Design and Evaluation of a Prototype for a Platform for AI algorithms in Medical Imaging: A Human-Centered Approach

Christof Schulz
s1934236

Master Thesis Human Factors and Engineering Psychology, 25 ECs
Department of Cognitive Psychology and Ergonomics
Faculty of Behavioural, Management and Social Sciences
University of Twente

First Supervisor: Dr. Simone Borsci

Second Supervisor: Dr. Martin Schmettow

UNIVERSITY OF TWENTE.

Abstract

Artificial intelligence promises to facilitate radiological tasks, like segmentation. However, the quality of interaction experienced by humans is minimally investigated. In this study, a multiphase work consisting of three research phases is presented to **1.** assess the quality of interaction between humans and AI in medical imaging, to **2.** propose a new implementation based on human-centered design methodology and to **3.** assess the quality of interaction between humans and the new prototype. In the first phase, a usability test (n=14) assessing the interaction between professionals and a segmentation AI model, embedded in open-source software, revealed that the implementation of the AI algorithm led to several usability problems (16 identified) and low satisfaction in use (50.2% (SD:10.88)). In the second phase, human-centered design methods were applied to develop a novel prototype for integrating medical imaging AI tools. In the third phase, a usability assessment of the new prototype with clinicians and other regular users of medical imaging software (n=13) showed that the new prototype seems to facilitate good usability with less usability problems (10 identified) and good satisfaction in use (89,6% (SD:4.55). Qualitative data were collected and analyzed for further feedback on the design of the prototype, and final suggestions for the further development of the new platform were provided.

Contents

1. Introduction

Due to the advancement of more capable graphics processing units (GPUs) and breakthroughs in deep learning (DL), artificial intelligence (AI) has made remarkable progress over the past years, including many new applications in radiology (Hosny, Parmar, Quackenbush, Schwartz, & Aerts, 2018). Various AI solutions have been developed for feature detection, disease classification, image segmentation, treatment planning, and many other tasks (Hosny et al., 2018; Panayides et al., 2020). If sufficiently developed, AI algorithms can provide support to these tasks within a fraction of the time which would be required by human operators. Thereby, AI in medical imaging could help to cope with the ever-expanding workloads on radiologists, enabling radiologists to focus on more complicated studies, and ultimately improve patient care (Kotter & Ranschaert, 2021; Omoumi et al., 2021).

Although AI in medical imaging has great potential, the adoption and deployment of AI in clinical practice remain limited. Some AI algorithms have been developed to the point in which their performance resembles or even surpasses the performance of human operators (Bejnordi et al., 2017; Esteva et al., 2017; Wang et al., 2017), but regardless they are not consistently used in the daily workflow of clinicians, yet (Hwang & Park, 2020; Lekadir et al., 2021). Scheetz et al. (2021) surveyed radiologists in Australia and New Zealand and reported that the vast majority of radiologists agree that the introduction of AI would improve their field. However, over 80% of the respondents have not yet used AI in their day-to-day practice. This raises the question of why AI, regardless of good performance, is not consistently used by

practitioners.

The implementation of AI algorithms into clinical practice remains challenging for a variety of reasons. Transforming healthcare with AI is not a trivial task, and it requires various stakeholders to come together and resolve ethical and legal concerns before AI tools can be utilized for patient care (Char, Shah, & Magnus, 2018; Schönberger, 2019). Furthermore, AI-based tools may not generalize beyond the data on which they were trained. Differences between the training sets and the real-world population can result in biases and erroneous predictions (Panayides et al., 2020; Willemink et al., 2020). The performance of DL models greatly depends on the availability of data on which the models can be trained. Sufficiently large, curated, and representative datasets have to be provided for the further development of AI models, and more advanced approaches for the sharing of datasets among different organizations is required. (Currie, Hawk, Rohren, Vial, & Klein, 2019; Panayides et al., 2020; Willemink et al., 2020). Sufficient raw image data typically exist in the databases of hospitals and clinics, but these images are not quite usable for the training and validation of AI algorithms. (Currie et al., 2019; Panayides et al., 2020). Often times, the imaging data needs to be curated by human experts to create datasets containing a "ground truth" from which newly developed AI models can learn and be validated (Currie et al., 2019). This procedure requires highly-skilled medical imaging experts to annotate a large number of medical images before they can be used for the further development of AI models. Thereby, possitbilities for annotating and sharing image datasets need to be optimized. Nevertheless, the availability of annotated datasets has improved throughout the last few years (Willemink et al., 2020; Zhang et al., 2021).

Another challenge for the consistent application of AI in medical imaging is the implementation of newly developed systems into existing work environments. Clinicians need to accept new systems, and they also need to comprehend how to interpret the information which is provided by such systems. However, the human workforce was not prepared for the AI evolution, and therefore, some physicians meet the implementation of AI with resistance (Briganti & Le Moine, 2020). AI algorithms can be perceived as complex, opaque, difficult to comprehend, or untrustworthy (Lekadir et al., 2021). Accordingly, clinicians need to be convinced that the benefits from the interaction with AI outweigh these concerns (Filice & Ratwani, 2020). Superior effectiveness and efficiency compared to the current modus operandi needs to be demonstrated and, most importantly, clinicians need to be satisfied with the interaction (Lekadir et al., 2021). Thus, the successful implementation of AI-based systems requires the careful consideration of the requirements and preferences of human operators in

order to facilitate the ease of use, satisfaction, and ultimately the acceptance of new AI technology.

The careful consideration of human capabilities and user requirements is scientifically rooted in the discipline of human factors and ergonomics (HFE). HFE is the application of knowledge about human capabilities (physical, sensory, emotional, and intellectual) and limitations to the design of artifacts to optimize the users' ability to accomplish their tasks error-free and in a reasonable amount of time and, therefore, to accept the system (Hegde, 2013; Wickens, Gordon, Liu, & Lee, 2004). A central element of HFE design is the usability of a system. It is the extent to which a product can be used efficiently, effectively, and satisfactorily by specified users in a specified context of use (ISO, 2018). Filice and Ratwani (2020) and Lekadir et al. (2021) argued that usability and the consideration of user requirements are important factors for the successful implementation of AI in clinical work environments. Applying HFE methods to maximize the usability of AI systems in medical imaging could ensure that tools are designed, developed, implemented, and used with a continuous focus on accentuating clinician performance from introduction through long-term use (Filice & Ratwani, 2020; Lekadir et al., 2021). Nevertheless, Lekadir et al. (2021) emphasized that usability and human factors are still not being adequately addressed when it comes to the design of AI systems. The authors emphasized that there are very few relevant scientific publications addressing the need and the approaches to facilitate usability of AI systems. The authors stated that poor usability of AI models might be one of the reasons for the limited translation of research in the field of AI to clinics (Lekadir et al., 2021). In line with Lekadir et al. (2021), Shneiderman (2020) argued that the acceptance and the adoption of AI technology depend on the human-centeredness of AI systems, and achieving human-centered AI will considerably increase human performance, while promoting self-efficacy, mastery, creativity, and responsibility. Thus, the aim of this research is to apply HFE methodology and human-centered design to maximize the usability and to promote the use of AI in medical imaging.

1.1 The Need for a Standardized Structure for a Human-Centered Approach to the Design of AI-based Systems in Medical Imaging

New AI technology in medical imaging is not supposed to replace clinicians, but instead it needs to augment clinicians in their profession (Liew, 2018; Pianykh et al., 2020; Thrall et al., 2018). Clinicians will operate with the assistance of AI, and the interaction between clinicians and AI becomes more important as new AI-based systems are implemented in existing workspaces. However, extensive studies of the interaction between AI systems and human operators in the field of medical imaging are still required (Felmingham et al., 2021;

Lekadir et al., 2021). The vast majority of the currently published studies focus on the technical development and the performance of AI models (Sujan et al., 2019). Thereby, existing evidence regarding the quality of interaction between AI tools and clinicians remains limited and proper design and integration of AI-based systems into existing workflows need to be investigated (Asan & Choudhury, 2021; Felmingham et al., 2021; Sujan et al., 2019). Thereby, extensive user testing with new AI systems is necessary to ensure that they meet the requirements of users and that they can be applied to support human goals, activities, and values (Shneiderman, 2020). Carayon et al. (2020) argued that one of the reasons for the poor real-world implementation of AI systems in clinical environments is the lack of usability and workflow integration.

The design of usable, visual, and interactive elements to support humans and to enhance the interaction with AI systems is not a trivial task. On one hand, AI systems need to be designed in such a way that they can be easily integrated into clinical workflows and, in most of the cases, the AI models need to be integrated in existing interfaces or systems (Filice & Ratwani, 2020; Lekadir et al., 2021; Omoumi et al., 2021). On the other hand, the information provided by these systems needs to presented in such a way that it facilitates the work of physicians. Suboptimal presentation of AI results to humans can lead to biases and fallacies (Alon-Barkat & Busuioc, 2023). For example, the immediate presentation of AI results can result in automation bias which leads to overcompliance and the actual decrease of the clinician's performance (Alberdi, Povyakalo, Strigini, & Ayton, 2004; Sujan et al., 2019). Presenting probabilistic information may confuse the reader in the interpretation of the results (Currie et al., 2019), or lead to selective adherence which causes selective adoption of AI advice in correspondence with one's own stereotypes and beliefs (Alon-Barkat & Busuioc, 2023). Conversely, presenting results in simplified terms neglects the uncertainties and complexities when working with AI-generated predictions, and may lead to inaccuracies in the assessment of the situation. To avoid fallacies and biases, the appropriate presentation of AI-generated results to human operators requires more research. If human-machine interfaces of AI systems are designed without taking human capabilities into account, they may oppose new impediments for clinicians instead of increasing the quality of provided healthcare services.

1.2 Theory of Human-Centered Design

To create human-centered AI systems which are tailored to the needs of users, methods such as human-centered design (HCD) could be employed. HCD is an approach to interactive systems development that aims to make systems usable and useful by focusing on the users, their needs and requirements, and by applying human factors/ergonomics, and usability knowledge and techniques (ISO, 2019). The term "human-centered design" is used rather than

"user-centered design" in order to emphasize that it also addresses impacts on a number of stakeholders, not just those typically considered as users. However, in practice, these terms are often used synonymously (ISO, 2019). The HCD process is typically iterative. In iterative design, the designer builds a usability-engineering life cycle around the concept of iteration (Nielsen, 1993). Phases of requirement analysis, design, testing, and evaluation will be repeated as often as necessary in order to produce an optimal version of the product. After completing a design, usability researchers note the issues they identified in the testing phase. They then propose fixes to these problems in a new iteration, which is then tested again to ensure that the "fixes" from the previous iteration did indeed solve the problems instead of creating new ones (Nielsen, 1993). Throughout multiple iterative cycles the designed product is evolved from a low-fidelity prototype to a readily usable product. This procedure ensures that user requirements are taken into account and ultimately promotes the usability of the system (Nielsen, 1993).

HCD is characterized by user involvement at all stages of the design (Abras, Maloney-Krichmar, & Preece, 2004). The active involvement of users and experts is especially important nowadays when designers are collaborating on increasingly complex projects (Maguire, 2001; Mao, Vredenburg, Smith, & Carey, 2005; Muratovski, 2021). Designers cannot be as knowledgeable as all the different types of users, and they cannot comprehend all the experiences of use they aim to create (Bruseberg & McDonagh-Philp, 2001). Consequently, multidisciplinary design approaches are becoming more common (Muratovski, 2021). For example, building a human-machine interface for the integration of AI technology in medical imaging does not only require expertise in human-computer interaction, but also knowledge from the domains of medicine, medical imaging, AI, data science, and software development. As designers typically do not have sufficient expertise in the required domains, collaboration with domain experts and intended users is essential for the success of the design project. In such cases, experts from different backgrounds share their knowledge and experience from the view of their own disciplines, and the result is a multidisciplinary team working towards a co-designed outcome (Muratovski, 2021). For instance, when designing for clinicians, medical practitioners could help designers to establish the parameters and the terminology of the problem and work closely with them through all stages of the design process by providing the necessary feedback (Muratovski, 2021).

