

4 Department of Medicine II, Division of Respiratory Medicine, Leipzig University Medical Center, Leipzig, Germany

5 Division of Nuclear Medicine, Medical University of Vienna, Vienna, Austria

6 Department of Nuclear Medicine, Leipzig University, Leipzig, Germany

7 Department of Veterinary and Animal Sciences, University of Copenhagen, Copenhagen, Denmark

Following Holm [8], we can describe the two approaches to xAI as the *Validation View* and the *Explanation View*. According to the former, validation is *sufficient* for justified use of black box AI tools in clinical decision-making [7, 11]. On the Explanation View, application of xAI is *mandatory* for justified clinical use of black box AI tools [6, 12].

In this short communication, we consider the need for explainability from the perspective of a large multi-disciplinary research project for predicting cachexia in cancer patients comprising medicine, data science, sociology, and philosophy. This study is embedded in recent studies that support a growing interest for positron emission tomography / computed tomography (PET/CT) in cancer-induced cachexia – not only for tumour assessment, but also for capturing systemic metabolic disturbances [13]. We find that reliable performance is sufficient in contexts where the tool performs a task, which the user of the tool can perform manually, whereas explanations are required when the user does not have the expertise or skills to do so (Fig. 1).

## Three AI tools

Subsequent deliberations stem from our insights as active collaborators within the LuCaPET research project that seeks to determine the role of [$^{18}$F]fluorodeoxyglucose (FDG)-PET/CT imaging in predicting cancer-induced cachexia in lung cancer patients [14, 15]. Within this project, we used three different AI-based tools for the analysis of [$^{18}$F]FDG-PET/CT images of lung cancer patients: a tool for automated segmentation of healthy tissues, another tool for automated segmentation of pathological lesions, and a classification tool based on extracted imaging biomarkers. Each tool has a unique purpose, and as a result, their reliance on xAI methods differs.

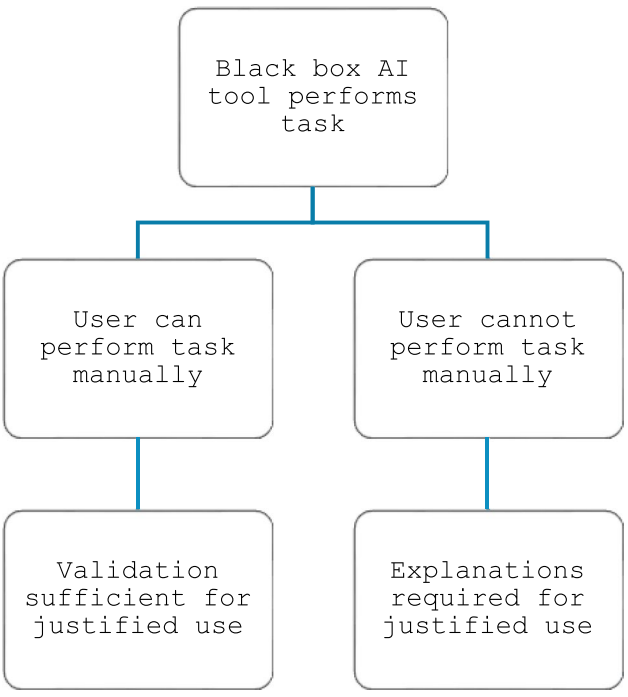


**Fig. 1** An overview of the way in which the ability of users of an AI tool to perform the task for which the AI tool is deployed determines whether explanations of individual outputs is required

### MOOSE: automatic segmentation of healthy tissues from CT images

In our project, we assessed the value of metabolic information derived from [$^{18}$F]FDG-PET/CT images of lung cancer patients to identify potential metabolic abnormalities related to the onset of cancer-associated cachexia. The first step in our analysis involved identifying the organs primarily involved in maintaining body homeostasis. The anatomical information in combined PET/CT imaging is provided by the CT component. Segmenting these anatomical regions is a tedious and labour-intensive task often performed manually and is subject to personal interpretation. In our research, we performed image segmentation with our in-house developed software MOOSE [16]. MOOSE is a deep learning-based software designed to automatically segment healthy organs from CT images. It utilizes the nnU-Net framework [17] to perform this task, significantly reducing the time and manual effort required for anatomical segmentation.

