HTA-mAIx 2.0 as a Decision Support System to support evaluation and adoption of medical artificial intelligence (AI): A mixed-method approach

By

Irene de Bruin

A Master's Thesis

Submitted to the Department of Science & Technology

Master program Health Sciences

University of Twente

In fulfilment of the requirements for a Master of Science (MSc) Degree

August 2025

Abstract

The availability of medical artificial intelligence (AI) applications grew exponentially during the last five years. However, little of these medical AI applications are implemented in the Dutch healthcare sector, specifically Dutch academic hospitals and top clinical hospitals. Several factors were identified in other studies and in previous research that could explain the low AI adoption (e.g. AI literacy, privacy concerns and algorithmic biases) and the difficulties in decision-making. This indicated the need for a comprehensive, evidence-based, and applicable AI evaluation framework. A previous study in 2024 resulted in the first concept of the Health Technology Assessment for medical AI applications (HTA-mAIx), which was further researched in this study to design the HTA-mAIx 2.0 as a Decision Support System (DSS). Therefore, this study's objectives were 1) the development of the DSS using multidisciplinary input, and 2) test the DSS' feasibility in terms of usability, initial acceptance, and perceived effects on AI literacy.

The overarching study design to address the objectives consisted of three phases: 1.1) moderated in-person card sorting, 1.2) unmoderated digital card sorting, 2) digital prototyping using a four-step approach, and 3) in-person feasibility and usability testing. The target population were identified using a power-interest grid which distinguished four key stakeholder groups: Healthcare staff, Decision-makers in Dutch academic and top clinical hospitals, IT staff including AI disciplines, and Medical technology specialists.

The results found that the relevant evaluation topics for medical AI for in the DSS were: 1) functional needs, 2) properties of medical AI, 3) decision-making context, 4) assessments, and 5) advice and recommendations. The feasibility of the digital prototype of the HTA-mAIx 2.0 as a DSS was perceived by the target population as an essential aid for hospitals to gain better understanding of medical AI, improve informed decision-making, and to cope with future AI-developments.

Table of Contents

1. Introduction

The last five years the availability of medical artificial intelligence (AI) applications for the healthcare sector grew exponentially, specifically for administrative support (e.g. Autoscriber (1) and Whisper (2)) and prediction models for efficiency in staff rostering and diagnostic support (e.g. Periscope (3)and Zeno AI (4)) (5, 6). AI is an umbrella term for different combinations of algorithms that try to imitate specific human intelligent behaviour involving logic and pattern-based thinking (7-9). AI has potential to be very useful aid in healthcare to diminish the issues pertaining the capacity of resources, accessibility, quality and improving the affordability of qualitative care to everyone (5). For example, a survey study found that generative AI technologies such as ambient listening to support healthcare staff in administrative tasks increased the reporting quality from 41.9% to 71%, improved healthcare staff's overall well-being by 32.3%, and 58.1% mentioned it to increase their productivity (10). Pantanowitz et al. (11) performed a blinded clinical validation and deployment in routine clinical practice of an AI-based algorithm to evaluate prostate core needle biopsies for probability of cancer, perineural invasion and calculation of cancer percentage. They found, amongst other findings, an area under the receiving operating characteristic curve (AUC) of 0.997 (95% CI 0.995 to 0.998) for accuracy of detecting cancer in the CNBs of internal test sets and external validation set (11). However, the AI-monitor of the Dutch Central Bureau for the Statistics (CBS) (12) showed that merely 18.4% of healthcare organisations in the Netherlands used at least one AI-technology in 2024.

In 2022 and 2024, the Dutch Association of Hospitals (NVZ) in collaboration with the Dutch integral care agreement (IZA) and the Dutch Federation of University Medical Centres (NFU) established several position papers indicating the need to improve healthcare efficiency and quality by increasing data-driven processes using medical AI (13-16). They assigned Dutch academic hospitals and top clinical hospitals to test and decide which medical AI applications add most value, because these hospitals are at the forefront of reshaping healthcare (13, 17).

The government provided the monetary support by instructing grants-organisations such as ZonMW (18) to divide 3.5 million euros to pilottest or scale up medical AI (14, 18). Regardless, academic and top clinical hospitals are not investing or implementing medical AI at the preferred pace of the NVZ, IZA and NFU.

1.1 Factors influencing AI adoption

The CBS statistics about reasons for not investing and implementing medical AI applications, found that lack of experience was the most important reason at 81.6%, followed by privacy concerns at 54.3% and regulatory consequences at 51.9% (12). Multiple studies researching decision-making about medical AI applications in academic and top clinical hospitals confirmed these reasons to be most relevant (5, 9, 17, 19-21). Esmaeilzadeh (5) and other researchers uncovered that the low adoption rate could be explained by seven overarching AI deployment challenges: 1) the inconsistency in time to production to experience the true benefits of medical AI applications due to the self-learning characteristics (time horizon) (5, 17), 2) trust issues due to lack of explainability in the internal working processes of medical AI applications (black box transparency) (5, 17, 19), 3) concerns in ensuring privacy and security due to the fine line of using patient data to improve health outcomes (e.g. quality of life) and protecting patients' rights and privacy (5, 9, 10, 21), 4) algorithmic biases due to lack of transparency in the development process (e.g. problem formulation, training data, testing data, model design, etcetera) of medical AI applications (5), 5) scarcity in data due to privacy barriers and poor data quality assurance (5, 20), 6) medical AI for general usage in hospitals is currently not plausible due to lack of data sharing standards (5, 11), and 7) potential pitfalls in the 'human-AI-collaboration' due to, for example, risks of becoming too dependent on medical AI and lack of awareness in accountability and responsibility of using medical AI (5, 21).

Research by Zary (17) and Kimiafar et al. (22) revealed that AI-literacy also affects the adoption of medical AI applications in hospitals. AI literacy reflects the knowledge, skills and experience someone has about AI (22). It primarily includes knowledge about the (internal) workings of AI applications, risk awareness of using AI applications, and evaluating AI and AI-generated content (17). A study by Liu et al. (19) found that the large amount of necessary knowledge about medical AI, anxiety due to lack of trust, and avoidance due to the perceived complexity of medical AI have a significant impact on AI literacy and adoption of medical AI applications.

These findings imply that Dutch academic and top clinical hospitals have multiple reasons (e.g. limited AI literacy and lack of decision-making support) to stop them implementing and using or studying medical AI, despite of financial incentives (12, 15). Therefore, a comprehensive framework to support decision-making in medical AI applications is needed, which also addresses the challenges of medical AI evaluation (e.g. tailored time horizon and data quality) (5).

1.2 Health Technology Assessment for medical AI applications: the HTA-mAIx

A HTA is a large, incremental framework that systematically and multidisciplinary carries out several assessments to predict the added value of medical technologies as accurately as possible (23). The assessments include systematic reviews, meta-analyses, clinical effectiveness and health economic modelling, and are meant to support decision-making about medical technologies (23). However, several studies found the standard HTAs to be inappropriate for assessing the value of medical AI applications to support decision-making (24). It does not incorporate the assessments and evaluation criteria of the essential characteristics of medical AI, for example, the adaptiveness and the effectiveness of medical AI vary depending on the algorithms, models, the development process and AI-literacy of the users (5, 24-29). This implies that the time horizon is complex to estimate for health economic evaluations (e.g. Markov modelling) (5, 30).

Multiple researchers and governmental bodies tried to develop a more suitable and comprehensive HTA framework for medical AI applications, however, were not entirely able to incorporate the essential medical AI characteristics (5, 11, 12, 29) due to various reasons. Fasterholdt et al. (31, 32) developed the Model for ASsessing the value of Artificial Intelligence (MAS-AI) with the intention to be applicable for all types of AI. Their approach consisted of first identifying the characteristics and properties of each type of medical AI starting with AI in visual diagnostics (radiology). During this process, they did not include all necessary perspectives (e.g. Data Scientists and the radiologists) and ended up with a framework that was only applicable to evaluate medical AI for radiology (31). The European Union network for HTA (EUnetHTA) developed two types of HTA frameworks: HTA Core Model (33) and the Next Generation HTA (HTx) (34, 35). These frameworks primarily included governmental and theoretical perspectives, and were focused on complex medical technologies, which also included robotics and eHealth (33, 34). Research of Grutters et al (36), Jiu et al. (29) and Van Haesendonck et al. (37) emphasize the importance of including a multidisciplinary approach in designing HTA frameworks for complex technologies, including medical AI.

This resulted in the creation of the first concept HTA-mAIx (Figure 1) (25). The content was collected through 1) literature research on the relevant evaluation aspects and criteria for medical AI applications, 2) semi-structured interviews with experts in the field of HTA, medical AI and regulations of medical AI, and 3) questionnaire to validate the identified evaluation aspects and criteria (25). Despite the multidisciplinary approach, the applicability of the first concept HTA-mAIx was considered too abstract by the researcher (25) and intended users to use in Dutch academic and top clinical hospitals. Recommendations for follow-up studies should verify the identified evaluation criteria, aspects and the first concept by key-stakeholder groups (25). Moreover, the effectiveness, usability and ease of use by the intended users should be iteratively tested and monitored through

multidisciplinary settings and User Experience methods (e.g. thinking aloud or cognitive walkthroughs) (25). Therefore, this study intended to follow these recommendations and design a prototype of the HTA-mAIx 2.0 as a Decision Support System (DSS).

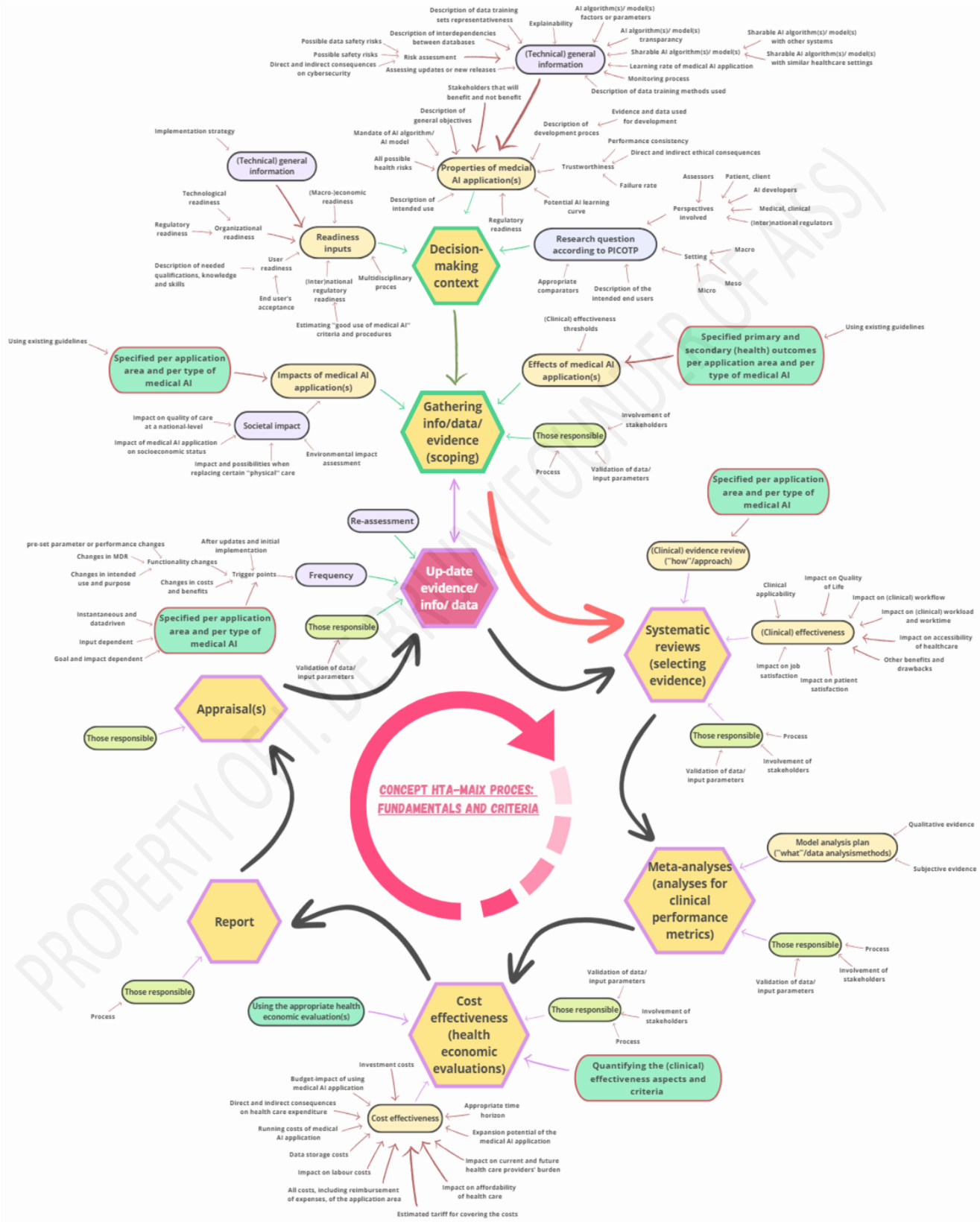


Figure 1. The first concept HTA-mAIx (25).

1.3 CeHReS roadmap

The first concept HTA-mAIx was designed based on the Human-Centred Design (HCD) principles to ensure the multidisciplinary approach (24, 25). To co-design a digital prototype of the HTA-mAIx as a Decision Support System (DSS) in this study and ensure a evidence-based structure for future research, the Centre for eHealth and Well-being Research (CeHReS) roadmap would be most suitable (38, 39). The CeHReS roadmap is an overarching framework that incorporates multiple models, including HCD, and other frameworks focusing on interdisciplinary and holistic development, implementation, and evaluation processes (38). The CeHReS Roadmap is made up of six main phases: 1) contextual inquiry, 2) value specification, 3) design, 4) operationalisation, 5) summative evaluation, and 6) formative evaluation (38). Contextual inquiry identifies the issues that need to be solved and the contextual factors, e.g. roles and activities of involved stakeholders (38). In the value specification phase, the identified issues and contextual factors are translated into requirements and prioritised by involved stakeholders to understand their expectations of the eHealth technology (38). Therefore, these requirements are the primary input to create prototypes, including the software, of the eHealth technology during the design phase.

The value specification phase and the design phase were the next steps to further develop the first concept HTA-mAIx as a DSS, due to the importance of identifying the mental models of relevant stakeholders to design the Information Architecture (IA) before prototyping (38, 40-42). Mental models reflect how users' interpret a system and how it influences their interactions with the system (41). This provides the needed input to design the flow of information that is presented to the users and indicates how to interact with the system, the IA (41, 42).

1.4 Usability

The second objective is to test the usability of the HTA-mAIx 2.0 as a DSS. In the context of this study, usability is defined according to Jakob Nielsen: *''a **quality attribute** that assesses how easy user interfaces are to use''* (43). It comprises at least five components to indicate the user experience of software/internet technologies: 1) learnability to indicate the ease of use, 2) efficiency of performing tasks, 3) memorability indicating the recognisability of how to use the system, 4) errors in using the system, and 5) satisfaction of using the system (40, 43). An important attribute to usability is utility, which shows how the intended users experience the functionality of the technology (43). To estimate the perceived usefulness of the prototype DSS, the usability is identified by the 10 usability heuristics of Nielsen (24, 40) and the expected utility is measured using the Non-adoption, abandonment and challenges to scale-up, spread and sustainability (NASSS) framework (38, 44, 45).

1.5 Non-adoption, abandonment and challenges to scale-up, spread and sustainability (NASSS) framework

As previously discussed, the third objective is to examine the utility of the HTA-mAIx 2.0 as a DSS. This study used the NASSS framework to understand the utility, because it aims to identify the factors that affect the successfulness of implementing technology-supported healthcare programs (38, 45, 46). This framework evaluates technologies as part of complex systems in healthcare, such as hospitals, and serves as an extension of the Diffusion of Innovation (DoI) framework (38). It consists of seven domains that each indicate a contextual factor in a complex system: 1) the illness or condition to indicate the level of complexity of the implementation context, 2) the technology to indicate the type of technology and what is necessary to use it, 3) the value proposition indicates what the values are to various involved stakeholders in the context, 4) the intended adopters focuses on the perceived intention to use and usefulness of the intended adopters, 5) the organisation addressing the capacity to innovate on different organisational levels, 6) the wider system addresses the external factors that influence the usage of the technology by the intended users, and 7) embedding and adoption over time indicating how future-proof the technology is (38, 45).

1.5 Aim and objectives

This study seeks to further research the content and redesign the first concept HTA-mAIx into the HTA-mAIx 2.0 as a DSS for top clinical and academic hospitals in the Netherlands. This DSS aims to support these hospitals in decision-making about medical AI to improve implementation, usage in practice and to cope with current and future AI developments in healthcare. The objectives of this study are:

1) Develop a prototype of the HTA-mAIx 2.0 as a DSS in multidisciplinary group settings;
2) Test the DSS' feasibility in terms of usability, initial acceptance, and perceived effects on AI literacy.

2. Methodology

The overarching study design was based on the CeHReS roadmap (38) and consisted of three phases: 1.1) moderated in-person card sorting, 1.2) unmoderated digital card sorting, 2) digital prototyping, and 3) in-person feasibility and usability testing. Phases 1.1 and 1.2 addressed the verifications of the evaluation criteria and categories in the first concept HTA-mAIx to develop the HTA-mAIx 2.0. Next, this redesigned HTA-mAIx was converted into a DSS based on a four-step approach in phase 2. In phase 3, the second objective was studied to determine the DSS' feasibility in terms of usability, initial acceptance and perceived effects on AI literacy. An interim conclusion is provided between the first phases and phase 2. Figure 2 shows the overarching study design in more detail.

