A survey on *good* AI: User-Centric AI Design in Healthcare

Andrea Berti[1,2*], Valentina Giannini[3,4], Simone Mazzetti[3], Maria Antonietta Pascali[2], Daniele Regge[3], Sara Colantonio[2]

[1*]Department of Information Engineering, University of Pisa, Via Caruso 16, Pisa, 56122, Pisa, Italy.
[2]Institute of Information Science and Technologies (ISTI), National Research Council (CNR), Via Moruzzi 1, Pisa, 56124, Pisa, Italy.
[3]Department of Radiology, Candiolo Cancer Institute, FPO-IRCCS, Strada Provinciale 142, Km 3.95, Candiolo, 10060, Turin, Italy.
[3]Department of surgical science, University of Turin, via Genova 3, Turin, 10126, Turin, Italy.

*Corresponding author(s). E-mail(s): andrea.berti@isti.cnr.it;
Contributing authors: gianninivalentina@gmail.com;
simone.mazzetti@ircc.it; maria.antonietta.pascali@isti.cnr.it;
daniele.regge@ircc.it; sara.colantonio@isti.cnr.it;

Abstract

The integration of Artificial Intelligence (AI) in healthcare has the potential to revolutionize patient care by enhancing diagnostic processes, treatment protocols, and overall healthcare delivery. However, the adoption of AI-powered tools and services is contingent upon establishing a robust foundation of trust among healthcare professionals. The ProCAncer-I project, informed by the FUTURE-AI framework, is at the forefront of this effort, promoting a user-centric design philosophy that prioritizes the needs and expectations of end-users, primarily clinicians and radiologists. This paper delves into the co-design methodology adopted by an interdisciplinary team, elucidating the collaborative efforts that underpin the customization of the FUTURE-AI principles to align with the clinical requirements of the project's partners. The introduction sets the stage for a comprehensive discussion on the significance of stakeholder engagement in the design and implementation of trustworthy AI systems within clinical settings.

Keywords: Artificial Intelligence in Healthcare, Artificial Intelligence, XAI, Trustworthiness, User-centric Design

1 Introduction

The integration of Artificial Intelligence (AI) in healthcare marks a significant transformation, enhancing diagnostic precision, treatment outcomes, and patient care [1–3]. AI tools can leverage large datasets and identify patterns to surpass human performance in several healthcare aspects, offering increased accuracy, reduced costs, and time savings when aiding humans. Trust among clinicians is pivotal for the successful adoption of AI tools [4–6], a principle that the ProCAncer-I European project, guided by the FUTURE-AI framework[1], endorses. The framework recognizes the urgency of user-centric design in fostering this trust and advocates for a multidisciplinary approach to design, involving stakeholders and end-users throughout the development process.

Co-design, a creative partnership that spans the entire design journey [7], is central to this approach. It involves stakeholders and end-users in a dynamic process that includes consultations, workshops, focus groups, and dedicated collaborative tools (e.g., Mirò, cards and games, simulated environments) to spur innovation and ensure inclusivity. The ProCAncer-I project aimed to assess how the FUTURE-AI principles could be tailored to meet the specific expectations and clinical demands of its partners. To this end, an interdisciplinary team crafted a survey to capture the clinical partners' perspectives, desires, and expectations for high-quality, trustworthy AI systems.

This team, consisting of experts in radiology, biomedical engineering, computer science, and mathematics, all with extensive experience in cancer imaging and AI, worked together from June to September 2022. They established a common language, defined the survey's focus, and collaborated on a shared document. The finalized survey, distributed via an online Google form, was distributed among the clinical partners to gather their insights. The findings from this survey and their implications are discussed in the following sections.

