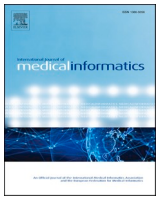
International Journal of Medical Informatics 206 (2026) 106127

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Validity and transferability of Model for ASsessing the value of Artificial Intelligence (MAS-AI)

Iben Fasterholdt [a,b,*], Julie S. Schrøder [a], Linda H. Hansen [a], James M. Bowen [b,c,d], Anne Gerdes [e], Kristian Kidholm [a], Tudor M. Haja [f], Francesco Calabrò [f], Rossana Cecchi [g], Alexandra Stanimirovic [b,c], Troy Francis [b,c], Valeria E. Rac [b,c], Benjamin S.B. Rasmussen [e,h]

[a] CIMT - Centre for Innovative Medical Technology, Odense University Hospital, Odense, Denmark
[b] Program for Health System and Technology Evaluation, Ted Rogers Centre for Heart Research, Peter Munk Cardiac Centre, Toronto General Hospital Research Institute, University Health Network, Toronto, Ontario, Canada
[c] Institute of Health Policy, Management & Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada
[d] Department of Health Research Methods, Evidence and Impact, Faculty of Health Sciences, McMaster University, Hamilton, Ontario, Canada
[e] CAI-X – Centre for Clinical Artificial Intelligence, Odense University Hospital, Odense, Denmark
[f] Department of Medicine and Surgery, Unit of Legal Medicine, University of Parma, Parma, Italy
[g] Department of Biomedical, Metabolic and Neural Sciences, Unit of Legal Medicine, University of Modena and Reggio Emilia, Modena, Italy
[h] Department of Radiology, Odense University Hospital, Odense, Denmark

ARTICLE INFO

Keywords:
HTA
Evaluation
Artificial intelligence
Transferability
Delphi

ABSTRACT

Objectives: In 2022, a multidisciplinary group of experts and patients published a Model for ASsessing the value of AI (MAS-AI) in medical imaging. MAS-AI is a critical tool for decision-makers, enabling them to make informed choices on the prioritization of AI solutions. The objective of this study was to assess the face validity and transferability of MAS-AI by investigating workshop participants' perceptions in Denmark, Italy, and Canada regarding the importance of its content.
Methods: A Delphi process was conducted, including inputs from four workshops with a sample of decision makers from hospitals or the healthcare sector, patient partners and various researchers and experts. The participants were asked to rate the importance of each of the domains and subtopics in MAS-AI on a 0–3 Likert scale.
Results: A total of 95 participants from three countries participated. The face validity of all MAS-AI domains was confirmed by Denmark, Canada, and Italy, with over 70 percent of the respondents in the first round rating the domains as moderately or highly important. Overall, the five process factors were considered moderately or highly important by between 93 percent and 87 percent of the respondents. All the individual subtopics under each domain were rated above the 70 percent cut-off, except five subtopics for Italy.
Conclusions: The study confirmed the validity of the MAS-AI domains in Denmark, Canada, and Italy. Several improvements in study design and data collection were identified. In the future, analyzing participants to understand which items were rated as important by whom could provide valuable insights.

1. Introduction

In recent years, the rapid advancements in artificial intelligence (AI) have brought transformative changes to various industries, including healthcare. As the applications of AI continue to expand, there is a growing need for robust and evidence-based assessment frameworks to evaluate their effectiveness. The concept of 'assessment' is grounded in the fundamentals of health technology assessment (HTA) and, in 2012, was tailored into a specific Model for Assessment of Telemedicine (MAST) [1]. The primary aim of the MAST framework was to elucidate the attributes and impacts of telemedicine in a manner comprehensible to those responsible for making decisions.

Abbreviations: AI, Artificial Intelligence; HTA, Health Technology Assessment; IT, Information Technology; MAS-AI, Model for ASsessing the value of Artificial Intelligence; MAST, Model for Assessment of Telemedicine; WS, Workshop.

* Corresponding author at: Centre for Innovative Medical Technology, Odense University Hospital, Sdr. Boulevard 29, Entrance 102, 4rd floor, 5000 Odense C, Denmark.
E-mail address: if@rsyd.dk (I. Fasterholdt).

https://doi.org/10.1016/j.ijmedinf.2025.106127
Received 27 June 2025; Received in revised form 24 September 2025; Accepted 25 September 2025
Available online 10 October 2025

Building on the accomplishment of MAST, a novel framework was developed: MAS-AI (Model of ASsessment of Artificial Intelligence) [2]. The primary objective of MAS-AI is to provide a comprehensive evaluation of the impact and outcomes of AI solutions across diverse HTA domains. As Fig. 1 shows, the following nine domains are included in MAS-AI: 1) Health problem and description of the application, 2) technology, 3 + 4) ethical and legal aspects, 5) safety, 6) clinical aspects, 7) economics, 8) organizational aspects, and 9) patient aspects. The development of MAS-AI was motivated by the importance of responsible and effective integration of AI technologies into healthcare. Drawing upon the insights gained from MAST and HTA via a comprehensive research-based approach and adapting them to AI-specific applications, MAS-AI aims to offer valuable insights into the potential benefits and challenges of AI implementation. With the demand for AI solutions rising, MAS-AI emerges as a critical tool for decision-makers, enabling them to make informed choices and prioritize patient-centered care. Other initiatives for holistically evaluate AI now exists like the EU-funded projects AI-MIND for dementia diagnosis [3], and EDiHTA – a HTA framework for digital health/AI [4]. Also, an overview from June 2025 of AI evaluation frameworks [5] exits.