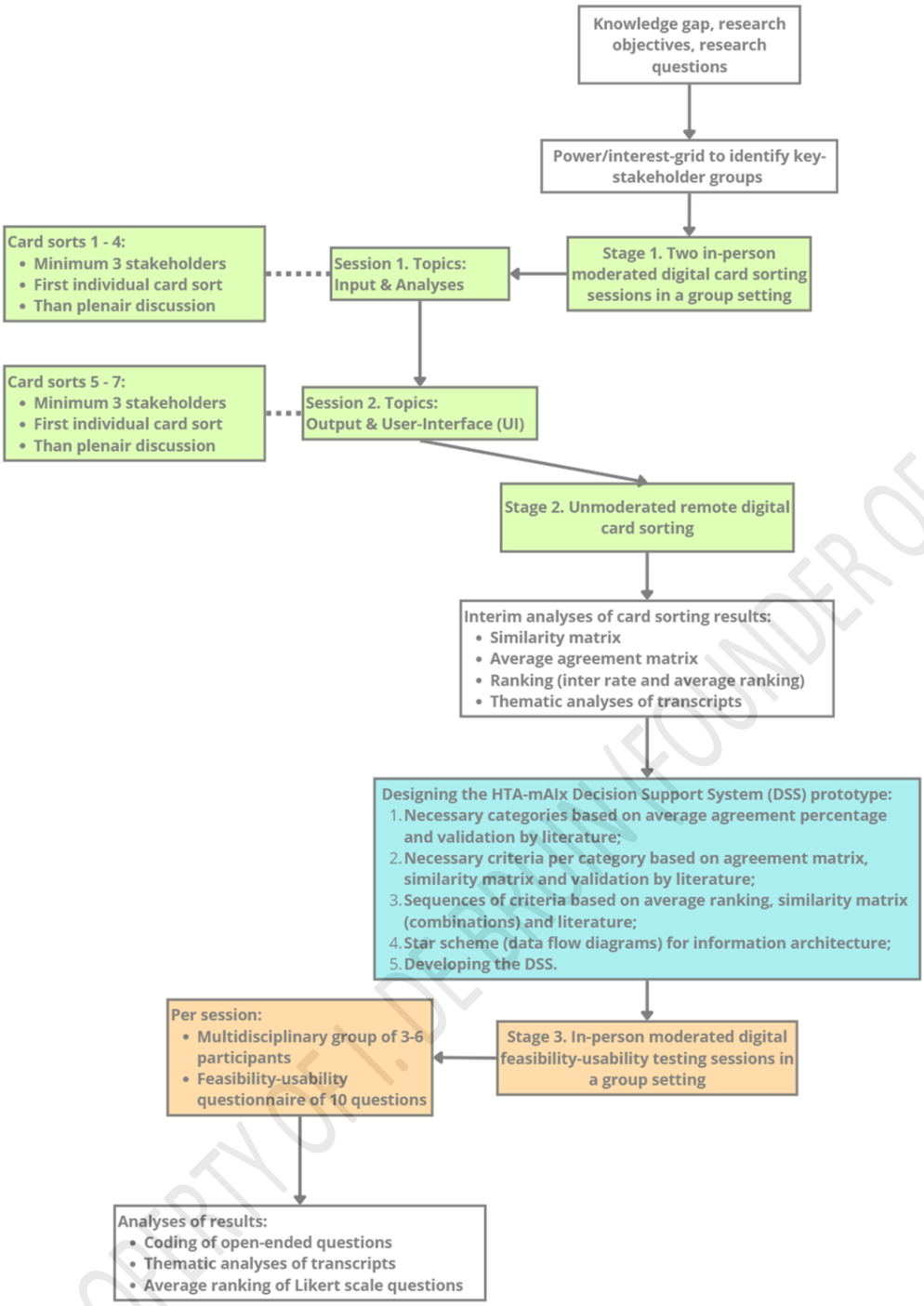


Figure 2. Overarching study design.

3. Phase 1.1 Value specification – in-person card sorting

3.1 Method

A qualitative card sorting task was used to explore an appropriate IA of the DSS that would comply with Human-Computer-Interaction (HCI) principles and the users' mental models of

the decision-making process surrounding medical AI (38, 41, 42). According to Chan (41) and Tankala (47) mental models can identify preferences in clusters and priorities by moderated in-person card sorting sessions. In-person card sorting is a user-centred design activity that involves sorting and prioritising cards to categories by the relevant stakeholders and plenary discussing their thought processes to gain insight into their mental models (39, 42, 48). The cards represented the evaluation criteria and the categories represented the evaluation subtopics.

Setting

Two in-person card sorting sessions took place in conference rooms at the Martini hospital in Groningen (NL): session 1 on Monday 26th of May between 8 – 9.30 am, and session 2 on Monday 26th of May between 4 – 5 pm. This hospital is a Dutch top clinical hospital that provides secondary healthcare services. The number of hospital beds is approximately 5000, making it a medium-sized hospital.

Participants

The target population were identified using the power-interest grid of the previous study by Bruin (25, 38). Four key stakeholder groups were distinguished:

- Healthcare staff of the emergency department (ED), surgery department (SD), radiology department (RD) and pathology department (PD) or internal medicine department (IMD) (3, 5, 7, 29);
- Decision-makers in Dutch academic or top clinical hospitals (29, 37, 49);
- IT staff including: data architects, privacy & security staff, IT advisors, AI specialists, computational pathologists or data scientists (50);
- Medical technology specialists, specifically clinical physicists (29).

Participants from these key-stakeholder groups were eligible if they met the following inclusion and exclusion criteria:

Inclusion	Exclusion
Either affected or involved in the evaluation and decision-making about medical artificial intelligence (AI).	Healthcare organisations other than academic and top clinical hospitals.
Work for Dutch academic and/or top clinical hospital.	Academic and top clinical hospitals not located in the Netherlands.
Work experience for more than 6 months in a profession that is representative for at least one key stakeholder group.	First-line hospitals.
	No experience in healthcare and evaluating medical AI.

The healthcare staff of the ED, SD, RD, and PD or IMD in Dutch academic or top clinical hospitals were selected because they are exposed most often to efficiency and capacity problems (28, 51, 52). These departments work according to guidelines and data-driven procedures due to the continuous collection of real-time, accurate patient-specific data (e.g. heartrate, blood pressure) (51, 52).

Twenty stakeholders were invited purposefully for both sessions by email and telephonic based on the inclusion/exclusion criteria. Seven stakeholders accepted the invitation for session 1, three of them actually finished session 1. Four stakeholders accepted the invitation for session 2, one of them actually finished session 2.

Materials

The moderated in-person card sorting was conducted in two group settings using Maze.co to automatically collect the data (53). Maze.co is a computer program for user research to support product development by analysing patterns and presenting it in agreement matrices and similarity matrices. It helps researchers to make user insights and mental models available through a range of different user experience studies, such as card sorting (53).

The card sorting study was in Dutch and consisted of seven sequential card sorts: 1) functional needs, 2) properties of medical AI, 3) organisational impact, 4) decision-making context, 5) effectiveness and impact assessments, 6) the advice and reasonings, and 7) report layout. Table 1 shows the number of categories and number of cards per card sort. Figure 3 shows an example of how a card sort was presented in Maze.co.

Card sort	Definition	Number of categories (N)	Number of cards (N)
1. Functional needs	Identifying the core of the problem and the requirements to solve the problem.	3	7
2. Properties of medical AI	Gaining insight into the characteristics and properties of the medical AI.	3	21
3. Organisational impact	Gathering the information to estimate the impact of the medical AI on the relevant aspects of the healthcare organisation.	6	21
4. Decision-making context	Identifying what specific evaluation criteria decision-makers find most relevant.	*Open card sort	20
5. Effectiveness and impact assessments	The methods and units of measurement that are	3	15

	required to evaluate the effectiveness and impact based on the previous card sorts.		
6. The advice and reasonings	The content and visualisations of the output.	3	15
7. Report layout	The content and layout of the report.	2	9

Table 1. Categories and cards per card sort.

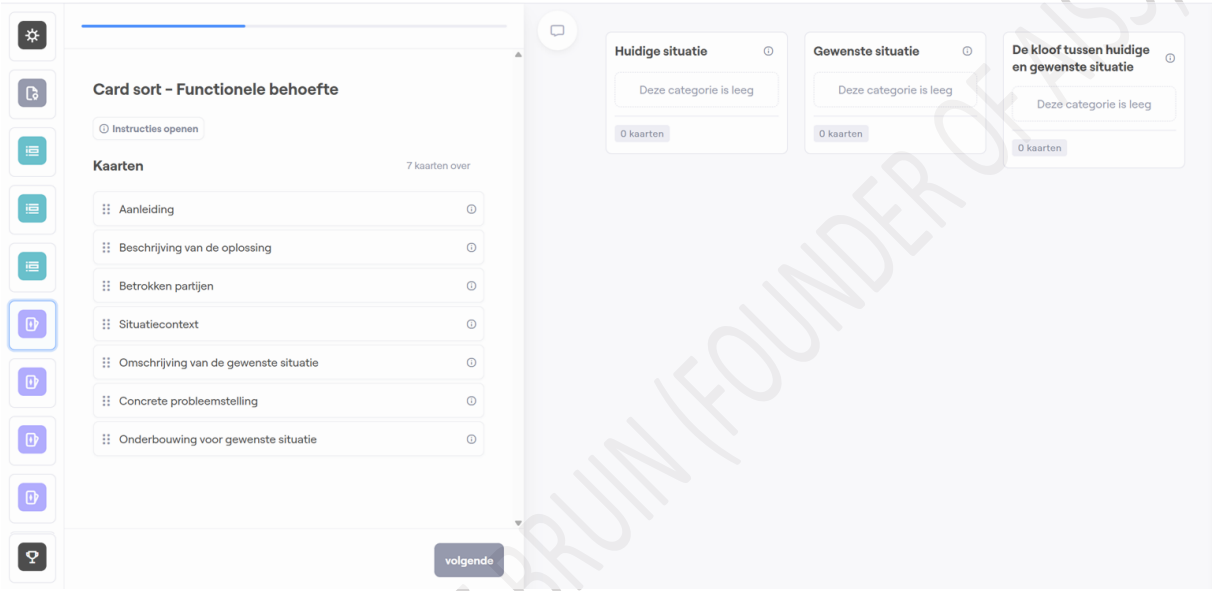


Figure 3. Example of a card sort presenting the cards with explanations (''i''-icons per card) and the categories.

The plenary group discussions were audio recorded using a Jabra microphone and Microsoft Teams on the researcher's laptop (54). Each participant was asked to bring their own laptop to gain access to the card sorting via Maze.co's website.

Procedure

To guide each moderated in-person card sorting session the researcher developed a playbook (Appendix A) to navigate the intervals between participants individually doing a card sort and the plenary discussions before going to the next card sort. Session 1 addressed the DSS categories and cards (evaluation criteria) related to input (card sorts 1 to 4). Session 2 addressed the assessment and output of the DSS (card sort 5 to 7).

Before the start of each session, the researcher send each participant an informed consent (IC) form to fill in and hand-in prior to the start of each session. Each participant gained accessibility to the card sorting task approximately 30 minutes before each session.

Each session began with an introduction explaining the aim of the card sorting and the structure of the HTA-mAIx 2.0, followed by an instruction addressing the procedure. Per card

sort, the participants were asked to first individually sort and prioritise the available cards on the left to the most suitable category on the right, according to their expertise. Next, the participants were asked to explain their answers by providing insights into their thought processes during the plenary discussion moderated and audio-recorded by the researcher. At the end of each interval, the researcher asked the participants their perception of how suitable the cards and the categories were for set card sort before going to the next card sort.

Analyses

The audio-recording of each session were automatically transcribed using the ''Transcribe'' function in Microsoft Teams. The transcripts were evaluated on accuracy by comparing the transcribed sessions to the audio-recording to filter out mistakes in the transcripts. This increased the reliability and validity of the analysis results (55, 56). The quantitative data (e.g. similarity matrix) was included in the data-analyses of the digital card sorting (phase 1.2) due to this card sorting task focused on understanding the participants' thought-processes (the 'why').

Next, each transcript was inductively analysed by separating the information into codes to identify main codes and subcodes per card sort that indicate the mental models of the participants (42, 48, 57). The researcher highlighted the quotes of participants that included relevant information. These quotes were used to identify the codes and grouped together to determine the main code. The main codes were paired to a card sort and were visualised in coding schemes (Tables 2 to 8 in Appendix B).

3.2 Results

Session 1 included three participants: two IT staff and one medical technology specialist. Session 2 included one IT staff, who also finished session 1. Each participant had more than 12 months experience and is influenced by medical AI applications in their profession. The participants work in a Dutch top clinical hospital. Session 1 lasted 81,5 minutes and session 2 took 49,1 minutes.

Functional needs

The analyses of the categories and evaluation criteria showed that all three participants considered the evaluation criteria to be inclusive and informative to understand the functional needs (Table 2). The participants suggested to incorporate the ''Gap''-category within the ''Preferred situation''-category as they were interpreted as having similar objectives. Two participants (IT staff) mentioned the need to provide clearer descriptions of the multidisciplinary approach.

Properties of medical AI

The analysis of the transcript of session 1 identified multiple patterns revealing the mental models of the participants (Table 3). The plenary discussion revealed that each participant

has a different interpretation of certain cards. This caused them to sort and prioritise the cards differently. For example, during the plenary discussion two participants (IT staff, Medical technology specialist) spurred a debate about the interpretation that some cards could be interpreted similarly. One participant (Medical technology specialist) states: *''Soms zie ik ook nog wat overlap, bijvoorbeeld data opslag en management en beschrijving van de informatiestromen en data uitwisselingen.''* Another participant (IT staff) replied: *''Nee, want data opslag en management gaat niet over het medische proces wat je wil gaan doen.''*

Two participants (IT staff) mentioned the importance to open the ''black-box'' and create transparency about the (patient) information that is required, including how it is processed by the supplier and the medical AI application.

Organisational impact

Analysis of the organisational impact revealed similar patterns in the participants' thought processes to the previous card sorts (Table 4). All stakeholders mentioned certain cards to have a shared allocation of categories, depending on how these cards are interpreted. For example, the card ''Hardware and software'' (Medical technology specialist): *'' Ik zie bijvoorbeeld benodigde hardware en software, ik denk natuurlijk als eerste aan technische gereedheid, Maar dat kan ook een verwachte financiële impact hebben, want blijkbaar heeft dat invloed op onze hele infrastructuur.''*

All participants considered the category ''Impact on partnerships'' too complex to include in the evaluation of medical AI. One stakeholder (Medical technology specialist) explained that the perspective is too broad and dependable on the willingness of multiple involved organisations (e.g. healthcare insurers, other regional hospitals) to form fair partnerships with equal say. Another participant (IT staff) pointed out that the range of influence a hospital has on involved organisations is strongly dependable on the financial considerations and the organisational readiness in terms of IT, workflows and workplace culture.

Decision-making context

This was an open card sort in which participants created their own categories of the cards they grouped together. Analysis of the transcript revealed similarities and differences in the participants' thought processes (Table 5). All three participants created four new categories that were differentiated similarly. All participants mentioned an 1) organisational vision and strategy category, 2) project value case of the medical AI category, 3) project implementation category, and 4) the medical AI application in practice category. One participant (IT staff) mentioned combining the financial category with the project value case category to represent the organisational readiness.

One participant (IT staff) explained that these categories were based on her perception of separating functional and operational criteria to provide an organisational perspective. Furthermore, two participants (IT staff) explained the project value case category as a means of estimating the real-world effectiveness of medical AI applications: *'' Kijk, we gooien het*

erin en dan wordt het een nieuwe standaard. Maar er wordt niet heel vaak echt een trial gedaan. Want We gaan theoretisch bedenken dat het beter is, dus praktisch ook beter en dat stukje mis ik.''

Two participants (IT staff, Medical technology specialist) suggested to combine the decision-making context and the organisational impact as these have similar objectives and require the same information.

Effectiveness and impact assessments

The identified patterns in the coding scheme showed multiple main codes and subcodes indicating how the participant perceived the categories and cards associated with effectiveness and impact assessments (Table 6). The participant (IT staff) mentioned that their understanding of each category dictated how they perceived its relevance in the evaluation of medical AI. The participant (IT staff) explained that the concept of health economic evaluations was too out of reach to comprehend: *'' Ik heb geen beeld wat die dingen precies betekenen, dus dan kan ik ze ook moeilijk scoren. Ik denk dat ze vooral allemaal gezondheid, economische iets doen, maar ik zou ze niet kunnen ranken.''*

The participant (IT staff) suggested that the Data Protection Privacy Assessment (DPIA) and the Impact Assessment Mensenrechten en Algoritmes (IAMA) are most relevant in the compliance assessment category. The participant (IT staff) explained that these assessments classify if the medical AI application is ethically and technically safe and secure to be used in Dutch hospitals. Furthermore, the participant (IT staff) mentioned certain cards to be more suitable in previous card sort topics, such as the cost-analyses often service as input for the assessment.

Advice and reasonings & report layout

Based on the analysis of session 2, the participant (IT staff) mentioned that the advice of the DSS should be a combination of visualisations and words to indicate the advice (Table 7). The three traffic-light colours (Red, Orange and Green) were considered to be the most appropriate and intuitive to present the advice. The reasonings to substantiate the advice should be written in bulletpoints, starting with the most important reasons. The participant (IT staff) emphasised the importance to base the order of the reasonings on answering the 'why': *'' Als je bijvoorbeeld een negatief advies ergens over zou schrijven, zeg maar, dan is de reden waarom je het negatieve advies schrijft, is het belangrijkste.''*

Furthermore, the participant (IT staff) suggested that the report should address the most relevant information and assessment results to substantiate the advice and potential recommendations for improvement (Table 8). The order of recommendations should be based on providing a weight to each criterium to indicate which recommendation is mandatory and/or needs to be prioritised before a final decision could be made.

4. Phase 1.2 Value specification – digital card sorting

4.1 Method

To address the relevant content and information flows of the IA to create the DSS, a quantitative approach was used (42, 58). The quantitative data were collected through unmoderated digital card sorting in Maze.co (53) to identify patterns and relevant criteria for the evaluation of medical AI (54). The patterns and relevant criteria were complementary to the mental models identified by the analysis of the two moderated in-person card sorting sessions (42, 54).

Participants

Participants had to meet the same criteria as before (phase 1.1), with the additional criterium that they could not have participated in the moderated in-person card sorting session.

Through convenience sampling thirty stakeholders were invited by email or face-to-face to fill-in the digital card sorting in Maze.co. Six stakeholders were purposefully invited by the researcher via LinkedIn between May 28th and June 9th. Eighteen of the thirty-six invited stakeholders accepted the invitation, seven of them actually finished the study.

Materials and procedure

The stakeholders that accepted the invitation were send the website link to access the digital card sorting study in Maze.co, which was in Dutch to adjust to the target population. In addition to the card sorting task in phase 1.1, participants were asked to sign a digital IC form to ensure their privacy and three background questions to identify the representativeness of the four key-stakeholder groups prior to the same seven card sorts. The three background questions were based on the inclusion/exclusion criteria: 1) their profession, 2) their months of work experience, and 3) affected by decision-making about medical AI in their profession. After, the same procedure of instructions per card sort and conditions to finish the digital card sorting study was applied as in phase 1.1.

The participants were able to fill-in the digital card sorting between May 28th and June 10th. The digital card sorting study was pseudonymous and took approximately 30 minutes to complete.

Data-analyses

The respondents' answers were collected in Maze.co and automatically analysed (53). The filled-in card sorts of the four participants in the moderated in-person card sorting (phase 1.1) were also included in the data-analyses, because the same card sorting study in Maze.co was used and this study mainly focused on collecting the quantitative data.

First, the relevance of the categories per card sort was calculated using the interrater reliability (average agreement, %) (56, 59). The interrater reliability per card indicates the percentage of participants who sorted a card to a certain category, and the interrater reliability per category shows the average agreement between respondents of all cards sorted to that category (59).

Next, the card-category combinations were calculated and presented in agreement matrices per card sort. This identified which cards were sorted the most to a certain category by the respondents (56, 59, 60). The percentages and colour indicate the average agreement of the respondents to sort a card to a category (59, 60). The higher the percentage, the darker the colour, and the stronger the card-category combination (59, 60).

Lastly, the preferred card-combinations and order of occurrence based on their perceived priority by the respondents per card sort, were calculated and visualised in similarity matrices (59, 60). The similarity matrix weighs the relationship between pairs of cards based on how the participants ranked each card sorted to a certain category (59, 60). This revealed the percentage which cards were most often clustered together (59, 60). The higher the percentage, the darker the colour, and the stronger the relationship between a pair of cards (59, 60).

4.2 Results

Twelve participants started the digital card sorting, eight of them completed the digital card sorting study. The two participants who were not able to complete this study were representative of the key stakeholder group 'healthcare staff'. They explained that the card topics, cards and categories after the first card sort became too complex and it was time-consuming to try to understand it.

Table 9 shows that all four key stakeholder groups were represented by at least one participant. All participants had at least three months work experience and are affected by medical AI in their work/profession.