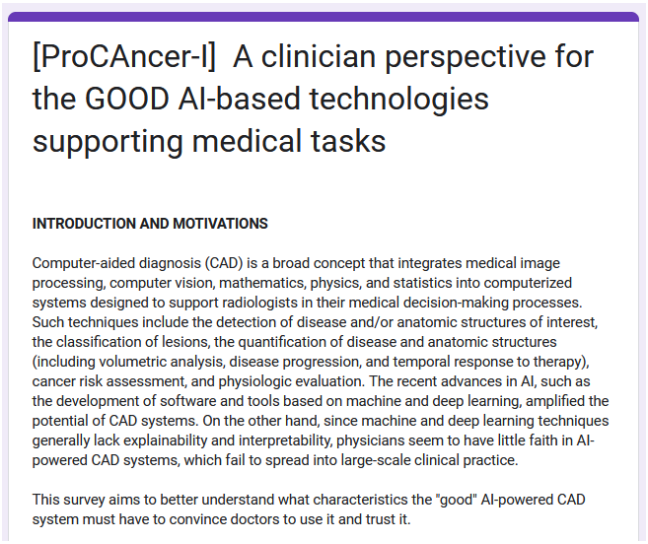
2 Survey content and structure

The survey was meticulously designed to cover a comprehensive range of topics crucial for the clinical application of AI. It aimed to gather both quantitative and qualitative insights on the clinicians' expertise with AI, their views on unreliable AI interventions, preferred methods of interaction with AI systems, performance expectations, and the attributes they consider most important in a trustworthy AI system.

A total of 26 questions were crafted, varying from multiple-choice to open-ended, to capture a wide array of data. The survey began with an introduction explaining its goals (Figure 1) and included a glossary to ensure participants fully understood the terms used (Figure 2). The topics addressed in the survey were:

- clinical expertise and current usage of AI tools (8 questions);
- opinions on unreliable AI-powered interventions (1 question);
- preferred reading modality and interaction with the AI system (3 questions);
- desired balance of sensitivity and specificity for different clinical tasks (4 questions);
- expected success rate for various clinical tasks (2 questions);
- most valued features of trustworthiness (4 questions);

[1] https://future-ai.eu/



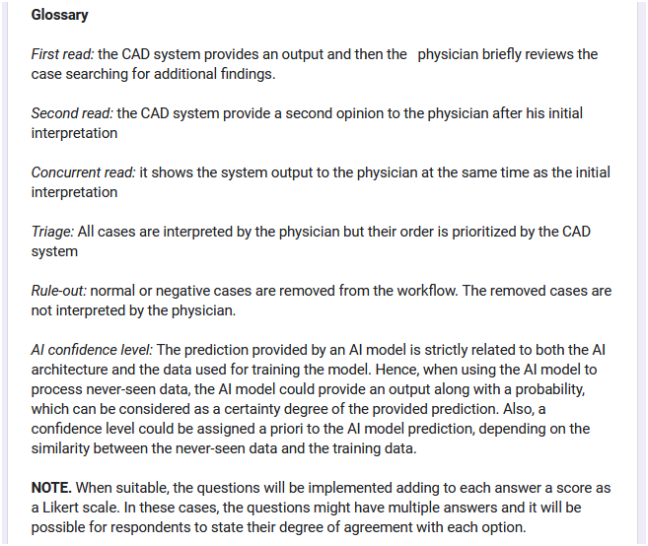
[ProCAncer-I] A clinician perspective for the GOOD AI-based technologies supporting medical tasks

INTRODUCTION AND MOTIVATIONS

Computer-aided diagnosis (CAD) is a broad concept that integrates medical image processing, computer vision, mathematics, physics, and statistics into computerized systems designed to support radiologists in their medical decision-making processes. Such techniques include the detection of disease and/or anatomic structures of interest, the classification of lesions, the quantification of disease and anatomic structures (including volumetric analysis, disease progression, and temporal response to therapy), cancer risk assessment, and physiologic evaluation. The recent advances in AI, such as the development of software and tools based on machine and deep learning, amplified the potential of CAD systems. On the other hand, since machine and deep learning techniques generally lack explainability and interpretability, physicians seem to have little faith in AI-powered CAD systems, which fail to spread into large-scale clinical practice.

This survey aims to better understand what characteristics the "good" AI-powered CAD system must have to convince doctors to use it and trust it.

Fig. 1 Entry page and motivations of ProCAncer-I survey on trustworthy and good AI



Glossary

First read: the CAD system provides an output and then the physician briefly reviews the case searching for additional findings.

Second read: the CAD system provide a second opinion to the physician after his initial interpretation

Concurrent read: it shows the system output to the physician at the same time as the initial interpretation

Triage: All cases are interpreted by the physician but their order is prioritized by the CAD system

Rule-out: normal or negative cases are removed from the workflow. The removed cases are not interpreted by the physician.

