AI and Ethics (2026) 6:53
https://doi.org/10.1007/s43681-025-00837-y

RESEARCH

Hyper-selective explainability: an empirical case study of the utility of explainability in a clinical decision support system

Shaul A. Duke[1] · Peter Sandøe[1] · Thomas Bøker Lund[1] · Elisabetta Maria Abenavoli[2] · Thomas Beyer[3] · Daria Ferrara[3] · Armin Frille[4] · Stefan Gruenert[3] · Osama Sabri[5] · Roberto Sciagrà[2] · Miriam Pepponi[2] · Hesse Swen[5] · Anke Tönjes[5] · Hubert Wirtz[5] · Josef Yu[3] · Lalith Kumar Shiyam Sundar[6] · Sune Holm[1]

Received: 20 September 2025 / Accepted: 12 November 2025 / Published online: 10 December 2025

Abstract
Explainability is a leading solution offered to address the challenge of AI's black boxing. However, a lot can go wrong when trying to apply explainability, and its success is far from certain. Moreover, there is insufficient empirical data regarding the effectiveness of concrete explainability efforts. We examined an explainability scenario for an AI decision support tool under development for the early detection of cancer-related cachexia, a potentially fatal metabolic syndrome. We conducted 13 interviews with clinicians who deal with cachexia, and asked about their prior experience with AI tools, their views on explainability, and presented an explainability scenario based on the Shapley Additive Explanations (SHAP) method. Most clinicians we interviewed had limited prior experience with AI tools, and a majority of them believed that the explainability of such an AI system for the early detection of cachexia is essential. When presented with the SHAP explainability scheme, they had limited familiarity with the features that contributed to the tool's ruling, and only a minority of the clinicians (nuclear medicine experts) stated that they could utilize these features in a meaningful manner. Paradoxically, it is the clinicians who come in contact with patients who cannot make use of this specific SHAP explanation. This study highlights the challenges of offering a hyper-selective explainability tool in clinical settings. It also shows the challenge of developing explainable-by-design AI systems.

Keywords Artificial intelligence · Explainability · xAI · Decision support · Clinical AI · Ethics · Cachexia

1 Introduction

The black box nature of artificial intelligence (AI) systems presents several ethical and practical challenges associated with this technology. Challenges include a lack of transparency that prevents identifying either biases or recurring errors in the model, and thus the ability to correct them [1–3]. In clinical settings, black boxing prevents doctors from understanding the reasoning behind the AI decision [4–7], from communicating this reasoning to patients [1, 3, 8–10], and from reaching an appropriate level of trustworthiness [1], which will avoid both over-reliance (automation bias; [11]) and under-reliance (algorithmic aversion; [12]).

One of the major tools devised to combat AI black boxing is post hoc explainability. This is a technique that allows "users to understand how the technology arrives at its predictions or recommendations" [3] by adding additional algorithms on top of the original model [5, 8, 13]. One of the advantages of this technique is that it can be applied to deep learning (DL) models, which, on the one hand, have proven to be quite successful in identifying patterns and making predictions, but on the other hand are also especially opaque, because they contain hidden layers that no one (including developers) can access [5, 6, 14–17].

However, explainability is a contested concept, as scholars do not seem to agree about how much black boxing is a problem [1, 9], if explainability is a suitable solution [5, 9, 13, 18–20], when it is required [20–23], what makes a good explanation [10, 24], and how effective explainability schemes are when applied in real life [3, 7, 8, 18, 21, 25, 26]. Scholars have thus repeatedly called for more empirical

Extended author information available on the last page of the article

research to examine the effectiveness of explainability [3, 9, 25].

Several scholars have also suggested that explanations are selective, in the sense that what constitutes a good explanation for one individual may not be a good explanation for another [13, 15]. They suggest that only certain individuals (notably certain experts) are in a good epistemic position—that is, they possess the correct domain knowledge and know-how—to understand the explanation and, for instance, identify mistakes [13]. Explainability should thus be tailored to the specific domain knowledge of recipients, e.g., clinicians [16, 27].

This article seeks to contribute to the debate about explainability in healthcare by focusing on the degree to which explainability can be domain-specific. Does domain specificity align with medical disciplines? Do clinicians who are less compatible with the explainability scheme only partially understand it, or are they even becoming more disoriented? Do these clinicians desire explainability? What are the implications of selective explainability schemes? This paper will argue that, in some instances, explainability schemes are not merely selective but hyper-selective, meaning that even experts from related or adjacent fields cannot make use of them. Beyond identifying this phenomenon, we will chart its implications for explainable AI (xAI) tools.

2 Black boxing and explainability

AI models tend to be black-boxed due to intentional opacity—for instance, for proprietary reasons—or because of their complexity and architecture [28, 29]. Deep learning techniques for training models have been found to be especially potent yet inherently black-boxed [3, 5, 14, 15]. The opacity of many AI models has been recognized by a large number of scholars as posing ethical, legal, and practical challenges. Black boxing has been associated, among other things, with the inability to identify bias and recurring errors [1–3, 30, 31], as eroding the concept of liability [3, 15], and with the inability to preserve autonomy and informed self-advocacy [6, 13, 32, 33].

In healthcare, and given that people's health and well-being are on the line, AI black boxing has been especially scrutinized. Among other things, it has been associated with preventing clinicians from understanding the reasoning behind the AI decision [4–7] and, therefore, from communicating that reasoning to patients or family members [1, 3, 8–10]. Moreover, black boxing in healthcare settings has been associated with issues of trustworthiness [1, 8, 10, 34]. In order for clinicians to use an AI tool, they need to trust it enough to use it instead of ignoring it [12], but not so much that they use it blindly and fall into overreliance [11]. All in all, black boxing is associated with detrimental effects on multiple stakeholders within healthcare, including clinicians, patients, family members, and healthcare providers.

While several remedies have been devised in order to address the AI black-boxing problem, explainability has been a central suggestion and a topic of heated debate in the field. Plainly put, the aim of explainability of an AI tool is that "an observer can understand the cause of a decision" [35]. Within AI, there are methods to train the model which are inherently explainable (or interpretable), such as decision trees [10, 13–15, 25, 30, 36], and others which are not, such as deep learning, and which require the production of post-hoc explainability [3, 5, 8, 9, 15, 18, 31]. Post-hoc explainability is a technique that allows "users to understand how the technology arrives at its predictions or recommendations" [3] by offering an interpretable, simplified model that approximates the original model [5, 8, 13]. Specifically, local post-hoc explainability provides an explanation for each single decision/prediction of the AI [3, 8, 15, 31].