Characteristic	Digital card sorting – n(%)	In-person card sorting – n(%)
Key stakeholder group		
- *Healthcare staff*	2(25)	1(25)
- *Decision-making*	1(13)	0(0)
- *IT staff*	2(25)	3(75)
- *Medical technology specialists*	1(13)	0(0)
- *Other*	2(25)	0(0)
Work experience		
- *< 3 months*	0(0)	0(0)
- *3 - 6 months*	1(13)	0(0)
- *6 - 9 months*	0(0)	0(0)
- *9 - 12 months*	1(13)	0(0)
- *> 12 months*	6(75)	4(100)
Affected by medical AI		
- *Yes*	8(100)	4(100)
- *No*	0(0)	0(0)
- *I don't know*	0(0)	0(0)
- *Other*	0(0)	0(0)
Total	8	4

Table 9. Participant characteristics (n = 12).

Functional needs

The interrater reliability (%) was calculated to identify the relevant categories to estimate the functional needs in the evaluation of medical AI, according to the participants. The categories 'current situation' (41%) and 'preferred situation' (42%) were found to be the most relevant, and the category 'gap between current and preferred situation' (20%) was found not to be relevant. This was also revealed in the agreement matrix (Figure 4). The card 'motive' appeared to be most relevant in the category 'current situation'. The most relevant cards in the category 'preferred situation' were 'substantiation of preferred situation' and 'description of preferred situation'. The similarity matrix reveals that these two cards were paired together most often, indicating their relationship and priority to estimate the functional needs.

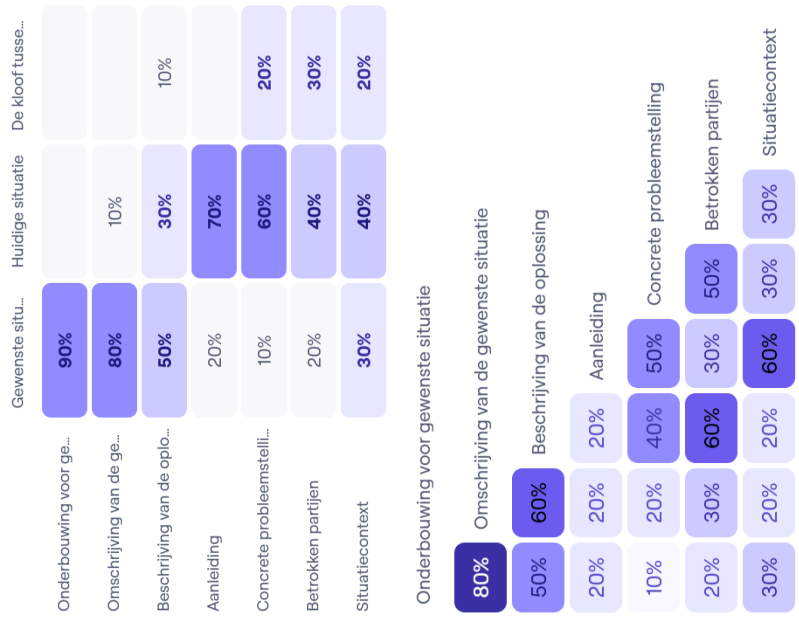
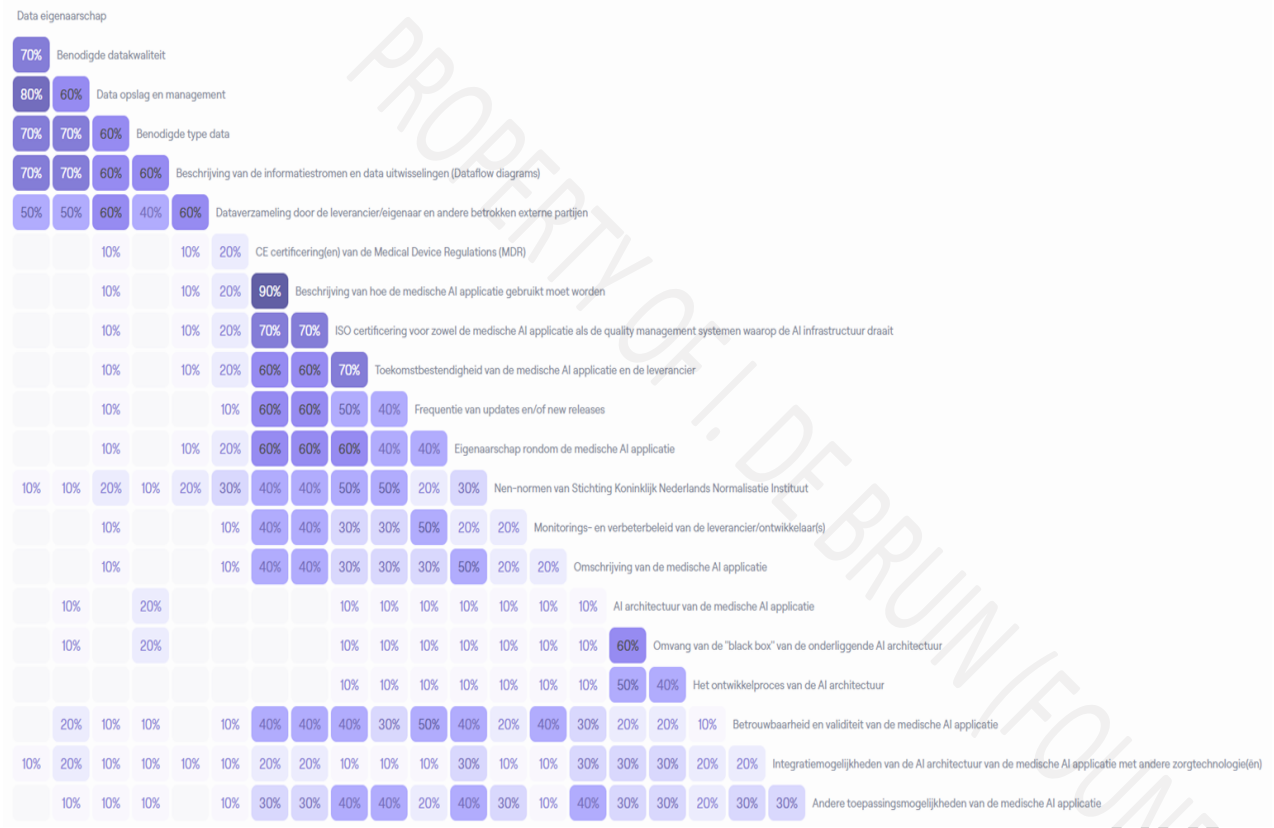


Figure 4. The agreement matrix (left) and similarity matrix (right) of functional needs (n = 12).

Properties of medical AI

The category ‘data characteristics’ had an interrater reliability of 57.5%, indicating that participants perceived this category as most relevant to estimate the properties of medical AI to support the evaluation and decision-making process. The second most relevant category was ‘product characteristics’ (45.3%) followed by ‘AI characteristics’ (27.3%). However, the agreement matrix (Figure 5) shows that over 60% of participants agreed that each category needed to contain certain cards which were only suitable in that particular category. The similarity matrix revealed that the card ‘CE certifications of MDR’ was paired most often to the card ‘description of how to use the medical AI application’ (90%). The card ‘AI architecture’ was paired the least to other cards, which does not reflect its perceived relevance considering the agreement matrix (70%).

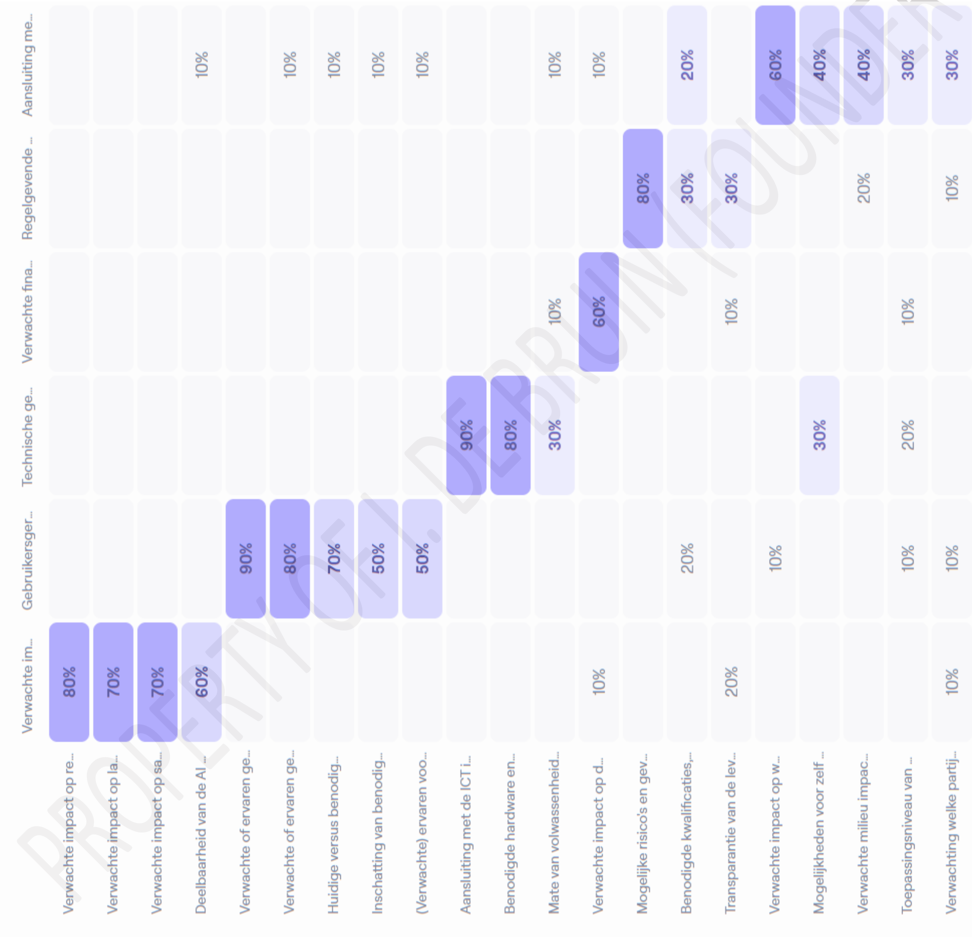


	Data karakteri...	Product karakt...	AI karakteristi...
Data eigenaarschap	**90%**		
Benodigde datakwaliteit	**80%**		**20%**
Data opslag en manage...	**80%**	10%	
Benodigde type data	**70%**		**20%**
Beschrijving van de infor...	**70%**	10%	
Dataverzameling door d...	**50%**	20%	
CE certificering(en) van ...		**90%**	
Beschrijving van hoe de ...		**90%**	
ISO certificering voor zo...		**70%**	**20%**
Toekomstbestendigheid...		**60%**	10%
Frequentie van updates ...		**60%**	10%
Eigenaarschap rondom ...		**60%**	**20%**
Nen-normen van Stichti...	10%	**40%**	10%
Monitorings- en verbete...		**40%**	10%
Omschrijving van de me...		**40%**	10%
AI architectuur van de m...			**70%**
Omvang van de "black b...			**60%**
Het ontwikkelproces va...			**50%**
Betrouwbaarheid en vali...		**40%**	**40%**
Integratiemogelijkhede...	10%	20%	**30%**
Andere toepassingsmog...		**30%**	**30%**

Figure 5. The agreement matrix (right) and similarity matrix (left) of properties of medical AI (n = 12).

Organisational impact

Out of the six categories to estimate the organisational impact of medical AI application, the category ‘technical readiness’ had the highest interrater reliability of 50%. The category ‘connection to the organisation’s vision and strategy’ had the lowest interrater reliability of 22.3%, indicating this category to be the least relevant. However, similar to the properties of medical AI, the agreement matrix (Figure 6) indicates differently. It shows here too that over 60% of the participant agreed that certain cards could only be sorted to a particular category. The similarity matrix reveals that the participants perceived most cards could not be paired to each other, except the cards about partnerships, the ease of use, AI literacy, and needed hardware and software.



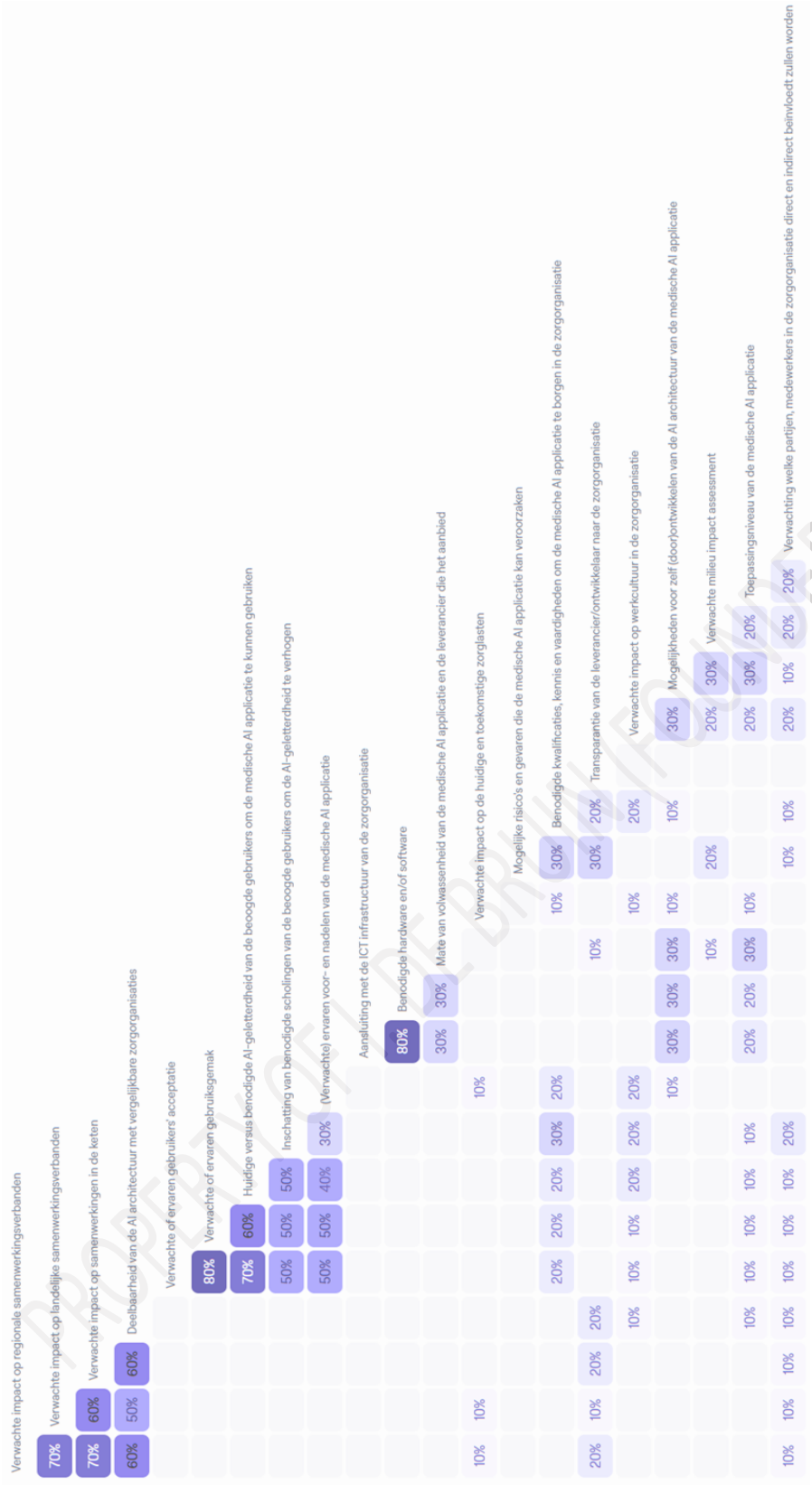


Figure 6. The agreement matrix (right) and similarity matrix (left) of organisational impact (n = 12).

Decision-making context

This was an open card sort in which participants were asked to first sort and prioritise the cards in clusters, after which they were asked to name each category (cluster). In total 34 categories were created by the participants. Four overarching evaluation topics were identified: 1) the organisational readiness, 2) the project value case, 3) Risk analyses, and 4) implementation assessments.

The similarity matrix (Figure 7) shows that the cards 'Life Years gained' and 'perceived clinical applicability' were clustered most often (70%) compared to the other cards. Furthermore, the similarity matrix identified multiple patterns in perceived card-combinations to estimate the decision-making context to evaluate medical AI.