To the best of our knowledge, guidelines and requirements for the design of human-centered AI systems in medical imaging are still lacking. Nevertheless, the application of classical HCD methodology could be useful for the design of the first generation of human-

centered AI systems. Because HCD is characterized by user involvement, it leads to the definition and the documentation of user requirements for new systems (Harte et al., 2017). Furthermore, usability testing is conducted in HCD cycles. Usability test take quantifiable measures of usability from the interaction between users and the prototype, and they allow for interference on the effectiveness, efficiency, satisfaction in use, learnability, and safety of newly designed systems (Lewis, 2006). Ensuring that newly designed AI systems have a good usability could show that these systems can be used as intended, efficiently, satisfactorily, and without causing unexpected issues for clinicians, and therefore, pave the way for a successful implementation of new AI systems.

1.3 Previous Applications of Human-Centered Design and Usability Testing in Medicine

HCD and usability testing are succesfully applied in medicine, already. Devices and systems which are used in critical work environments such as clinics need to be designed with careful consideration of the users operating these systems, and thus HFE have been relevant to the design of medical devices for decades. The safe and reliable use of medical devices is especially important in healthcare systems, because flaws in the design can lead to patient harm, and poorly designed medical devices have reportedly been causing harm to patients in the past (Hegde, 2013; Schmettow, Schnittker, & Schraagen, 2017). Several examples for the application of HFE methodology to the design of medical devices exist in the literature. In the following, three studies in which HFE methods are applied to the design and evaluation of medical devices are summarized.

Harte et al. (2017) made a use case for the application of a HCD procedure to enhance the usability, human factors, and the user experience (UX) in a connected health system. They provided a structured methodology to ensure that user needs are taken into consideration during the design process while maintaining a rapid pace of development. They proposed a three-phase approach for HCD: In the first phase, a use case is created and user requirements should be defined. This phase involves methods to elicit and visualize the user requirements for the design of the product, such as user interviews, storyboards, paper prototypes, or mockups. The second design phase involves expert inspections. Different types of evaluation methods could be carried out here. For instance, the authors described that a multidisciplinary expert group could review the materials which were designed in the previous phase by conducting a heuristic evaluation or cognitive walkthroughs. In the third phase, a usability test is conducted. It can feature various methods such as task assessment, think-aloud, and several validated scales for quantified measurements such as satisfaction or workload, depending on the goals of the project and the readiness of the prototype. The authors applied the proposed methodology to their use

case and showed that the methodology is indeed useful for the rapid development of a human-centered system in medicine. They further emphasized that their methodology offers a structured design approach that aids with the documentation of goals and requirements, while taking the user needs into account, as well.

Schmettow et al. (2017) provided a detailed protocol for the usability testing of medical infusion pumps. The authors reported that a commonly used infusion pump was equipped with a poorly designed interface. Consequently, a combination of HCD methods was applied to propose a re-design for the interface of the infusion pump. In a subsequent usability test, they aimed to compare how participants performed on both interfaces when conducting a set of representative tasks. A within-subject design with three testing sessions was employed, and a combination of usability measures such as task completion, deviations from the optimal pathway, and time on task were assessed. Furthermore, the performance of the participants was compared during each of the three testing sessions to account for learning effects. Concludingly, The authors reported that already after the second testing session the participants performed better when using the new design of the infusion pump. The authors showed also that the usability of medical devices such as infusion pumps can be improved by applying HFE design methodology, and furthermore they provided a structured approach for profound usability testing to compare different designs (Schmettow et al., 2017).

To organize and arrange the different data resources needed for the development of AI in medical imaging, García-Peñalvo et al. (2021) proposed a design for a user-friendly platform to edit medical images and to apply available AI algorithms to stored medical images. They conducted a requirement analysis, made use cases, and performed a heuristic evaluation based on Nielsen's ten heuristics for the design of user interfaces (UIs) (García-Peñalvo et al., 2021; Nielsen, 2020). The results of the heuristic evaluation showed overall very high scores for the design of the new platform, as the application of design heuristics for the design of the novel platform seemed to facilitate the quality of interaction. The experts also identified issues in the proposed design such as violations of design principles in the image editor and a suboptimal integration of the AI algorithms. For instance, to increase the ease of use of the integrated AI algorithms, the designers decided only to present the name of the AI algorithms, but the functionality and description of the AI algorithm were not shown. This confused the expert evaluators as it violated usability principles, like for instance the visibility of the system status. García-Peñalvo et al. (2021) stated that one of the main challenges in designing medical imaging platforms with integrated AI tools lies within finding the right balance in providing all the relevant information to the user without overwhelming them.

Harte et al. (2017), Schmettow et al. (2017), and García-Peñalvo et al. (2021) showed how different HFE methods can be used for human-centered and usable design, and they proposed methods for the evaluation of newly designed systems. HFE theory provides a wide range of methodologies which can be applied when designing or testing new systems. Based on the design state and the goal of the project, different approaches can be chosen. Among other publications, these previous applications of HFE theory and HCD can serve as a methodological foundation for the design of the first generation of human-centered UIs for the integration of AI in medical imaging.

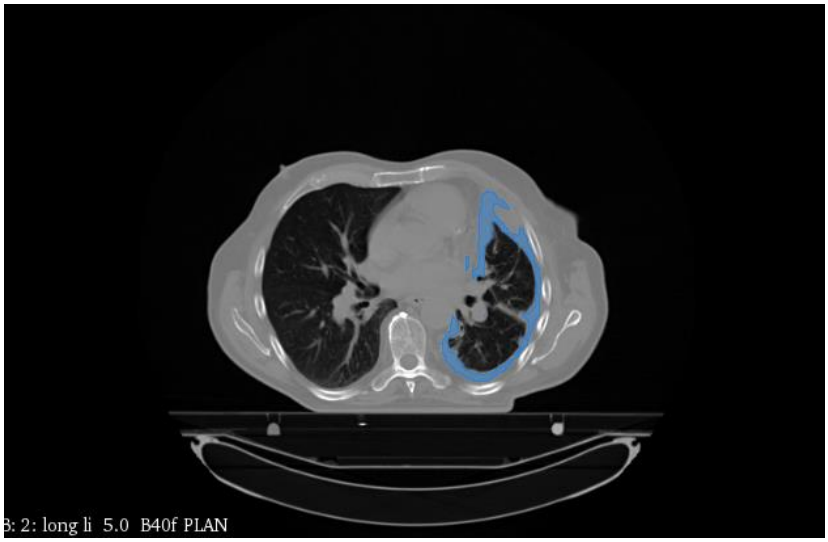
1.4 Making a Use Case for the Design of an AI-Based Human-Machine Interface in Medical Imaging

A use case was made in order to design an AI-based medical imaging system based on HFE methodology. Use cases are a commonly used method to analyze user requirements and user preferences, and they address all foreseeable aspects of use. They can be viewed as a reference point for the further development of the system (Harte et al., 2017). For the use case in this study, we selected an AI algorithm which was developed at the Netherlands Cancer Institute (NKI). The NKI is a large research institute which employs more than 50 research groups. Proceeding innovative research reveals new possibilities of treatment, and the aim is to increase the prospects of patients. For this reason, the NKI highlights the importance of translational research to get a thorough understanding of the disease but also to have an impact on the patient well-being (NKI, n.d.).

For the present use case, an AI algorithm for the automatic segmentation of mesothelioma from CT scans was selected. Mesothelioma is a malignant cancer type of the mesothelium: a thin tissue layer that covers internal organs (Robinson, Nowak, Robinson, & Creaney, 2008). The trained algorithm estimates which areas in a CT scan-series are recognized as mesothelioma and then performs automatic image segmentation on the estimated tumor areas (see Figure 1).

Figure 1

Automated segmentation of the estimated tumor area on a CT scan, created by an AI algorithm



The automatically generated segmentation could be used to estimate the volume of the tumor. AI generated volumetric assessment of mesothelioma is promising because it could be faster and more accurate than the current gold standard evaluation which is called “modified response evaluation criteria in solid tumors” (mRECIST) (Murphy & Gill, 2017). In the mRECIST procedure, the clinician measures two diameters of tumors and estimates tumor growth or recession based on these two individual measurements (Lencioni & Llovet, 2010). This procedure is the current standard in the clinics as it is quick and easy, but it leads to high inter-observer variability (Murphy & Gill, 2017). Volumetric assessment of the tumor could solve this issue, but the manual calculation of the tumor volume takes too much time for human operators and is therefore not applicable in clinical practice. An AI algorithm, however, could carry out this task within seconds.

Although the technological development of the AI algorithm for the automatic segmentation of mesothelioma from CT scans was successful, it is currently used for research purposes only. The AI can produce remarkable results and could prove useful for real world implementation. However, it is not clear if clinicians who typically do not have the same technical understanding as the developers of the algorithm would be able to apply and utilize the AI algorithm as intended, and the AI algorithm is not readily implemented in clinical systems, yet. Extensive user tests and usability assessments for the utilization of the algorithm have not been conducted. The requirements of clinicians and other end users for appropriate utilization of the algorithm are not defined. Furthermore, the AI algorithm is integrated into a

specific software application and can only be initiated by users when working with the UI of this program. Therefore, the usability of the UI in which the algorithm is embedded also plays an important role for the usability of the AI algorithm. To investigate the usability of the AI algorithm, a state-of-the-art assessment in form of a usability test was conducted to assess the interaction between users and the AI algorithm embedded in the UI.

1.5 Overall Goals and Research phases for the Development and Testing of a Human-Centered AI System in Medical Imaging

In this study, we report a multiphase work consisting of multiple research phases which build up on another. The overall goal of this work was to assess the current implementation of the AI tool for the automatic segmentation of mesothelioma and, based on the outcome of the initial usability assessment, to provide an evidence-based methodology for the generation of data, user requirements, stakeholder requirements, and design concepts for a human-centered system for the integration of AI tools in medical imaging in the future. To achive this goal, three research phases were conducted:

- ***Phase 1:*** In this phase a usability assessment was conducted to analyze the current integration of an AI tool in an open-source software. The results of the initial test resulted in a recommendation for a more usable design of the software application in which the AI algorithm is integrated.
- ***Phase 2:*** In the second phase, a new prototype for a novel UI for AI tools in medical imaging was proposed. HCD methods such as focus group reviews and heuristic design were utilized for the iteration of a new prototype to develop the design for a new platform for the integration of AI tools in medical imaging.
- ***Phase 3:*** In the third and final phase, the usability and functionality of the new prototype were assessed by clinicians and other intended users, and final recommendations for the further development of the system based on the findings were provided.

2. Phase 1. Initial Assessment of the AI Tool Embedded in Open Source Software

The AI algorithm which was selected as a use case is currently integrated into a software application called 3DSlicer (slicer.org). 3DSlicer can be used for various purposes in medical image analysis. It is currently applied for the integration of the AI algorithm, because it has advantages for developers and researchers: It is easy to integrate AI-models, free, open-source, easy to customize, and officially supported by NVIDIA. However, 3DSlicer is not explicitly

designed for the integration of AI algorithms and it is rather an additional feature in the versatile software application. Thereby, the aim of this assessment was to find out whether the integration of 3DSlicer contributes to good usability of the AI algorithm. The usability test could furthermore prove useful to find out if users are able to make use of the AI algorithm as intended and it could help to elicit user requirements for future implementation of AI algorithms.