AI confidence level: The prediction provided by an AI model is strictly related to both the AI architecture and the data used for training the model. Hence, when using the AI model to process never-seen data, the AI model could provide an output along with a probability, which can be considered as a certainty degree of the provided prediction. Also, a confidence level could be assigned a priori to the AI model prediction, depending on the similarity between the never-seen data and the training data.

NOTE. When suitable, the questions will be implemented adding to each answer a score as a Likert scale. In these cases, the questions might have multiple answers and it will be possible for respondents to state their degree of agreement with each option.

Fig. 2 The glossary included in ProCAncer-I survey

- most valued features of reliability (2 questions);
- preferred elements and format of AI output (2 questions).

The survey was concise to respect the clinicians' limited time, and it concluded with an open-ended question to give the opportunity for additional

feedback. We report in Figures 3 – 6 some of the most relevant questions included in the questionnaire; the complete version of the questionnaire can be found at the following link: https://docs.google.com/forms/d/e/1FAIpQLScvtWfzJcRg0c7Hcu4vqQjdYOLk6ooLTFXq8xj6XXBjNfsFhw/viewform

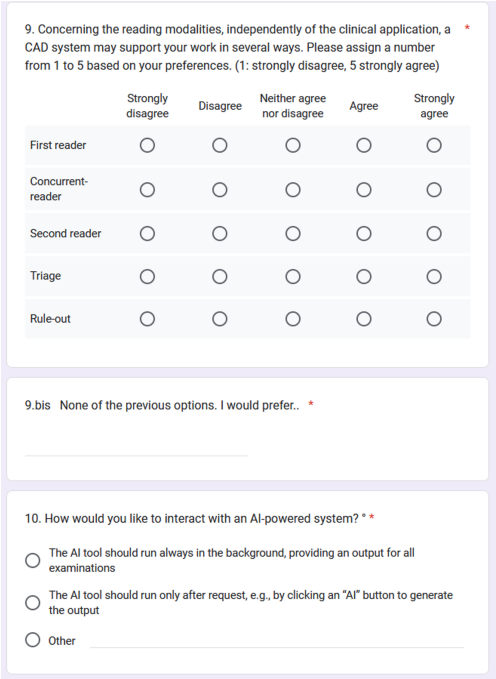


Fig. 3 Questions on preferred reading and interaction modalities

3 Analysis of survey findings

The survey findings provide a rich variety of data, reflecting the diverse perspectives of the participants. We collected a total of 38 responses from October 2022 and March 2023. We carried out the analysis by utilizing Python scientific libraries. Many of the survey questions utilized a Likert scale format, allowing the responses to be analyzed as sentiment scores by assigning an integer value ranging from -2 (complete disagreement) to $+2$ (complete agreement) to the possible answer values. With that notation, a sentiment score of 0 would represent neutrality, a positive score indicates affirmative feedback, and a negative score reflects unfavorable feedback.

Demographic data of the survey participants are concisely presented in Figure 7. Typically, the respondents were European radiologists over the age of 35, predominantly employed in public healthcare settings, with a specialization in abdominal imaging and a basic understanding of AI.

The survey revealed that 56% of participants currently employ AI tools, predominantly in the context of research, education, or for clarification purposes. Conversely,