Although input came from a literature review of international studies on evaluating AI in healthcare [6], MAS-AI was primarily developed in a Danish context. Thus, to ensure the quality and generalizability of the model, the validity and transferability need to be addressed. Validity refers to the extent to which a research study accurately measures or assesses what it claims to measure, and transferability refers to the extent to which the findings of a research study can be applied or generalized to other settings, populations, or contexts beyond the specific conditions of the study. The objective of this study was to assess the validity and transferability of MAS-AI by investigating workshop participants' perceptions in Denmark, Italy, and Canada of the importance of the information included in MAS-AI.

2. Methods

2.1. Design

The Delphi technique is a widely accepted method for consensus-building concerning a specific topic. This is done by gathering data from respondents within their expertise [7]. In this study, a modified Delphi technique was conducted to validate MAS-AI by 1) developing a well-structured questionnaire about the importance of the different domains and topics in MAS-AI for data collection; and 2) conducting one round of asking respondents at workshops to answer the Delphi questionnaire. The structural logic of the questionnaire was based on previous published work with a similar research goal of validating a HTA framework with the exception that the technology was telemedicine and also they did not mix roles and aggregated results in their study [8]. Data were collected in three different countries to assess the international transferability of MAS-AI.

2.2. Structured questionnaire

An overview of the three main parts of the MAS-AI model and their respective content is shown in Fig. 1. Two steps cover nine domains and process factors for a MAS-AI assessment.

For the validity and transferability analysis, data were collected using a structured questionnaire on the importance of the nine domains in the MAS-AI model and their corresponding topics. The questionnaire also contained questions about the importance of the five process factors for the MAS-AI assessment. Demographic data, specifically the respondents' place of territory, was also included in the questionnaire. The original four questionnaires can be seen in Online Resource 1–4, which contains the complete questionnaires regarding the validity of MAS-AI. Note that the questionnaire used in the first workshop was slightly modified, with some domains renamed and questions reworded for the following three workshops, see the elaboration in Online Resource 6.

The importance of each domain, topic, and process factor was rated on a 0–3 Likert scale (0 = not important, 1 = somewhat important, 2 = moderately important, and 3 = highly important). The data collection process encompassed both paper-based versions of the questionnaire and digital versions utilizing the questionnaire tool SurveyXact [9]. If a respondent answered "0″ in the online version, they were provided with the opportunity to state the reason for this response.