We find that using MOOSE to automate a task, which can be undertaken manually, does not require the application of xAI methods to ensure the tool's clinical acceptance. The primary reason MOOSE does not demand xAI methods is that it is always possible for a qualified human to look at the CT images and ascertain whether the system has made a correct segmentation. In this scenario we notice that for clinical acceptance, users mandate overall accuracy and reliability of the system rather than detailed explainability.

### LION: automatic lesion segmentation in PET images

The second tool, LION [18], also uses the nnU-Net framework [17] but is focused on the automatic segmentation of lesions in PET images. LION is trained with [$^{68}$Ga]PSMA and [$^{18}$F]FDG PET images. However, LION faces greater challenges than MOOSE due to the inherent variability of lesions in terms of shape, size, and location/distribution, as well as the need to differentiate lesions from anatomical regions with naturally high tracer uptake.

Lesion segmentation is a critical component of oncological imaging, directly impacting treatment planning and prognosis. However, as with MOOSE, LION's task is primarily to automate a segmentation task that can be done

manually by qualified humans. Like MOOSE, we also find that the requirement for LION for clinical applicability is reliability and accuracy and not detailed explainability. Nonetheless, users have voiced that it would be valuable to have a risk assessment tool that highlight difficult cases or regions of lower confidence (e.g., inflamed tissues rather than tumours) in need of human validation.

Although such segmentation tasks can, in principle, be manually performed and verified, explainability may still provide added value when the model's performance is imperfect. Understanding why the AI system missegments or shows uncertainty can help clinicians interpret outputs, identify systematic errors, and improve the model's reliability. In this way, even for tasks that are manually verifiable, xAI can contribute to trust and refinement of the tool.

### Cachexia classification tool

The third component of our project is a binary classification model designed to analyse volumetric and metabolic imaging data—specifically, Standardized Uptake Values (SUVs) and volumes from segmented PET/CT regions—in a retrospective cohort of lung cancer patients. This model aims to identify organ-level metabolic signatures associated with cachexia, a severe wasting syndrome marked by the loss of muscle and fat, which correlates with poor prognosis in cancer patients [14].

The model integrates imaging-derived parameters (mean SUV normalized to body weight and aorta uptake, as well as CT-derived tissue volumes) from segmented regions without visible malignant lesions. These data are obtained from non-cachectic patients, cachectic patients, and individuals in early stages of cachexia development. Further, the model incorporates demographic and available blood-based parameters to classify patients into one of three cachexia progression stages. Currently, the model remains under development and does not yet meet the threshold for clinical utility, reflecting limitations inherent to our retrospective dataset [15]. Alongside model refinement, we are actively evaluating the need for explainability in a fully developed cachexia classification tool, and these considerations are discussed in this section.

In contrast to MOOSE and LION, the cachexia classification tool's predictions – once meeting the threshold for clinical utility—would be used for direct clinical decision-making, aiming to support diagnosis and treatment planning. Moreover, the prediction of cancer-induced cachexia is based on the assessment of metabolic patterns in multiple organs and founded on similar tracer distribution patterns in several hundred patients with and without proven cachexia (so called "training data"), which – together – represents a data volume that cannot be comprehended and analyzed easily by a single user. Hence, the only way the user can validate the system's predictions is if an explanation of how the model arrived at its prediction is offered. Hence, it seems highly relevant to apply xAI methods to the third tool.

Of the many xAI methods available, we have chosen Shapley Additive Explanations (SHAP) analysis. SHAP identifies the contribution of each input feature to a prediction, offering a quantifiable view of model reasoning. However, its interpretability and clinical relevance depend heavily on the quality and consistency of the underlying data. In our work [14], we found that heterogeneity in imaging acquisition protocols from three European study sites, patient information reporting, and dietary records limited the ability to identify precise metabolic patterns associated with early cachexia development. Better-curated and standardized datasets, including longitudinal clinical parameters, nutritional assessments, and harmonized imaging features, would allow SHAP and similar xAI methods to generate explanations that are both more robust and clinically meaningful.

An important question is whether SHAP alone is sufficient to provide clinical users with explanations they can understand and trust. This highlights a broader point regarding xAI methods: if the purpose of explanations is to foster user trust and support deployment, then the explanations must be tailored to the decision context and accepted as relevant by the field of expertise. Individual user preferences alone are insufficient. Without careful assessment of what users require from an explanation, using xAI to promote trust may be ineffective. Therefore, determining whether a use context calls for explainability is only one step; a second crucial step involves assessing how the tool will integrate into clinical workflows and which explanation methods will best communicate meaningful information to users. Without these steps, even reliable and accurate AI tools may fail to translate into patient benefit.