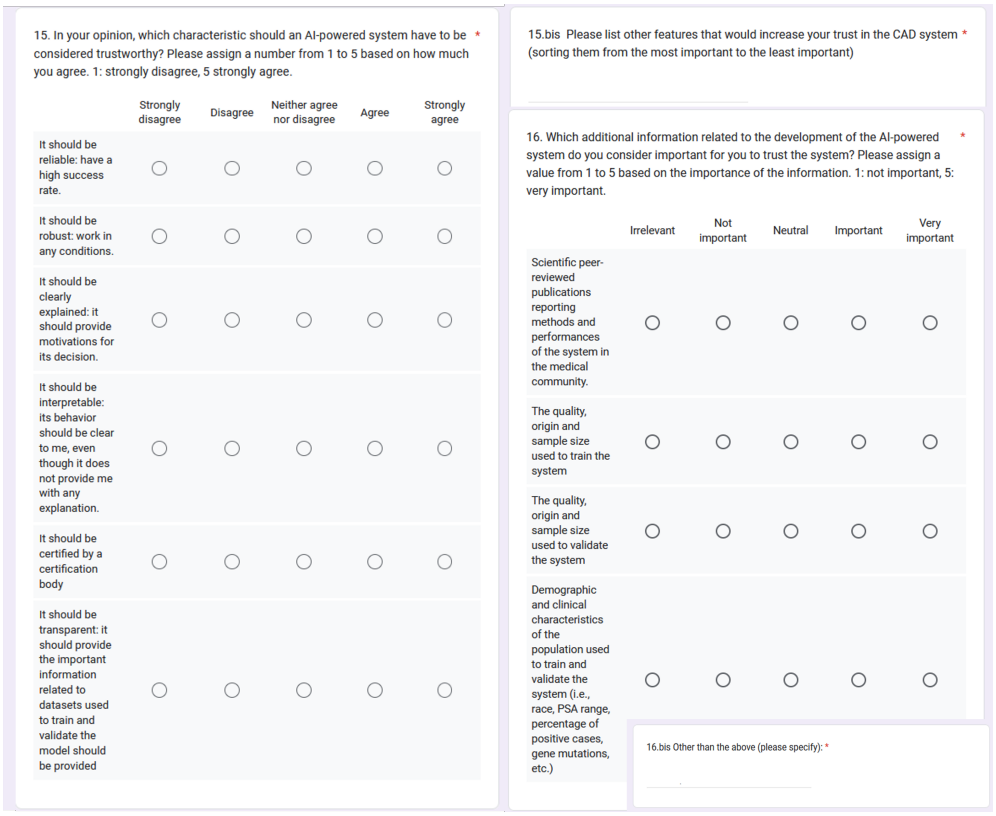


Fig. 4 Questions on most valued trustworthiness features and additional information

41% of respondents indicated they do not utilize AI tools due to lack of access. Only 3%, which corresponds to a single respondent, expressed distrust in AI tools (Figure 8 left). For users who have adopted AI, its primary application has been in the detection of diseases (Figure 8 right).

Regarding the preferred reading modality, the "second reader" option emerged as the most favored, with a sentiment score of 0.68. In contrast, the "rule-out" option was met with disfavor, reflected in a negative sentiment score of -0.22, which aligns with expectations given that such systems entirely bypass radiologist consultation. Notably, all reading modalities received scores under 1, as can be seen in Figure 10. This outcome, coupled with the fact that nearly all participants use, or would like to use, AI, implies a need for clearer, more task-specific options. This is partly supported by the findings from the question about distrust, where no distrust cases were reported by the majority of respondents (Figure 9). Additionally, open-ended feedback highlighted ongoing concerns regarding lesion characterization among clinicians.

The analysis also showed that the preferred integration modality in the examination workflow was for the AI tool to always run in the background, providing an output for all examinations. This can be seen in Figure 10 right.