2.1 Methods

2.1.1 Design

A preliminary usability test was conducted to assess the quality of interaction and the satisfaction of users with the AI algorithm and its implementation. Triangulation of methods was priotized to assess the interaction in multiple ways and to produce valid and coherent results of the current level of usability. The test was conducted in a controlled lab-setting. It involved various tasks which participants were asked to solve while concurrently thinking aloud. Task performance measures, satisfaction, post-task surveys, as well as eye-tracking were employed to evaluate the interaction between users and the AI-algorithm embedded in the UI of 3DSlicer.

2.1.2 Participants

Participants were recruited via convenience sampling. In total, 14 participants completed the study. The age of the participants ranged from 22-42 (m=30.5). Regarding the background of the participants, nine participants were PhD candidates or master's students. Another five participants were employed as radiologsts at the NKI. For the eye-tracking study, nine participants were included in the analysis. Due to poor quality of the gaze samples from the eye-tracker, five participants had to be excluded.

2.1.3 Materials

Workflow & User Goals. To test the AI algorithm, user goals for the utilization of the algorithm in 3DSlicer were defined. The user goals were derived by observing the workflow of clinicians who were applying the AI algorithm to annotate medical image data for research purposes. It needs to be emphasized that this workflow is only relevant for research purposes and does not resemble a potential workflow in a clinical work environment. By observing the annotation experts, four user goals were defined for applying the AI algorithm in this context of work:

1. Loading image data into the program (LD)
2. Viewing the data (VD)

3. Initiating the AI algorithm (iAI)
4. Modifying the AI segmentation (mAI)

Subsequently, the user goals were analyzed and split into tasks for assessment in the usability test. A set of eight tasks was created in total (see Table 1).

Table 1

Description of the eight main tasks for the usability testing, description of each task and its associated user goal i.e., LD (loading image data into the program); VD (viewing the data); rAI (running the AI algorithm); mAI (modifying the AI segmentation).

Task	Description of the task	User goal
1	Load DICOM (image) data into the UI	LD
2	Select desired CT scan	LD
3	Scroll + zoom	VD
4	Change window level (contrast)	VD
5	Start AI algorithm	rAI
6	Erase segmentation	mAI
7	Paint segmentation	mAI
8	Use threshold tool	mAI

Each task was divided into individual steps (e.g. mouse clicks or key strokes) for task completion (Appendix A). Alternative pathways were considered for each task. The tasks were paraphrased into realistic task-scenarios. A task-scenario describes what the test user is trying to achieve by providing some context and the necessary details to accomplish the goal. Crafting task scenarios is a balance between providing just enough information so users are not guessing what they are supposed to do and not too much information so the discovery and nonlinearity of real-world application usage can be simulated (Sauro, 2013). The adjustment of the difficulty was especially challenging, because experienced and inexperienced users participated in the usability test. A pilot test (n=3) was employed with inexperienced users to adjust the difficulty and the time limit of the task-scenarios so that it would be possible for inexperienced participants to comprehend the task-scenarios and to complete the tasks.

Concurrent Think Aloud Protocol. The participants were asked to think aloud while performing the task. Think aloud refers to the constant formulation of thoughts, intentions, emotions, or other internal processes while solving a task (Van Den Haak, De Jong, & Jan Schellens, 2003). It is useful for a better understanding of the user's perspective when interacting with the testing materials. Due to time limitations in the research setup, it was

decided to employ concurrent think aloud instead of retrospective think aloud. However, concurrent think aloud can have an impact on the task performance, especially on the time needed to solve a task (Van Den Haak et al., 2003).

Eye-Tracking & Screen Recording. For the eye-tracking a Tobii Pro Fusion 250hz was used. It is a stationary eye-tracker which was attached to the bottom of a 1920x1200 monitor. Tobii Pro Lab 1.181 was used to record the screen and to capture the eye-tracking data during the usability test. Tobii Pro Lab is a software to design, conduct, and analyze eye-tracking studies. The software was simultaneously used for the screen recording.

System Usability Scale. Towards the end of the study we integrated the System Usability Scale (SUS). The SUS is a "quick and dirty" questionnaire to measure the satisfaction in use (Brooke, 1996). Each of the ten items is a statement and participants can agree or disagree with each statement by rating it on a 5-point Likert scale. The ratings of each item are added up and multiplied with 2.5 to produce a final score between 0 and 100. The SUS has been used for more than two decades now. Analyses of a large number of SUS scores have shown that it is a highly robust and versatile tool for usability research (Bangor, Kortum, & Miller, 2008).

2.1.4 Procedure

Preparation and execution of the study followed a pre-written study protocol (Appendix B). After the participants entered the testing room, they were welcomed and introduced to the study with an information sheet. Then, the informed consent form was presented to the participants. After reading the informed consent, the participants were asked to tick the statements on the form and sign the consent form.

After the consent form was signed, the test administrator explained to the participants how concurrent think aloud works and what they should do if they were unable to solve a task. Then, the eye-tracker was calibrated. Screen-recording and eye-tracking devices were started and the task scenarios were presented. Tasks were presented one-by-one. After reading one task, the participants attempted to complete it. When participants successfuly completed the task or after the time limit of three minutes was exceeded, the participants were asked to move on to the next task. Task eight was an exception to this rule, as it had a higher time limit of five minutes due to a higher complexity of the task. The procedure lasted between 20 and 40 minutes per participant.

2.1.5 Data analysis

To model the current level of usability we observed participants performance using the following parameters:

Success (Effectiveness/ efficiency). A task was considered successful if participants were able to fully complete it. To manage the time constratints of the study, we put a time limit of 180s for task 1-7 and 300s for task 8. A pilot test (n=3) showed that these time constratints would be more than enough time to complete the tasks, even if participants were unfamiliar with the interface. We decided to increase the maximum time for T8 because the pilot also showed that this task was perceived as more complex than the other tasks by the participants. Due to the time constraints, it could be argued that this is rather a measure of efficiency than effectiveness. However, the pilot test showed that the time limit was large enough to assume that participants would not be able to accomplish the task without external help after the time limit elapsed and that they often gave up trying before the time limit was reached.

Efficiency. Another measure that was used is deviations from the normative pathway. Deviations refer to the difference in clicks or keystrokes between the optimal pathway and the pathway chosen by a participant. Prior studies have shown that although users might eventually complete a task successfully, they reveal a high number of deviations from the normative pathway. Every deviation from the optimal way of doing a task increases the risk of suboptimal outcomes, even if operators are able to do corrective actions much of the time. Deviations are likely to cause additional cognitive workload, interruptions, and time-constraints (Schmettow et al., 2017).

The time on task (ToT) is another measurement which is typically taken in usability tests. However, ToT is influenced by the concurrent think-aloud which was employed in this study. Talking while performing a task seems to have an influence on the ToT and it makes ToT vulnerable to individual differences among the participants since there are some participants who do not talk much and others who talk a lot (Olsen, Smolentzov, & Strandvall, 2010). Therefore, ToT is not the most robust measure in this study and it was decided to exclusively focus on deviations as a measure of efficiency.

To identify usability problems in the current design, we focused on experienced (observed) and verbalized interactive issues. To identify common problems and to rate their severeness, the following qualitative methods were used:

Incident coding and usability problem breakdown. For in-depths analysis of interaction sequences, incident coding was applied. Incident coding is a method in which events that hint towards the existence of a usability problem are noted in a structured report (Schmettow et al., 2017). Especially ineffective or inefficient operations were taken into account, but also communication events such as positive or negative comments were noted. The following codes were used for the analysis recordings of the interaction:

- *E=Ineff**E**ctive operation, e.g. "subject fails to load data into the program"*
- *I=Ineff**I**cient operation, e.g. "subject searches various menus before finding the desired tool"*
- *C=Communication event, e.g, "subject indicates that he/she would have expected to find the function elsewhere"*
- *P=Positive comment, e.g, "subject indicates that he/she found it easy to load data into the program"*
- *N=Negative comment e.g, "subject indicates that he/she finds the layout of the UI confusing"*

Findings of the incident coding were used to identify recurring patterns of ineffective or inefficient operations. Each pattern was labelled, the corresponding user behavior was described, and causes of the problems were identified. Grouping the findings of the qualitative analysis into patterns helps to identify which systematic issues in the design were observed among multiple participants. Thereby, the severeness of a usability problem can also be determined by assessing the percentage of people who encounter each problem.

Eye-tracking and heatmaps. The eye-tracking data was processed into aggregated heat maps to visualize on which elements of the interface the participants directed their gaze. Thereby, the gaze data of each participant was mapped on screenshots of the interface. This resulted in a screenshot with a heatmap for each task. Heatmaps can help to understand how users perceive an interface and it can give a quick and intuitive understanding of how people interact with a stimulus (Djamasbi, 2014). The Tobii Pro Lab manual mapping function was used. A screenshot of the interface for each task was uploaded into Tobii Pro Lab. Then, the gaze points from the screen recordings were manually mapped to the screenshot of the task. This procedure was repeated for each participant and each task. The data of all participants was aggregated per task and then a heatmap was produced to visualize the average gaze of all participants per task.

Analysis of quantitative measurements in R-studio. R-Studio 2021.09.4 was used for the quantitative data analysis. For the dataset see Appendix C, for the syntax of the data analysis, see Appendix D.

2.2 Results

2.2.1 Current level of usability of the system

The results showed high task completion rates with exception of tasks 4 and 8 which

showed distinctively lower completion rates. Task 4 was only completed by half of the participants and task 8 was completed by nine participants. The mean deviations per task showed that task 4 and task 8 also caused a high number of average deviations per participant. Task 4 caused an average of 6.93 deviations per participant and task 8 had an average of 4.21 deviations per participant. Moreover, tasks number 1, 3, 5, and 6 had more than one deviation on average per participant (see Table 2).