Within healthcare, the most high-stakes AI systems that are being used and rolled out are clinical decision-support systems (CDSSs), which are intended to help doctors reach a specific decision regarding a patient [3, 8, 9, 16, 37]. This insistence in healthcare on making such tools decision-support tools rather than autonomous tools, which is also regulated by agencies such as the FDA [38], implies that the AI is there to help the decision rather than dictate it. As Amann and colleagues state, "CDSSs cannot currently be considered to support clinical decisions without additional explainability" [8]. Explainability can thus offer these clinicians an indication of why this system ruled the way it did. Armed with this knowledge, clinicians can, in turn, validate each specific ruling, consider additional variables to inform their decision, and use it to communicate the rationale of the decision to patients.

Two of the most popular techniques for explaining AI black-box decisions to clinicians are heatmaps and feature lists. Heatmaps (or saliency maps or saliency masks) work for image-based decisions, with the explainability indicating, using a color overlay, the areas that weighed most and least on the decision of the AI [15, 18, 21, 39, 40]. In Feature explainability, the clinician receives a list of features and an indication of the degree to which each of them weighed on the decision, including counter-indications [6, 15, 16, 18, 31]. There are several practical tools to produce this feature-explainability, and one of them is the Shapley Additive Explanations (SHAP) scheme. As Srinivasu and colleagues explain, "SHAP is a standardized method for determining the relative significance of features in different models" [16:5]. As such, it produces textual output and can offer explainability independent of any image.

3 Explainability as a contested concept

While explainability represents a possible solution to the AI black-boxing problem, it is also a highly contested concept, with scholars often challenging its necessity, suitability, and efficacy. First, scholars do not seem to agree on the degree to which black boxing is a problem [1, 6, 9]. London, for instance, uses the example of drugs and the fact that clinicians prescribe drugs without knowing their mechanism of action to drive the point that black boxing is widespread and acceptable within healthcare [6]. Similarly, Durán and colleagues argue that, since clinicians operate non-AI technologies that they cannot explain, the same can apply to AI systems [1]. Indeed, such examples amount to an argument that a double standard has been created within healthcare for AI technologies. On the one hand, there is much black-boxing that is considered legitimate in medical practice prior to AI, while on the other hand, AI clinical systems are not allowed to be black-boxed [9].

The next type of challenge is whether explainability is a suitable solution to black-boxing. For instance, Babic and colleagues argue that explainability may foster error in perception among clinicians: "post hoc rationalizations of a black box are unlikely to contribute to our understanding of its inner workings. Instead, we are likely left with the false impression that we understand it better" [5:285]. Ghassemi and colleagues, on their part, argue that "using post-hoc explanations to assess the quality of model decisions adds an additional source of error—not only can the model be right or wrong, but so can the explanation" [18:474]. While going over the entire range of arguments of why explainability may be an unsuitable solution for AI black boxing is beyond the scope of this article, we will state that there are a variety of arguments made by scholars, beyond the two examples given here [9, 13, 19, 20].

Most importantly, scholars have raised questions about the efficacy of explainability schemes when applied in real-world settings [3, 7, 8, 18, 21, 25, 26]. Does the explainability actually help the clinician? For instance, Cutillo and colleagues argue that "Explainable outputs are not necessarily actionable or informative outputs" [21:4]. One reason for this skepticism seems to be that many scholars agree that explainability is context-bound [3, 8, 9, 30, 40–42]. This makes the question of efficacy an empirical one: under which settings are explainability schemes effective and under which they are not? A central problem in answering this question is the limited empirical research on explainability schemes in concrete contexts, which explains the repeated calls in the field for more such empirical examinations [3, 9, 25, 43].

Another implication of the highly contextual nature of explainability is that what constitutes a good explanation for one individual may not do so for another [13, 15]. For instance, Cortese and colleagues state "it is known that explanations are dependent on the receiving user; for example, the explanation for a medical diagnostic varies if the target is a patient, a doctor, or a hospital auditing authority" [35: 3]. Combi and colleagues go a step further in depicting this dependence on possessing the correct domain knowledge and know-how to understand a given explanation and argue for selectivity within clinicians: "Specific domains as cardiology, oncology, neurology, healthcare policy, and so on, have their own vocabulary, specific shared knowledge about diagnosis, treatments, and so on […] XAI systems have, thus, to deal with jargon, abbreviations and terminological heterogeneity, idiosyncratic usage habits, and different kinds of knowledge" [27:7]. Explainability should thus be tailored to the specific domain knowledge of recipients, if it to be effective [16, 27].

However, such tailoring has another ramification. It may cause an epistemic shift within clinical practice. Clinicians and healthcare scholars have for a few years now recognized that new forms of AI may "undermine the epistemic authority of clinicians" [44: 205; see also 1]. Yet less attention is given to the epistemic shift from one type of expert to another [33]. Decisions seem to be shifting from treating clinicians to clinicians who work with images, since many of these AI systems use images as input. As such, scholars of concrete cases should try to examine whether such an epistemic shift is likely.

4 Methods and the examined tool

This paper attempts to address this need to examine explainability in the context of a concrete case by analyzing a system under development. It is based on a qualitative study of an AI clinical decision support system currently under development for the early detection of lung cancer-related cachexia [45]. Cachexia is a potentially fatal syndrome [46–49] characterized by the loss of muscle and sometimes also fat mass [49, 50]. While currently cancer-induced cachexia can be identified only after its symptoms have appeared, and the wasting of the body is underway, a consortium of Western European partners has set out to develop a deep learning AI model which will detect cachexia related to lung cancer in earlier stages of the disease, based on the processing of [18F]FDG-PET/CT images [46]. Moreover, the team, from its inception, has set out to also examine the explainability aspects of this model, along the lines of explainability-by-design principles [46].