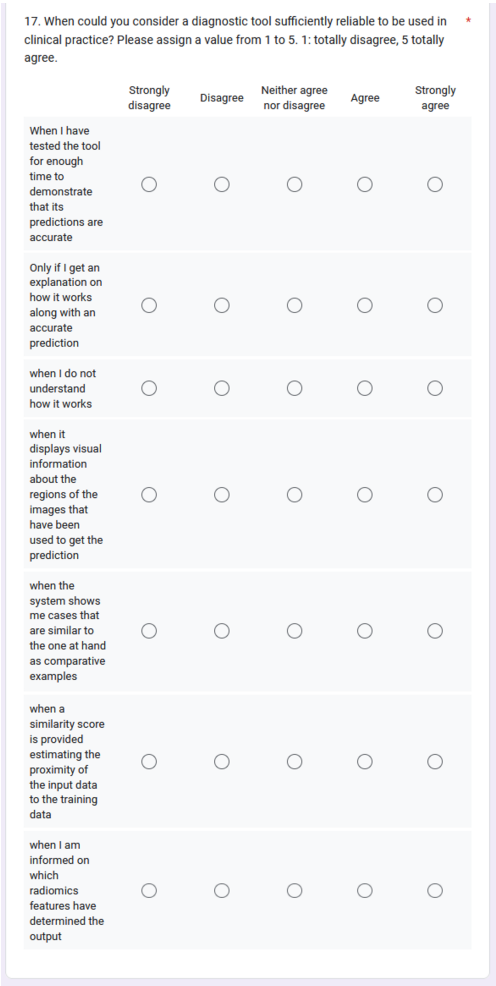


Fig. 5 Questions on most valued reliability features

Regarding the expected detection rates for detection tasks, such as for lesion detection, among positive cases, neither of the proposed options were particularly favored by the respondents. Only one option was distinctly unpopular, i.e., the one regarding the balance between false positives and false negatives. Figure 11 shows the sentiment score and the statistics for this question for positive (up-left) negative cases (down-left) and those on performance metrics for positive (up-right) and negative (down-right) cases. One supplementary free-text answer was also of interest on this topic:

> *Lesion detection is a tricky subject. False negative must first be avoided; however, false positives are also very detrimental to workflow when significant in number. After determining the highest possible sensitivity, a minimum degree of specificity must also be maintained*

In terms of trustworthiness, the respondents appreciated all the provided options, with a particular inclination towards aspects such as reliability, certification, and

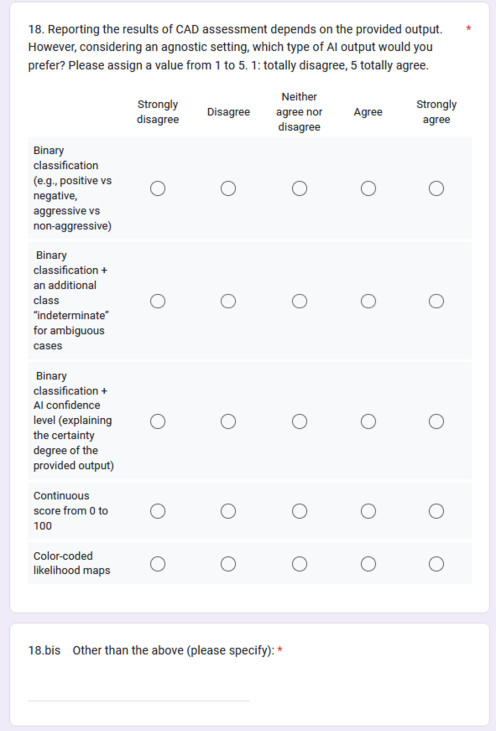


Fig. 6 Questions on preferred way to receive AI models' output

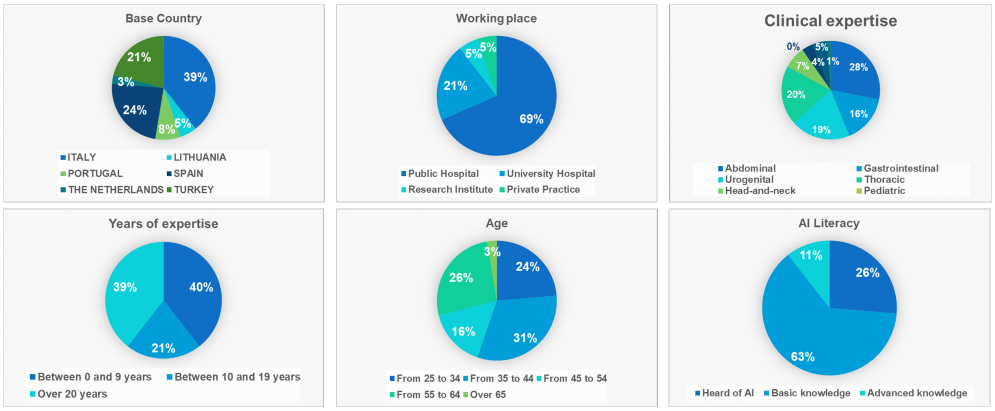


Fig. 7 Demographic and professional profile of respondents (in the "clinical expertise" chart, 0% correspond to pediatric specialisation)

openness, as shown on the left of Figure 12. When it comes to supplementary details that could enhance the credibility of an AI instrument, the participants once again expressed a favorable opinion for all listed options, showing a marked preference for