Table 2

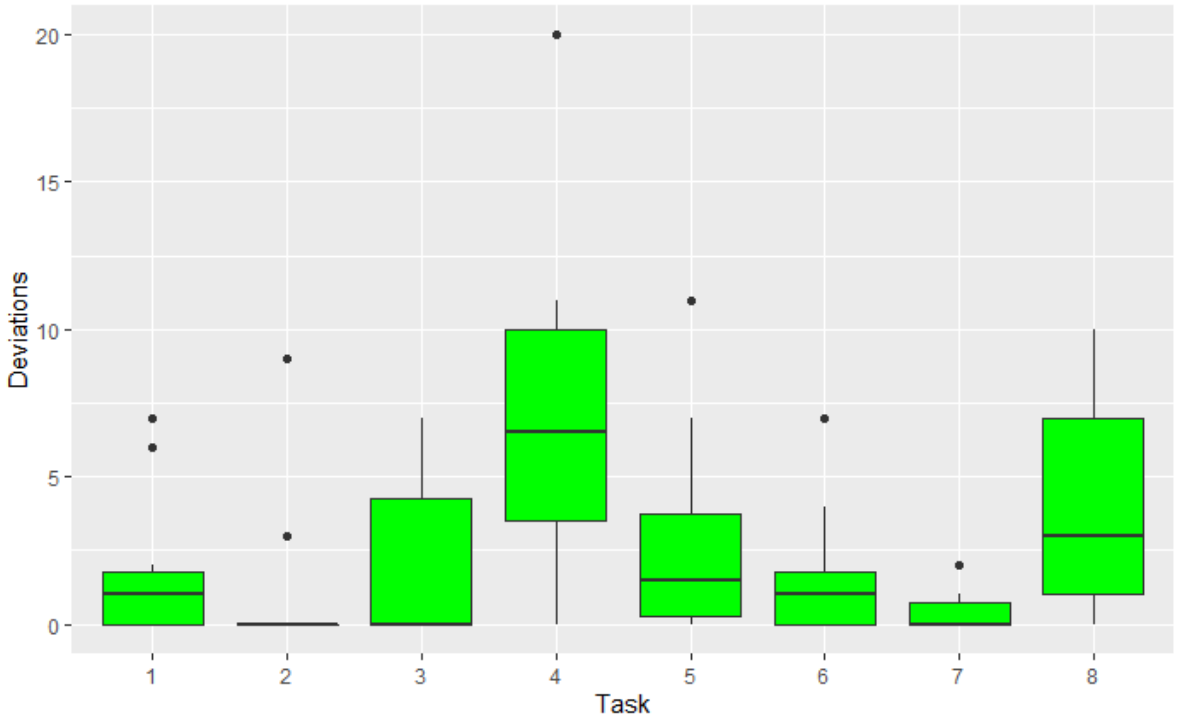
Task completion rates and mean deviations from the normative pathway per task

Task	Task completion rate	Mean deviations from the normative pathway
1: Load DICOM data	86% (12/14)	1.57
2: Select CT scan	86% (12/14)	0.86
3: Scroll + zoom	86% (12/14)	1.79
4: Change window level	50% (7/14)	6.93
5: Start AI algorithm	93% (13/14)	2.79
6: Erase segmentation	93% (13/14)	1.57
7: Paint segmentation	100% (14/14)	0.43
8: Use threshold tool	64% (9/14)	4.21

Boxplots were used to visualize the distribution of the deviations from the normative pathway (see Figure 2).

Figure 2

Boxplot of deviations from the normative pathway (Y), grouped by the eight tasks (X)



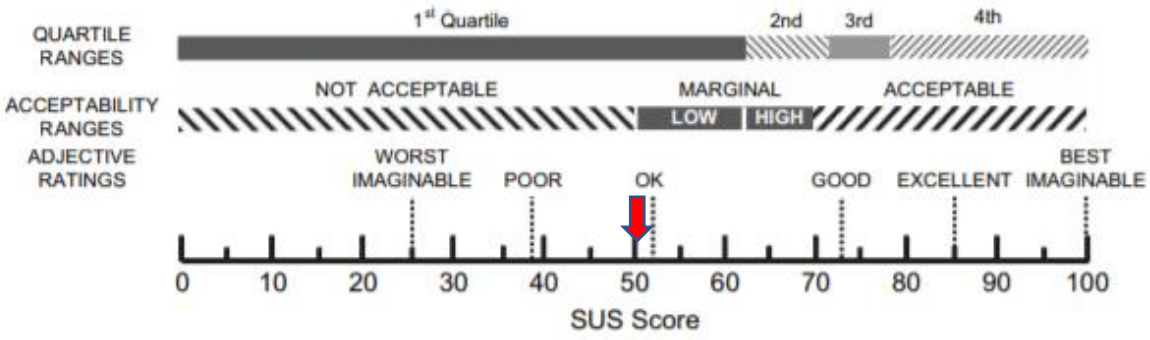
Some of the tasks seemed to cause a considerable amount of deviations. Especially task 4 seems to be problematic with a high median in deviations (*md = 7*). Task 8 also seemed to cause a high number of deviations from the normative pathway (*md=3*), as the upper quartile reaches 7 deviations. Tasks 1, 3, 5, and 6 caused medium levels of deviations from the normative pathway. Only tasks 2 and 7 seemed intuitive and did not seem to cause many deviations from the normative pathway.

2.2.2 SUS Scores (Satisfaction)

The SUS results showed that the participants were rather dissatisfied with the interaction. The average satisfaction score among all participants was m = 50.21 on a scale from 0 to 100 (SD=10.52). Bangor et al. (2008) released a large-scale comparison of mean SUS scores to classify SUS results (see Figure 3).

Figure 3

A comparison of mean System Usability Scale (SUS) scores by quartile, adjective ratings, and the acceptability of the overall SUS score (Bangor et al., 2008).



Note. ⬇ = SUS average score in the current usability study

Comparing the results from this study to Bangor et a. (2008), the SUS score in this test is located in the lowest quartile of SUS scores and it is on the brink of not being acceptable anymore. To identify what caused the low survey score, the mean ratings of the individual items of the questionnaire were analyzed. Question 2 of the SUS "*I found the application unnecessarily complex"* received the lowest average rating with 3.75/10 points. Other distinctively low ratings have been found in question 4 *"I think that I would need the support of a technical person to be able to use this application"* and in question 9 *"I felt very confident in using this application"* with average ratings of 4.5/10 points each. Thereby, it seems that the

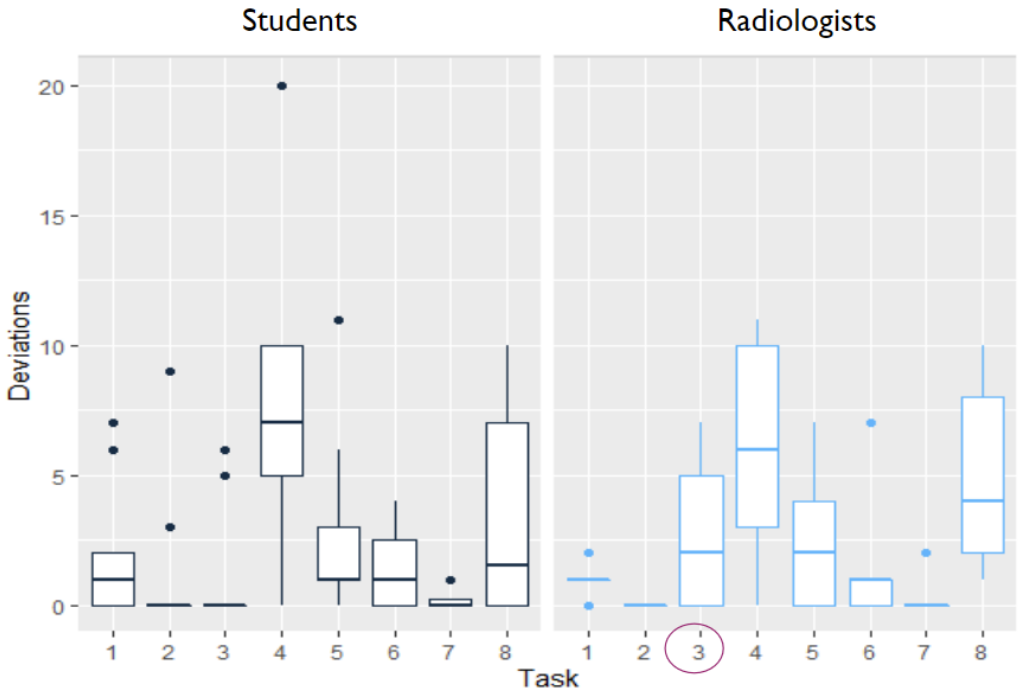
participants were dissatisfied with the usage of the program because it was perceived as too complex or confusing.

2.2.3 Comparison Between Groups

Radiologists and Students – Deviations. A comparison of the deviations between students and radiologists showed very similar results between the two groups (see Figure 4).

Figure 4

This figure shows a comparison between students and radiologists in terms of deviations from the normative pathway per task, visualized in form of boxplots

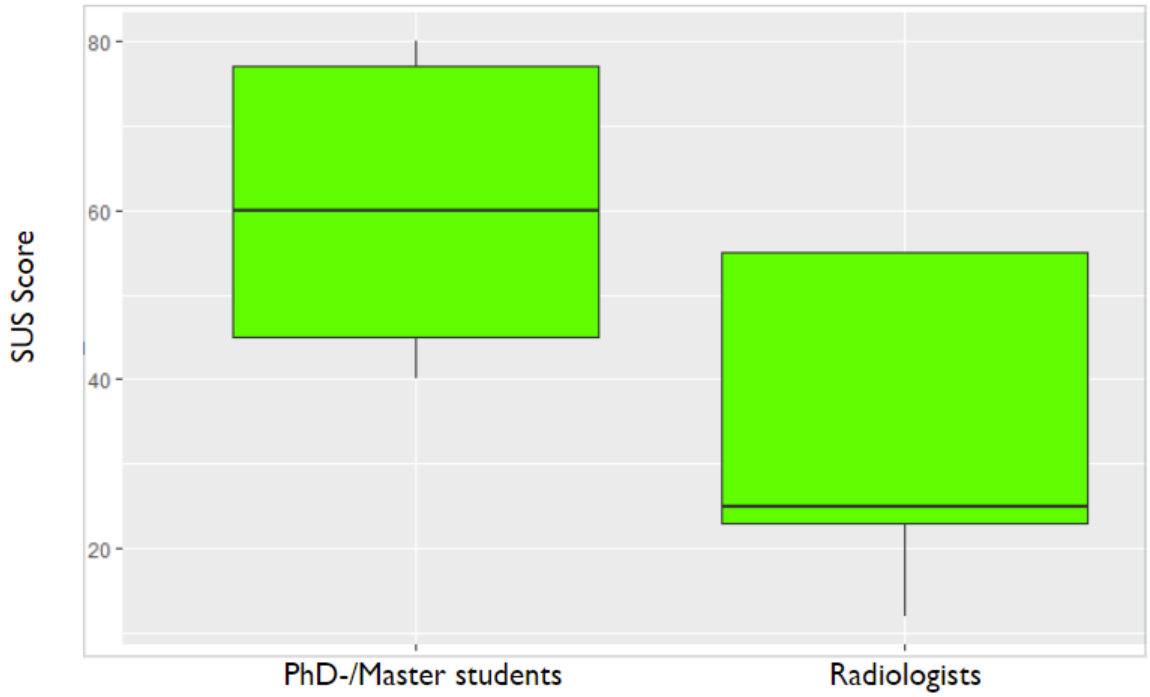


The comparison of deviations from the normative pathway between groups showed that the deviations of students and radiologists looked similar. A major difference can only be observed in task 3. Task 3 was about scrolling images and zooming in and out. Radiologists are used to zooming and changing between slices in clinical picture archives and communication systems (PACS). The software application which was assessed in this usability test does not adhere to the conventional controls of PACS and therefore the radiologists struggled with basic functionalities such as scrolling and zooming.

Radiologists and Students – Satisfaction. An average SUS score comparison between radiologists and students showed that it seems like students were more satisfied with the interaction than radiologists (see Figure 5).

