From the clinic: A survey on trustworthy AI in breast cancer*

Smriti Joshi[a], Anais Emelie[a], Maciej Bobowicz[b], Gianna Tsakou[c], Stefanie Charalambous[c], Zohaib Salahuddin[d], Oliver Diaz[a], Karim Lekadir[a]

***Abstract*— With the fast-growing applications of artificial intelligence (AI) in healthcare, it is essential to keep track of their credibility and reliability. We conducted a survey with healthcare practitioners to obtain requirements for developing reliable AI tools for breast cancer. We share our findings with the healthcare community, hoping that our work serves as a resource to build extensively validated and trustworthy solutions.**

I. Introduction and Methods

Recent developments in machine learning-based techniques show evidence of improving the current standard of care in the healthcare sector. There is growing interest in research for breast cancer, which surpassed lung cancer in 2020 as the most common cancer worldwide [1]. Here, we report results from a survey conducted with breast radiologists, oncologists, and surgical oncologists to get an insight into various aspects required for building trustworthy artificial intelligence (AI) models for breast cancer treatment. This survey was divided into eight sections with each section presenting 4-5 questions (free text, multiple choice, ranking-based, as necessary). The reported statistics are presented from 23 responses to the survey, including practising clinicians from Europe (65%), Asia (13%), Africa (17%) and South America (5%).

II. Results

The survey started with an open question regarding clinicians' concerns about AI deployment in hospitals. It was followed by tailored questions on breast cancer reflecting six dimensions of trustworthy AI defined by the FUTURE-AI guidelines [2]. The results are summarized in Fig. 1. The detailed explanations are as follows:

A. Concerns regarding AI deployment in hospitals

1) Clouding human judgment and interpretation, usage in the absence of human supervision, legal framework
2) Unreliability of the AI solutions, lacking validation and bias mitigation strategies, lacking evidence of additional improvement over the current standard-of-care

*Supported by Europe research and innovation programme under grant agreement No 101057699 (RadioVal) & by Horizon 2020 grant agreement No 952103 (EuCanImage), the project FUTURE-ES (PID2021-126724OB-I00) from the Ministry of Science and Innovation of Spain

[a] Departament de Matemàtiques i Informàtica, Universitat de Barcelona, Spain

[b] 2nd Department of Radiology, Medical University of Gdansk, Poland

[c] Maggioli SPA, Research and Development Lab, Greece

[d] Maastricht University, Department of Precision Medicine, the Netherlands

B. Fairness

The principle of *Fairness* states that AI algorithms should maintain the same performance across all subgroups of individuals, including under-represented groups.

Under-represented groups: Very young or old populations, males, high breast density, previous cosmetic surgery, pregnant patients, individuals without health insurance and low-resource backgrounds, high socio-economic class, minority ethnic groups, groups not included in the screening programs.

Additional sources of bias: Patient history (occurrences in the family, breast injuries, surgeries, hormonal treatment), clinical variables (presence of germline mutation, menopausal status), lifestyle choices (emotional health, smoking habits).

C. Universality

This principle states that a medical AI tool should be reproducible and generalizable outside the controlled environment where it was built.

Specific clinical and technical settings recommendations: presence of standardized protocols, affordability, infrastructure (internet access, appropriate equipment and dedicated workstations), presence of IT support, responsible training of healthcare professionals, software integration, tool's ability to work with varying degree of input information.

Obstacles to universal deployment: Unknown cost-to-benefit ratios, integration with existing systems, software performance and stability, differences in legislative requirements (for example, privacy and data protection laws, liability etc.), lack of adaptability to new research and heterogeneity in workflows, lack of appropriate training (e.g., considering local language, level of expertise, quality control etc.)

Differences in high- and low-resource settings: Technical support and infrastructure, absence of key clinical input required (e.g. magnetic resonance images, expensive drugs for treatment etc.)

D. Traceability

The Traceability principle states that medical AI tools should be developed together with mechanisms for documenting and monitoring the lifecycle of the AI tool.