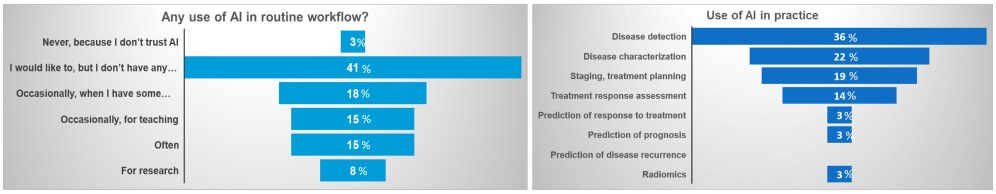


Fig. 8 Left: the use of AI in clinical workflow. Right: the clinical task for which AI is used

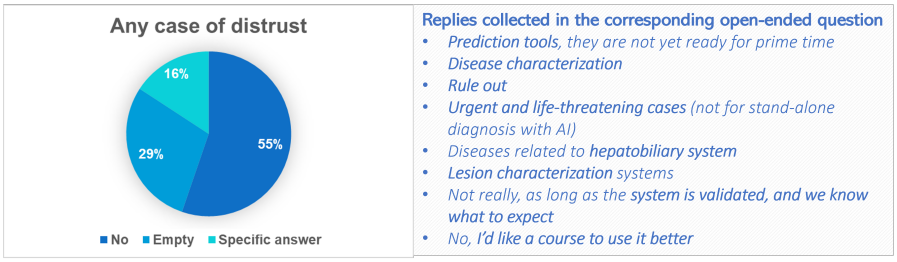


Fig. 9 Cases of distrust

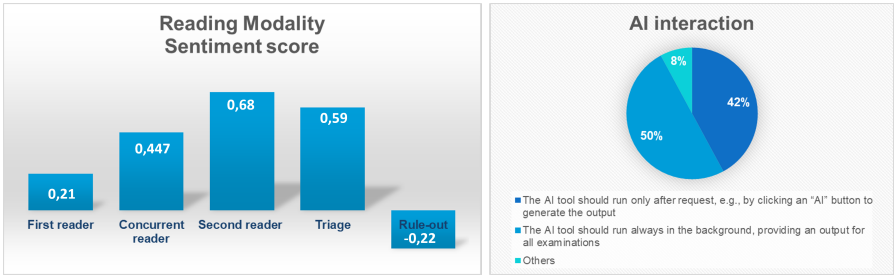


Fig. 10 Left: preferences for the reading modality. Right: preferred workflow integration

revealing details concerning the quality and the provenance of the data employed in training the AI model in question (Figure 12 on the right). This emphasizes the importance of being transparent about the data used.

In terms of reliability, participants placed as the highest valued feature the practical on-field demonstration of the AI tool's accuracy in everyday usage. This was closely followed by the importance of having explanations for the AI model's outputs and its precision (Figure 13). The preference for the AI tool to be understandable and its operations to be transparent to radiologists seemed less important. These observations imply that the primary concern is the AI tool's accuracy. Ideally, the results produced should be explainable. An in-depth access to the AI tool's mechanics is not deemed essential, provided that the outcomes are dependable and supported by a logical explanation.

Lastly, regarding the way results should be presented, respondents showed a preference for an output also including a confidence score of the prediction. This is reported in Figure 14.