Figure 5

Average SUS scores of PhD-/Master students compared to average SUS scores of radiologists



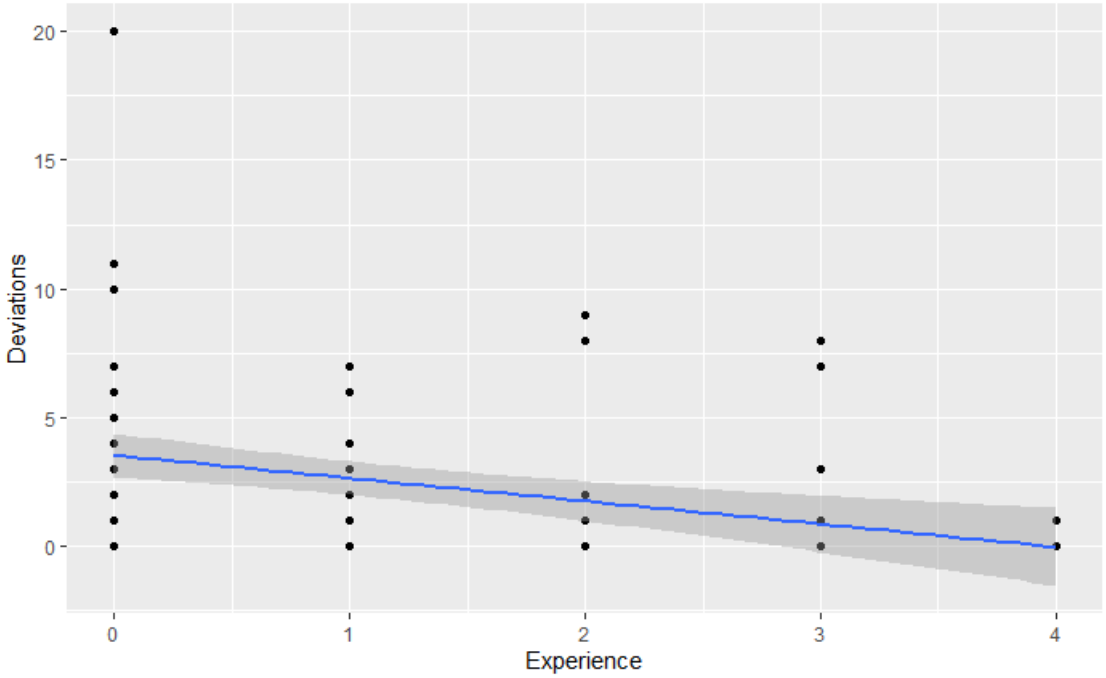
The boxplots indicate that there is a profound difference in satisfaction between students and radiologists. Students seem to almost have a decent level of satisfaction with a median SUS score at 60 while the boxplot of the radiologists seems to be much lower with a median SUS score at 25.

2.2.4 Prior experience and deviations

Furthermore, the influence of prior experience with the UI on the amount of deviations was analyzed. Experience was measured on a 5-point Likert scale, based on self-assessment. Participant's predicted deviations is equal to 3.52 – 0.89 per experience level. The 95% CI is [-1.36, -0.43]. Especially for fist time users with of 0, the UI seemed difficult to use and to navigate, and therefore they caused a much higher amount of deviations (see Figure 6).

Figure 6

Linear relation between experience (X) and average deviations (Y)



2.2.5 Usability Problem Breakdown

The findings of the incident coding showed that various usability problems existed in the current interface. A total of 16 usability problems were found and analyzed (see Table 3).

Table 3

Usability issues identified in the analysis of the recordings

Number	Description	Visibility
01	Window levels are too hard to find	79% (11/14)
02	The threshold function is too hard to use	79% (11/14)
03	The sub-menus are too complex	43% (6/14)
04	Pathway of the data is hard to find	36% (5/14)
05	The segment editor has too many tools	36% (5/14)
06	“Import DICOM” button is easily confused with “add data” button,	29% (4/13)
07	It is not clear how to zoom in	23% (3/13)
08	Functionality of the view menu is misunderstood	23% (3/13)
09	Navigation structure is not clear	23% (3/13)
10	Drag & Drop is not obvious	14% (2/13)
11	Changing window levels is not possible with hotkeys	14% (2/13)
12	Segmentations module is easily confused with the segment editor module	14% (2/13)
13	Import data button is highlighted but not clickable	7% (1/13)
14	User gets lost and can’t recover	7% (1/13)

15	NVIDIA icon is unknown	7% (1/13)
16	Draw function is not intuitive to use	7% (1/13)

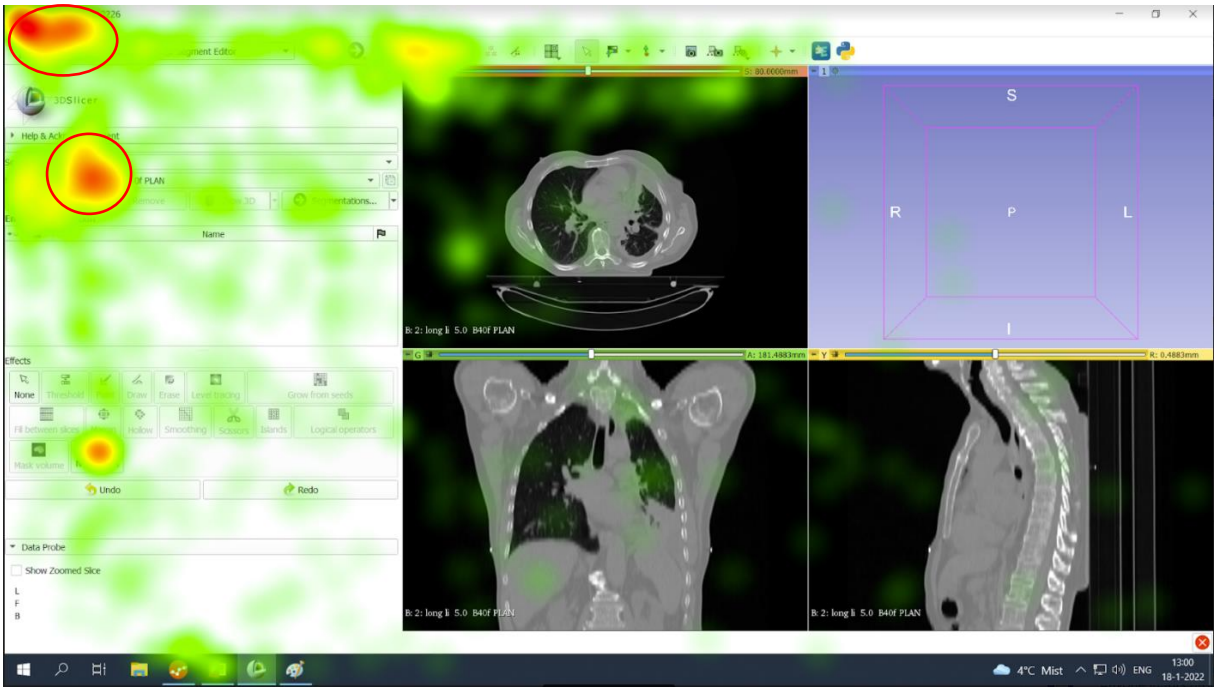
Two of the identified issues stood out, usability problem (UP) 01 (*"window levels are hard to find"*) and UP 02 (*"threshold function is hard to use"*) were experienced by 11 out of 14 participants which hints towards the existence of two substantial design problems that were encountered by the vast majority of the participants. These usability problems relate to task 4 (*"change the window level"*) and task 8 (*"apply the threshold function"*) which also had worse task completion rates and caused the most normative path deviations. The findings seem to align and it underlines that these two functions were especially hard to use for the participants. Other UPs were less prominent, but still noticeable. A full analytical breakdown with description, causes, and outcomes of each individual UP can be found in Appendix E.

2.2.6 Eye-tracking results

Heatmaps for each of the tasks with comments are available in Appendix F. Especially the heatmap of task 5 "*initiating the AI algorithm*" delivered interesting results. The heatmap of task 5 showed that participants focused their gaze on locations that did not have anything to do with the task completion (see Figure 7).

Figure 7

Heatmap for task 5 displaying where participants focused their attention, red color highlights more focus of the gaze while green color shows that the areas only received superficial attention



Note: Upper red circle is on the top-left menu for general settings, the lower red circle is on the specifications for the segmentations

The participants thought they can access the AI model in the top-left menu of the interface, although it was suggested in the task-scenario that the AI model needs to be accessed via the segment editor. The analysis of the heatmap highlighted that the layout and the functionalities of the menus was not clear to the participants and that the segment editor seemed hard to find.

2.3 Discussion

The current implementation of the AI algorithm seemed to contribute to mediocre usability with some tasks causing no or only minor diffulties and some tasks causing major difficulties. Due to the low usability measures in some of the tasks, the implementation of the AI into efficient workflows could become problematic. A lack of usability and workflow implementation hinders the adoption of AI algorithms in clinical environments (Lekadir et al., 2021). The utilization of the AI algorithm in itself caused few problems, but the interface in which it was embedded seemed to cause most of the problems during this usability test. The initiation of the AI algorithm was one of the easier tasks (task 5, 93% completion rate, 2.79 mean deviations). However, the software in which the AI algorithm is currently integrated seems to cause several usability issuess. For instance, adjusting the view by changing the window level or changing the threshold in the annotation problems for the participants. These functionalities are necessary to make use of AI-generated results in the workflow for medical image annotation, but these are not funcationalities of the AI algorithm in itself. Several other issues related to the software application were identified in the usability problem breakdown and in the analysis of the eye-tracking data. Issues often related to a cluttered and confusing structure of the UI, and participants especially struggled with finding functionalities and tools. To improve the usability of the AI implementation, changes in the design of the UI should be made and several usability issues need fixing.

Furthermore, the satisfaction scores generated by the SUS were low. Especially the scores of the radiologists were low, and it seemed like they were profoundly dissatisfied with the interaction. The low SUS score could be caused by prior experience of radiologists with other UIs. The current design consistently fails to adhere to conventionalities and industry standards for image viewers. For instance, the radiologists who participated in this study usually use Picture Archiving and Communication Systems (PACS), but many functionalities in the tested UI are bound to different keys than in PACS. The systems in which AI algorithms are integrated should be specifically designed to augment the work of professionals to make data

annotation procedures more efficient and to encourage usage of the UI with embedded AI. This does not seem to be the case at the moment.

The observed usability issues paired with rather low satisfaction scores underlined that the current implementation of the AI algorithm is suboptimal in the context of the pre-defined user goals and tasks. The UPs which were found may hinder an efficient workflow and could cause issues in the application of the AI algorithm in the medical image annotation procedure. The low satisfaction in use could also be problematic. Filice and Ratwani (2020) described that clinicians need to be convinced that the benefits of AI usage outweigh the disadvantages and that they need to be satisfied with the systems which they are using. Lekadir et al. (2021) and Shneiderman (2020) emphasized that the acceptance of AI systems is what will determine their success and their long-term implementation. The system in which the AI is embedded should have good usability, and it should satisfy and convince users of the interaction. Thereby, the current usability issues need to be tackled and new possibilities for the integration of AI need to identified in order to promote a better implementation of AI algorithms in medical image annotation in the future. Conclusively, we advised to change the integration of the AI algorithm in the future and to look for new possibilities for implementation which support the user goals in a better manner.

3. Phase two. Designing a New UI for Integrating AI: Requirements, Prototyping, and Expert Evaluation

The goal of this phase was to propose the design of a new UI for the implementation of AI algorithms in medical imaging. One aim of this phase was to improve the shortcomings which were identified in the previous usability test, another goal was to generate new ideas and concepts for the proposed software architecture for the implementation of AI in a medical image annotation program. We created an interactive prototype to simulate workflows for utilizing the new UI and to assess the usability of the proposed design. To ensure that the new UI would fit the user requirements, HCD methodology was applied and various users and stakeholders were involved during the development and the testing phases. For the newly designed prototype the AI algorithm for the automatic segmentation of mesothelioma from CT scans served as a case example which could be integrated into the newly proposed software architecture. This phase covers a brainstorming session, the initial design of a preliminary prototype, an expert evaluation in a focus group setting, and the design of the interactive prototype.

3.1 Brainstorming Design Requirements and New Ideas

Before initiating the design process, the goals and expectations associated with this

research project had to be sufficiently defined. Therefore, a brainstorming session with stakeholders was conducted to define the goals and to align the expectations of the stakeholders which were involved in the project.