Between 2022 and 2024, we carried out 13 semi-structured interviews with clinicians from three hospital centers (in Austria, Italy, and Germany) who, at the time of

Table 1 Interviewed clinicians according to medical discipline

Category	Number of clinicians
Endocrinologist	1
Oncologist	3
Oncologist + outpatients	1
Nuclear medicine	3
Respiratory expert	1
Radiotherapy expert	1
Palliative care	3
Total	13

the interviews, were involved with lung cancer cachexia (see Table 1). Clinicians were recruited through a convenience sample, with a focus on diversity across medical disciplines at each medical center. The interviews were conducted remotely via video conference, held in English, and followed an interview guide (see Appendix Table 4). The interview had two parts: the second, relevant to this paper, and the first, outside its scope. In this second part, clinicians were asked about their prior experience with AI tools and their views on explainability, and they were presented with an explainability scenario based on the SHAP (Shapley Additive Explanations) method. Interviewees were shown a vignette depicting a SHAP output for an imaginary patient (Patient X) with values of features that supposedly contributed to the AI's decision regarding the onset of early stages of cachexia (see Fig. 1). They were asked to evaluate their ability to read and make use of such an explainability scheme, specifically to monitor for signs of anomalies, dysfunctions, and unexpected performance that might alert to a low-credibility scenario at play, as mandated by the new European Artificial Intelligence Act [9].

Interviews were transcribed and anonymized using both software and human handling. Transcriptions were coded using a hybrid method that combines deductive and inductive processes, in several rounds in an iterative form [51]. A number of authors also discussed the coding between rounds in order to improve analysis and minimize bias. Informed consent was secured from all interviewees, and the research was approved by the University of Copenhagen's research ethics review board (Ethics approval number: 504–0358/22–5000).

5 Results

Our first set of questions to clinicians, relevant to the issue of explainability, pertained to their prior experiences with automated healthcare tools and specifically AI. This line of questioning was aimed at determining their familiarity with similar systems, in order to understand whether their expectations are based on experience and the standards they observed or on their preconceptions about AI. Of the 13 interviewed clinicians, 11 had prior experience with automated clinical tools, while two did not (see Table 2). Most of these clinicians with experience (11/13) have had experience in their research capacity, rather than in daily diagnosis or treatment. For instance, one interviewee stated that "we don't use artificial intelligence for routine, but we use it for

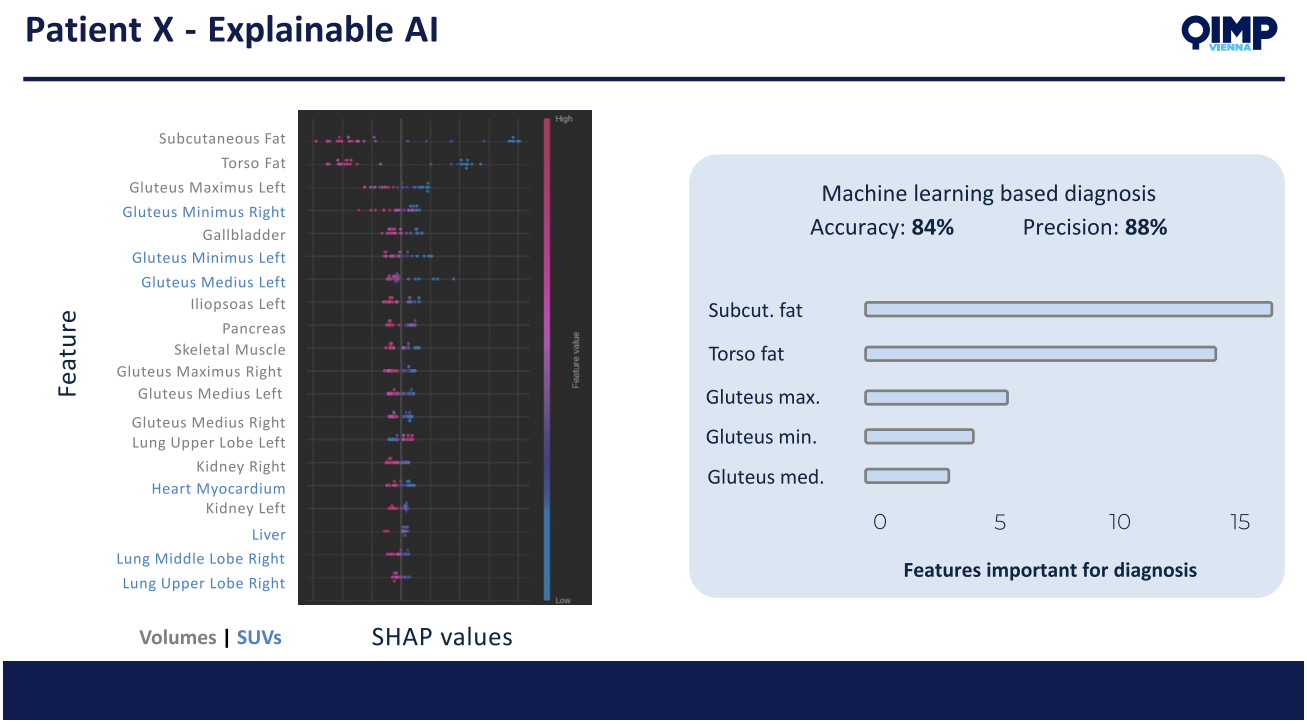

Fig. 1 The SHAP values-based explainability scheme presented in a vignette form to interviewees

Table 2 Questions about prior experience and explainability

Question	Yes	No	Did not answer
Prior experience with automated healthcare tools	11	2	–
Prior experience with AI healthcare tools	7	6	–
Is AI explainability essential?	7	4	2

research study" (2a.02). At the same time, another answered to the same question "in everyday practice? No, because it's just not used for decision-making in oncology. I use it in research" (2b.12). Within this mostly research experience, 6 of the 11 experienced proper clinical AI tools, and not just automated tools (such as simple clinical calculators). For instance, one clinician stated, "I was a member of the Dragon Project. It's a European project on the application of big data and artificial intelligence on COVID patients" (2c.05). Another clinician stated "I only have the experience from automated tools from my studies in dermatology. There is the melanoma detecting AI what I know okay but not from my daily work" (2b.10). Moreover, some of the research experience that our interviewees reported on was from tools that are under development, and not yet operational even in the research capacity. For instance, one clinician described their involvement in training a model: "In this specific work, the AI we learn AI with some images, okay, and we teach, you trained the AI, and we say to the mathematical algorithm that this is a Hodgkin lymphoma, this is a non-Hodgkin lymphoma" (2c.04).