2.3. Selection of respondents

Convenience sampling was used while including participants at the

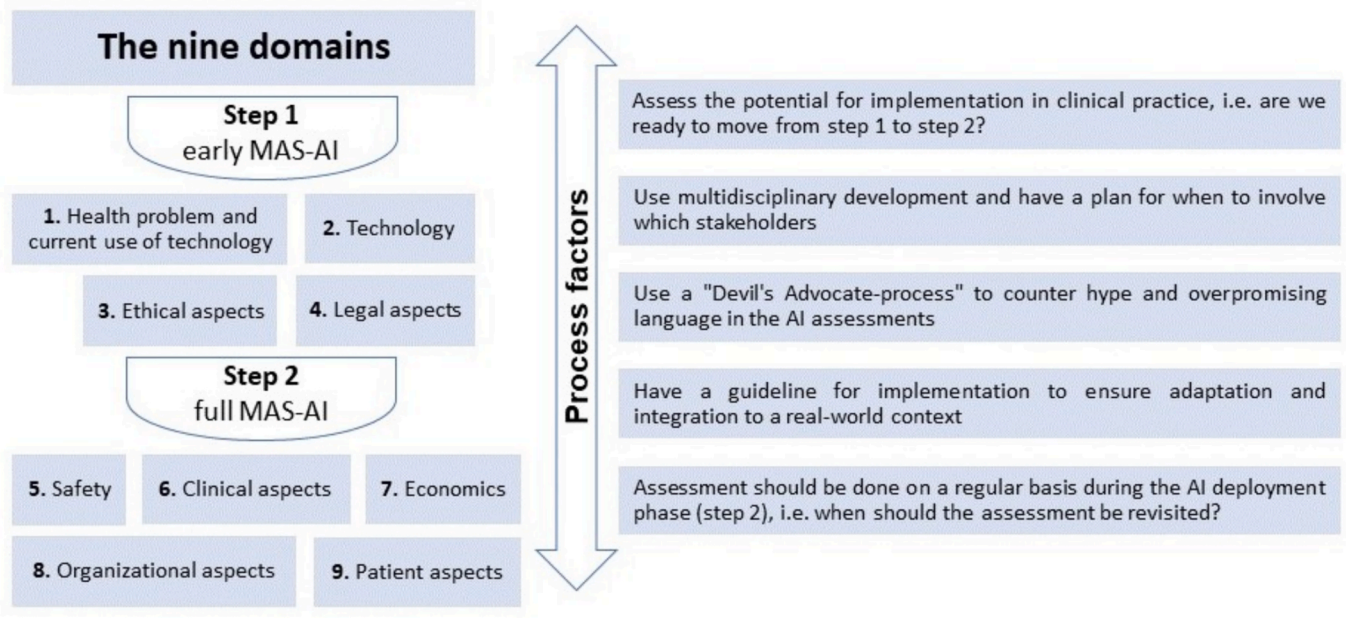


Fig. 1. Overview of MAS-AI Abbreviations: MAS-AI, Model for ASsessing the value of Artificial Intelligence; AI, Artificial Intelligence.
Source: Reproduced from 2

four workshops. While workshop 1 (WS1) and workshop 3 (WS3) used formal invitations issued by the workshop team, workshop 2 (WS2) was part of the AI track on the WHINN (Week of Health and INNovation) conference and hence participant selection could not be controlled. We advertised that the main target group for the workshops was stakeholders from hospitals or the healthcare sector, e.g., decision-makers, patient partners, clinicians, AI experts and researchers. Recruitment for the Italian workshop 4 (WS4) took into account the professional qualifications of the participants, and sought to involve as many health professionals, lawyers and engineers working in the field of AI as possible.

2.4. Workshop format

The aim of the four workshops was to get the participants' feedback on what MAS-AI looks like, which elements are most important (validation with a questionnaire), as well as to collect change requests for MAS-AI. All workshops started out with a presentation of the three parts in the MAS-AI model. However, small workshop variations arose, because the country comparisons was not planned from the very onset.

WS1 and WS3 participants had the opportunity to prepare for the session with pre-work (reading a case and the main article) which was not the case in WS2. Next, the Delphi questionnaire was distributed and completed. Participants were able to ask the workshop presenters any clarifying questions or get help if they had difficulty completing the questionnaire. WS1 and WS3 were full-day workshops, while WS2 was a shorter session of about 1.5 h. In WS4, in an online meeting lasting approximately 1.5 h, the participants were sent the main article and the questionnaire, and it was explained to them what was contained in both documents and what the researchers' aim was in completing the questionnaire. Once any doubts had been clarified, participants were asked to complete the questionnaire and return it by e-mail.

WS2 and WS3 each had two moderated group sessions following the questionnaire to supplement the Delphi results with more quantitative input. WS3 had a special focus on adjustment to a Canadian context. Results of this group work is not reported here (see perspectives).

2.5. Data analysis

Median score and minimum and maximum values were estimated for Likert scale questions, i.e. the questions on the importance of the domains, topics, and process factors for the MAS-AI model, in accordance with guidelines [7]. The quantitative data analysis was performed using SurveyXact [9] and Excel [10]. Fisher's exact test was used to assess the differences in the proportion answering 'moderately important' or 'highly important' between Denmark, Canada, and Italy since some categories had less than five observations. Bonferroni Correction was used to correct for multiple testing.

Missing answers were accepted. However, if a respondent left the response on the importance of the overall domain blank but provided answers on the importance of the underlying topics within that domain, an average of the individual respondent's answers regarding the underlying topics within the domain was imputed. This average was considered the answer to that domain's importance, and the usual rounding rules were applied. In addition, if two answers were provided for the same question without clearly indicating which answer was valid, the lowest answer was chosen as the response. WS4 systematically did not collect answers on the importance at the overall domain and hence these were all imputed.

In the analysis of the questions, a pre-defined stopping rule from the grant proposal was applied. We consider the participants' agreement to be stable in terms of face validity if 70 % or more of them rate each domain as 'moderately important' or 'highly important' in Round 1, as recommended [11]. As the study only collected data by questionnaire and did not involve patients, no approval of the study from the ethical committee was required in Italy or Denmark [12]. For the qualitative Canadian study, the Canadian team has received the University Health Network institutional REB (23–5787).

3. Results

A total of 95 respondents participated in the study, with 27 respondents from Denmark (WS1 had 12 participants and WS2 had 15 participants), 33 respondents from Canada (WS3) and 35 respondents from Italy (WS4). The two Danish workshops were held in Odense, with WS1 on November 22, 2021, and WS2 on November 10, 2022. WS3 was a hybrid workshop in Toronto, Canada, on May 24, 2022, with online participation as well. WS4 was held entirely online in Italy on October 23, 2023. Details regarding the respondents' place of territory or region can be found in Fig. 2. The respondents consisted of decision-makers from hospitals or the healthcare sector, patient partners, experts and researchers from the public sector, participants from companies, IT architects, and individuals with other similar backgrounds. In addition, engineers dealing with AI, and law graduates participated in Italy.

Tables 1 and 2 present the participants' answers to the Likert scale questions on the importance of each of the nine domains and five process factors, including their subtopics in the MAS-AI model. Overall, there was a significant difference between the three countries in the proportion answering either 'moderately important' or 'highly important' in four domains and four topics (see Table 1), and no significant difference among advices or process factors (see Table 2).