3.1.1 Participants

Five stakeholders participated in the brainstorming session. The participants had diverse backgrounds (see Table 4).

Table 4

Participants of the brainstorming session

Participant	Profession
1	AI expert, postdoc
2	AI expert, PhD student
3	Radiologist, M.D.
4	Usability expert, assistant prof.
5	HFE student, master's student

3.1.2 Materials

A set of six questions was presented to the participants of the brainstorming session. The questions were selected to further define which users would be required for the subsequent user testing phase and to get a better understanding which additional workflows, goals and tasks should be supported in the new system. The questions were selected based on a paper of Maguire (2001) on human-centered design methodology. The following questions were selected:

1. What are the objectives of this research project?
2. Who are the intended users?
3. What are the main goals/tasks of the users?
4. What key functionalities/tools are needed to support the goals of the users?
5. What are the usability goals?
6. How will users obtain assistance in learning how to use the system?

3.1.3 Procedure

The session was conducted remotely in a digital environment. The questions were presented to the participants one by one. The participants had several minutes to write individual answers on digital sticky notes. Multiple answers to each question were allowed. After each participants submitted their individual answers, a discussion round was initiated. The participants now looked at the answers of the other participants and grouped similar

answers into clusters. Each cluster was rapidly discussed and the group members elaborated their answers. This procedure was repeated for each of the six question. The whole session lasted for about one hour in total.

3.1.4 Results

The results of the brainstorming session were summarized and a short summary of the results for each of the questions is presented below:

1. Objectives of the research project

Goal of the project is to create a platform to analyze medical image data with AI models and to combine multiple functionalities in an all-in-one approach. The platform should support functionalities such as viewing data, editing segmentations, uploading AI models, and interacting with AI-output. The design of the platform should consider user needs and requirements, and evidence about the usability and the user experience should be collected to provide user-friendly solutions. A new standardized workflow for medical image annotation with AI support and data sharing should be introduced.

2. Intended users

The intended users for the platform are experts in medical imaging (e.g. radiologists or pulmonologists), AI researchers in medical imaging, and students who are studying in a field related to medicine or AI.

3. Main goals of the users

The main goal users want to achieve is to apply novel AI technology to medical image data for research purposes. To apply AI algorithms to medical image data, users need to be able to upload, share and view medical images. AI algorithms need to be applicable to stored medical image data, and editing AI-generated output is necessary for further medical image annotation and the rapid generation of new training datasets. Furthermore, users might want to create radiological reports, analyze cohorts of patients, or run quantifiable AI-based studies for research purposes.

4. Functionalities and tools needed to support user goals

Data import and export of medical image data is essential. New data sharing approaches could be featured in the new system to facilitate the circulation of medical image data and annotated datasets. The platform needs an image viewer for the inspection of medical images and annotations. Readily usable and integrated AI algorithms should be implemented into the systems and the usage of AI tools should be possible without any technical background knowledge. Tools for annotating medical image data with AI support should be provided to

make the data annotation process more efficient. Furthermore, analysis tools should be integrated to further dissect the AI results and to offer more depths in the analysis when working with AI results. Finally, the UI should be customizable based on the user preferences, and therefore hotkeys and quick access functionalities are required to promote expert usage.

5. Usability goals

The overarching goal is to to facilitate medical image annotation tasks and to support the interaction between users and AI algorithms in a user-friendly manner. To reach this goal, the platform should adhere to usability standards to ensure effective and efficient workflows. The UI needs to be intuitive and easy to learn for inexperienced users, but it also needs to support expert users who are making use of the platform on a regular basis. Furthermore, the satisfaction in use and the likelihood to recommend need to be high to facilitate the acceptance and the use of the system, therefore the design should be appealing and functional.

6. Assistance in using the system

Assistance in using the system could be arranged by creating an online page with written instructions, FAQ, and manual. Furthermore, video training and tutorials could be provided online. It should be possible to use the platform without manual help and just with the materials which are provided online.

3.1.5. Conclusion

The brainstorming session was useful to define the priorities, goals, and preliminary requirements for this design project. The stakeholders had similar ideas and expectations for the project, but also some different expectations were discussed in the discussion phases. The results of the brainstorming session can be viewed as a starting point for the design phase.

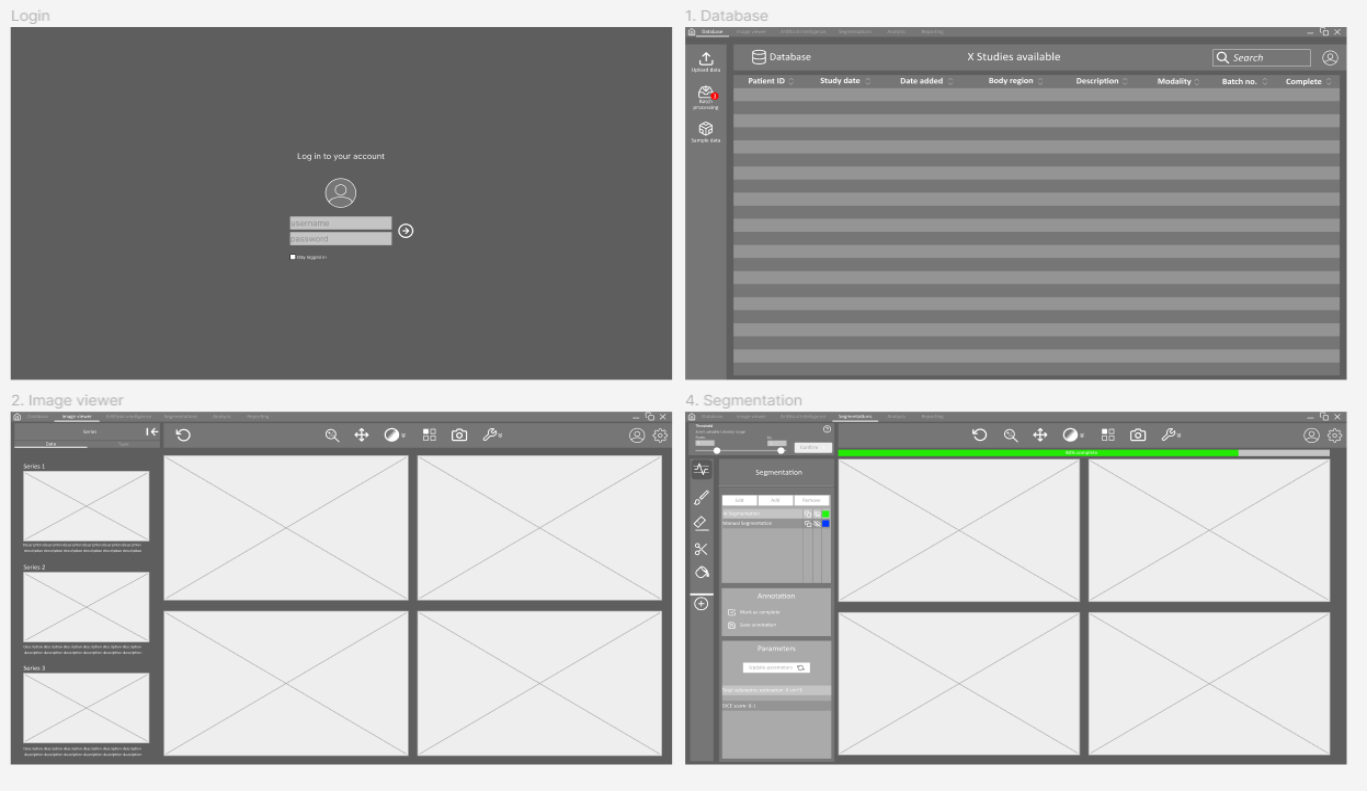
3.2 Design of the initial prototype

After the brainstorming session, a preliminary lo-fi prototype was created to visualize first concepts and ideas for the preliminary design. Prototypes are means for exploring and expressing design intents for interactive computer artifacts (Houde & Hill, 1997). Prototypes differ in fidelity, which refers to the detail and realism in the design of the prototype (Babich, 2017). Low-fidelity (lo-fi) prototypes are quick and easy mockups, allowing for a rapid turnaround of early design concepts. Lo-fi prototypes can be simple paper drawings or digital sketches to visualize ideas for the design (Babich, 2017; Esposito, 2018). Initiating the design process with a simple prototype bears the advantage that high-level concepts can be visualized and tested as tangible artifacts without causing much effort (Babich, 2017). The more sophisticated the prototype becomes, the more tedious it will be to make changes again.

Accordingly, a lo-fi prototype was created as a first design iteration. The prototype consisted of four slides, designed in a minimalistic and simplified fashion. The prototype was created in Figma, a free tool for collaborative UI design and prototyping (Figma.com). One slide each was presented for the login screen, the database, the image viewer, and segmentation workspace (see Figure 8).

Figure 8

Four digital slides as a lo-fi prototype, designed to visualize layout and general design concepts



Note: Top-left is a login screen, top-right is the database, bottom-left is the image viewer, bottom-right is the segmentation workspace

The lo-fi prototype did not include all intended sections for the future design, for instance the AI section was not conceptualized, yet. The slides of the lo-fi prototype can be inspected in Appendix G.

3.3 Focus group review of the lo-fi prototype with experts

To assess the lo-fi prototype, we organized a focus group with experts to review the initial design. The aim of the focus group was to gather feedback for the improvement of the

lo-fi prototype and to generate new ideas and concepts for the further development of the prototype in the next iterations. Focus groups are especially useful in early stages of the design process when they are utilized for expert evaluations and concept generation (Bruseberg & McDonagh-Philp, 2001). Specifically, we performed a mini focus group with five participants (n=5). This type of focus group is useful when the topic needs to be explored in greater depths and when participants have long and substantiated experiences to share (Krueger, 2014). The members of the focus group were purposefully sampled from the working staff of the NKI, and only members with expertise in medical imaging or AI research were selected (see Table 5).

Table 5
Mini focus group participants and their backgrounds

Participant	Profession	Abbreviation
1	Radiologist	R1
2	Radiologist	R2
3	Pulmonologist	P
4	AI researcher	A1
5	AI researcher	A2

3.3.1 Procedure of the focus group

The participants were welcomed and introduced by the moderator. Before the session started, the participants agreed that the session would be audio recorded. The purpose of the design project was explained to the group. It was emphasized that discussion and criticism were necessary and appreciated. The slides of the lo-fi prototype were presented one-by-one to stimulate the discussion. The moderator showed each slide to the group of participants, and he asked questions about the layout, concepts, buttons, and functionalities which were implemented in the lo-fi prototype. Furthermore, the moderator encouraged the participants to pitch their own ideas and concepts for an improvement of the lo-fi prototype. To balance more and less talkative members of the focus group, the moderator encouraged more quite participants to voice their opinion specifically. The session lasted for 60 minutes in total.

3.3.3 Results of the focus group

After the analysis of the focus group transcription, 33 suggestions for improvement were identified. We distinguished between changes for the next design iteration and long-term changes for future releases of the product. Some changes which were suggested by the participants were not feasible for the next design iteration due to a limited level of readiness and realism of the next prototype iteration. Nevertheless, the remarks of the participants could be useful for future releases of the product and these suggestions were noted as long-term design

suggestions. After the distinction between suggestions for the next design iteration and suggestions for the final version of the program were made, 21 suggestions for the immediate improvement of the prototype were left. To view all suggestions which were made (including suggestions for the long-term) and the transcription of the focus group, see Appendix H. In the following, a summarized overview is presented to highlight what suggestions were concerned with.