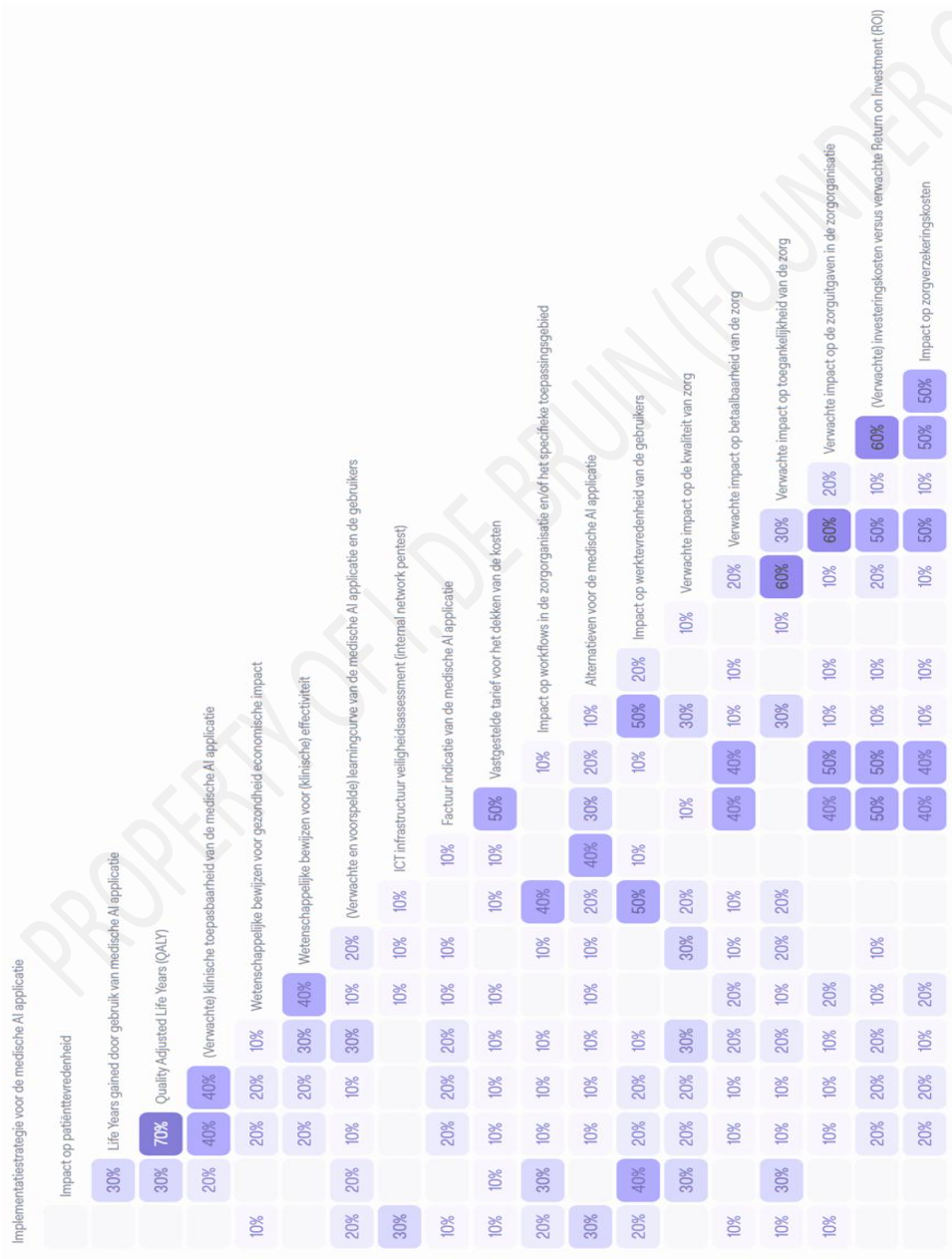
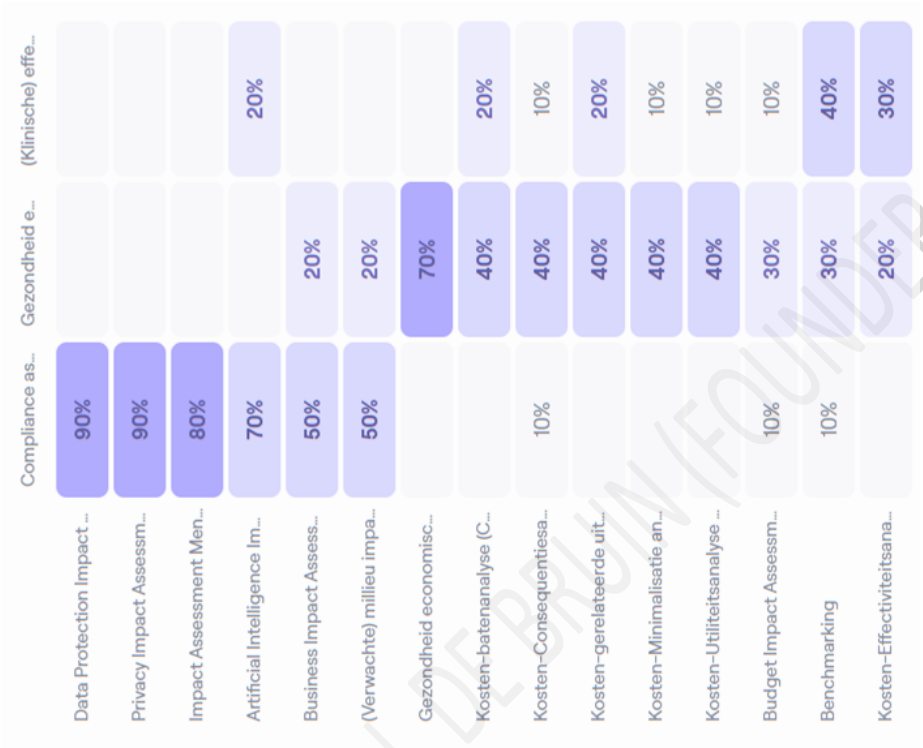


Figure 7. The similarity matrix of decision-making context (n = 12).

Effectiveness and impact assessments

The interrater reliability of the category ‘compliance assessments’ was the highest (51.1%), followed by the category ‘health economic evaluations’ (35.5%). The category ‘(clinical) effectiveness’ was perceived as irrelevant by the participants, having an interrater reliability of 18.9%. This was also revealed in the agreement matrix (Figure 8).

The similarity matrix revealed multiple patterns in pairs of combinations perceived by which cards were clustered and ranked together most often by the participants.



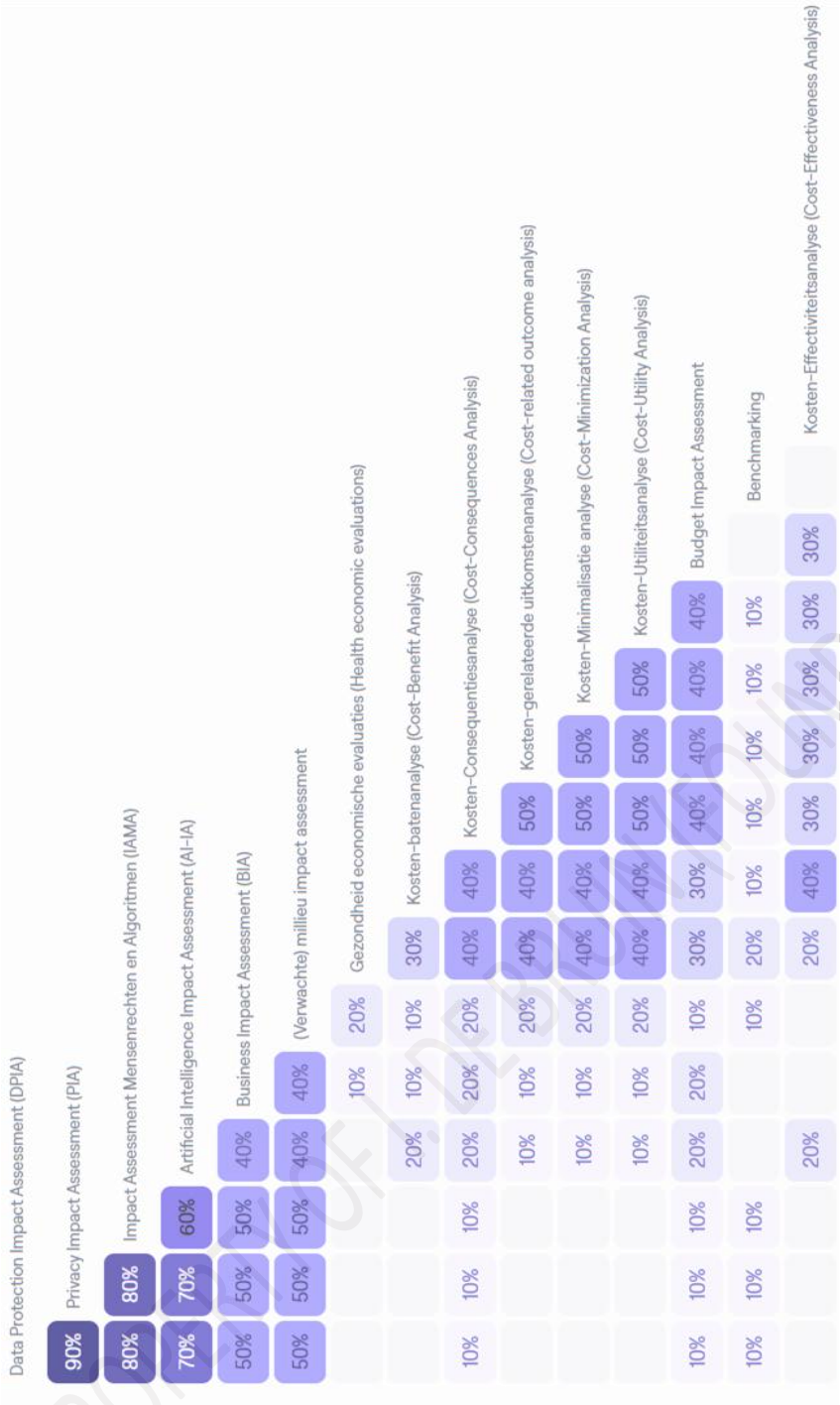


Figure 8. The agreement matrix (right) and similarity matrix (left) of effectiveness and impact assessments (n = 12).

The advice and reasonings & Report layout

The agreement matrices and similarity matrices reveal the preference to indicate the advice using visualisations and words (Figure 9). The reasonings were preferred in a bullet-points format, including recommendations to further look into and results of the impact assessments. Figure 10 shows the preference to include all evaluation topics in the report and the report being developed in Microsoft Word.

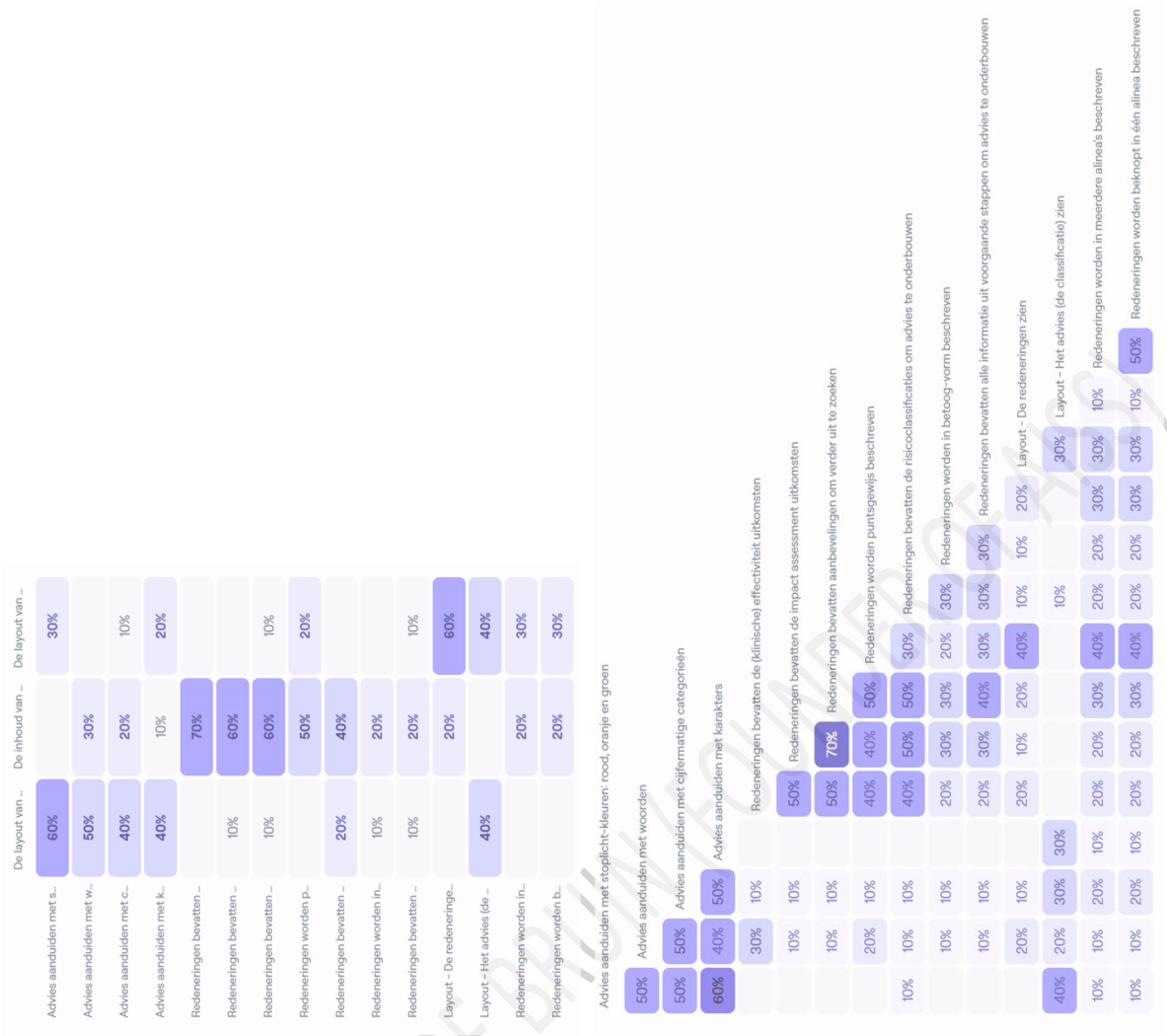


Figure 9. The agreement matrix (left) and similarity matrix (right) of the advice and reasonings (n = 12).

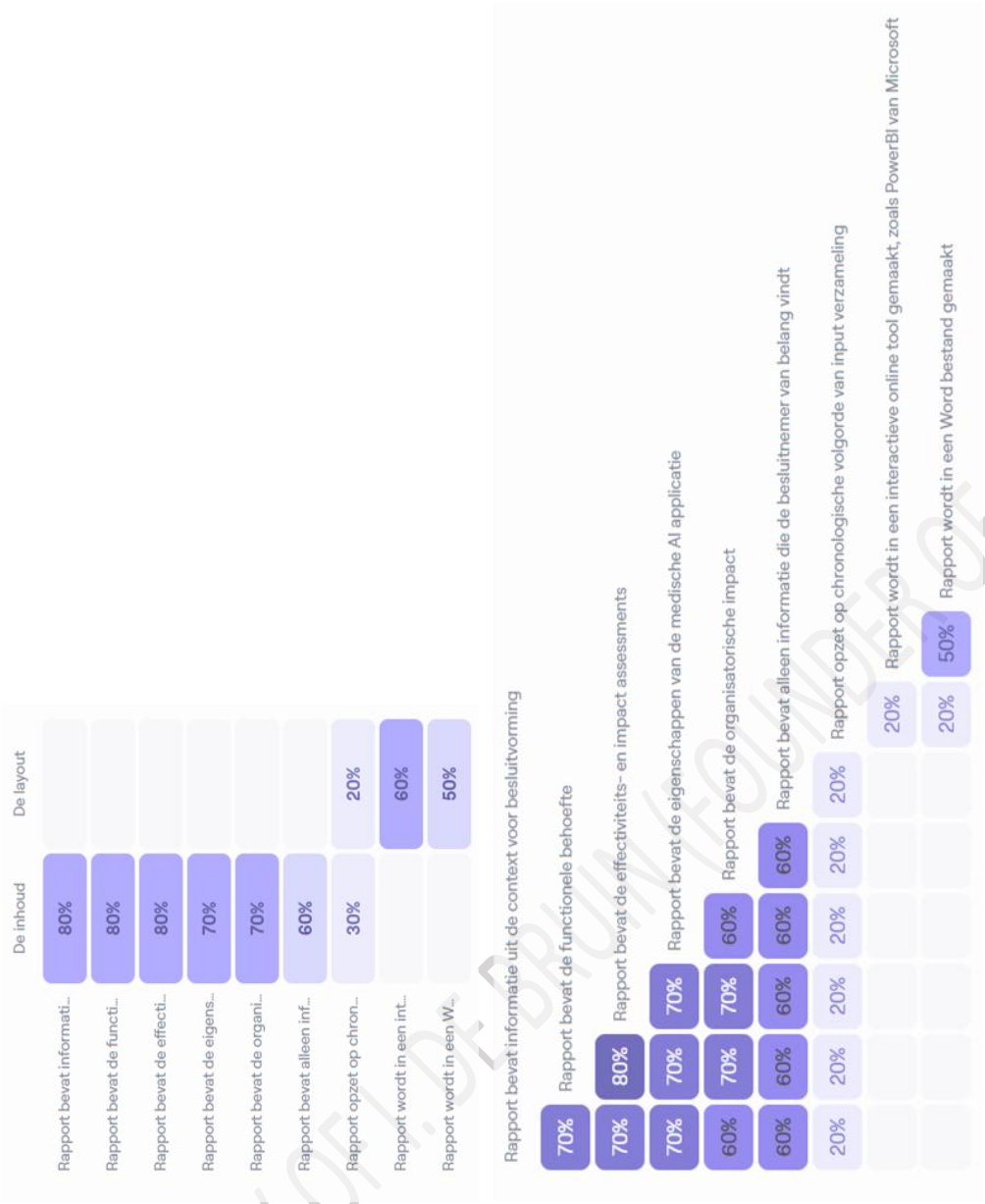


Figure 10. The agreement matrix (left) and similarity matrix (right) of report layout (n = 12).

4.3 Interim conclusion phase 1.1 and phase 1.2

The results of phase 1.1 and phase 1.2 both revealed that the evaluation elements in order of appearance should be 1) the functional needs, 2) the properties of medical AI, 3) the decision-making context, 4) the effectiveness and impact assessments, and 5) the advice and recommendations for improvements. Considering the evaluation categories within each these elements, multiple similarities were also found between the results of phase 1.1 and phase 1.2. For example, both concluded the ''gap''-category did not bring in any new and relevant information to better understand the functional needs. Furthermore, both considered four overarching evaluation categories to address the decision-making context and merge the organisational impact and decision-making context.

5. Phase 2 Design – digital prototyping

5.1 Method

A four-step approach was used to develop the IA and the design of the prototype DSS, combining the qualitative data of the in-person card sorting and the quantitative data of the digital card sorting (39, 41, 42). These four steps are based on the digital prototyping principals of the design phase of the CeHReS roadmap (e.g. co-design) (38) combined with the design process of creative technology (58). Each step consists of certain requirements for value specifications of the HTA-mAIx 2.0 as a DSS to comply with the principles of the CeHReS roadmap (e.g. holistic and intertwined) (38).

Materials and procedure

The digital prototype of the DSS was designed using Canva.com (61) to develop the wireframes of the IA and Figma.com (62) to connect the wireframes and program the IA to design the HTA-mAIx 2.0 as a DSS (38, 42, 58).

Step 1. Identifying the relevant categories per evaluation topic (card sort)

The interrater reliability (average agreement, %) was used to identify which categories were considered as most relevant to include in the evaluation of medical AI by the participants (56, 59). The main codes and subcodes of the moderated in-person card sorting study (phase 1.1) were used to identify which relevant category suited to which relevant evaluation topic. Next, the literature and regulatory documentations (e.g. AI-act (63)) were utilized to validate the outcomes of this step before going to the second step, to increase the reliability and validity of the DSS. The literature was identified within scientific journals (e.g. the lancet, Pubmed), government databases (e.g. Rijksoverheid) and government websites (e.g. IAMA, DPIA and AI-act) using the search strategy of previous research (25).

Step 2.Identifying the relevant evaluation criteria per category in each topic

The agreement matrices were used to identify which evaluation criteria (cards) were perceived as most relevant to each relevant evaluation category per relevant topic in estimating the value of medical AI (59, 60). Similar to step 1, literature (e.g. AI business model canvas (64)) and regulatory documentations served to validate the outcomes of this second step to decrease potential bias in the card sorting results and incorporate all relevant criteria (9, 11, 55). This increased the reliability and validity of the DSS (55).

Step 3. Identifying the combinations and order of appearance of evaluation criteria

To identify the order of appearance of the outlined relevant evaluation topics, categories and criteria, the similarity matrices served as primary input to indicate the pairs of evaluation criteria and order of relevance (42, 59, 60). To validation of these outcomes were compared

to the main codes and subcodes per relevant topic and regulatory documents (e.g. the DPIA (65) and IAMA (66)) to increase the reliability and validity of the outcomes before step 4 (55, 67, 68).

Step 4. Developing the IA & prototype: workflows and User Interface

This step involved developing the IA of the digital prototype of the DSS of the HTA-mAIx. Star schemes were developed per evaluation topic and connected to each other to show the appropriate IA. A star scheme is a method of structuring data to increase explainability of software systems' IA (e.g. DSS) and increase performance efficiency of the system (69). Each star scheme included a fact table showing all the evaluation criteria. A dimension table for each criterium was added to provide the relevant context (e.g. the information or question to answer the criterium) (59, 60, 69, 70). This created a multi-dimensional model representing the flows of information based on the results of phase 1.1 and phase 1.2 (69). The results of phase 1.2 was also used to design the UIs of each page in the prototype DSS, because the plenary discussions identified preferences in UIs (58, 68).