With less than half of the interviewees having experience with clinical AI tools, and most of that experience stemming from their research capacity rather than their treatment or diagnostic capacity, it is safe to say that our cohort had limited experience with such tools. Moreover, having mostly research experience with healthcare AI tools is relevant, as such tools rarely include explainability during the academic development stage. Initially, academic developers focus primarily on training and validating the model. This is also reflected in the interviews, in which several clinicians appeared unfamiliar with the concept of explainability and offered confused answers to questions about whether they think such explainability should be included. For instance, one clinician answered "Maybe if it's good to substitute proteins in quickly or not so quick maybe to make it more specific what part of the nutrition needs to be substituted very quickly" (2b.08). Hence this clinician initially understood the explainability to be suggestions of what to do with the diagnosis offered by the AI, rather than an explanation about the likely contributors to the AI decision. Another interviewee, when asked if explainability is important, answered: "I think at the beginning, how to use the tool and probably" (2c.06). Hence, before being corrected, they perceived explainability as training regarding how to use the tool. A third example also reveals a lack of awareness of what AI explainability is: "it would be an explanation. Why this is necessary. So that the people are kind of alerted. That it's an important issue" (2b.09).

Next interviewees were asked if having an explanation that will accompany the AI ruling regarding early stages of cachexia is 'essential' or 'nice to have'? If they had misconceptions about what explainability is (as some of them did), they were corrected and asked again. Out of our 13 interviewees, seven answered that it is essential, four that it is 'nice to have', and 2 gave non-relevant answers. The following excerpts constitute examples of essential utterances.

> "I would like to have some information about the alert, and not only a simple alert. I would like to have some information about how they get to the decision that this is a case of early cachexia" (2c.05).

> "Because it's also important for me as a physician to check again the information upon the decision or the finding was made, like if I maybe put in the wrong weight and then this is one of the reasons why the decision was made, I would want to like check it again or be sure that I put in the right…" (2b.10).

> "So, whatever I include in my everyday practice, I need to understand it. So, something that is just there, and, I just say, okay, I don't know how this information comes up or what I can do with it. I would not include. So, if there is an AI-based early alert, this needs to be explained to me" (2b.12).

These are examples of nice-to-have:

> "It would be nice to have some kind of explanation there. […] It would be nice to know not only that there was alert, but some kind of explanation there. And that he can see how it fits together with what the patient's history and everything" (2a.14).

> "I think it's nice to have. So, if I know the system or, let's say, the algorithm behind the system, which gives the alert, it's not necessary to further explain. But it's helpful on which exact points the alert was made. But it's for my feeling, it's a very good point. It's only nice to have because it does not really influence procedure following. So, it's just for my interest, but for which intervention I will decide, it is not dependent exactly on the reason" (2a.01).

With regards to this last example, it seems that this clinician opted for 'nice to have' conditioned on being able to "know the system, or let's say the algorithm behind the system", which is impossible in the case of deep learning systems. Thus, while we included this clinician in the 'nice to have' column, this may actually be a case of 'essential'.

Overall, a majority of interviewees commented that explainability is pertinent to their use of the tool. One of the reasons given for this is the need to explain to patients the reason they will now be treated for cachexia. For instance, one clinician stated, "We have to explain to the patient what is our choice. The choice to take the medicine" (2c.07).

Interviewees were also asked what kind of explanation they would like to receive alongside the AI's ruling, and the most common response was a desire to see the features that went into the decision. For instance, one clinician expressed the desire to see "on which component or which is the most important component in your model you used for the alarm generation" (2a.01). Another plainly stated "I need the explanation of the type of feature analyzed" (2c.04).

The latter part of the interview was dedicated to examining the clinicians' understanding of the explainability scheme presented to them (based on SHAP values) and their ability to comprehend it. SHAP values serve as a method for explaining the output of a machine learning model, by displaying a recreation of how much each feature (each input variable) contributed to the model's prediction [16, 26]. As described in the Methods section, we used a vignette that offers an explainability screen for a decision about the early stages of cachexia (see Fig. 1). Interviewees were asked to examine it and were provided with clarifications on its meaning. Specifically, the interviewer explained this is an explainability screen for a particular ruling of the AI algorithm about a patient (patient X), that the list on the left side of the screen were the top five features that weighed into the decision (with the percentage reported), and that the list on the left side contained a much longer list arranged by the same order of weight on the decision.

Clinicians were first asked to describe their familiarity with these features. Overall, seven clinicians stated they were familiar with the features (three of whom were from the nuclear medicine department), five stated they were partially familiar, and one stated they were not familiar with the features (see Table 3). For instance, belonging to the familiar group one clinician from nuclear medicine stated with confidence their familiarity with these features "Yes yeah, the gluteus muscle and other muscles yes, I'm definitely familiar with these regions" (2a.14). Another clinician went a little further and stated they are not only familiar with these features, but all clinicians should be "I think these are, so let's say the terms you used [the features] should be familiar to each clinician" (2a.01). Going to partial familiarity, here is an example of a clinicians who understood all but one feature: "I'm not familiar with that [torso fat] but otherwise the other ones iliopsoas pancreas yeah the all other terminologies are already self-explaining but the torso fat is not something which we're using" (2b.11). At the other pole there was only one clinician which stated from the start that they are not familiar with these features. Answering the question of familiarity of these features, he answered "No, not particularly" (2c.05).

Table 3 Questions about the features appearing in the explainability scheme

Question	Yes	No	Partially
Are you familiar with the features?	7	1	5
Are you in a position to evaluate the validity of these features?	2	10	–

Of course, familiarity is a somewhat abstract term that can mean both a deep understanding of what a feature represents or just knowing the name of a bodily region. Indeed, when digging deeper, it emerges that many of those who stated they are familiar or partially familiar do not understand these features beyond their names. For instance, one clinician in our familiar group said: "I recognize the name, but I don't know how they can be used […] as connected to cachexia, I can recognize only subcutaneous fat" (2c.a3). Another clinician in the same category answered to the question if they are familiar with these features "Yes. Every feature. I know all of them […] But really, in my work, I don't evaluate every muscle" (2c.07). This same clinician was asked in response if it means that some of these features they only know by name, to which they answered affirmative.