- ***User control:*** The experts in the focus group emphasized that it is essential to have user control over certain settings. For instance, the opacity of the segmentation needs to be adjustable based on the user's preferences. Annotators need to be able see the image below the area which is segmented (R1, R2).
- ***Comprehensibility:*** Some of the labels were not intuitive enough, and they were hard to understand for the participants. For instance, the label "batch processing" should be changed to "my cases" to make it more clear that cases which are appointed to the user can be found here (R1, R2, P).
- ***Accessibility of documentation functions:*** Annotation and documentation were allocated in different menu tabs. However, it would be beneficial if the users could document the cases while annotating in the same window, so the users can see the documentation and the annotation on the same screen. Therefore, documentation should be available in the annotation workspace (R1, R2, A1, A2).
- ***Screen size vs. visibility of the annotation progress:*** A large progress bar for the annotation progress was implemented in the lo-fi prototype, but a radiologist voiced an issue that the progress bar is taking away valuable space from the image viewer. Therefore, it was suggested to move the progress bar to a different position where it does not decrease the space of the viewer, as the viewer space should always be as big as possible (R1).
- ***Removing redundant functions*:** A tool for taking a screenshot in the viewer is redundant, as conventional ways of taking screenshots suffice (R1, R2).

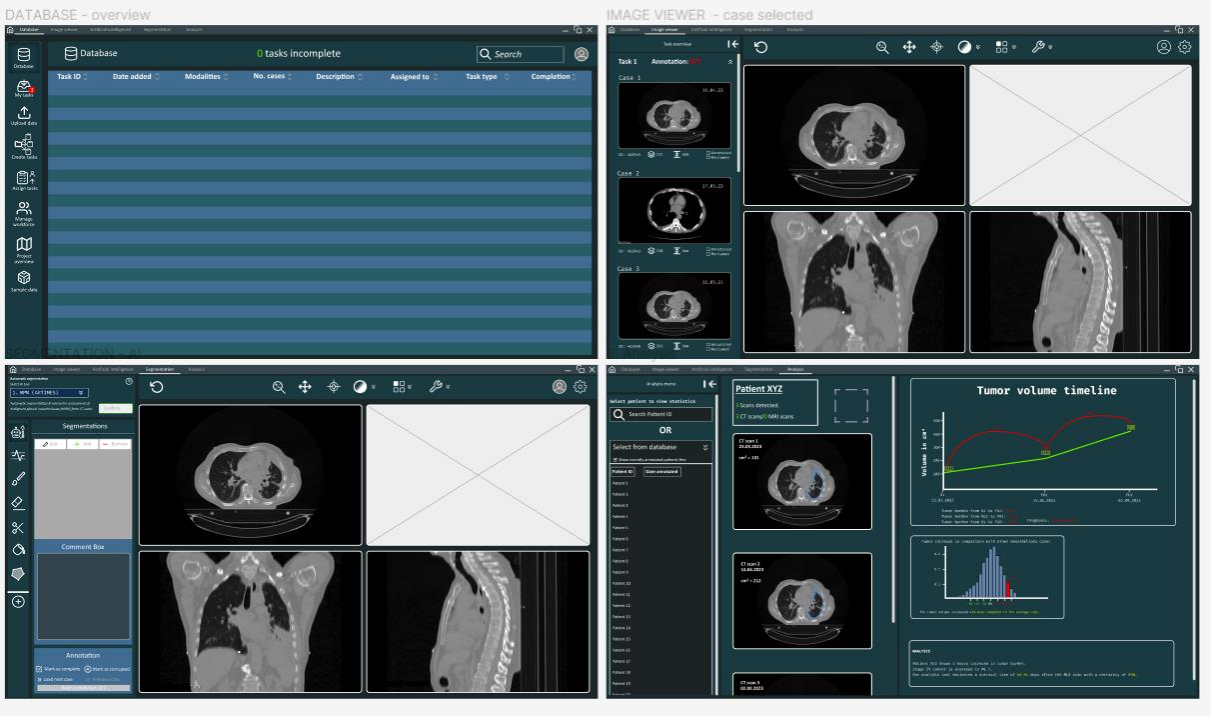
3.6. Design of the Interactive Prototype

The revised lo-fi prototype in tune with the indications of the focus group was used as to design a high-fidelity prototype. High-fidelity (hi-fi) prototypes are highly functional and interactive and they are very close to the final product, with most of the necessary design assets and components developed and integrated (Esposito, 2018). The hi-fi prototype was created with Figma. Various elements were added in the new design iteration. For instance, visual

concepts were designed with more attention to detail, colors were added, additional icons and functionalities were implemented, and more descriptive text was added to the UI. Furthermore, additional sections such as AI implementation and analysis were conceptualized and designed (see Figure 9).

Figure 9

Slides of the hi-fi prototype, designed with Figma and interlinked to simulate functionality



The prototype was designed in consideration of UI design principles and heuristics (Molich & Nielsen, 1990). In particular, the graphical design of the prototype was created to provide a realistic UI; in such sense icons, graphs and images were created by hand in Figma and were designed so that they could be adopted when building a real software application. It was intended to design the UI in an appealing fashion, because aesthetic design can have a large influence on user preferences and the perceived usability of systems (Lindgaard & Dudek, 2003; Tractinsky, Katz, & Ikar, 2000). The layout and the structure were inspired by other modern software applications (see Appendix I), and the layout structure remained similar among the different sections to promote consistency within the application. Icons were designed in a minimalistic but distinct fashion and we chose icons and fonts big enough to be recognized easily, even for users who might have visual impairments. Moreover, a dark color design for the UI was picked. Dark interfaces could be beneficial for medical image experts such as radiologists, because radiologists typically work in dim rooms to spot details on medical images. Thereby, a dark-them ed UI is easier on the eyes, especially when users are working in

a dark environment, staring at the screen for many hours.

Addressing the functionality of the prototype, it simulated workflows for importing and sharing medical image data, viewing data, selecting and initiating AI, annotating medical image data, assigning annotation tasks to other users, and analyzing AI-generated results. Users were able to click many but not all of the supposedly clickable elements on the interface to simulate the interaction and the pathways for utilizing the application. The menu structure of the UI and the different sequences in a workflow were simulated in a realistic manner. However, the interactivity of the prototype was limited to some degree. The actual application and the handling of the different tools and functionalities could not be simulated in Figma. Furthermore, additional interactive features such as hotkeys or customizations were not integrated.

The prototype consisted of more than 150 individual slides, which were interlinked to simulate the proposed workflows. To interact with the prototype, see Appendix J.

4. Phase three. Conducting a Usability Test with the New Prototype

The goal of the last phase was to assess the interaction between users and the new interactive prototype and to elicit additional user feedback for further design iterations of the application. User tests are an integral part of HCD methodology and should be conducted after each design iteration (Abras et al., 2004; Maguire, 2001; Norman & Draper, 1986). Thus, a usability test was conducted to assess the proposed design with members of the intended user groups. One goal of the testing phase was to assess if the newly proposed workflows were functioning as intended, and whether the users can reach their goals effectively and efficiently, while also being satisfied with the interaction. Another goal of this phase was to elicit additional user feedback and to assess the reaction of users towards the individual sections and design elements which were added since the previous expert evaluation. Retrospective user interviews were conducted after the usability test to gather additional verbalized feedback on the specific sections and design elements. The results of the different measurements were analyzed, compared and discussed in this section and we proposed final suggestions for the further development of the platform.

4.1 Methods

4.1.1 Design

The prototype was assessed remotely in an online environment for user testing and a withing-subject design was employed. Task-scenarios were crafted to assess usability variables such as task completion (efficiency), deviations from the normative pathway (efficiency), and SUS (satisfaction). Moreover, the usability test was conducted to identify and analyze usability

problems in the current design by applying incident coding as an observation technique. Moreover, post-task semi-structured interviews were conducted to elicit additional feedback and deeper insights into the users' thoughts on the proposed design.

4.1.2 Participants

The participants were selected via convenience sampling. Exclusively participants with experience in medical imaging were selected because they could presumably give more substantiated feedback on the subject matter. 13 participants assessed the prototype in total. Seven of the participants were employed as medical doctors (MDs) with a specialization in medical imaging (six radiologists, one pulomonolgist). The other six participants were technical medicine students who were familiar with medical image annotation. 10 participants were female, and three were male. The mean age of the participant group was rather young, ranging from 23 to 43 years (m = 30.77).

4.1.3 Ethical Approval and Consent

The study has been approved by the ethics committee of the University of Twente (Request-no. 220821). Participants filled out an informed consent and agreed to the terms of the study before the usability test. They were allowed to cancel before and during the study, or to have their data deleted anytime after the study without any further questions asked.

4.1.4 Materials

The interactive prototype which was described in the previous phase was the object of assessment. The user test was conducted remotely, and therefore the prototype had to be easily accessible for the participants from their computers at home. We integrated the prototype in UseBerry (useberry.com) which is an online service for remote user testing. Prototypes designed with popular prototyping tools such as Figma can easily be plugged in UseBerry. They also offer options for the customization of the user test. For instance, the researcher can choose in which order slides are presented, and tasks and prompts can be linked to specific prototype slides. Researcher and participant communicated via Microsoft (MS) Teams (teams.com), and the session was also recorded with the open broadcasting software (OBS)(obsproject.com), which is a free and open source software for video recording and live streaming. The informed consent and study information which were presented to the participants can be found in Appendix K.

To guide the users in their interaction two introductory scenarios were written to simulate a realistic usage of the prototype. The first scenario was written from the perspective

of a medical image annotation expert who received a task to annotate medical image scans for an AI research project. The second scenario was written from the perspective of an AI researcher who is responsible for managing an AI research project, and thereby, wants to distribute medical image data and annotation tasks to annotation experts. To complete the scenarios users were expected to perform certain tasks (see Table 6) and steps in a sequential order.

Table 6

Tasks created for the usability assessment with the corresponding intro scenario

Task	Description of the task	Scenario
1	Import assigned medical image data	Annotator
2	Load CT scan into viewer	Annotator
3	Change the layout	Annotator
4	Change the window level	Annotator
5	Select and initiate the AI	Annotator
6	Load the next case	Annotator
7	Upload image data into the UI	Project manager
8	Create annotation tasks	Project manager
9	Distribute tasks to annotators	Project manager

To view the task scenarios, tasks, and sequential steps for task completion, see Appendix L. Users were asked to verbalise during the interaction (concurrent think aloud) and after the test they were interviewed following a protocol for a semi-structured interview (see Appendix M). At the end of the usability test, as part of the debrief, users were also asked to fill a questionnaire composed by the following parts:

- **System Usability Scale:** Simililar to the previous usability assessment, the SUS survey was used to determine the satisfaction in use.
- **Net Promoter Score:** Furthermore, we implemented the Net-Promoter Score (NPS). The NPS is a one-item survey introduced by Reichheld (2003) to assesses the likelihood to recommend a company or a product. On a scale from 0 to 10 the participants rate the item. “How likely is it that you would recommend […] to a friend or colleague?”.
- **Demographics Survey:** A demographics survey was included, as well. The survey assessed age, gender, profession and previous use of medical image annotation software.