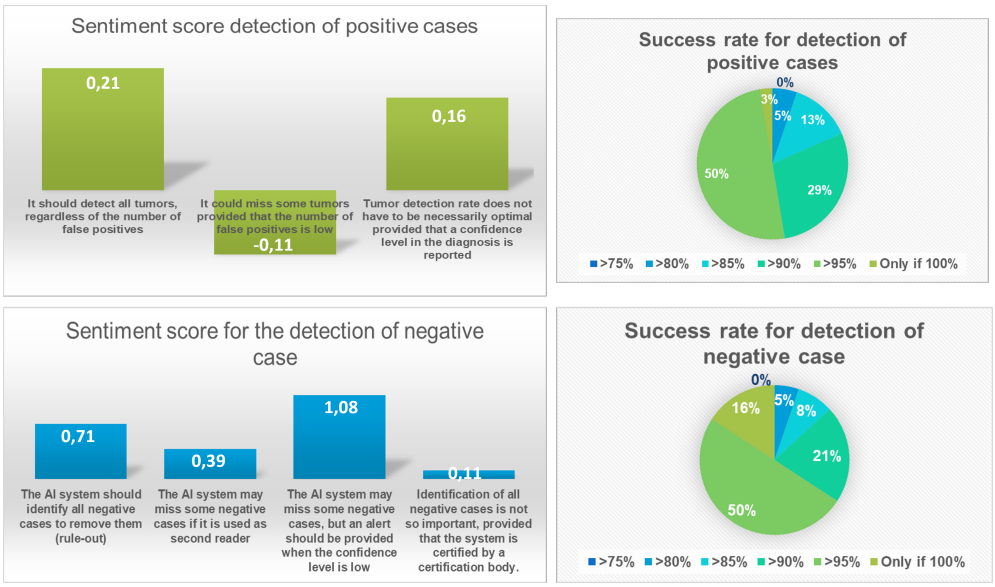


Fig. 11 Expectations in terms of sensitivity and specificity balance and success rates of detection AI tools