The next question clinicians were asked is "Using this explanation, would you feel you are in a position to evaluate the validity of the recommendation?" To this question, 10 clinicians stated they were not in a good position to validate these features, while 2 (both from nuclear medicine) said they were in a good position, and one interviewee did not comment on the issue. For instance, one oncologist answered: "whether this is correct or not is something that I can't evaluate with the knowledge I have" (2b.12), while a respiratory expert answered similarly "I would not be able to validate it because I'm not an expert in these issues" (2b.09). Another clinician answered to the same question: "I can validate only retrospectively, but not, so it's, I'm based at this time point on other parameters I have available" (2a.01). In other word, they are saying they would be able to validate using their own knowledge only some time afterwords, when they would be able to see if cachexia has set or not. Moreover, when asked who is in a good position to evaluate the ruling using the explainability, virtually all clinicians stated that nuclear medicine and maybe radiology were the best departments to do so. Most of them gave a short

answer, such as: "Well, I think it could be the Department of Nuclear Medicine, probably. So, radiologists and physicians were involved in nuclear medicine, I think" (2c.06). Yet we also got some 'thinking out loud' answers, such as: "we need a physician who strictly follows the patient every day or very frequently, and the physician able to assess subcutaneous fat, for example, just measuring the diameter of the abdomen, just to have an idea, and also able to assess pancreas or able to assess gallbladder. So, a physician who had particular experience in imaging" (2c.05).

6 Discussion

Our interviews with clinicians who are involved directly or indirectly with lung cancer cachexia have shown a limited prior experience with AI tools, especially when it comes to their diagnosis and treatment capacities, as opposed to their research capacity. It is thus safe to say that their expectations about explainability come mostly from preconceptions, rather than experience. When asked, a majority among them think having explainability alongside the AI ruling is essential. Surprisingly, studies rarely ask clinicians to what degree they need the explainability alongside the ruling, which makes this finding important. Among other things, they feel they require explainability in order to explain to patients the rationale behind the ruling and the measures it entails. The literature also supports this as a strong motivation to require explainability [10].

The idea that a list of features, along with their weight on the AI's decision, can constitute an appropriate explainability scheme was raised by several interviewees. As such, it aligns with some of the practices carried out in the healthcare field [16]. Once presented with this tool's features, more than half of our cohort stated they were familiar with these features, but upon digging deeper, it appeared that much of this familiarity was very shallow and not something that could really help them utilize the explainability scheme. We associate this result, at least partially, with interviewees feeling they are being 'tested', and wanting to show competence; i.e., with social desirability bias [52]. Ultimately, the fact that ten out of the thirteen interviewees stated they could not use this explainability scheme in order to assess the AI's ruling is a strong indication of their ability to utilize this specific scheme.

This comes as no surprise, since these features constitute measurements of internal organs as recorded by a PET/CT imaging examination, and further processed by an AI segmentation tool called MOOSE. These are organ measurements and imaging methods that most lung cancer clinicians do not typically encounter, whereas nuclear medicine experts do. This clarifies why all three nuclear medicine interviewees stated they are familiar with these features, and why two of them have also said they could validate the AI ruling using this explainability tool. It also explains why the rest of the clinicians suggested that nuclear medicine experts (together with radiologists) would be suitable for validating the AI scheme.

In other words, the relevance of feature-based explainability for a group of experts boils down to the input variables the AI model is fed. Even within adjacent medical disciplines that address the same conditions (in our case, lung cancer and cachexia), explainability may require different domain knowledge, which will bar all but certain clinicians from understanding and utilizing an explainability scheme. This means that developers who wish to incorporate feature-based explainability into a model they are creating should be aware of who is equipped to make use of it based on these particular features.

Paradoxically, in our case, this feature-based explainability scheme does not cohere with doctors belonging to the clinical disciplines that come in contact with patients prone to lung cancer cachexia. This entails that the wish to use this particular explainability in order to communicate effectively with patients will be stifled. It also means that patients' information about the AI ruling will come secondhand, which might erode their understanding of what led to this decision. This is especially true, as nuclear medicine experts admitted to us, because clinicians tend to read only the conclusions of reports sent by nuclear medicine departments. For instance, one nuclear medicine expert stated about how much the reports they send to other clinicians (about specific cases) are actually read: "I think it can be expected that mainly, only the last part of the report, like the conclusion, is read" (2a.14). Thus, even if the nuclear medicine expert would choose to include information taken from the explainability scheme, since clinicians tend to read only the conclusion, they would likely miss this information.

7 Conclusion

The idea that what constitutes a good explanation in an AI explainability scheme depends on the user, and that different types of users (e.g., different types of experts) require different explanations, is not new in itself [13, 15]. However, the study of explainability is lacking in empirical examinations, especially regarding its effectiveness [3, 9, 25], a gap this article aims to address. Using 13 semi-structured interviews with clinicians from different disciplines who are involved with the diagnosis or treatment of lung cancer cachexia, we examine the explainability scheme of one tool under development. This tool, designed to detect the early stages of cachexia, is planned to include a feature-based

explainability scheme. Our analysis shows that only one type of clinician—nuclear medicine experts—are equipped to utilize this explainability scheme, while other types of clinicians who deal with this syndrome are not. This means that understanding an explainability scheme can be highly selective with regard to users' prior professional training.

Before exploring the ramifications of the hyper-selectivity of explainability, we should first address the question of generalization. This study is based on interviews with 13 clinicians about a specific tool under development for a particular syndrome (cancer-related cachexia). As mentioned above, explainability is highly context-dependent and may thus not be hyper-selective across different conditions. Our argument is thus not that all explainability schemes (we only analyzed feature explainability) in all situations are hyper-selective, but that there is such a phenomenon, the breadth and scope of which we are unsure of.