In WS1 and WS2, all respondents considered five out of nine domains to be moderately or highly important. The remaining four domains also received high ratings for their importance. Domain 2 and Domain 8 were rated as moderately or highly important by 88 percent of the respondents, while Domain 7 and Domain 9 were rated as moderately or highly important by 92 percent and 93 percent of the respondents, respectively. The respondents in WS1 and WS2 generally considered the underlying topics of the domains as moderately or highly important, with no ratings falling below 70 percent. Ratings on the importance of three topics stood out from the rest. This concerned the topic 'Maturity of the AI technology' (in Domain 2), considered moderately or highly important by 19 out of 26 respondents (73 percent). In the same way, the topics 'Is the AI application integrating Ethics by Design?' and 'Autonomy' (in Domain 3) were considered as moderately or highly important by 20 out of 26 respondents (77 percent). It is worth noting, that the number of respondents in WS1 and WS2 varies due to missing data or incomplete responses. For the responses on the importance of process factors, along with their underlying topics, most respondents in WS1 and WS2 rated them as moderately or highly important, ranging from 84 percent to 100 percent.

For WS3, the only three domains not considered moderately or highly important by all 33 respondents were Domain 1, Domain 4, and Domain 8. Domain 1 and Domain 8 were rated as moderately or highly important by 97 percent of the respondents, and Domain 4 was rated as moderately or highly important by 91 percent of the respondents. All of the underlying topics were rated moderately or highly important by over 70 percent of the respondents, ranging between 82 percent and 100 percent. All of the responses on the importance of advice and process factors, along with their underlying topics, were assessed as moderately or highly important by over 70 percent of the respondents as well, and the lowest score concerned the topic 'Assess the maturity', which was rated as moderately or highly important by 25 out of 33 respondents (76 percent).

Regarding WS4, none of the domains had a score of 100 percent, meaning zero domains were considered to be moderately or highly important by all of the 35 respondents. The estimates showed that all domains were rated moderately or highly important by over 70 percent of the respondents, ranging between 71 percent and 91 percent. The WS4 respondents generally considered the underlying topics of the domains moderately or highly important, with no ratings falling below 70 percent in 33 out of the 38 topics. The topics 'Equity' (in Domain 3) and

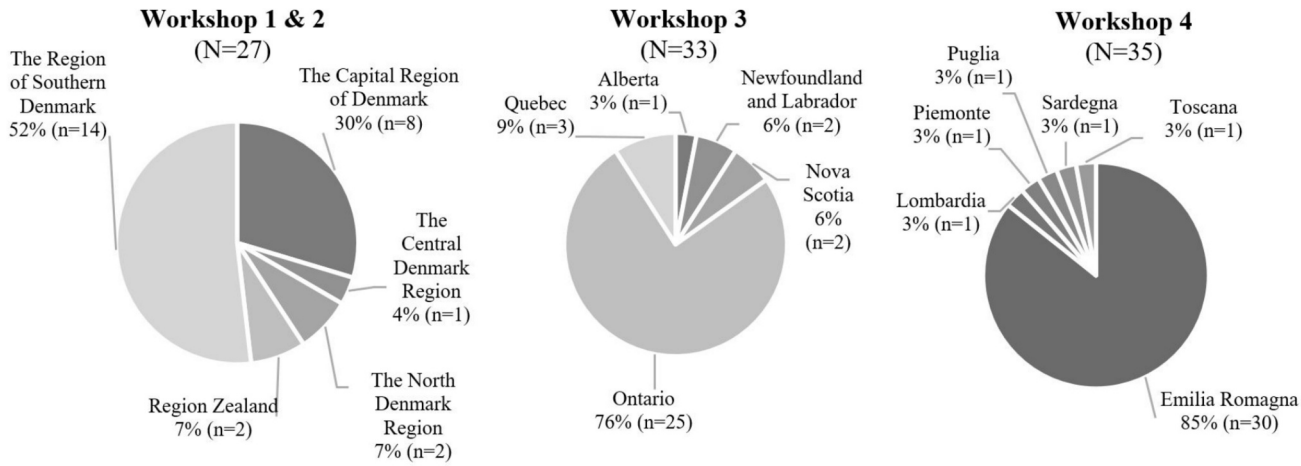


Fig. 2. Region or territory of the respondents. The proportion of responses at the workshops based on the region or territory of the respondents.

'Legal requirements transformed into functionalities in the application' (in Domain 4) were considered moderately or highly important by 22 out of 35 respondents (63 percent). Likewise, the topics 'Implementation requirements and culture,' 'Consequences for roles' (both in Domain 8), and 'Patients' willingness and satisfaction' (in Domain 9) were considered moderately or highly important by 24 out of 35 respondents (69 percent).

The majority of the respondents in WS4 rated the importance of process factors as moderately or highly important, ranging from 80 percent to 91 percent.