5.2 Results

Figure 11 shows the digital prototype of the DSS of the HTA-mAIx 2.0 based on the four-step approach. The most relevant features are the overviews of progressions of projects (pages 2 and 3) and the steps in the evaluation process (pages 3-11), because these features show the IA of the HTA-mAIx 2.0. The evaluation elements in order are: 1) functional needs excluding the 'gap'-category (page 5), 2) properties of medical AI including the data characteristics, AI characteristics and product characteristics (pages 6 and 7), 3) decision-making context including the organisational readiness as a category instead of a separate topic (page 8), 4) assessments in which two categories were renamed and the ethical impact category is added (pages 9 and 10), and 5) presenting the advice and reasonings through traffic-light-colours (pages 11 and 12).

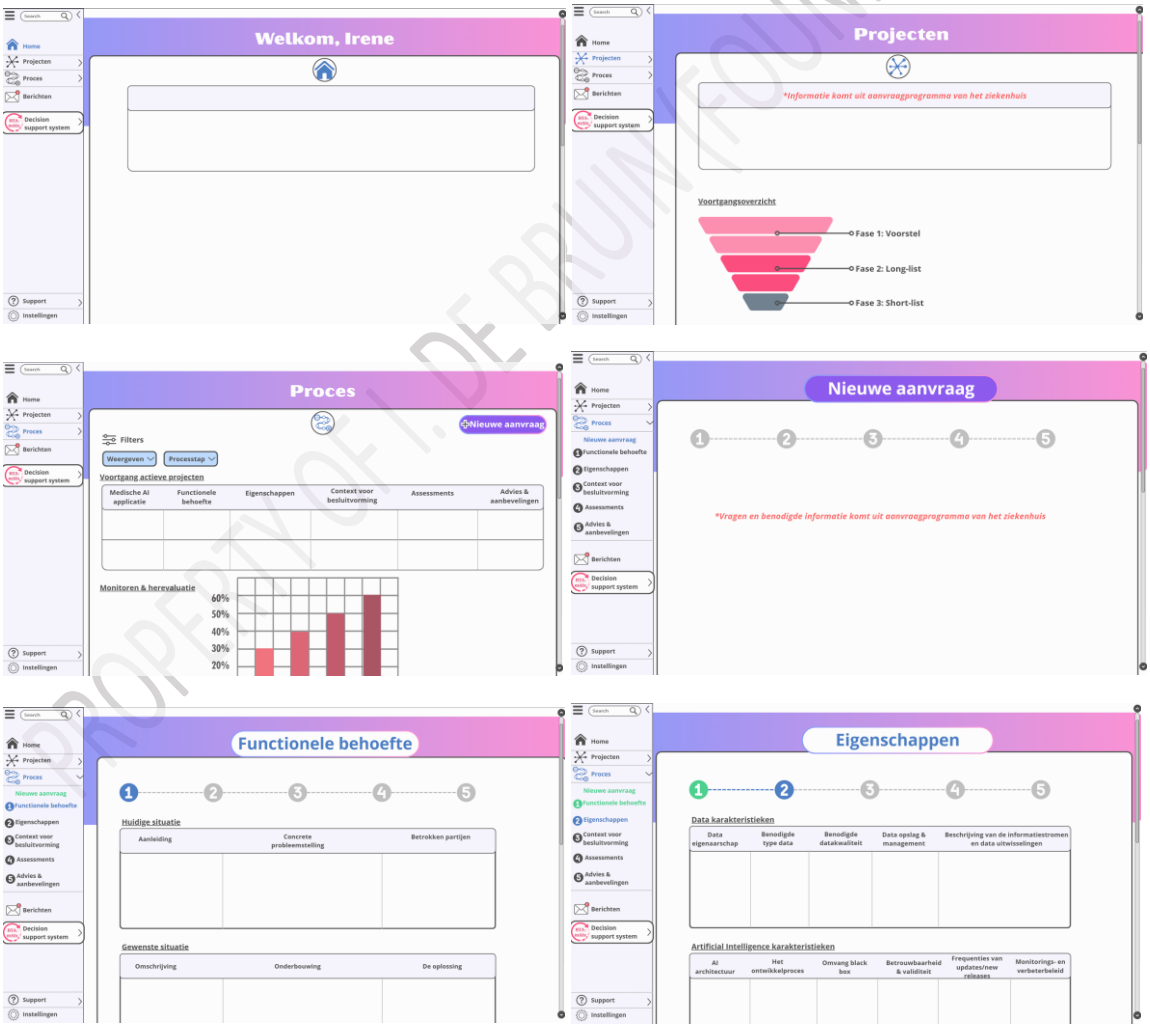
The evaluation process, indicating the IA, goes as follows. When a medical AI application is requested by healthcare staff to potentially be used into practice, the healthcare professional and IT staff go to the icon representing the process in overarching the menu on the left (page 1). This icon will change its colour from black to blue indicating that the users are located at this page (page 3), which starts by providing an detailed and customizable overview of all ongoing projects (medical AI applications being evaluated). The users are able to see if similar medical AI applications are being evaluated or are already evaluated to prevent potential miscommunications and double workload for hospital staff. A new medical AI application for evaluation is added by pressing the purple button on the top-right (page 3). This button brings the users to the page (page 4) showing the application questions, which are similar to the application programmes hospitals use (e.g. Zenya, Topdesk) due to the information input in the IA of this page being connected to these programmes. Once the new

medical AI application is registered in the DSS, the five sequential topics in the evaluation process appear below the process-icon.

The topic colours blue to indicate the current progression in the evaluation process, which will colour green if all relevant criteria within the topic are provided with enough information to go to the next topic (pages 5 to 12). The categories within each evaluation element are visually separated in rectangular blocks to create an aesthetic design and follow the standards of similar programmes. To increase the understanding of the evaluation process for medical AI, a page containing explanations about the HTA-mAIx, including the evaluation topics, categories and criteria, is added (page 14). The users can go to this page any time during the evaluation process by clicking the icon of the HTA-mAIx (white rectangular button) in the left menu.

The prototype DSS included only the evaluation elements and the categories to decrease participant burden in the feasibility-usability testing (phase 3).

The separate pages can be found in appendix C.



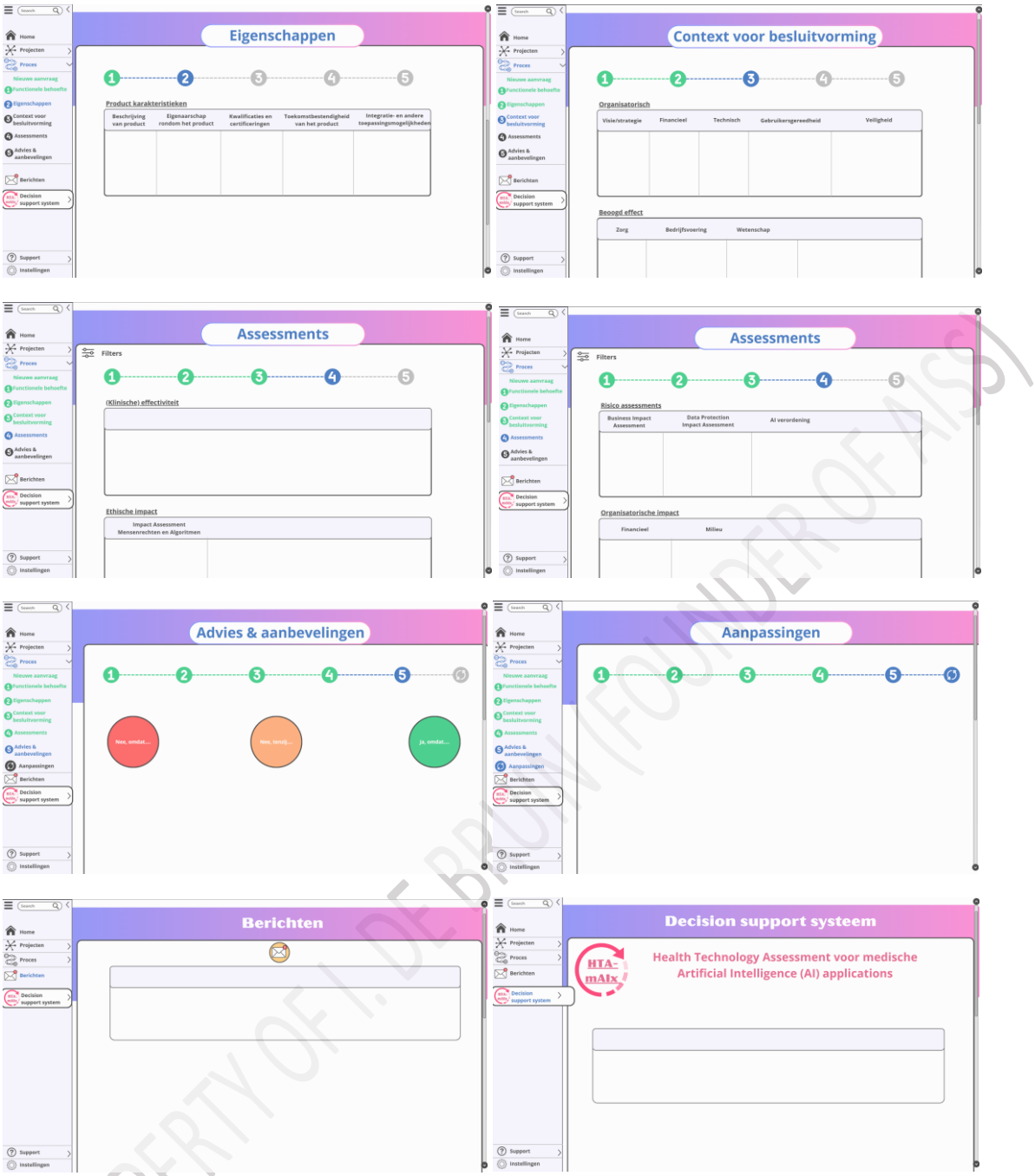


Figure 11. The digital prototype of the Decision Support System of the HTA-mAIx (14 pages).

6. Phase 3 Design – feasibility and usability testing

6.1 Method

A qualitative approach was used to test the feasibility of the prototype DSS in terms of usability, initial acceptance, and perceived effects on AI literacy (5, 40, 43). The qualitative data were collected through a feasibility questionnaire and thinking aloud method to gain contextual insights of the user experience (UX) (38, 40, 68, 71). Thinking aloud method was

used to estimate the usability by instructing the participants to verbalize their thoughts whilst moving through the system's UIs (38, 70, 71).

Setting

Similar to the in-person card sorting sessions, the feasibility and usability testing sessions took place at the Martini hospital in Groningen (NL). A conference room was booked for the feasibility-usability session on June 19th (8 – 9.15 am), and a conference room was booked for a session on July 14th (2 – 3 pm).

Participants

The target population included the same four key stakeholder groups as the card sorting studies (phase 1.1 and 1.2). However, the included stakeholders were not allowed to participate if they took part in the moderated in-person card sorting sessions in addition to the aforementioned inclusion/exclusion criteria.

Thirty stakeholders were purposefully invited by email or telephonic by the researcher between May 28th and July 11th to partake in either the session on June 19th or schedule a separate session at their earliest convenience (before July 15th). Eight stakeholders accepted the invitation for the session on July 19th, three of them actually finished the study. Twenty stakeholders accepted the invitation and were willing to plan a separate session, one of them actually planned the separate session on July 14th and completed the study.

The stakeholders that accepted the invitation received an informed consent (IC) form by email. The researcher instructed them to fill-in the IC form and hand it in (digitally or printed) before the start of each session.

Materials and procedure

Similar to the in-person card sorting sessions, a playbook was used to guide the feasibility-usability testing sessions. It is build up out of the same three columns as the playbook for moderated in-person card sorting (phase 1.1, Appendix A and D): the first column indicates the timeframes (cumulatively) per activity mentioned in the middle column, and the third column addressed the checklist-items per activity. It started with fifteen minutes of preparation, followed by the introduction and instructions to explain the aim, procedure and hand the printed prototype to the participants (Figure 11, Appendix C). After, the participants had 60 minutes to think aloud and fill-in the feasibility questionnaire.

The feasibility questionnaire was in Dutch, created in Microsoft Forms by the researcher and consisted of an introduction and instruction, three background questions, and 10 questions to measure the feasibility in terms of usability, initial acceptance and perceived effects on AI literacy (Appendix E). The introduction explained the objectives of feasibility-usability testing, followed by a set of instructions was presented to explain the procedure and how the prototype DSS would be reviewed. The participants were instructed to fill-in the questionnaire together to semi-mimic the multidisciplinary approach in the evaluation

process and decision-making process of medical AI (24). The participants were asked to think aloud when looking through the prototype DSS and answering the questions (24, 71). The thinking aloud data was audio-recorded by the researcher using a Jabra microphone and Microsoft Teams.

The three background questions based on the inclusion/exclusion criteria: 1) their profession, 2) their months of work experience, and 3) affected by decision-making about medical AI in their profession. The 10 questions were separated into two parts: part 1 consisted of four open-questions and one 5-point Likert scale question about the design of the HTA-mAIx 2.0, part 2 had four open-questions and one 5-point Likert scale question about the content of the HTA-mAIx 2.0.

The design questions were based on the 10 usability heuristics of Jakob Nielsen (40), because these heuristics reflect general principles to navigate the evaluation of the usability of the User Interface (UI) of a software/internet technology (40, 72). The 10 usability heuristics are: 1) visibility of system status, 2) match between the system and the real-world, 3) user control and freedom, 4) consistency and standards, 5) error prevention, 6) recognition rather than recall, 7) flexibility and efficiency of use, 8) aesthetic and minimalist design, 9) help users recognise, diagnose and recover from errors, and 10) help and documentation (40). To test the usability, five heuristics focusing on the design were incorporated into the feasibility questionnaire. Question 1 addressed the aesthetic and minimalist design heuristic by asking about the colour scheme of the prototype. Question 2 asked the perceived consistency in design to cover the consistency and standards heuristic. Question 3 focused on the visibility of the system heuristic by asking about the location of navigational aids in the prototype. Question 4 addressed the match between system and real-world heuristic by asking the understandability of the linguistics. Lastly, question 5 asked the perceived intuitiveness on a scale of 1 to 5 to measure the flexibility and efficiency of use heuristic (40).

The five questions about the content of the prototype were based on five domains of the NASSS framework, because it indicated the contextual factors necessary to be successful (38, 45) Question 1 focused on the technology domain by asking the expected type of data generated. Question 2 asked the expected demand-side value to address the value proposition domain. Question 3 covered the adopter system domain by asking the expected impact on staff. Question 4 asked the expected readiness for this prototype/change to address the organisation domain. Lastly, question 5 focused on the embedding and adoption over time domain by asking the expected scope for adoption over time on a scale of 1 to 5.

Each session took approximately 75 minutes to complete.

Analyses

The transcripts of the audio-recordings were transcribed in Microsoft Teams using the ''transcribe'' function. These generated transcripts were compared to the audio-recording by the researcher to filter out any type of translated mistakes. Next, the transcripts and open-

questions were thematically analysed using both deductive and inductive analyses approaches due to the theories on which the 10 feasibility questions were based (40, 57, 73). By also applying the inductive approach, points of improvement and other new contextual information about the perceived feasibility of the prototype DSS became insightful (45, 73). The design part and the context part were analysed separately by the researcher to collect the relevant information to answer the research objectives. The results were presented in thematic analyses mappings in which the themes based on the theories, the deductively and inductively identified codes per theme, and subcodes to explain the codes were visualised (57, 73). The entire thematic analysis and mappings (Figures 12 and 13) can be found in Appendix F.

6.2 Results

Two feasibility-usability testing sessions were done, one session included three participants and the other session included one participant. Two participants were representative for the healthcare staff and worked at either a Dutch academic hospital or a Dutch top clinical hospital. The other two participants were representative for the IT staff, both working at a Dutch top clinical hospital. All participants had at least six months work experience and were affected by medical AI in their profession (Table 10).

Characteristic	**Feasibility-usability testing – n(%)**
Key stakeholder group	
- *Healthcare staff*	2(50)
- *Decision-making*	0(0)
- *IT staff*	2(50)
- *Medical technology specialists*	0(0)
- *Other*	0(0)
Work experience	
- *< 3 months*	0(0)
- *3 - 6 months*	0(0)
- *6 - 9 months*	1(25)
- *9 - 12 months*	0(0)
- *> 12 months*	3(75)
Affected by medical AI	
- *Yes*	4(100)
- *No*	0(0)
- *I don't know*	0(0)
- *Other*	0(0)
Total	4

Table 10. Participant characteristics (n = 4).

6.2.1. Content

Technology

The analyses of the transcripts and questionnaires revealed three subcodes that indicated how the participants perceived the expected type of data generated by the prototype DSS (Figure 13). All stakeholders suggested to include an introduction and instruction page when opening the prototype DSS: *'' zoiets gewoon kort en Misschien kan je nog een subkopje doen met meer informatie of zo Als je er meer over wilt weten. Maar ik denk als een soort van intro pagina met wat deze tool is, waarom het er is, voor wie het is en wat het doet of zo. Dus hoeft niet heel lang te zijn,…''* (IT staff & Healthcare staff).

All participants mentioned to understand for whom the prototype DSS is intended and in what context it needs to be used: *'' de inzet is dat uiteindelijk is om inzetbaarheid van AI in de praktijk te meten.''* (IT staff).

The value proposition

Analysis (Figure 13) identified that all participants perceived the prototype DSS to be of added value to evaluate medical AI and support decision-making about medical AI: *'' nou van wat ik er nu over weet, kiezen welke AI je moet gebruiken. Dus je ramt alles erin en dan druk je op de grote knop en dan zegt hij, Ja doen.''* (IT staff & Healthcare staff).

The adopter system

The analysis uncovered multiple contextual insights into the expected effects on staff of the participants have of the prototype DSS (Figure 13). All participants mentioned that they expect the prototype to be of added value in their current work processes: *'' Het is een verrijking van het huidige proces.''*(IT staff).

One participant (IT staff & Healthcare staff) suggested to add 'history' and 'user-tag' features: *'' Maar ik denk dat dat wel een goeie is dat je ziet, Wie, welke gebruikers, zeg maar met een gebruikerstag of zo, en hoe laat Het is toegevoegd, of welke dag en tijd, timestamp erbij.''*

The organisation

The analysis showed that all participants expect the prototype DSS to be beneficial to support decision-making about medical AI (Figure 13). One participant (IT staff) mentioned that usage of this system would strengthen evidence-based decision-making and provide the necessary specification to evaluation processes of medical AI applications. Another participant (IT staff & Healthcare staff) noticed the connection to improving the AI-literacy: *'' Het is eigenlijk niet Alleen een mooi overzicht, Maar het leert ze ook wat. Dus dat zijn natuurlijk allemaal voordelen, dat het gewoon het zorgt ook voor nieuwe kennis.''*

One participant (IT staff & Healthcare staff) mentioned that potential disadvantage of the DSS could be the negative attitude of an intended user to using new systems in to do their work: *'' Dat vinden mensen misschien irritant. Dat ze dat moeten gaan leren, maar dat is natuurlijk altijd bij alles wat nieuw is.''*

Embedding and adoption over time

Analysis revealed that on average (4) the participants expected the DSS to be supportive in future medical AI developments in hospitals (Figure 13). One participant (IT staff) explained that it helps to understand and identify the entire scope of medical AI: *'' Alles zit straks in dit systeem verweven. Ik weet niet welke groep mensen bij elkaar op dit moment AI zo goed en feitelijk zou kunnen beoordelen.''*

One participant (IT staff & Healthcare staff) mentioned that the DSS would be helpful in preventing evaluation mistakes by providing explanations with the advice: *'' Stel, je hebt alles erin geknald en dan vraag je advies. En dan zegt hij, nee, want dat klopt niet. Dan denk je, oh, oeps, ik heb het verkeerde ding erin gezet en dan zou je dat gewoon even wisselen. En dan patsboem nieuw advies.''*

6.2.2. Design

Aesthetic an minimalist design

The analysis of the transcripts and open-questions identified multiple codes and subcodes explaining how the participants perceived the aesthetics and minimalist design of the prototype DSS (Figure 12). The colour schemes used were experienced differently by the participants. Two participants (IT staff, Healthcare staff) suggested that the colour transitions and colour interpretations were not aesthetic, whilst the two other participants (IT staff) indicated the opposite. Three participants (IT staff, Healthcare staff) mentioned to consider safeguarding the inclusiveness by adjusting the colours and design to make differentiation between pages and process steps understandable for all users, including colour-blindness. Lastly, three participants (IT staff, Healthcare staff) stated that visually indicating the progression using colours is useful but need to be neutral colours which could not be interpreted as 'good' or 'bad'. However, one participant (IT staff & Healthcare staff) stated the opposite.