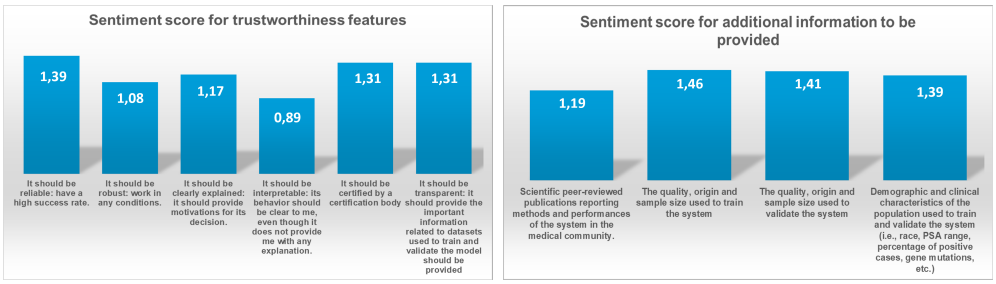


Fig. 12 Desiderata for trustworthiness

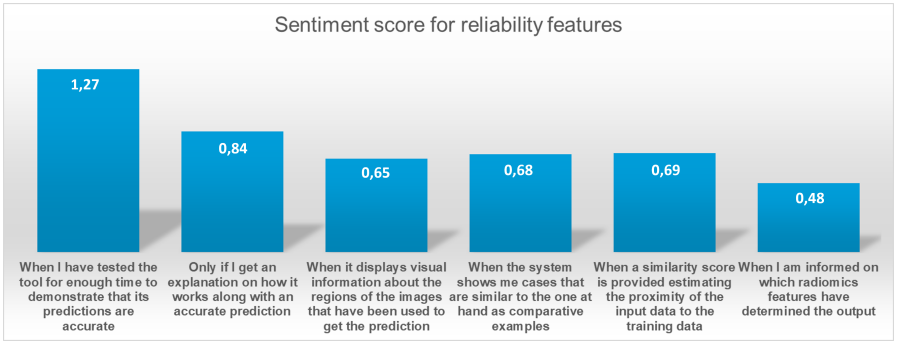


Fig. 13 Desiderata for reliability

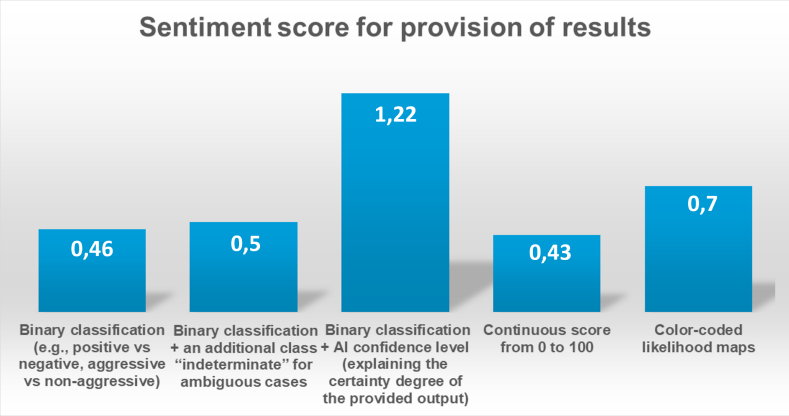


Fig. 14 Desiderata for AI outcome delivery

4 Discussion and conclusion

The findings we gathered were of interest and indicated a need for improvement in certain questions. As we prepared the survey and analyzed the results, it became clear that the trustworthiness and critical acceptance of an AI tool are heavily influenced by the specific clinical task for which it is employed. Consequently, we are of the opinion that tasks such as segmentation, detection, characterization, and outcome prediction ought to be investigated individually with tailored questions. This approach is in line with the standards set by international regulatory and certification agencies, which categorize Software as Medical Devices into different classes (for instance, FDA's CADx, CADt, CADe, and EMA's Class I, II, III).

Overall, the key take-away messages of the survey findings can be summarized as follows:

1. Rule-out and lesion characterization models seem to be considered less desirable, reliable and trustworthy.
2. Reliability and robustness, along with certification and transparency are the most valued features to AI models' trustworthiness. Reliability is verified with on-field usage.
3. Explanations and data transparency are the most desired features. The model's outcomes should be motivated, but the model itself does not have to be necessarily interpretable.
4. The prediction outcomes of an AI model should be accompanied by a confidence value and an alert should me sent when the confidence is low.

According to these findings, our guidance of the ProCAncer-I AI framework will focus on ensuring transparency, robustness, an accurate estimation of model's uncertainty, and explanation facilities. In response to input from ProCAncer-I clinical partners and result evaluations, the survey underwent revisions. These revisions included the introduction of new queries, such as those identifying the optimal point on the ROC curve, and minor adjustments to current questions, like the one concerning

reading and integration methods. The revised survey was submitted to the Scientific Board of the European Society of Oncologic Imaging (ESOI), which endorsed its distribution to the members of the society. Additionally, the survey was shared with the AI4HI "AI Validation Working" Group to synchronize efforts with fellow members and establish more defined criteria for assessing trust in clinical validation activities.

The internal survey was instrumental in tailoring the FUTURE-AI guidelines and pinpointing key areas for compliance efforts. The collaborative design activities of the project are still in progress, as the identification of risk sources and points of vulnerability in clinical settings is ongoing.

Acknowledgements. This study has been partially carried out under the European Union's Horizon 2020 research and innovation program under grant agreement No 952159 (ProCAncer-I) and the Tuscany Region project NAVIGATOR funded and supported by Bando Ricerca Salute Regione Toscana 2018 (DD 15397/2018).

References

[1] Al Kuwaiti, A., Nazer, K., Al-Reedy, A., Al-Shehri, S., Al-Muhanna, A., Subbarayalu, A.V., Al Muhanna, D., Al-Muhanna, F.A.: A review of the role of artificial intelligence in healthcare. Journal of personalized medicine **13**(6), 951 (2023)

[2] Bajwa, J., Munir, U., Nori, A., Williams, B.: Artificial intelligence in healthcare: transforming the practice of medicine. Future healthcare journal **8**(2), 188 (2021)

[3] Alowais, S.A., Alghamdi, S.S., Alsuhebany, N., Alqahtani, T., Alshaya, A.I., Almohareb, S.N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H.A., *et al.*: Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC medical education **23**(1), 689 (2023)

[4] Henry, K.E., Kornfield, R., Sridharan, A., Linton, R.C., Groh, C., Wang, T., Wu, A., Mutlu, B., Saria, S.: Human–machine teaming is key to ai adoption: clinicians' experiences with a deployed machine learning system. NPJ digital medicine **5**(1), 97 (2022)

[5] Alanazi, A.: Clinicians' views on using artificial intelligence in healthcare: opportunities, challenges, and beyond. Cureus **15**(9) (2023)

[6] Choudhury, A.: Factors influencing clinicians' willingness to use an ai-based clinical decision support system. Frontiers in Digital Health **4**, 920662 (2022)

[7] Robertson, L.J., Abbas, R., Alici, G., Munoz, A., Michael, K.: Engineering-based design methodology for embedding ethics in autonomous robots. Proceedings of the IEEE **107**(3), 582–599 (2019)