4. Discussion

Our study found an overall high level of agreement on the importance of the content of the MAS-AI model across all four workshops. Based on the results, the face validity of all MAS-AI domains was confirmed by Denmark, Canada and Italy, since all domains were considered moderately or highly important by more than 70 percent of the respondents in the first round. Overall, the five process factors were considered moderately or highly important by between 93 percent and 87 percent of the respondents in the three countries. When rating the individual subtopics under each domain, all were above the 70 percent cut-off, except for five subtopics for Italy. The assessment of the MAS-AI model's transferability from a Danish context to a Canadian context and an Italian context appears positive based on the gathered results from the workshops. Compared to the Delphi study of the MAST [1] and a Delphi study of an AI framework for dementia [3], this study showed similar results regarding the same domains, suggesting that all three models are quite valid. However, it is possible that the Delphi process lacks sufficient sensitivity in order to detect issues regarding the chosen domains.

4.1. Strengths and limitations

A key strength of our approach lies in its use of the Delphi process. It provides a systematic and iterative means of achieving consensus among experts. Minimizing individual biases and synthesizing input from a wide range of participants, the Delphi method enhances the robustness and reliability of the findings. Furthermore, incorporating multidisciplinary input from diverse stakeholders through international workshops in Italy and Canada underscores MAS-AI's adaptability and relevance across healthcare systems. However, the findings of this study has some limitations. Firstly, in the data collection process, this study utilized a paper version and a digital questionnaire at the Canadian workshop, which offered an advantage by allowing everyone to participate. However, using paper could have introduced uncertainty in the answers, as some respondents may have missed or neglected to answer specific questions, resulting in missing data. Secondly, it was not possible to determine the number of individuals invited to the four workshops, as well as the number of individuals who participated. Thus, no inferences was made regarding the participation rates of this study. Thirdly, it was not possible to analyze potential differences between respondents' respective roles e.g., decision makers and experts, etc. This would have made it possible to understand which items were rated as important by whom, as the respondents represent different stakeholders with different values and preferences. It was a goal to target decision-makers (persons who participate in making decisions on investment in AI) and to a lesser degree, experts and patients. Without knowing the exact numbers, we do know that we ended up having a high proportion of experts, which may have affected our findings. This stakeholder imbalance should be avoided in future iterations by recruiting participants more narrowly. In addition, it may be considered a limitation that the industry and developers were not included as respondents in this study, with the exception of a few in WS2 and WS4, as their view may have been beneficial as well. Fourthly, when data from WS1 and WS2 were not combined, some of the subtopic responses fell below 70 percent (see Online Resource 5). Fifthly, this study only includes "round 1" Delphi data, which is why one could argue that this study resembles a validation survey rather than a traditional Delphi study. However, since this study was planned and conducted with the Delphi methodology in mind and thus was communicated in this way to the participants, we still regard this study as a Delphi study. It is not unheard of to modify the Delphi method and in a study featuring agricultural and extension education research, authors identified three frequent modifications to the conventional Delphi method: mode of delivery, number of rounds or iterations, and threshold for reaching consensus of agreement [13]. Also Best et al., argue that a Delphi method can employ as many iterations as the investigator(s) deem necessary for consensus of agreement to be reached [14]. They further argue that single-iteration Delphi approach may be useful when time or resources are limited (hence reduce the costs and time inherent in undertaking a large-scale comprehensive Delphi studies) or when a quick overview of expert opinions is sufficient. Although surprisingly very few single-iteration studies have appeared in the literature, our study created feasible and favorable conditions for the application of single-round Delphi approach while considering the appropriate short and defined time period alongside a readily assembled group of experts. Also, given the results (the high level of agreement on most subtopics), we considered that performing two rounds was unnecessary. However, it might have been beneficial to do so for Italy, as the responses of WS4 showed a lower level of agreement than the other workshops. The lower level of agreement in WS4 on questions about certain ethical and legal aspects and organizational aspects is probably

Table 1
Response to questions about importance of domains and topics as part of the basis for decisions on investment in AI.