Understanding this hyper-selectivity in terms of being competent to utilize a given AI explainability scheme has several implications. First, it highlights that when designing an explainable-by-design AI tool, one must consider the variables that train the machine learning model and the type of expert capable of understanding and possibly validating them. These experts will become the de facto decision-makers of this tool; otherwise, you may designate a decision-maker who is divorced from the explainability offered. This makes developers' tasks much more complex, since at that stage of the process, they are usually focused only on deciding which variables will help produce a model that makes the best prediction.

The next question should be: Does it make sense that this group of experts would be the decision-makers for this clinical decision? In our case, if this AI tool for identifying early-stage cachexia becomes operational in its current format, nuclear medicine experts will become the decision-makers, who will determine whether a patient is in this early stage. Nowadays, there is no such decision maker since there is no way to detect these early stages of the syndrome, and the decision maker about mid- or late-stage cachexia is lung cancer treating and diagnosing clinicians (oncologists, respiratory experts, etc.). Hence, this scenario will cause a shift in the locus of decision-making and may have significant epistemic consequences, since now much of the knowledge about cachexia diagnosis is shifting from diagnosing/treating clinicians to nuclear medicine experts. This may cause an erosion in these diagnosing/treating clinicians' understanding of the cachexia syndrome.

Second, if a given feature-based explainability can be understood and utilized only by a particular group of experts, one should ask whether, after identifying which group it is, the logic for providing explainability still holds. As discussed above, explainability has known drawbacks [5, 9, 13, 18–20], and in such cases, the "costs" may outweigh the benefits. In our case, once we understand that this group consists of nuclear medicine experts and that they do not come into direct contact with patients, some of the rationale for explainability —helping relay information to patients —is eroded. In such a situation, does offering explainability still make sense? For validation purposes, and since nuclear medicine experts do seem to be in a position to identify errors, there still seems to be logic in having an explainability schemer; however, a discussion needs to be carried out here. We should also ask, what happens to patient autonomy if the clinicians who come in contact with them cannot explain to them why the decision that they are in the early stages of cachexia was made? It is hard to avoid the conclusion that their autonomy is severely eroded.

Lastly, this notion that, in some cases, there is hyper-selectivity in explainability schemes—in which only one particular type of clinician can understand it while the others get confused by it—underscores the fragility of explainability when applied to deep learning models. It joins a long list of known vulnerabilities in explainability, such as the fact that explainability itself constitutes an additional source of errors [6, 18] or that explainability schemes may foster errors in perception among clinicians [5, 26]. This is not to say that an effective xAI is impossible, but rather that it requires the correct settings and considerable attention.

Appendix

See Table 4.

Table 4 Interview guideline

Introduction	
Consent	Thank you for participating
Professional details	We are conducting this interview for our use at the University of Copenhagen in Denmark. We, and the rest of our team, want to use the interview with you to understand healthcare specialists' attitudes toward the use of artificial intelligence in medical decision-making (for educational and research purposes). For this reason, it is necessary for us to audio tape our discussion, and subsequently to transcribe what you tell us on paper. The audio file will only be available to our team, and to the person that transcribes the interviews. The transcription will be completely confidential: any reference to you or to anyone you talk about will be removed when the audio is transcribed. The audio recording and the transcription will be sent to the University of Copenhagen, as the university is responsible for the project and ensures your confidentiality. The interviews will be quality-checked, and then your contact-information will be deleted. For GDPR reasons, we have to be able to prove informed consent from informants. Therefore, we will keep a copy of the audio file in an encrypted server at the University of Copenhagen for 5 years which only I and the project responsible researcher have access to At any time during or after the interview, you can withdraw your participation from this project, and demand that your interview not be used in the project. Of course, once your contact information has been deleted, we cannot withdraw your interview **Do you consent that…** if "no" to any of the bullet points below, stop the interview You understand the procedure we just went through? if no: read again or clarify We record this interview, and that the interview is transcribed That the recording is shared with the person that will transcribe our talk, and with the data responsible researcher at the University of Copenhagen That the University of Copenhagen is responsible for the data, and will use transcripts from the interviews in the current research project and future research and educational projects Could you state your position, and how long you have been practicing medicine?
Contextual questions	
Current experience with cachexia	Can you say what your role is with regards to the diagnosis and treatment of lung cancer? What is your current role with regards to the diagnosis and treatment of cachexia?
Cachexia diagnosis	Do you ever diagnose lung cancer patients with cachexia (when relevant)? Do you know whether your colleagues diagnose lung cancer patients with cachexia? Has the section/department that you work in ever set up guidelines for the diagnosis of cachexia? (yes: what are they?) How do you define cachexia? How do you separate cachexia from malnutrition? Who decides if a patient has cachexia?
Initial thoughts	
Presenting the project Initial reaction Influence on treatment Explainability	As you probably know, an AI-driven early cachexia detection tool is currently being developed, by an international collaboration that includes a team in this hospital. This tool is created by machine learning algorithms. We are interviewing clinicians who may come in direct contact with this tool in the future in order to understand their needs and desires regarding the explainability of the AI. Explainability is an AI feature that tries to clarify why the AI made a certain recommendation. It is needed since most machine learning statistical calculations are opaque ("black box"), even to those who develop them Do you have any initial thoughts about Explainable AI? If you were to receive an alert from such an AI system that tells you that a certain cancer patient has early stages of cachexia, would it be helpful? *If does believe* What steps would you take if you received such an alert? Would it be essential that this alert include an explanation or is it just nice to have? *If does need explanation* What would you need the explanation for? What kind of explanation or explanations would you need? How would you like this information to be presented? What wouldn't you want to be presented with? What do you think makes a good explanation?
Current use of automated tools	
Experience history Known problems	Do you have experience with automated tools as part of your medical work? *If does have experience* Can you describe these systems? Were there problems with any of these systems? Were you sometimes skeptical regarding these systems' results? Could you evaluate the validity of their results? Do you think other clinicians might have difficulties with these systems?
Vignette	

Table 4 (continued)

Introduction	
The list of features	I will now show you a list of variables that are used by the early detection cachexia system. *Show the list*
Familiarity	Are any of them familiar to you?
Direct contact	*Unless not familiar with any of them*
Competency	I want us to go over each of the ones you are familiar with, and for you to tell me if you currently have any direct way to know if this variable is incorrect
Suitable target of explainability	Are there variables you think should not be included in the tool?
	Are there variables that are not in the list you think should be included?
	Assuming that this is the list of variables that will go into the recommendation regarding cachexia; would you feel you are in a position to evaluate the validity of the recommendation?
	Now that you know the list of variables, which department is best suited to validate this data?