## Conclusion

Based on real-life use cases, explanations are likely required when a black box AI tool undertakes tasks that users cannot perform or validate themselves. If the user can verify outputs manually, documented reliability and accuracy may suffice, but explainability can still add value when outputs are uncertain or errors occur. More generally, close collaboration among physicians, AI developers, and other stakeholders is crucial to ensure that AI tools are trustworthy and useful in clinical practice (Fig. 1).

**Author contributions** SH wrote the first draft of the manuscript and all authors provided input for the manuscript and/or commented on previous versions of the manuscript. All authors read and approved the final manuscript.

**Funding** This work received funding under Project number ERA-PERMED2021-324.

**Data availability** No data were collected and used for this article.

## Declarations

**Competing interests** LSKS and TB are co-founders of Zenta GmbH without conflicts with this manuscript. No other authors have relevant financial or non-financial interests to disclose.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

## References

1. Maleki Varnosfaderani S, Forouzanfar M. The role of AI in hospitals and clinics: transforming healthcare in the 21st century. Bioengineering. 2024;11:337. https://doi.org/10.3390/bioengineering11040337.
2. Ribeiro MT, Singh S, Guestrin, C. "Why should i trust you?" Explaining the predictions of any classifier. In Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 1135–1144. https://doi.org/10.1145/2939672.2939778
3. Bedoya AD, et al. Minimal impact of implemented early warning score and best practice alert for patient deterioration. Crit Care Med. 2019. https://doi.org/10.1097/CCM.0000000000003439.
4. Guidi JL, et al. Clinician perception of the effectiveness of an automated early warning and response system for sepsis in an academic medical center. Ann Am Thorac Soc. 2015. https://doi.org/10.1513/AnnalsATS.201503-129OC.
5. Tonekaboni S, et al. What clinicians want: contextualizing explainable machine learning for clinical end use. Proc Mach Learn Res 2019;106:359–380. https://proceedings.mlr.press/v106/tonekaboni19a.html
6. Amann J, et al. To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. PLoS Digit Health. 2022;1:e0000016. https://doi.org/10.1371/journal.pdig.0000016.
7. Babic B, et al. Beware explanations from AI in health care. Science. 2021;373:284–6.
8. Holm S. On the justified use of AI decision support in evidence-based medicine: validity, explainability, and responsibility. Camb Q Healthc Ethics. 2023;32:1–7. https://doi.org/10.1017/S0963180123000294.
9. Holzinger A et al. Explainable AI methods—a brief overview. In *xxAI—Beyond Explainable AI*(eds. Holzinger, A. et al.) Lect Notes Comput Sci 13200, (Springer, Cham, 2022). https://doi.org/10.1007/978-3-031-04083-2_2
10. Lipton ZC. The mythos of model interpretability. Commun ACM. 2018;61:36–43. https://doi.org/10.1145/3233231.
11. Ghassemi M, Oakden-Rayner L, Beam A. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3:e745–50.
12. van der Velden BHM, et al. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal. 2022;79:102470. https://doi.org/10.1016/j.media.2022.102470.
13. Frille A, Ferrara D, the LuCaPET Consortium. [18F]FDG-PET/CT imaging in assessing cancer-induced cachexia. Curr Opinion Clin Nutr Metab Care. 2025;28(5):373–8. https://doi.org/10.1097/MCO.0000000000001149.
14. Ferrara D, et al. Detection of cancer-associated cachexia in lung cancer patients using whole-body [18F]FDG-PET/CT imaging: a multi-centre study. J Cachexia Sarcopenia Muscle. 2024. https://doi.org/10.1002/jcsm.13571.
15. Frille A, et al. "Metabolic fingerprints" of cachexia in lung cancer patients. Eur J Nucl Med Mol Imaging. 2024;51:2067–9.
16. Sundar LKS, et al. Fully automated, semantic segmentation of whole-body 18F-FDG PET/CT images based on data-centric artificial intelligence. J Nucl Med. 2022;63:1941–8.
17. Isensee F, et al. Nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18:203–11. https://doi.org/10.1038/s41592-020-01008-z.
18. Pires M et al. Fully automated FDG and PSMA lesion segmentation in PET imaging via deep learning. Eur J Nucl Med Mol Imaging 2024;51. Abstract EANM'24 Congress Oct 19–23, 2024.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.