	Results from Workshop 1 and Workshop 2 Denmark (N = 27)			Results from Workshop 3 Canada (N = 33)			Results from Workshop 4 Italy[d] (N = 35)		
Domains and topics	Median	Min-Max	Proportion answering 'moderately important' or 'highly important'	Median	Min-Max	Proportion answering 'moderately important' or 'highly important'	Median	Min-Max	Proportion answering 'moderately important' or 'highly important'
Domain 1: Health problem and description of the application	3	2–3	100 %	3	1–3	97 %	3	1–3	89 %
Health problem of the patients	3	1–3	96 %[a]	3	1–3	97 %	3	0–3	89 %
Description of the application	3	2–3	100 %[a]	3	0–3	94 %	3	1–3	91 %
The objectives, design and aim of the study	2	1–3	89 %[b]	3	0–3	94 %	2	0–3	86 %
Domain 2: Technology	2	1–3	96 %	3	2–3	100 %	2.5	1–3	86 %
Development, performance and validation of the AI model	3	1–3	96 %[a]	3	1–3	97 %	3	0–3	86 %
Maturity	2	0–3	73 %[a]	2	1–3	85 %	3	1–3	80 %
Compatibility & Adaptability	3	1–3	92 %[a]	3	0–3	94 %	3	0–3	89 %
Manageability	2	1–3	87 %[c]	2	1–3	97 %	3	0–3	83 %
Security	3	2–3	100 %[c]	3	0–3	97 %	3	0–3	89 %
Usability	2	1–3	85 %[a]	3	1–3	97 %	3	1–3	91 %
Domain 3: Ethical aspects *	2.5	2–3	100 %	3	2–3	100 %	3	0–3	79 %
Is the AI application integrating Ethics by Design?	2	1–3	77 %[a]	3	0–3	91 %	2	0–3	86 %
Beneficence and patient integrity	3	2–3	100 %[a]	3	2–3	100 %	3	1–3	89 %
Privacy *	2	2–3	100 %[a]	3	1–3	97 %	3	0–3	74 %
Equity *	2	0–3	85 %[a]	3	0–3	94 %	2	0–3	63 %
Trust, transparency, accountability, and responsibility	3	1–3	88 %[a]	3	0–3	94 %	3	0–3	83 %
Autonomy	2	1–3	77 %[a]	3	0–3	91 %	3	0–3	77 %
Domain 4: Legal aspects *	3	2–3	100 %[a]	3	0–3	91 %	2	0–3	71 %
Map the legal landscape for the entire lifecycle of the application	2	1–3	87 %[c]	2	0–3	85 %	2	0–3	71 %
Legal requirements transformed into functionalities in the application *	2.5	1–3	87 %[c]	2	0–3	82 %	2	0–3	63 %
Application reviewed or approved by regulatory authorities	3	2–3	100 %[c]	3	0–3	91 %	3	0–3	80 %
Domain 5: Safety	3	2–3	100 %	3	2–3	100 %	2.5	0–3	84 %
Clinical safety	3	2–3	100 %[a]	3	2–3	100 %	3	1–3	89 %
Technical safety	3	1–3	92 %[a]	3	1–3	97 %	3	1–3	89 %
Continues monitoring of safety and new practice	3	1–3	96 %[a]	3	1–3	97 %	3	0–3	83 %
Upcoming challenges regarding safety assurance	2	1–3	81 %[a]	2	1–3	91 %	2	0–3	74 %
Domain 6: Clinical aspects	3	2–3	100 %	3	2–3	100 %	3	1–3	91 %
Sensitivity, specificity, and receiveroperating characteristic curve (ROC)	3	2–3	100 %[a]	3	0–3	91 %	3	1–3	91 %
Effects on morbidity	3	2–3	100 %[b]	3	0–3	88 %	3	1–3	86 %
Effects on mortality	3	2–3	100 %[b]	3	0–3	85 %	3	1–3	97 %
Time to event	2	1–3	80 %[b]	3	1–3	94 %	3	1–3	91 %
Effects on quality of life	2	1–3	92 %[b]	3	1–3	97 %	3	1–3	91 %
Domain 7: Economic aspects *	2	1–3	92 %[a]	3	2–3	100 %	2	0–3	83 %
Societal economic evaluation	2	1–3	88 %[b]	3	1–3	97 %	2	0–3	83 %
Business case	2	1–3	80 %[b]	2	1–3	85 %	2	0–3	77 %
Use of health service	2	1–3	84 %[b]	3	1–3	91 %	2	0–3	89 %
Domain 8: Organizational aspects *	3	1–3	88 %[a]	3	1–3	97 %	2	0–3	73 %
Consequences for the workflow	2	1–3	81 %[a]	3	1–3	94 %	2	0–3	77 %
Consequences for the user	3	2–3	100 %[a]	3	1–3	97 %	2	0–3	77 %
Implementation requirements and culture	3	1–3	88 %[a]	2	1–3	91 %	2	0–3	69 %
Consequences for roles	2	1–3	85 %[a]	3	1–3	91 %	2	0–3	69 %
Domain 9: Patient aspects	2.5	1–3	93 %	3	2–3	100 %	3	2–3	89 %
Patients' willingness and satisfaction*	2	0–3	80 %[b]	3	1–3	97 %	2	1–3	69 %

(*continued on next page*)

Table 1 (*continued*)

Domains and topics	Results from Workshop 1 and Workshop 2 Denmark (N = 27)			Results from Workshop 3 Canada (N = 33)			Results from Workshop 4 Italy[d] (N = 35)		
	Median	Min-Max	Proportion answering 'moderately important' or 'highly important'	Median	Min-Max	Proportion answering 'moderately important' or 'highly important'	Median	Min-Max	Proportion answering 'moderately important' or 'highly important'
Technical improvement during imaging process	2	1–3	80 %[b]	2	0–3	88 %	2	1–3	86 %
Clinical-based patient benefits	3	1–3	92 %[b]	3	2–3	100 %	3	2–3	100 %
Overall patient and social benefits	3	1–3	92 %[a]	3	0–3	97 %	3	2–3	100 %

Note: Separate data from WS1 and WS2 can be found in Online Resource 5 and methodological considerations in Online Resource 6.
*p-value < 0.05, Fisher's exact test with Bonferroni corrected p-values.
[a] Denmark: One respondent from WS1 did not answer this question (paper version), therefore N = 26; [b] Denmark: Two respondents from WS1 did not answer this question (paper version), therefore N = 25; [c] Denmark: The topic was not included in the WS1 questionnaire, therefore N = 15; [d] Italy: Mean of proportion answering 'moderately important' or 'highly important' was calculated for each domain based on the associated topics.