Consistency and standards

Analysis revealed seven codes and multiple subcodes to indicate the consistency and standards of the prototype DSS perceived by the participants (Figure 12). The design consistency in the layout between pages was perceived as cohesive by all participants. Two participants (IT staff, Healthcare staff) perceived the messages feature as a risk due to adding another communication medium to the other available media, e.g. Outlook and Microsoft Teams. Their suggestion was to rename the messages feature to prevent any adverse attitudes and misinterpretations. Three participants (IT staff, Healthcare staff) mentioned that the interconnections between process steps were not clear enough. However, one participant (IT staff & Healthcare staff) did not experience this: *'' Ik vind het heel sowieso heel helder en overzichtelijk dit systeem. Precies weet waar je wat in terugziet, zeg maar.''*

Visibility of system status

The analysis of the transcripts and questionnaires shows that all participants were able to recognise navigational aids (Figure 12). However, they suggested to add 'back' and 'next' buttons to make the workflow of the process more inclusive to all users despite their age and digital skills (IT staff & Healthcare staff): *'' Ik denk dat dat logisch is, misschien dat het voor boomers heel verwarrend is, hoe ze dan terug moeten.''*

Match between system and real-world

Based on the analysis, each participant perceived the understandability of the linguistics differently (Figure 12). Two participants (IT staff, Healthcare staff) mentioned that the terminology to indicate the evaluation criteria were too difficult to be interpreted by the intended users. One participant (Healthcare staff) suggested to indicate all information in the prototype DSS in just one language, e.g. either English or Dutch, that is most suitable for all intended users. Contrary to these participants' perceptions, one participant (IT staff & Healthcare staff) indicated that the linguistics were entirely understandable to the intended users: *'' Je moet ook gewoon bepaalde woorden gebruiken, toch? Ik snap ze allemaal. Het is niet dat ik nu een woord heb gezien en denk, huh? Wat bedoel je daarmee?''*

Flexibility and efficiency of use

The analysis revealed that the participants considered the intuitiveness on average a 4 out of 5, indicating that they perceived the design of the prototype DSS as intuitive to interpret. One participant (IT staff & Healthcare staff) suggested the need to incorporate a form of system status feedback to communicate any operational problems, e.g. when the system does not respond when a user clicks on a feature (Figure 12).

7. Discussion

This study aimed to further research the first concept HTA-mAIx as a DSS to support Dutch academic and top clinical hospitals in making well-substantiated decisions about medical AI applications and improve implementation of these technologies. The objectives were to develop the HTA-mAIx 2.0 as a DSS in multidisciplinary groups and to test this prototype's feasibility in terms of usability, initial acceptance, and perceived effects on AI literacy. A multi-method approach was used to collect the data to address the aim and objectives: 1) in-person moderated and digital card sorting to understand the mental models and priorities of the intended users, 2) digital prototyping to develop the IA and design the UI of the HTA-mAIx 2.0 as a DSS, and 3) in-person feasibility and usability testing to assess the usability, initial acceptance and perceived effects on AI literacy.

This study found that Dutch academic and top clinical hospitals consider a DSS to assess and evaluate medical AI applications to be helpful for making well-informed decisions about investing and implementing medical AI into their hospitals. Furthermore, a DSS would provide a comprehensive list of all relevant evaluation topics and criteria based on technical, ethical, regulatory, and medical requirements for estimating the added value of medical AI.

The relevant evaluation topics in order of appearance are: 1) functional needs, 2) properties of medical AI, 3) decision-making context, 4) assessments, and 5) advice and recommendations. Lastly, the feasibility of the HTA-mAIx 2.0 as a DSS is perceived as an essential aid for hospitals to gain more insight into the comprehensiveness of medical AI, and cope with changing medical AI regulations and up-and-coming AI developments in healthcare. This is based on the initial acceptance and practicality of the HTA-mAIx 2.0 as a DSS by the target population due to, amongst others, its perceived usefulness in guiding the users through the evaluation process. This study achieved to design a low-fidelity prototype of the HTA-mAIx 2.0 as a DSS.

7.1 Interpretations

7.1.1 Card sorting results

The identified overarching evaluation topics could be considered in line with findings of similar studies in developing HTA for medical AI. Boverhof et al. (24) performed a systematic review and identified four overarching domains to include in the HTA Core Model for medical AI assessment: 1) Technology & performance, 2) Human & organizational, 3) Legal & ethical, and 4) Transparency & usability. However, the overarching evaluation topics of this study are differently paired and are addressed in a specific order based on the identified mental models and prioritisations of all relevant perspectives (e.g. regulatory and medical) and intended users. Boverhof et al. (24) and other currently known studies about developing a HTA framework for medical AI used theoretical approaches (e.g. systematic reviews) including HTA experts and/or other academic researchers to validate the identified data from literature (31, 49, 74-76). No follow-up and validation studies of these findings were found to address the practical implications and include all intended users, even though these researchers did mention the targeted population to be hospital decision-makers, healthcare professionals and medical AI developers (24, 74).

The current study suggests that the omission of all intended users may have been a particularly impactful oversight. The results found multiple conflicts of perceived relevant information to evaluate medical AI between each key-stakeholder group and between each involved stakeholder. For example, the need to assess the developmental process of medical AI applications (e.g. model-training/testing methods) was considered as irrelevant according to the participants even though this information deemed to be a priority according to regulatory policies (e.g. AI-act (9, 63), IAMA (66)). Farah et al. (74) also emphasized the importance of integrating the privacy and security considerations. Additionally, the similarity and agreement matrices implied that most stakeholders found the evaluation criteria to assess clinical effectiveness of medical AI to be irrelevant, however, they agreed that assessing the clinical effectiveness is a priority in the assessments-topic. Based on these results it would not be possible to perform this assessment due to not having the necessary input information, such as Quality of Life and Quality Adjusted Life Years (24, 74).

Another conflict was found in the relevance to address the explainability criterium, which could be explained by differences in levels of AI literacy (5, 17). Within the IT staff stakeholder group and between the IT staff and Medical technology specialist stakeholder groups disagreements occurred about the depth of understanding the development process for every involved party. Interestingly, these stakeholder groups seemed to speculate their perceptions of these depths from their own professional perspective and opinions without considering obligations from regulatory policies. For example, in February 2025 both the AI-act (9, 63) and the Autoriteit Persoonsgegevens (AP) (77) obliged all European organisations, including healthcare, wanting to integrate or already integrated any type of AI into their practices to increase employees' AI literacy. This includes understanding the developmental process: the motive for development (context), type and quality of data used for development, quality of input, understanding risks and consequences of misuse, responsibilities, how to use, etcetera. Farah et al. (74) and Esmaeilzadeh (5) identified similar conflicts suggesting that equal knowledge about the explainability positively affects understanding the black box and its clinical applicability and perceived performance. Therefore, lack of AI literacy could explain conflicts in perceived relevance to include a holistic perspective and comprehensive approach to evaluating and decision-making about medical AI (5, 17).

Finally, this study also found differences in perceived levels of complexity of the comprehensiveness in evaluating medical AI between healthcare and IT staff of academic hospitals compared to staff of top clinical hospitals. For example, three healthcare providers working in a top clinical hospital were not able to finish the study due to difficulties in understanding the evaluation topics and criteria presented. On the contrary, healthcare providers working in academic hospitals were able to finish the study and expressed liking their participation. A possible explanation could be the differences in experience using medical AI, and differences in their involvement in the developmental process of medical AI. The AI-monitor of the CBS also identified the effect of these contextual factors (12). Academic hospitals are often equipped with AI-departments (e.g. DASH in UMCG) that test, develop, etc. different types of medical AI in clinical practice, which is not always the case in top clinical hospitals (78, 79). This is mostly due to financial support, the academic context and accessible resources to innovate. The HTA-mAIx 2.0 (25) as a DSS could provide a more balanced level of knowledge and experience between academic and top clinical hospitals.

7.1.2 Digital prototyping

The HTA-mAIx is fundamentally based on basic elements of a standard HTA framework and regulatory documents, which is in line with similar studies (24, 32, 74). The four-step approach combining the requirements of the involved stakeholders, intended users, literature and regulatory documents to design the IA and UI, as has however not previously been used to the author's knowledge. Most HTA studies stopped their development process when a conceptual framework was created based on theories, regulations and HTA-expert opinions (24, 34, 49, 74). User verification, testing, and further development was not an

aspect considered in these HTA frameworks. Jakob's law states the effect of User Experience dictates the successfulness of embedding innovations into current practice (43, 68, 70). This could explain why the other HTA frameworks for medical AI were not implemented in practice (38, 68, 80).

The IA is based on Machine Learning (ML) models and will be ML-driven to ensure explainability and transparency of the DSS to comply with regulatory policies (9, 65, 66). Ramezani et al. (81) suggest that AI-driven HTA frameworks could be more effective compared to ML-driven HTA frameworks, due to potential high-quality results and improve decision-making about medical AI applications. This suggest that these researchers consider AI to be a reliable and appropriate evaluator of medical AI in clinical practice. However, the AI-act (9, 63) prohibits the use of AI to evaluate AI because this is comparable to ''marking your own exam'' and ethically not allowed. Moreover, the performance of AI is dependent on accessible information and the quality of the input used to create the output (a.k.a ''garbage in is garbage out''). Currently, the required data and technical standards to implement this way of evaluation is merely a 'wish' and not comparable to reality, which Ramezani et al. (81) reflect upon in their discussion.

7.1.3 Feasibility and usability testing results

The results found that the participants already considered the current state of the prototype DSS to be very helpful in decision-making about medical AI. However, the division of responsibilities during the evaluation process was perceived as unclear. Due to lack of similar studies it was difficult to explain this finding. Esmaeilzadeh (5), Zary et al. (17) and the AP (77) speculate that AI literacy and experience affects clarification on task-divisions of every involved party concerning investing and implementing the appropriate medical AI application. This might provide a possible explanation. However, these types of comprehensive HTA frameworks as a DSS to evaluate medical AI and support decision-making are new areas of investigation. This indicates unexplored territory and needs further investigation.

7.2 Limitations

The generalizability of the results is limited due to the small number of participants actually finishing the studies. This decreased the reliability and validity of the results due to the lack of representativeness of the four key-stakeholder groups (56, 57). The multidisciplinary approach caused higher participant burden for the healthcare staff and IT staff stakeholder groups in this study. Consequently, potential new insights into the requirements and points of improvement for the HTA-mAIx 2.0 to be perceived as useful and better indicate the feasibility according to the NASSS framework remained unidentified (44-46, 56, 67). As mentioned before, similar studies merely developed a conceptual HTA framework for medical AI without further developing, testing and validating it. Therefore, despite the limited generalizability of the results the HTA-mAIx 2.0 has a greater chance of success.

Moreover, the online tools used for the card sorting studies (phases 1.1 and 1.2) and to design the prototype DSS were not appropriate enough for this research. Maze.co was intuitive and easy to use when the card sorting study was made, however, the user friendliness by the participants was perceived as low and the automatic analyses were very difficult to manually adjust after the study was completed. This lowered the reliability and validity of the results due to lack of transparency (56).

7.3 Recommendations

To address the limited generalizability of the results, a redo of this research on a larger scale is necessary to better represent the targeted population and validate the prototype DSS. This will provide more accurate results indicating the feasibility in terms of usability, initial acceptance and perceived effects on AI-literacy of the prototype DSS. Consequently, the reliability and validity of the research and prototype DSS will increase. Further research into the IA per type of medical AI (e.g. ambient listening, diagnostics) and its AI architecture (e.g. the algorithms, models) is needed to develop a clickable, high-fidelity prototype DSS (24, 38). Additionally, three participants in the IT staff and medical technology specialists stakeholder groups indicated the perceived usefulness to include a digital room for testing medical AI applications based on Federated Learning. This may avoid the need to pilottest every medical AI application to estimate its effectiveness, which takes a lot of time due to the planning and preparations, and spending (74, 81).

References

1. van Buchem MM, Kant IM, King L, Kazmaier J, Steyerberg EW, Bauer MP. Impact of a digital scribe system on clinical documentation time and quality: usability study. JMIR AI. 2024;3(1):e60020.
2. OpenAI. Introducing Whisper: OpenAI; 2022 [Available from: https://openai.com/index/whisper/
3. van der Meijden SL, van Boekel AM, Schinkelshoek LJ, van Goor H, Steyerberg EW, Nelissen RG, et al. Development and validation of artificial intelligence models for early detection of postoperative infections (PERISCOPE): a multicentre study using electronic health record data. The Lancet Regional Health–Europe. 2025;49.
4. Wellenstein DJ, Woodburn J, Marres HA, van den Broek GB. Detection of laryngeal carcinoma during endoscopy using artificial intelligence. Head & Neck. 2023;45(9):2217-26.
5. Esmaeilzadeh P. Challenges and strategies for wide-scale artificial intelligence (AI) deployment in healthcare practices: A perspective for healthcare organizations. Artificial Intelligence in Medicine. 2024;151:102861.
6. Maleki Varnosfaderani S, Forouzanfar M. The Role of AI in Hospitals and Clinics: Transforming Healthcare in the 21st Century. Bioengineering. 2024;11(4):337.

7. van Kolfschooten H. Towards an EU Charter of Digital Patients' Rights in the Age of Artificial Intelligence. Digital Society. 2025;4(1):6.
8. Sun K, Roy A, Tobin JM. Artificial intelligence and machine learning: Definition of terms and current concepts in critical care research. Journal of Critical Care. 2024;82:154792.
9. Butt J. Analytical study of the world's first EU Artificial Intelligence (AI) Act. International Journal of Research Publication and Reviews. 2024;5(3):7343-64.
10. Galloway JL, Munroe D, Vohra-Khullar PD, Holland C, Solis MA, Moore MA, et al. Impact of an Artificial Intelligence-Based Solution on Clinicians' Clinical Documentation Experience: Initial Findings Using Ambient Listening Technology. Journal of General Internal Medicine. 2024;39(13):2625-7.
11. Pantanowitz L, Quiroga-Garza GM, Bien L, Heled R, Laifenfeld D, Linhart C, et al. An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. The Lancet Digital Health. 2020;2(8):e407-e16.
12. Centraal Bureau voor de Statistiek AI-monitor 2025: Centraal Bureau voor de Statistiek; 2025 [Available from: https://www.cbs.nl/nl-nl/longread/aanvullende-statistische-diensten/2025/ai-monitor-2024/2-gebruik-van-ai-technologie-door-nederlandse-bedrijven
13. ICT&health Samen de digitaliseringsambitie van het IZA waarmaken. ICT&health. 2024.
14. Peeters R, Westra D, Gifford R, Ruwaard D. Wie, wat, waar? De invloed van het Integraal ZorgAkkoord op bestaande regionale netwerken in de zorg. TSG-Tijdschrift voor gezondheidswetenschappen. 2024;102(2):59-66.
15. Nederlandse V, van, Ziekenhuizen. AI in de zorg. Nederlandse Vereniging van Ziekenhuizen; 2024. p. 40.
16. Nederlandse V, van , Ziekenhuizen , . Digitalisering: Nederlansde Vereniging van Ziekenhuizen; n.d. [Available from: https://nvz-ziekenhuizen.nl/standpunten/digitalisering.
17. Zary N. AI Literacy Framework (ALiF): A Comprehensive Approach to Developing AI Competencies in Educational and Healthcare Settings. Preprints: Preprints; 2025.
18. ZonMw Methodologie n.d. [Available from: https://www.zonmw.nl/nl/methodologie.
19. Liu Y, Su Y-Y, Alhur AA, Naeem SB. Factors influencing artificial intelligence (AI) literacy in the age of generative AI chatbots for health information seeking. Information Development.0(0):02666669251343030.
20. Aldosari B, Aldosari H, Alanazi A. Challenges of Artificial Intelligence in Medicine. Stud Health Technol Inform. 2025;323:16-20.
21. van Smeden M, Moons C, , Hooft L, , Kant I, , van Os H, , Chavannes N, , . Leidraad kwaliteit AI in de zorg. In: Ministerie van Volksgezondheid WeS, editor. OSFHome: OSFHome; 2023. p. 81.
22. Kimiafar K, Sarbaz M, Tabatabaei SM, Ghaddaripouri K, Mousavi AS, Mehneh MR, et al. Artificial intelligence literacy among healthcare professionals and students: a systematic review. Frontiers in Health Informatics. 2023;12(0):168.
23. Organization WH. Health technology assessment of medical devices: World Health Organization; 2025.
24. Boverhof B-J, Redekop WK, Visser JJ, Uyl-de Groot CA, Rutten-van Mölken MPMH. Broadening the HTA of medical AI: A review of the literature to inform a tailored approach. Health Policy and Technology. 2024;13(2):100868.