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s43681-025-00837-y.

Author contributions S.D. wrote the main manuscript text. T.B. and S.H. contributed to the text. S.D. interviewed clinicians. P.S., T.B., E.M.A, A.F., R.S., and J.Y. facilitated the research. All authors reviewed the manuscript.

Funding The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by ERA PerMed [grant number 2021–324], Innovation Fund Denmark, the Austrian Science Fund [grant number https://doi.org/10.55776/I5902], Bundesministerium für Bildung und Forschung [grant number FKZ 01EO1501], Mitteldeutsche Gesellschaft für Pneumologie [grant number 2018-MDGP-PA-002], European Commission Research Directorate-General [grant number 779282], the Saxon State Ministry for Science, Culture and Tourism, Progetto PETictCAC, Regione Toscana.

Data availability Due to data privacy, we are not allowed to share interview data.

Declarations

Conflict of interests The authors declare no competing interests.

Ethics approval and consent to participate The University of Copenhagen research ethics review board approved our interviews (approval: 504–0358/22–5000) on October 10, 2022. Respondents gave recorded consent before starting interviews.

Consent for publication Not applicable.

References

1. Durán, J.M., Jongsma, K.R.: Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. J. Med. Ethics **47**, 329–335 (2021)
2. Funer, F., et al.: Responsibility and decision-making authority in using clinical decision support systems: an empirical-ethical exploration of German prospective professionals' preferences and concerns. J. Med. Ethics **50**, 6–11 (2024)
3. Hildt, E.: What is the role of explainability in medical artificial intelligence? a case-based approach. Bioeng **12**, 1–25 (2025)
4. Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V.I.: Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med. Inform. Decis. Mak. **20**, 1–9 (2020)
5. Babic, B., Gerke, S., Evgeniou, T., Cohen, I.G.: Beware explanations from AI in health care. Science **373**, 284–286 (2021)
6. London, A.J.: Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Cent. Rep. **49**, 15–21 (2019)
7. Muralitharan, S., et al.: Machine learning–based early warning systems for clinical deterioration: systematic scoping review. J. Med. Internet Res. **23**, 1–22 (2021)
8. Amann, J., et al.: To explain or not to explain?: artificial intelligence explainability in clinical decision support systems. PLOS digit. health **1**, 1–18 (2022)
9. Freyer, N., Groß, D., Lipprandt, M.: The ethical requirement of explainability for AI-DSS in healthcare: a systematic review of reasons. BMC Med. Ethics **25**, 1–11 (2024)
10. Tonekaboni, S., Joshi, S., McCradden, M.M., Goldenberg, A.: What clinicians want: contextualizing explainable machine learning for clinical end use. Paper presented at: machine learning for healthcare conference. (2019) August 9 2019, Ann Arbor, MI
11. Challen, R., et al.: Artificial intelligence, bias and clinical safety. BMJ Qual. Saf. **28**, 231–237 (2019)
12. Dietvorst, B.J., Simmons, J.P., Massey, C.: Overcoming algorithm aversion: people will use imperfect algorithms if they can (even slightly) modify them. Manage. Sci. **64**(3), 1155–1170 (2018). https://doi.org/10.1287/mnsc.2016.2643
13. Vredenburgh, K.: The right to explanation. J Polit Philos **30**, 209–229 (2022)
14. Busuioc, M.: Accountable artificial intelligence: holding algorithms to account. Public Adm. Rev. **81**, 825–836 (2021)
15. Guidotti, R., et al.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**, 1–42 (2019)
16. Srinivasu, P.N., Sandhya, N., Jhaveri, R.H., Raut, R.: From blackbox to explainable AI in healthcare: existing tools and case studies. Mob. Inf. Syst. **2022**, 1–20 (2022)
17. Zech, J.R., et al.: Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. **15**, 1–17 (2018)