Table 2
Response to questions about importance of advices or process factors as part of the basis for decisions on investment in AI.

	Results from Workshop 1 and Workshop 2 Denmark (N = 27)			Results from Workshop 3 Canada (N = 33)			Results from Workshop 4 Italy (N = 35)		
	Median	Min-Max	Proportion answering 'moderately important' or 'highly important'	Median	Min-Max	Proportion answering 'moderately important' or 'highly important'	Median	Min-Max	Proportion answering 'moderately important' or 'highly important'
Advices or process factors	3	1–3	93 %	3	1–3	91 %	2	0–3	87 %
1) Assessment should be done on a regular basis during the AI deployment phase, so when should the assessment be revisited?	3	1–3	92 %[a]	3	1–3	94 %[c]	3	0–3	91 %
2) Use multidisciplinary development with active participation across all stakeholders – have a plan for when to involve which stakeholders.	3	1–3	84 %[a]	3	1–3	94 %	3	0–3	89 %
3) Use a "Devil's Advocate-process" to counter hype and overpromising language in the assessments of AI, e.g., by having people in the assessment team who are skeptical towards the AI application.	3	1–3	84 %[a]	2	0–3	88 %	3	0–3	86 %
4) The organization should have a guideline for implementation to ensure adaptation and integration to real-world existing workflows and context.	3	2–3	100 %[b]	3	0–3	91 %	3	0–3	89 %
5) Assess the maturity: Judge the potential for clinical practice implementation through classification in development phases, i.e., are we ready to move from step 1 (project phase) to step 2 (operation phase)?	2.5	1–3	96 %[a]	2	0–3	76 %	2	0–3	80 %

Abbreviations: AI, Artificial Intelligence.
[a] Denmark: Two respondents from WS1 did not answer this question (paper version), therefore N = 25; [b] Denmark: Three respondents from WS1 did not answer this question (paper version), therefore N = 24; [c] Canada: One respondent did not answer this question (paper version), therefore N = 32.

related to the large number of engineers working with AI and legal experts not specifically involved with AI answering the questionnaire. The former are probably very focused on the possibilities offered by AI, to the extent that they underestimate the importance of ethical and legal issues, while the latter are probably less interested in organizational aspects. However, a more accurate assessment would require knowledge of the experts who responded to the individual questionnaires. Lastly, the workshops were performed over a two-year timeframe, which could affect the answers, given the rapid change in the AI environment.

4.2. Perspectives

MAS-AI is primarily designed for healthcare decision makers—such as medical directors, hospital department heads, treatment councils, procurement organizations, policymakers, and HTA bodies. It helps guide decisions on AI technology implementation in healthcare but additionally, developers and researchers can use MAS-AI for development, data collection, or research purposes. The high agreement on relevance we find does not indicate a huge demand for changes and the positive results will hopefully make it easier to use the MAS-AI in the future. Already we have applications of the MAS-AI framework in several countries and in both medical imaging and other areas, e.g. organizational AI [15] and VR treatment of social anxiety in psychiatry [16]. However, in addition to the closed-ended questions, the Delphi questionnaire contained three open text fields where respondents could comment on missing aspects, redundant or irrelevant aspects, and the two steps in the MAS-AI model. The free text comments we received, suggest several refinement needs and these qualitative data are currently

being analyzed for key themes and patterns in the joint action project EUCanScreen [17]. We will publish these additional qualitative data analysis in connection with the upcoming article on the updated MAS-AI. In EUCanScreen the MAS-AI will specifically be validated in the area of AI technology used in cancer screening in Europe.

5. Conclusions

The study confirmed the validity of the MAS-AI domains in Denmark, Canada, and Italy. Several improvements in study design and data collection were identified. In the future, analyzing participants to understand which items were rated as important by whom could provide valuable insights.

Ethical approval

Not applicable.

CRediT authorship contribution statement

Iben Fasterholdt: Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Julie S. Schrøder:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation. **Linda H. Hansen:** Writing – review & editing, Software, Formal analysis, Data curation. **James M. Bowen:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis. **Anne Gerdes:** Writing – review & editing, Investigation. **Kristian Kidholm:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization. **Tudor M. Haja:** Writing – review & editing, Writing – original draft, Investigation, Data curation. **Francesco Calabrò:** Writing – review & editing, Writing – original draft, Investigation, Data curation. **Rossana Cecchi:** Writing – review & editing, Writing – original draft, Supervision, Investigation, Formal analysis. **Alexandra Stanimirovic:** Writing – review & editing, Investigation, Funding acquisition. **Troy Francis:** Writing – review & editing, Investigation. **Valeria E. Rac:** Writing – review & editing, Supervision, Resources, Investigation, Funding acquisition. **Benjamin S.B. Rasmussen:** Writing – review & editing, Writing – original draft, Supervision, Investigation, Funding acquisition, Conceptualization.