25. Bruin I. Eerste aanzet voor een Health Technology Assessment (HTA) raamwerk voor medische Artificial Intelligence (AI): een kwalitatieve en kwantitatieve aanpak: University of Twente; 2024.
26. Vermeulen RJ, Govers TM, van Leeuwen KG. Early health technology assessment: the value of valuing AI applications. European Radiology. 2024.
27. Elvidge J, Hawksworth C, Avşar TS, Zemplenyi A, Chalkidou A, Petrou S, et al. Consolidated Health Economic Evaluation Reporting Standards for Interventions That Use Artificial Intelligence (CHEERS-AI). Value in Health. 2024.
28. Roppelt JS, Kanbach DK, Kraus S. Artificial intelligence in healthcare institutions: A systematic literature review on influencing factors. Technology in Society. 2024;76:102443.
29. Jiu L, Hogervorst MA, Vreman RA, Mantel-Teeuwisse AK, Goettsch WG. Understanding innovation of health technology assessment methods: the IHTAM framework. International Journal of Technology Assessment in Health Care. 2022;38(1):e16.
30. Voets M M. VJ, Slump C H., Siesling S., Koffijberg H., . Systematic Review of Health Economic Evaluations Focused on Artificial Intelligence in Healthcare: The Tortoise and the Cheetah. Value in Health. 2022;25(3):340-9.
31. Fasterholdt I, Naghavi-Behzad M, Rasmussen BSB, Kjølhede T, Skjøth MM, Hildebrandt MG, et al. Value assessment of artificial intelligence in medical imaging: a scoping review. BMC Medical Imaging. 2022;22(1):187.
32. Rasmussen B S HMG, Fasterholdt I, Kidholm K, . Model for ASsessment of Artificial Intelligence Centre for Clinical Artificial Intelligence; [Available from: https://cai-x.com/projects/previous-projects/mas-ai.
33. EUnetHTA. HTA core model - Guiding principles on use 2015 [Available from: https://www.eunethta.eu/wp-content/uploads/2018/01/The-HTA-Core-Model_Guiding-principles-on-use_20151218.pdf.
34. Garcia-Saez G, Goettsch W, Driessen J H M, Nemeth B, Petrova G, Siirtola P, Röning J, Zemplenyl A T, Hernando M E, . Next Generation Health Technology Assessment to support patient-oriented, societally oriented, real-time decision-making in Diabetes. The HTx Consortium; 2020.
35. HTx. Next Generation Health Technology Assessment European Union; n.d. [Available from: https://www.htx-h2020.eu/.
36. Grutters JPC, Kluytmans A, van der Wilt GJ, Tummers M. Methods for Early Assessment of the Societal Value of Health Technologies: A Scoping Review and Proposal for Classification. Value in Health. 2022;25(7):1227-34.
37. Van Haesendonck L, Ruof J, Desmet T, Van Dyck W, Simoens S, Huys I, et al. The role of stakeholder involvement in the evolving EU HTA process: Insights generated through the European Access Academy's multi-stakeholder pre-convention questionnaire. J Mark Access Health Policy. 2023;11(1):2217543.
38. Kip H, Beerlage de Jong N, Gemert-Pijnen Lv, Sanderman R, Kelders SM. EHealth research theory and development : a multidisciplinary approach. Abingdon, Oxon: Routledge; 2024. Available from: https://www.taylorfrancis.com/books/9781003302049.
39. Paul CL. A modified delphi approach to a new card sorting methodology. Journal of Usability studies. 2008;4(1):7-30.
40. Nielsen J. Ten usability heuristics. 2005.
41. Chan M. Mental Models: NNgroup; 2024 [Available from: https://www.nngroup.com/articles/mental-models/.

42. Tankala S, Sherwin, K., . Card Sorting: Uncover Users' Mental Models for Better Information Architecture: NNgroup; 2024 [Available from: https://www.nngroup.com/articles/card-sorting-definition/#:~:text=Card%20sorting%20is%20a%20research,navigate%20to%20on%20your%20website.
43. Nielsen J. Usability 101: Introduction to Usability: NNgroup; 2012 [Available from: https://www.nngroup.com/articles/usability-101-introduction-to-usability/.
44. Fernando M, Abell B, McPhail SM, Tyack Z, Tariq A, Naicker S. Applying the Non-Adoption, Abandonment, Scale-up, Spread, and Sustainability Framework Across Implementation Stages to Identify Key Strategies to Facilitate Clinical Decision Support System Integration Within a Large Metropolitan Health Service: Interview and Focus Group Study. JMIR Medical Informatics. 2024;12(1):e60402.
45. Abell B, Naicker S, Rodwell D, Donovan T, Tariq A, Baysari M, et al. Identifying barriers and facilitators to successful implementation of computerized clinical decision support systems in hospitals: a NASSS framework-informed scoping review. Implementation Science. 2023;18(1):32.
46. Alami H, Lehoux P, Papoutsi C, Shaw SE, Fleet R, Fortin J-P. Understanding the integration of artificial intelligence in healthcare organisations and systems through the NASSS framework: a qualitative study in a leading Canadian academic centre. BMC Health Services Research. 2024;24(1):701.
47. Tankala S, Sherwin, K.,. Card Sorting vs. Tree Testing: NNgroup; 2024 [Available from: https://www.nngroup.com/articles/card-sorting-tree-testing-differences/
48. Conrad LY, Tucker VM. Making it tangible: hybrid card sorting within qualitative interviews. Journal of Documentation. 2019;75(2):397-416.
49. Karargyris A, Umeton R, Sheller MJ, Aristizabal A, George J, Wuest A, et al. Federated benchmarking of medical artificial intelligence with MedPerf. Nature Machine Intelligence. 2023;5(7):799-810.
50. Unsworth H, Wolfram V, Dillon B, Salmon M, Greaves F, Liu X, et al. Building an evidence standards framework for artificial intelligence-enabled digital health technologies. The Lancet Digital Health. 2022;4(4):e216-e7.
51. Chenais G, Lagarde E, Gil-Jardiné C. Artificial Intelligence in Emergency Medicine: Viewpoint of Current Applications and Foreseeable Opportunities and Challenges. J Med Internet Res. 2023;25:e40031.
52. Olawade D, Clement David-Olawade A, Wada O, Asaolu A, Adereni T, Ling J. Artificial Intelligence in Healthcare Delivery: Prospects and Pitfalls. Journal of Medicine Surgery and Public Health. 2024:100108.
53. Maze.co. Card sorting n.d. [Available from: https://maze.co/features/card-sorting/
54. Tchivi E, Sharma B, Paea S. A Systematic Review of the Comparison of Different Types of Card Sorting. IEEE Access. 2025;13:52334-52.
55. Noble H, Smith J. Issues of validity and reliability in qualitative research. Evidence Based Nursing. 2015;18(2):34.
56. Righi C, James J, Beasley M, Day DL, Fox JE, Gieber J, et al. Card sort analysis best practices. Journal of Usability Studies. 2013;8(3):69-89.
57. Alhojailan M I. Thematic Analysis: A critical review of its process and evaluation. 2012.
58. Mader AH, Eggink W, editors. A design process for creative technology. 16th International Conference on Engineering and Product Design, E&PDE 2014; 2014: The Design Society.

59. Cunha L. Understanding your card sorting results: Maze.co; 2024 [Available from: https://help.maze.co/hc/en-us/articles/5813625557139-Understanding-your-card-sorting-results.
60. Boag W, Hasan A, Kim JY, Revoir M, Nichols M, Ratliff W, et al. The algorithm journey map: a tangible approach to implementing AI solutions in healthcare. NPJ Digital Medicine. 2024;7(1):87.
61. Canva. Over Canva n.d. [Available from: https://www.canva.com/.
62. Figma Denk groter. Ontwikkel sneller. n.d. [Available from: https://www.figma.com/nl-nl/.
63. van Leeuwen KG, Doorn L, Gelderblom E. The AI Act: responsibilities and obligations for healthcare professionals and organizations. Diagnostic and Interventional Radiology.
64. Metelskaia I, Ignatyeva O, Denef S, Samsonowa T, editors. A business model template for AI solutions. Proceedings of the international conference on intelligent science and technology; 2018.
65. Georgiou D, Lambrinoudakis C. Data protection impact assessment (DPIA) for cloud-based health organizations. Future Internet. 2021;13(3):66.
66. Gerards J, Schaefer M, Vankan A, Muis I. Impact assessment mensenrechten en algoritmes. 2021.
67. Lantz E, Keeley JW, Roberts MC, Medina-Mora ME, Sharan P, Reed GM. Card sorting data collection methodology: how many participants is most efficient? Journal of Classification. 2019;36(3):649-58.
68. Nugraha WAP. The Power of UX Laws: Enhancing User Experience Research and Design Processes. Medium; 2024.
69. Jensen C, S., Back Perdersen, T., Thomsen, C., . Multidimensional Databases and Data Warehousing: Morgan & Claypool; 2010.
70. Barrera LF, Ramos AC, Florez-Valencia L, Pavlich-Mariscal JA, Mejia-Molina NA, editors. Integrating Adaptation and HCI Concepts to Support Usability in User Interfaces-A Rule-based Approach. WEBIST (2); 2014.
71. Nielsen J. Thinking Aloud: The #1 Usability Tool: NNgroup; 2012 [Available from: https://www.nngroup.com/articles/thinking-aloud-the-1-usability-tool/
72. Arhippainen L. Ten User Experience Heuristics. Chichester, UK: John Wiley & Sons Ltd; 2013. p. 1-8.
73. Proudfoot K. Inductive/Deductive Hybrid Thematic Analysis in Mixed Methods Research. Journal of Mixed Methods Research. 2023;17(3):308-26.
74. Farah L, Borget I, Martelli N, Vallee A. Suitability of the current health technology assessment of innovative artificial intelligence-based medical devices: scoping literature review. Journal of medical Internet research. 2024;26:e51514.
75. Merlin T, Street J, Carter D, Haji Ali Afzali H. Challenges in the Evaluation of Emerging Highly Specialised Technologies: Is There a Role for Living HTA? Appl Health Econ Health Policy. 2023;21(6):823-30.
76. Gyldmark M, Lampe K, Ruof J, Pöhlmann J, Hebborn A, Kristensen FB. IS THE EUNETHTA HTA CORE MODEL® FIT FOR PURPOSE? EVALUATION FROM AN INDUSTRY PERSPECTIVE. International Journal of Technology Assessment in Health Care. 2018;34(5):458-63.
77. AI-geletterdheid, (2025).
78. UMCG Data Science Center in Health (DASH): UMCG; n.d. [Available from: https://umcgresearch.org/w/dash

79. Radboudumc AI for Health n.d. [Available from: https://www.ai-for-health.nl/
80. Devi RS, Widagdo PB. MOBILE APPLICATION USER INTERFACE & USER EXPERIENCE DESIGN WITH GAMIFICATION AS A SOLUTION TO GADGET DEPENDENCY. Abdi Dosen: Jurnal Pengabdian Pada Masyarakat. 2025;9(2):807-18.
81. Ramezani M, Bakhtiari A, Daroudi R, Mobinizadeh M, Fazaeli AA, Olyaeemanesh A, et al. Applications of artificial intelligence and the challenges in health technology assessment: a scoping review and framework with a focus on economic dimensions. Health Economics Review. 2025;15(1):46.

Appendix A: Playbook moderated in-person card sorting sessions (DUTCH)

Sessie 1

Tijd (time)	**Acties (actions)**	**Checklist**
-15 minuten	Conference room preparation	- Link delen naar Maze.co; - Tafels in de juiste vorm zetten (vierkant/rechthoek); - Naambordjes plaatsen; - Laptops klaarzetten/ participanten nemen eigen laptop mee; - Opladers klaarzetten, indien nodig; - Audio opname apparatuur controleren;
-5 minuten	Inloop	Presentielijst bij af gaan en informed consent controleren.
+15 minuten	Introductie en korte toelichting over het onderzoek	- Controleren of iedereen in Maze.co kan komen via de gedeelde link; - Welkom en voorstellen; - Het onderzoek en de aanleiding; - Card sorting toelichten: wat is het en het doel; - De HTA-mAIx toelichten: proces, evaluatieonderdelen, categorieën, criteria (cards); - Instructies hoe deze sessie gaat verlopen.
+5 minuten (totaal: 20 minuten)	Participanten doen individueel de card sort – functionele behoefte	Timer zetten voor 5 minuten;
+7 minuten (totaal: 27 minuten)	Plenary group discussions	Timer zetten voor 7 minuten; Audio opname starten; Questions:

		- Hoe ervaarde jullie het sorteren en prioriteren van de kaarten? - Kunnen jullie iets vertellen over de afwegingen die hebt gemaakt bij het sorteren en prioriteren? - Zijn er kaarten of categorieën die jullie missen en/of anders zouden doen?
+10 minuten (totaal: 37 minuten)	Participanten doen individueel de card sort – Eigenschappen van de medische AI applicatie	Timer zetten voor 10 minuten; Audio opname pauzeren;
+10 minuten (totaal: 47 minuten)	Plenary group discussions	Timer zetten voor 10 minuten; Audio opname hervatten; Questions: - Welke afwegingen hebben jullie gemaakt bij het sorteren en prioriteren van de kaarten bij de categorieën? - Wat vinden jullie van de kaarten? - Zijn er kaarten, categorieën die jullie anders zouden zien of toevoegen? - Op een schaal van 1 (niet dekkend) tot 5 (volledig dekkend), hoe dekkend zijn de benoemde kaarten en categorieën om de eigenschappen van een medische AI te weten?
+10 minuten (totaal: 57 minuten)	Participanten doen individueel de card sort – Organisatorische impact	Timer zetten voor 10 minuten; Audio opname pauzeren;
+10 minuten (totaal: 67 minuten)	Plenary group discussions	Timer zetten voor 10 minuten; Audio opname hervatten; Questions: - Hoe ervaarde jullie het sorteren en prioriteren van de kaarten? - Welke afwegingen hebben jullie gemaakt bij het sorteren en prioriteren van de kaarten bij de categorieën? - Op een schaal van 1 (niet dekkend) tot 5 (volledig dekkend), hoe dekkend zijn de benoemde kaarten en categorieën om de organisatorische impact van een medische AI te schatten?

+10 minuten (totaal: 77 minuten)	Participanten doen individueel de card sort – Context voor besluitvorming	Timer zetten voor 10 minuten; Audio opname pauzeren;
+10 minuten (totaal: 87 minuten)	Plenary group discussions	Timer zetten voor 10 minuten; Audio opname hervatten; Questions: - Welke categorieën hebben jullie gemaakt en waarom? - Welke afwegingen hebben jullie gemaakt in de sorteringen en prioriteringen van deze card sort?
	Afsluiting sessie 1	Closing questions: - Op een schaal van 1 (niet dekkend) tot 5 (volledig dekkend), hoe dekkend zijn de benoemde kaarten en categorieën om medische AI te evalueren op meerwaarde? - Zijn er nog evaluatiecriteria die nog missen en wel van belang zijn? - Zijn er nog vragen, onduidelijkheden en/of opmerkingen na deze card sorting sessie? Audio opname stoppen; Closing statements: - Participanten bedanken voor hun deelname en input;

Sessie 2

Tijd (time)	**Acties (Actions)**	**Checklist**
-15 minuten	Conference room preparation	- Link delen naar Maze.co;

		- Tafels in de juiste vorm zetten (vierkant/rechthoek); - Naambordjes plaatsen; - Laptops klaarzetten/ participanten nemen eigen laptop mee; - Opladers klaarzetten, indien nodig; - Audio opname appratuur controleren;
-5 minuten	Inloop	Presentielijst bij af gaan en informed consent controleren.
+15 minuten	Introductie en korte toelichting card sorting	- Checken of iedereen in Maze.co kan komen via de gedeelde link; - Welkom en voorstellen; - Het onderzoek en de aanleiding; - Card sorting toelichten: wat is het en het doel; - De HTA-mAIx toelichten: proces, evaluatieonderdelen, categorieën, criteria (cards); - Instructies hoe deze sessie gaat verlopen.
+10 minuten (totaal: 25 minuten)	Participanten doen individueel card sort – Effectiviteits- en impact assessments	Timer zetten voor 10 minuten
+10 minuten (totaal: 35 minuten)	Plenary group discussions	Timer zetten voor 10 minuten; Audio opname starten; Questions
+10 minuten (totaal: 45 minuten)	Participanten doen individueel card sort – Output classificaties en redeneringen	Timer zetten voor 10 minuten
+10 minuten (totaal: 55 minuten)	Plenary group discussions	Timer zetten voor 10 minuten; Audio opname hervatten; Questions
+5 minuten (totaal: 60 minuten)	Participanten doen individueel card sort – Rapport layout	Timer zetten voor 5 minuten
+5 minuten (totaal: 65 minuten)	Plenary group discussions	Timer zetten voor 5 minuten; Audio opname hervatten; Questions
	Afsluiting sessie 2	Closing questions: - Op een schaal van 1 (niet dekkend) tot 5 (volledig dekkend), hoe dekkend zijn

	de benoemde kaarten en categorieën om medische AI te evalueren op meerwaarde? - Zijn er nog evaluatiecriteria die nog missen en wel van belang zijn? - Zijn er nog vragen, onduidelijkheden en/of opmerkingen na deze card sorting sessie? Audio opname stoppen; Closing statements: - Participanten bedanken voor hun deelname en input;

Appendix B: Entire coding processes and schemes (Tables 2 to 8).

Main code and subcodes	Definition of code	#IT staff	#Medical technology specialists
Gap - **Identifying missing information** - **Identifying affected parties** - **Similar to preferred situation**	- Identifying the gap between current situation and preferred situation - Identifying the involved parties that will be affected by closing the gap - The gap could be interpreted the same as the preferred situation	2	1
Completeness - **Criteria are inclusive**	- The mentioned criteria entirely cover the needed information to identify the functional needs	1	1
Evaluation perspective - **Unclear on multidisciplinary approach**	- The understanding of how to decide the importance of criteria is unclear	1	1

- **Broadness of perspective** - **Considering personal data processing** - **Process owner decides data processing**	- To what extent the users need to consider a micro, meso, or macro level - To consider the necessity to use personal data for medical AI - Process owner decides data processing

Table 2. Functional needs (Dutch: Functionele behoefte) – card sort 1 (n = 3).

Main code and subcodes	Definition of code	#IT staff	#Medical technology specialist
Shared allocation - **Difficulty assigning criteria** - **Various interpretations**	- A criterium that fits not just one category - The definition and interpretation of certain criteria differ per stakeholder group	2	1
Range of importance - **Varying cross-compliance**	- The required qualifications and certifications differ in priority per type of medical AI	1	1
Regulatory - **Compulsory requirements per stakeholder** - **Differing division of responsibilities** - **Differing motives for suppliers in technology** - **Non-compliance consequences** - **Verifying authorities** - **Role of data processing agreement** - **Hospital responsibilities**	- The responsibility to create comply with regulatory differs per stakeholder group - The involved parties each have their on responsibilities in complying with regulatory requirements - Suppliers have different reasons for operating in a certain way - The consequences that might occur when one or more involved parties are non-compliant to regulatory requirements - Authorities that validate the ISO certificate(s) of medical technologies - The role of data processing agreement in being regulatory compliant	2	1

- **Creation of measurement instructions**	- The responsibilities a hospital has to be regulatory compliant		
AI versus medical technologies		2	1
- **Regulatory responsibilities** - **Role of medical device regulations** - **Industrial partnerships** - **Current obligations to regulations**	- The regulatory responsibilities to the involved parties differ for AI and medical technologies - The medical device regulations (MDR) apply for all medical technologies that want to be available on the market - The suppliers anticipate to the knowledge of new regulatory requirements and work together to set the standards - The importance to comply to current regulatory requirements to use technologies in healthcare		
Data processing		2	1
- **Data usage** - **Data quality requirements differ** - **Ownership of medical AI**	- How the suppliers are going to use hospital data - The necessary requirements of the input and output differs per involved perspective - The owner of the medical AI application is responsible for all AI evaluation criteria		
Data evaluations		2	1
- **Equally important** - **Similar interpretations** - **Organisation of ownership**	- The data characteristics evaluation criteria are equally important due to safety and privacy risks - Characteristics and properties perceived similar - The responsibility of organising the ownership of medical AI applications is an organisational characteristic		
Decision-making requirements		2	1
- **Input and output**	- Input and output of medical AI is essential for decision-making		
Hospital responsibilities		1	1

- **Patient expectations** - **Monitoring performance and safety**	- Patients expect the best possible care and data privacy from hospitals - The hospital needs to monitor the reliability and validity of the output to ensure safety		
Black-box evaluation - **Importance of understanding** - **Hospital size dictates requirements** - **Knowledge differs per stakeholder** - **Task differences in chain**	- Understanding the black-box is crucial to maintain quality of care - The hospital size and capacity should be taken into account to determine the importance of understanding the black-box - Not all users need to understand the black-box - Each organisation in the healthcare chain has different tasks	1	1

Table 3. Properties of medical AI (Dutch: Eigenschappen van de medische AI) – card sort 2 (n = 3).