Documentation: Clinicians expect to receive relevant information summarising the most important aspects of the AI system, including (in order of importance) clinical uses,

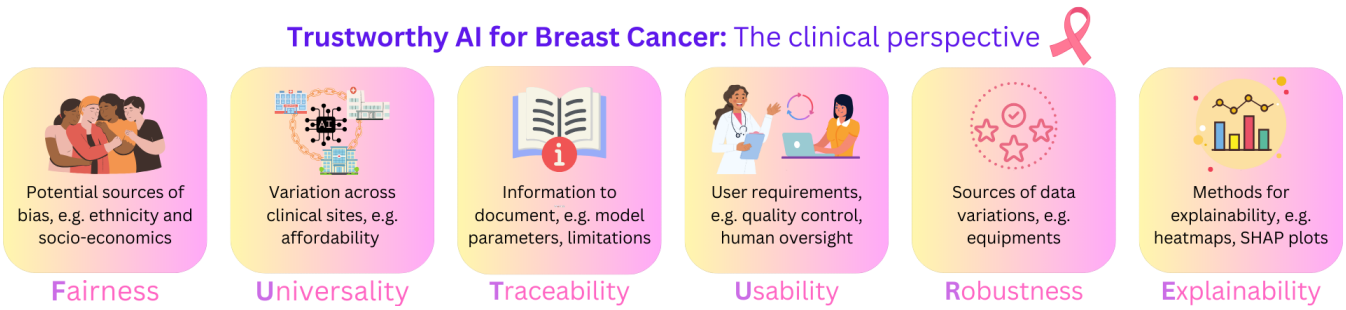

Fig. 1. Summary of the contents

limitations, counter-indications, model parameters, evaluation results, potential biases, ethical considerations, potential benefits, training material, data characteristics, and licenses.

Auditing AI tool: Once the AI tool is deployed, the tool should be audited periodically. 35% of clinicians prefer annual audits, 13% prefer more than one audit a year, 30% recommend auditing when there is a significant update to the tool (for example, retraining), and 22% when there is a noticeable drop in performance of the tool. Most responders believe that AI developers within the developing team (26%) and clinicians actively using the tool (24%) should participate in the auditing activities. Furthermore, individuals from outside the tool development and usage such as third-party auditing authority (16%), AI developers outside the developing team (13%) and clinicians within (13%) and outside (8%) the user institute can also be involved.

E. Usability

Usability refers to the ability of the end-users to properly use AI tools in their real-world environment to achieve a clinical goal efficiently and satisfactorily.

Human oversight: To ensure that the AI errors are reported, clinicians show interest in manual correction of segmentation, flagging AI errors at any stage of the pipeline, and acceptance/rejections of outputs with the reasoning.

Frequency of feedback: 40% of clinicians indicated that they are willing to provide feedback every time the output is rejected, while 27% can report it every time they use the AI tool. The rest of the clinicians would like to provide feedback every 10, 20 or 50 uses.

F. Robustness

Robustness refers to the ability of the tool to maintain its performance and accuracy when it is applied under varying conditions in the real world.

What affects robustness?: Variability in equipment and protocols, quality of imaging and clinical data, technical issues during imaging, patient heterogeneity and missing data.

Annotation: In addition to the imaging data, manual annotations are required to train/validate the AI models.

What should be annotated?: Region-of-interest, size or diameter, the pattern of enhancement, signal intensity, type of uptake, diffusion, oedema, lymph nodes, histoarchitectural surrounding changes, metallic tissue markers, and injuries.

How to segment?: We learnt that clinicians prefer to segment in axial view (69%) as opposed to other views (8%) or plane with the highest resolution (23%). To reduce the time required for segmenting large datasets, 62% clinicians prefer to correct segmentations delivered by an AI-based segmentation algorithm instead of manual segmentation (15%) or correction of segmentation by trained non-experts (23%).

G. Explainability

This principle states that medical AI tools should provide clinically meaningful information about the logic behind the AI decisions.

What kind of visualizations: We asked the clinicians to evaluate the popular ways of providing explanations of a model and we found that they consider visualizations such as heatmaps (69%) and SHAP plots (77%) very helpful in interpreting the results of AI models.

Assessing impact of providing explanations: To evaluate the benefit of adding explanations to the model, clinicians recommend collecting statistics on the final agreement between clinicians and the AI tool, user confidence, the number of AI tool usage as a percentage of managed patients, and efficiency of the workflow (e.g., the time required per case).

III. Conclusion and Discussion

In this work, we strived to provide a clinical perspective on the development of trustworthy AI models for breast cancer. In the future, we would like to extend this work by collecting more responses from clinicians worldwide[1] as well as by including other stakeholders who are involved in the development, deployment, and monitoring of the AI tool. Additionally, we would share our findings with the clinicians to perform a fine-grained analysis with respect to imaging modalities and treatment protocols.

References

[1] Global Cancer Observatory, "The global cancer observatory (gco) is an interactive web-based platform presenting global cancer statistics to inform cancer control and research." https://gco.iarc.fr/ (2023). Accessed: 2023-08-07

[2] Lekadir, Karim, et al. "FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare." arXiv preprint arXiv:2309.12325 (2023).

[1]Survey for clinicians: `https://form.jotform.com/223202140983346`