Informed consent

Not relevant.

Funding

This work was supported by an in-house fund at Odense University Hospital (Denmark) named “Konkurrencemidler” in Danish. Also, the international fund at Odense University Hospital (Denmark) provided funding for the workshop travelling for the Danish research team and the article processing costs is covered by University of Southern Denmark. The Canadian workshop was funded by the TRANSFORM-Heart Failure Collaboration Starter Grant. The funders had no role in the design of the study and collection, analysis, or interpretation of data, in writing the manuscript or in the decision to submit the manuscript for publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors thank the people who participated in the four workshops and contributed valuable inputs for MAS-AI. Also, a special thanks to Mette B. Horup who provided very valuable help with the use of SurveyXact when setting up and distributing the electronic questionnaires for the hybrid workshop.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2025.106127.

Data availability

We are in the process of publishing our raw data online for easy access. Until then, the workshop material used during the current study are available from the corresponding author on reasonable request.

References

[1] K. Kidholm, A.G. Ekeland, L.K. Jensen, J. Rasmussen, C.D. Pedersen, A. Bowes, et al., A model for assessment of telemedicine applications: mast, Int. J. Technol. Assess. Health Care 28 (1) (2012) 44–51.
[2] I. Fasterholdt, T. Kjølhede, M. Naghavi-Behzad, T. Schmidt, Q.T.S. Rautalammi, M. G. Hildebrandt, et al., Model for assessing the value of artificial intelligence in medical imaging (MAS-AI), Int. J. Technol. Assess. Health Care 38 (1) (2022) e74.
[3] R. Di Bidino, S. Daugbjerg, S.C. Papavero, I.H. Haraldsen, A. Cicchetti, D. Sacchini, Health technology assessment framework for artificial intelligence-based technologies, Int. J. Technol. Assess. Health Care 40 (1) (2024) e61.
[4] EDiHTA (European Digital Health Technology Assessment) Project. EDiHTA: Bridging Digital Health and Health Technology Assessment [Internet] place of publication unknown [Available from: https://edihta-project.eu/.
[5] J.F. Hammer, Aleksandra Herberg, S. Stephan, Teuteberg Frank. Navigating the complexity of evaluating artificial intelligence in healthcare. ECIS 2025 Proceedings. 2025;11.
[6] I. Fasterholdt, M. Naghavi-Behzad, B.S.B. Rasmussen, T. Kjølhede, M.M. Skjøth, M. G. Hildebrandt, et al., Value assessment of artificial intelligence in medical imaging: a scoping review, BMC Med. Imaging 22 (1) (2022) 187.
[7] C.-C. Hsu, B.A. Sandford, The Delphi technique: making sense of consensus, Pract. Assess. Res. Eval. 12 (2007) 10.
[8] K. Kidholm, L.K. Jensen, T. Kjølhede, E. Nielsen, M.B. Horup, Validity of the model for assessment of telemedicine: a Delphi study, J. Telemed. Telecare 24 (2) (2018) 118–125.
[9] SurveyXact: Ramboll Management Consulting; [Available from: https://www.surveyxact.com.
[10] Microsoft Excel: Microsoft Corporation; [Available from: https://office.microsoft.com/excel.
[11] P.J. Green. The content of a college-level outdoor leadership course for land-based outdoor pursuits in the Pacific Northwest a Delphi consensus: Thesis (Ed. D.)–University of Oregon, 1981. 1982..
[12] The Regional Committees on Health Research Ethics for Southern Denmark. Forsøgstyper uden anmeldelsespligt 2016 [Available from: https://komite.regionsyddanmark.dk/wm428123.
[13] R. Best, J. Campbell, M. Edwards, L. Cline, An overview of the Delphi method’s origin, modifications, and use an overview of the Delphi method’s origin, modifications, and use to augment instrument development and data collection: a research note, J. Int. Agric Extens. Educ. 32 (2025).
[14] S. Pan, M. Vega, A.J. Vella, B.H. Archer, G. Parlett, A mini-Delphi approach: an improvement on single round techniques, Prog. Tour. Hosp. Res. 2 (1996) 27–39.
[15] V. Bellini, F. Calabrò, E. Bignami, T.M. Haja, I. Fasterholdt, B.S. Rasmussen, et al., Applying the model for assessing the value of AI (MAS-AI) framework to organizational AI: a case study of surgical scheduling assessment in Italy, J. Med. Syst. 49 (1) (2025) 108.
[16] P.T. Ørskov, M.B. Lichtenstein, M.T. Ernst, I. Fasterholdt, A.F. Matthiesen, M. Scirea, et al., Cognitive behavioral therapy with adaptive virtual reality exposure vs. cognitive behavioral therapy with in vivo exposure in the treatment of social anxiety disorder: a study protocol for a randomized controlled trial, Front. Psych. 13 (2022) 2022.
[17] CAI-X. EUCANSCREEN [Internet] 2024 (cited 2025 Sep 24) [Available from: https://cai-x.com/projects/current-projects/eucanscreen.