Main code and subcodes	Definition of code	#IT staff	#Medical technology specialist
Ease of use - **Preventing mistakes**	- The ease of use of the medical AI is important for the users to prevent making mistakes	2	1
Shared allocation - **Hardware and software** - **Architecture and financial impact connected** - **Shareability of AI**	- The hardware and software impact technical readiness and financial readiness - The impact on the IT architecture is also connected to the financial impact - The shareability of AI fits multiple categories	2	1
Suitability - **Category and criteria mismatch** - **Interpreting functional versus**	- Criteria that do not match with the available categories - The criteria need to have a clear formulation to interpret is properly	2	1

operational criteria - **Environmental impact assessment importance**	- The perceived importance of measuring the environmental impact of medical AI		
Criteria comprehensibility - **Level of applicability**	- Interpreting the level of applicability difficulties	1	1
Perceived complexities - **Broadness of considerations** - **Purchasing approaches** - **Perceived usefulness** - **AI knowledge and usage**	- Considering a broad perspective on impact on collaborations in the chain is perceived as too complex - How a medical AI application is purchased determines the impact on external collaborations - The perceived usefulness differs per stakeholder group in the chain - Understanding the medical AI determines how useful it is to healthcare	2	1
Workflow changes - **Perceived impact on workload**	- Changing workflows could negatively impact the workload in the chain	2	1
Partnerships - **No national collaborations** - **Partnerships versus impositions** - **Perceived level of control** - **Regional collaborations** - **Considering future opportunities** - **Role awareness in partnerships**	- No existing national partnerships - Some collaborations are voluntary and others are obligatory - The say of a hospital in collaborations is influenced by mandate and favouritism - Regional collaborations more likely - Prioritizing partnerships with hospitals in medical AI might change if they start to work together instead of individually - Each hospital has a certain role or task in partnerships	2	1

- **Demanding requirements** - **Contributions in partnerships** - **Financial and capacity considerations**	- Within partnerships hospitals can demand certain requirements from each other to compare, which is complex - The contributions from each hospital affects the contributions from other hospitals - Contributions are dependent on the financial and capacity of staff		
Financial considerations		1	1
- **Impact of human support and organisation**	- Perceived financial impact of human support and the organisational implications need to be included		
Different interpretations		1	1
- **Supplier maturity & regulatory readiness**	- Maturity of medical AI supplier could be considered regulatory		
Hospital size		1	1
- **Tertiary versus secondary hospitals**	- Possibilities for further AI development differs per hospital size		
Completeness		2	1
- **Criteria-category fitting** - **Organisational perspective instead of user perspective** - **Staying organisation-minded**	- The framing of the criteria affects the coverage of the categories - To determine the organisational impact of medical AI, an organisational perspective is preferred - It is important to keep the hospital or organisation at the centre in decision-making about medical AI		
Organisational readiness		2	1
- **Helicopter view** - **Organising ownership**	- Important to look at the organisational readiness as the overall picture - Organising ownership about medical AI is a crucial first step		

- **Acquiring knowledge and skills**	- The available knowledge and skills about medical AI is important for a hospital to ensure long-term embedding	

Table 4. Organisational impact (Dutch: Organisatorische impact) – card sort 3 (n = 3).

Main code and subcodes	Definition of code	#IT staff	#Medical technology specialist
New categories		2	1
- **Organisation vision/strategy** - **Application/ project value case** - **Application/ project implementation** - **Application/ practice** - **Financial** - **Distinguishing organisation and AI**	- Separating functional from operational criteria and categories - The evidence that medical AI applications are beneficial to the hospital and are value for money - The impact on healthcare insurance costs - Importance to distinguish the criteria organisationally and AI-specific		
Range of influence		2	1
- **Healthcare insurance out of range** - **Negotiating return on investment**	- The influence a hospital has on healthcare insurance decision-making - Possibility to influence healthcare insurers' support by estimating return on investment		
Intertwining		2	1
- **Organisational impact as subcategory** - **Prior knowledge**	- Organisational impact could be considered a subcategory of decision-making context - Knowledge about the organisational impact before decision-making		
Missing criteria		2	1
- **Impact on staff** - **Current versus new functionalities**	- The impact medical AI has on the staff - Comparing current functionalities of workflows to the estimated functionalities		

	of the workflow in which AI is integrated		
Real-world effectiveness - **Theory versus reality**	- Important to differentiate between the theoretical effectiveness and the practical effectiveness	1	0
Decision-making requirements - **Risk classifications** - **Control of suppliers** - **Transparency of suppliers**	- The risk assessments are important to even start new processing agreements - The control suppliers have on healthcare and integrating medical AI - The level of transparency of a medical AI supplier affects the decision-making	2	1

Table 5. Decision-making context (Dutch: Context voor decision-making) – card sort 4 (n = 3).

Main code and subcodes	Definition of code	#IT staff
Intuitiveness - **Understanding health economic evaluations** - **Understanding cost-benefit analyses** - **Understanding assessment differences** - **Comprehensibility affects prioritisation**	- Interpreting health economic evaluations differs per perspective - Sorting cost-benefit analyses to health economic evaluations differs per stakeholder - Interpreting the assessments differs per stakeholder group - Knowledge and expertise about assessments influences their prioritisation	1
Different allocation - **Knowledge for decision-making**	- Certain assessment-criteria need to be known before starting the assessments	1
Tradeoffs - **Privacy measures versus costs**	- Importance of weighting the necessary privacy measures to the measures' costs.	1
Assessment prerequisites - **Comply or explain** - **Compliance assessments**	- For a privacy risk to be acceptable, a substantiation is needed to explain and determine whether it is valid or not	1

	- Risk analyses are prerequisite in hospitals	
Process ownership		1
- **Cost awareness**	- Process owner needs to be aware of all hospital costs	
Benchmarking		1
- **Clinical effectiveness measures**	- Comparing clinical effectiveness outcomes between hospitals	
- **Data output**	- Comparing clinical effectiveness outcomes between hospitals	
Compliance prerequisites		1
- **Similar purposes**	- Some privacy assessments have the same definition and aim	
- **Combining assessments**	- Assessments could be combined to one if the risks are similar	
- **Similarities in compliance measurements**	- Assessments could be combined if the compliance measures are similar	
- **Privacy compliant**		
- **Ensuring conformity**	- Being privacy compliant is most important in the decision-making about medical AI	
	- Using the same privacy guidelines for all medical technology that uses data	
Renaming		1
- **Risk analyses instead of compliance**	- Renaming the category risk analyses to fit with all types of assessments	

Table 6. Effectiveness and impact assessments (Dutch: Effectiviteit en impact assessments) – card sort 5 (n = 1).

Main code and subcodes	Definition of code	#IT staff
User Interface explainability		1
- **Understanding advice indications**	- Explainability of the output of the decision support system needs to be clear	
Reasoning content requirements		1
- **Answering the why**	- Importance to address why a certain advice is given by the decision support system is crucial	
- **Considering complexity of subject**		
- **Considering necessary adjustments**	- Importance to consider lowering the complexity of the medical AI application by stating the	

- **Weighting criteria for ranking recommendations** - **Conformation to hospital formats**	reasonings in bulletpoints form most important to least important information - Well-substantiating the advised adjustments based on previous steps and level of importance to do each adjustment - Indicating the importance of each criterium based on weights - Importance of matching the formats to the hospital formats	

Table 7. Advice and reasonings (Dutch: Advies en redeneringen) – card sort 6 (n = 1).

Main code and subcodes	Definition of code	#IT staff
Visualisations		1
- **Combining words and visuals** - **Intuitiveness**	- Starting with visualisations to words is most suitable for everyone - Using traffic lights is intuitive and universal to comprehend	
Customizations		1
- **Customizable layout** - **Chronologically content** - **Including rationales** - **Prioritising advice substantiation** - **Only relevant information** - **Decision-making perspective**	- The user's preferences in the report layout should be customizable in the most appropriate software program (e.g. Word or PowerBI) - Report layout should be chronologically - The report content needs to include rationales and be logically written, which does not have to be chronologically - Well-substantiating the advice of the decision support system is most important to address in the report - The report must differentiate between primary information and secondary information to substantiate the advice	

	- The content needs to match with the information needs of the decision-maker(s)	
Periodic evaluations - **Remain in control** - **Changing privacy risks** - **Re-assessing first**	- Evaluating periodically to remain in control and support decision-makers - Re-evaluation necessary when privacy risks change in the implemented medical AI application - Consider the changes to decide to what extent the re-assessment needs to be done	1
Multidisciplinary approach - **Multidisciplinary input necessary** - **Multidisciplinary workflow**	- Due to the complexity of hospitals, multiple and different expertise is required to evaluate medical AI applications - Multiple and different experts need to collaborate in the evaluation of medical AI applications	1

Table 8. Report layout (Dutch: Rapport layout) – card sort 7 (n = 1).

Appendix C: Prototype Decision Support System of the HTA-mAIx

*Not included in online document.

Appendix D: Playbook moderated in-person feasibility-usability testing sessions

Tijd (time)	**Acties (actions)**	**Checklist**
-15 minuten	Zaal voorbereiden	- Tafels in de juiste vorm zetten; - Laptops en audio-installatie klaarzetten; - Het prototype DSS en Maze.co openzetten op de laptops van de deelnemers; -

-5 minuten	Inloop	- Presentielijst bijhouden; - Informed consent controleren of eventueel laten invullen voor de start van de sessie;
+10 minuten	Introductie	- Het onderzoek en aanleiding toelichten; - Het prototype DSS kort toelichten: hoe ontwikkeld, doel en hoe bedoeld is om te gebruiken; - Het programma benoemen met kort de instructies en tijdsindicaties, inclusief uitleg thinking aloud; - Aangeven dat de deelnemers mij gedurende de sessie vragen mogen stellen als ze iets niet begrijpen;
+25 minuten (totaal: 35 minuten)	Feasibility testing	- Multidisciplinaire activiteit voor de deelnemers; - Deelnemers bladeren door het prototype DSS en vullen vragenlijst in Maze.co in; - Thinking aloud.
+5 minuten (totaal: 40 minuten)	Afronding feasibility testing en usability testing introduceren	Afronding feasibility testing: - Vragen of iedereen alle vragen heeft kunnen invullen; - Vraag: hoe ervaarden jullie het prototype DSS wat betreft hoe het eruit ziet? Intro usability testing: - Elke deelnemer papier met taken en benodigde informatie geven; - Eén laptop voor de deelnemers overlaten waar ze samen de usability test uitvoeren; - De AI case toelichten; - Het proces, doel en tijdsindicatie benoemen;
~~*+30 minuten (totaal: 70 minuten)*~~	~~Usability testing~~	~~- Screen recording en audio recording starten;~~ ~~- Op eigen laptop meekijken met deelnemers;~~

		- ~~Vragen of onduidelijkheden beantwoorden wanneer nodig;~~
+5 minuten (totaal: 75 minuten)	Afsluiten	Afsluitende vragen: - Hoe ervaarden jullie het uitvoeren van deze testen? - Zou het DSS ter ondersteuning bruikbaar zijn in de praktijk? - Zijn er nog vragen of opmerkingen? Afsluiting onderzoek: - Screen recording en audio recording stoppen; - Deelnemers bedanken voor hun inzet en bijdragen; - Contactgegevens meegeven;
10 minuten	Ruimte herinrichten naar originele staat	- Laptops uitschakelen; - Tafels en stoelen terugzetten; - Eventuele rommel opruimen; - Laptops terugbrengen naar Martini Academie; - Zaal afsluiten (indien nodig);

Appendix E: The feasibility questionnaire

*Not included in online document.

Appendix F: The entire deductive/inductive hybrid thematic analysis results and mappings (Figures 12 and 13)

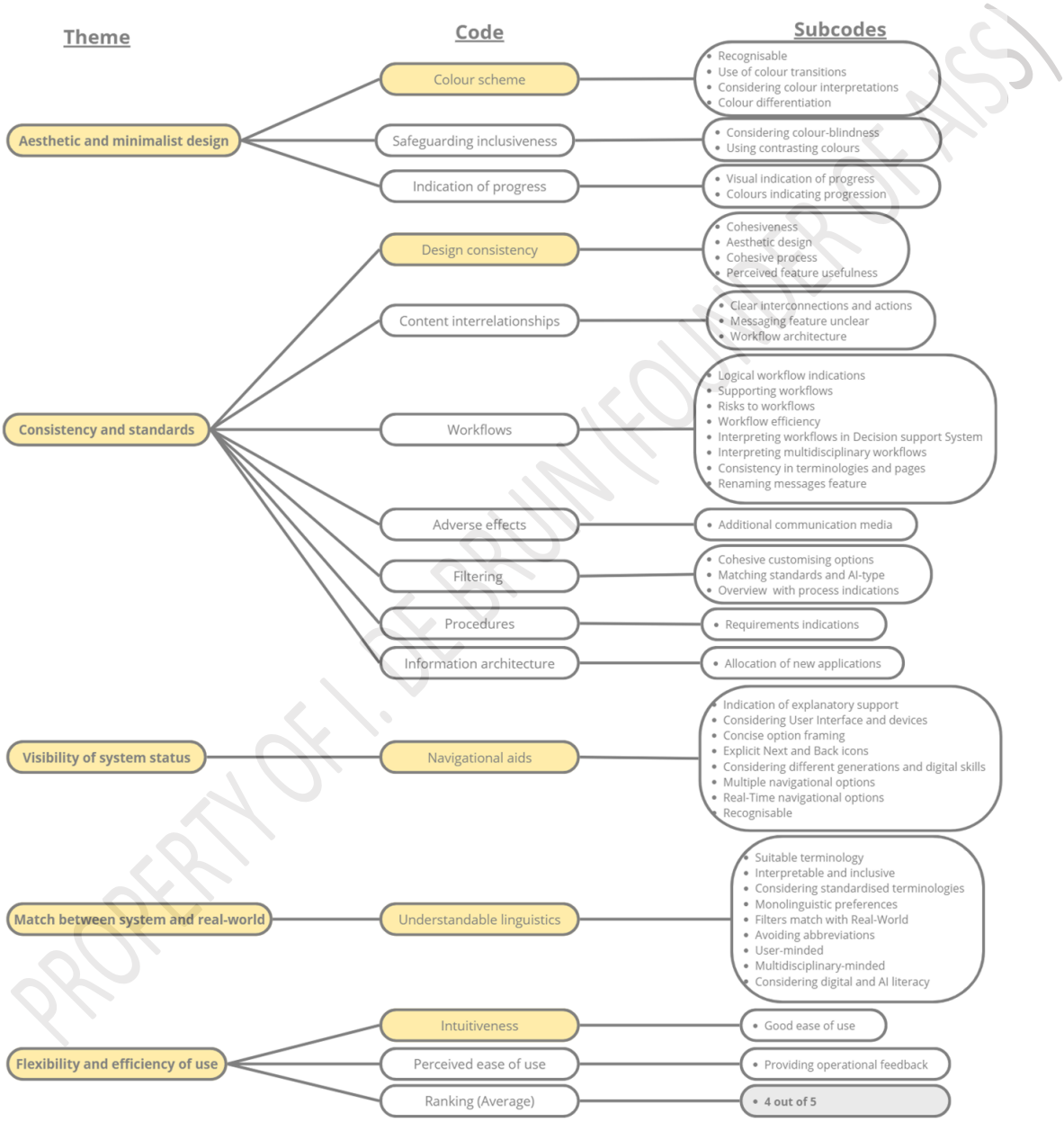


Figure 12. Map of themes, codes and subcodes to the design of the prototype DSS.

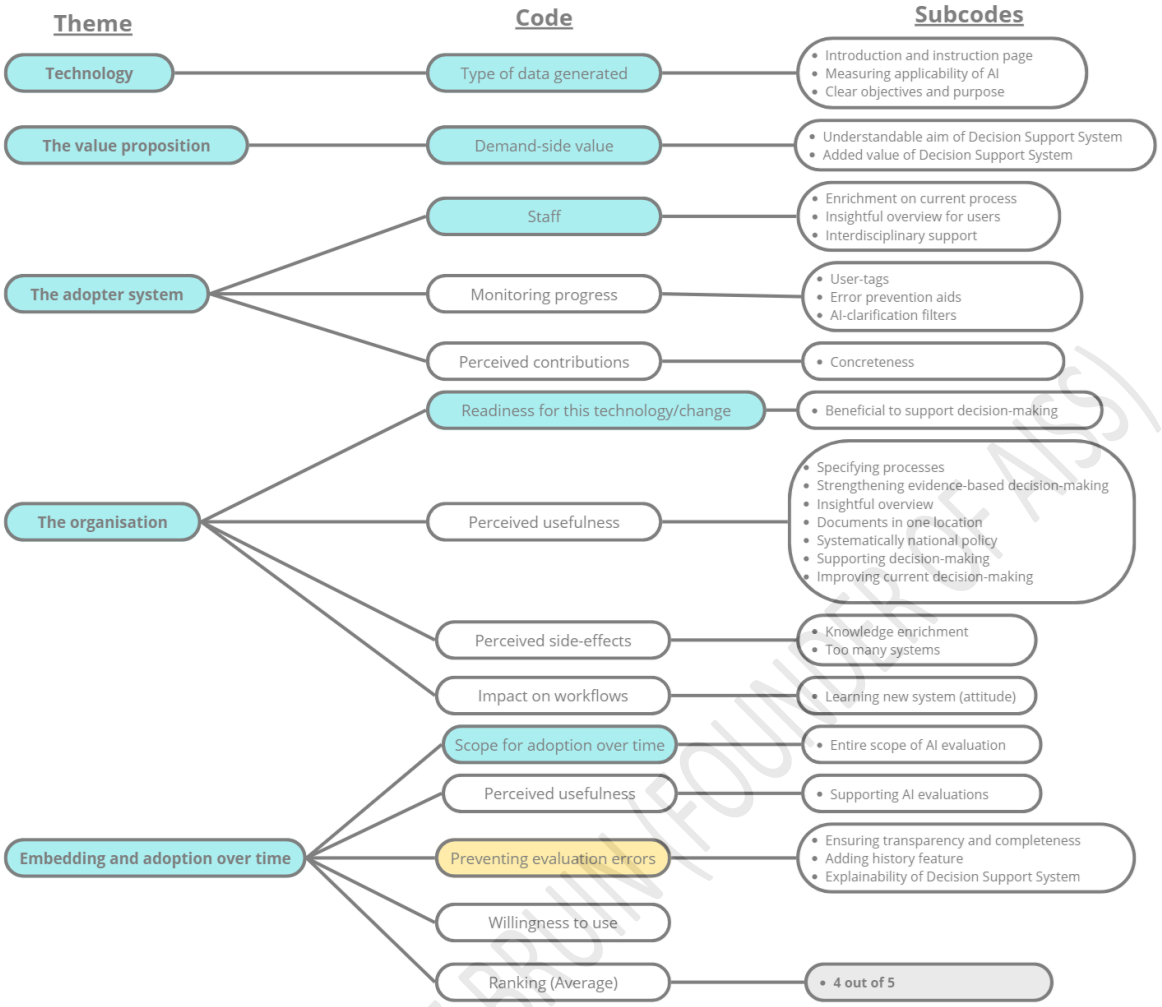


Figure 13. Map of themes, codes and subcodes to the content of the prototype DSS.