18. Ghassemi, M., Oakden-Rayner, L., Beam, A.L.: The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit. Health. **3**, 745–750 (2021)
19. Kroll, J.A., et al.: Accountable algorithms. U. Pa. L. Rev. **165**, 633–705 (2017)
20. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**, 206–215 (2019)
21. Cutillo, C.M., et al.: Machine intelligence in healthcare: perspectives on trustworthiness, explainability, usability, and transparency. npj Digit Med. **3**, 1–5 (2020)
22. Ferrario, A., Loi, M., Viganò, E.: Trust does not need to be human: it is possible to trust medical AI. J. Med. Ethics **47**, 437–438 (2021)
23. Holm, S.H.: On the justified use of AI decision support in evidence-based medicine: validity, explainability, and responsibility. Camb Q Healthc Ethics. 1–7 (2023). https://doi.org/10.1017/S0963180123000294
24. Gerlings, J., Jensen, M.S., Shollo, A.: Explainable AI, but explainable to whom? An exploratory case study of xAI in healthcare. In: Lim, C., Chen, Y., Vaidya, A., Mahorkar, C., Jain, L.C. (eds.) Handbook of artificial intelligence in healthcare: Vol 2: practicalities and prospects volume 2, pp. 169–198. Springer-Verlag, London (2022)
25. Mishra, A.: Transparent AI: reliabilist and proud. J. Med. Ethics **47**, 341–342 (2021)
26. Wysocki, O., et al.: Assessing the communication gap between AI models and healthcare professionals: explainability, utility and trust in AI-driven clinical decision-making. Artif. Intell. **316**, 1–17 (2023)
27. Combi, C., et al.: A manifesto on explainability for artificial intelligence in medicine. Artif. Intell. Med. **133**, 102423 (2022)
28. Barocas, S., Hardt, M., Narayanan, A.: Fairness and machine learning: limitations and opportunities. ABC-CLIO, Santa Barbara, CA (2019)
29. de Laat, P.B.: Algorithmic decision-making based on machine learning from big data: can transparency restore accountability? Philos. Technol. **31**, 525–541 (2018)
30. Izumo, T., Weng, Y.-H.: Coarse ethics: how to ethically assess explainable artificial intelligence. AI Ethics **2**, 449–461 (2022)
31. Vale, D., El-Sharif, A., Ali, M.: Explainable artificial intelligence (XAI) post-hoc explainability methods: risks and limitations in non-discrimination law. AI Ethics **2**, 815–826 (2022)
32. Fritz, J.: On the scope of the right to explanation. AI Ethics. **5**, 2735–2747 (2025)
33. Shin, D.H.: Debiasing AI: rethinking the intersection of innovation and sustainability. Routledge, London (2025)
34. Hatherley, J.J.: Limits of trust in medical AI. J. Med. Ethics **46**, 478–481 (2020)
35. Cortese, J.F., Cozman, F.G., Lucca-Silveira, M.P., Bechara, A.F.: Should explainability be a fifth ethical principle in AI ethics? AI Ethics **3**, 1–12 (2023)
36. Kemper, C.: Kafkaesque AI? Legal decision-making in the era of machine learning. Intell. Prop. L. J. **24**, 251–294 (2019)
37. Blomberg, S.N., et al.: Effect of machine learning on dispatcher recognition of out-of-hospital cardiac arrest during calls to emergency medical services: a randomized clinical trial. JAMA Netw. Open **4**, 1–10 (2021)
38. Duke, S.A.: Deny, dismiss and downplay: developers' attitudes towards risk and their role in risk creation in the field of healthcare-AI. Ethics Inf. Technol. **24**, 1 (2022). https://doi.org/10.1007/s10676-022-09627-0
39. Fernandez-Quilez, A.: Deep learning in radiology: ethics of data and on the value of algorithm transparency, interpretability and explainability. AI Ethics. **3**, 257–265 (2023)
40. Holzinger, A., Müller, H.: Toward human–AI interfaces to support explainability and causability in medical AI. Computer **54**, 78–86 (2021)
41. Alpsancar, S., Buhl, H.M., Matzner, T., Scharlau, I.: Explanation needs and ethical demands: unpacking the instrumental value of XAI. AI Ethics. **5**, 3015–3033 (2025)
42. Herzog, C.: On the risk of confusing interpretability with explicability. AI Ethics. **2**, 219–225 (2022)
43. Borch, C., Hee Min, B.: Toward a sociology of machine learning explainability: human–machine interaction in deep neural network-based automated trading. BD&S **9**, 1–13 (2022)
44. Grote, T., Berens, P.: On the ethics of algorithmic decision-making in healthcare. J. Med. Ethics **46**, 205 (2020)
45. 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
46. Frille, A., et al.: "Metabolic fingerprints" of cachexia in lung cancer patients. Eur. J. Nucl. Med. Mol. Imaging (2024). https://doi.org/10.1007/s00259-024-06689-8
47. Muscaritoli, M., et al.: ESPEN practical guideline: clinical nutrition in cancer. Clin. Nutr. **40**, 2898–2913 (2021)
48. Ni, J., Zhang, L.: Cancer cachexia: definition, staging, and emerging treatments. Cancer Manag. Res. **12**, 5597–5605 (2020). https://doi.org/10.2147/CMAR.S261585
49. Suhag, V., Sunita, B.S., Sarin, A., Singh, A.K.: Cancer, malnutrition and cachexia: we must break the triad. Int. J. Med. Phys. Clin. Eng. Radiat. **4**, 64–70 (2015)
50. Fearon, K., et al.: Definition and classification of cancer cachexia: an international consensus. Lancet Oncol. **12**, 489–495 (2011)
51. Fereday, J., Muir-Cochrane, E.: Demonstrating rigor using thematic analysis: a hybrid approach of inductive and deductive coding and theme development. Int J Qual Methods **5**, 80–92 (2006)
52. Bergen, N., Labonté, R.: "Everything is perfect, and we have no problems": detecting and limiting social desirability bias in qualitative research. Qual. Health Res. **30**, 783–792 (2020)

Authors and Affiliations

Shaul A. Duke[1] · Peter Sandøe[1] · Thomas Bøker Lund[1] · Elisabetta Maria Abenavoli[2] · Thomas Beyer[3] · Daria Ferrara[3] · Armin Frille[4] · Stefan Gruenert[3] · Osama Sabri[5] · Roberto Sciagrà[2] · Miriam Pepponi[2] · Hesse Swen[5] · Anke Tönjes[5] · Hubert Wirtz[5] · Josef Yu[3] · Lalith Kumar Shiyam Sundar[6] · Sune Holm[1]

✉ Shaul A. Duke
saduke@gmail.com

Peter Sandøe
pes@sund.ku.dk

Thomas Bøker Lund
tblu@ifro.ku.dk

Elisabetta Maria Abenavoli
elisabettabenavoli@gmail.com

Thomas Beyer
thomas.beyer@meduniwien.ac.at

Daria Ferrara
daria.ferrara@meduniwien.ac.at

Armin Frille
Armin.Frille@medizin.uni-leipzig.de

Stefan Gruenert
stefan.gruenert@meduniwien.ac.at

Osama Sabri
Osama.Sabri@medizin.uni-leipzig.de

Roberto Sciagrà
roberto.sciagra@unifi.it

Miriam Pepponi
miriam.pepponi@unifi.it

Hesse Swen
Swen.Hesse@medizin.uni-leipzig.de

Anke Tönjes
Anke.Toenjes@medizin.uni-leipzig.de

Hubert Wirtz
Hubert.Wirtz@medizin.uni-leipzig.de

Josef Yu
josef.yu@meduniwien.ac.at

Lalith Kumar Shiyam Sundar
Lalith.shiyam@med.uni-muenchen.de

Sune Holm
suneh@ifro.ku.dk

[1] University of Copenhagen, Copenhagen, Denmark

[2] Azienda Ospedaliero-Universitaria Careggi, Florence, Italy

[3] Medical University of Vienna, Vienna, Austria

[4] Department of Medicine II (Division of Respiratory Medicine), Leipzig University Medical Center, Leipzig, Germany

[5] University Hospital Leipzig, Leipzig, Germany

[6] DIGIT-X lab, Department of radiology, LMU university hospital, Munich, Germany