4.1.4 Procedure

The test was conducted online. Participants were greeted and introduced at the beginning of the session. Then, the participants opened a Qualtrics link which included the information sheet and the informed consent. After consent was given, the participants started sharing their screen and the recording of the session with OBS began. Concurrent think-aloud was explained by the researcher before the start of the usability test. Then, a link to the online user test with the prototype integrated in UseBerry was shared. The first introductory scenario was presented, and subsequently, the first set of tasks was presented to the user. The participant moved on to the next set of tasks after the task was completed successfully or after a time limit of 180 second was exceeded. When the first set of tasks was completed, the second introductory scenario was presented and the participants were asked to complete another set of tasks. After all tasks were presented, the interview was conducted. The questions were asked in accordance with the interview protocol. After the interview was conducted, a link to the questionnaire with demographics, SUS, and NPS was shared. Finally, the participant was debriefed and the meeting was closed. In total, the whole session lasted between 30 and 60 minutes.

4.1.5 Data Analysis

Quantitative Data Analysis

Usability metrics for efficiency and effectiveness were collected as in the first phase. Specifically, we used task completion as a measure of effectiveness, and deviations from the normative pathway as a measure of efficiency. Moreover, satisfaction ratings of the participants were collected with the SUS.

To calculate the NPS score, the responses are grouped into “promoters” (9-10 rating), “passively satisfied” (7-8 rating), and “detractors” (0-6 rating). The percentage of detractors (0-6 rating) is simply subtracted from the the percentage of the promoters (9-10 rating) (Reichheld, 2003). The passively satisfied are ignored in the calculation.

The video recordings and the concurrent think-aloud audio were analyzed and coded. The same incident coding scheme and the same usability breakdown method were applied as in the first phase. The dataset was created in MS Excel (Appendix N) and the quantitative data analysis was conducted with R-studio 2022.12.0 (Appendix O).

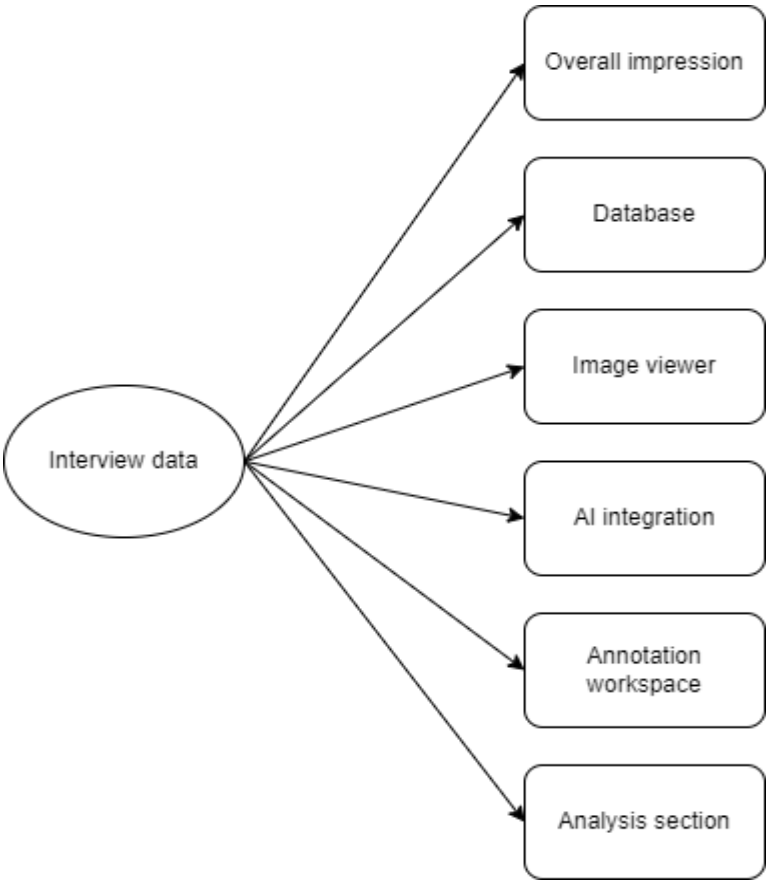
Qualitative Data Analysis

The video recording and the concurrent think aloud were analyzed using incident coding and usability problem breakdown analysis similar as in the first phase.

For the analysis of the semi-structured interviews, the verbalized information was transcribed (Appendix P). After the transcription, the interview data was structured and divided into six categories. Categories were created based on the section of the prototype, resulting in six categories associated with different sections of the prototype:
i.) Overall impression (e.g., when participants referred to the prototype in general without referring to a specific section); ***ii.) Database*** (e.g., when participants were referring to the database properties, characteristics or functionalities); ***iii.) Image viewer*** (e.g. when participants referred to the image viewer properties or functionalities associated with the image viewer); ***iv.) AI integration*** (e.g. when participants discussed the implementation of the AI algorithm); ***v.) Annotation workspace*** (e.g. when participants referred to the annotation section or any tool or functionality associated with editing segmentations); ***vi.) Analysis Section*** (e.g. when participants referred to any design element or characteristic which can be found in the analysis section) (see Figure 10).

Figure 10

In the first round of analysis, the interview data were divided based on the section which was discussed



In the second round of analysis the categorization was further divided and the information in each section was divided between positive feedback, negative feedback, and neutral feedback/suggestions for improvement. For instance, when a participant would mention that "the image viewer looks great", then this would be categorized as *image viewer – positive feedback*. The division between positive, negative, and neutral statements was made to assess which aspects of the prototype were perceived as positive and to understand sections require more improvement.

In the final round of the analysis, interview codes were created. Similar statements were grouped together and they were given a fitting label, also referred to as a code. For example, if one participant said "the image viewer was easy to use" and another participant said that he/she "did not encounter any difficulties when using the viewer", then both of these statements were coded as *image viewer – positive feedback – easy to use.* Although these statements are voiced differently, both statements have an equivalent meaning. Thereby, both statements would be labelled as the same code. A coding tree was used for the documentation of the categories, groups, and codes (see Appendix Q).

To determine which codes were more prominent, the frequency of each code was counted. Moreover, the frequency of positive, negative, and neutral statements in each category was assessed to determine which sections received more positive feedback and to determine which specific sections require improvement. The interviews were analyzed and coded with ATLAS.ti 22 (atlas.ti).

4.2 Results

4.2.1 Usability of the Prototype

The tasks were completed successfully by the vast majority of the participants. In terms of effectiveness, seven out of nine tasks were completed by all 13 participants. Task 1 was completed by 11 participants, and task 3 was completed by 12 participants (see Table 7).

Table 7

Mean task completion rate and mean deviations from the normative pathway grouped by task

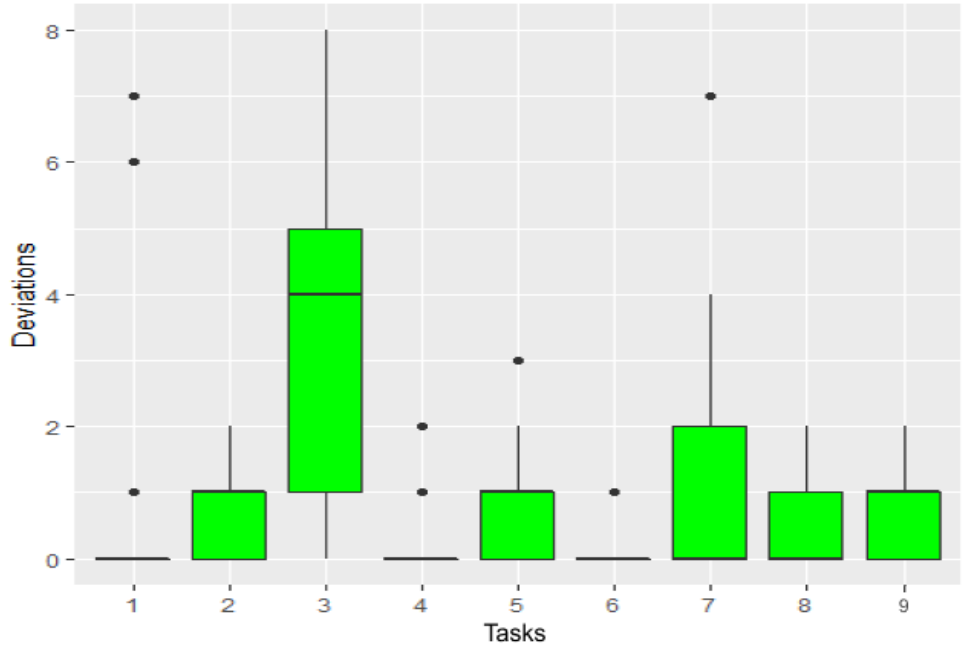
Task	Description of the task	Task completion	Mean deviations
1	Importing medical image data	84.6% (11/13)	1.08
2	Load CT scan into viewer	100% (13/13)	0.77
3	Change the layout	92.3% (12/13)	3.54
4	Change the window level	100% (13/13)	0.31
5	Select and initiate AI	100% (13/13)	0.92

6	Load the next case	100% (13/13)	0.08
7	Upload image data	100% (13/13)	1.38
8	Create annotation tasks	100% (13/13)	0.64
9	Distribute tasks to annotators	100% (13/13)	0.92

In terms of efficiency, as suggested by Figure 11, task three (changing the layout) seemed to produce a high number of average deviations compared to the other tasks (m = 3.54).

Figure 11

Boxplots of the deviations from the normative pathway, one boxplot per task



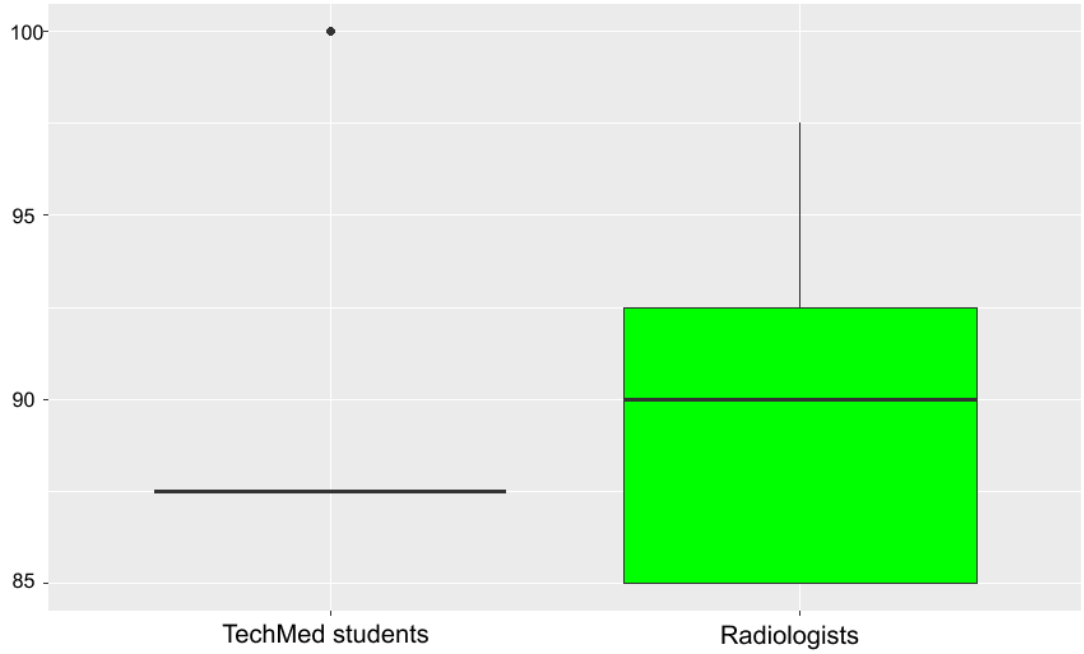
The rest of the tasks caused minimal to no deviations from the normative pathway. In fact, the median amount of deviations is either 0 or 1 for all tasks except for task 3 in which the median is 4 deviations. The upper quartile of all boxes is at 0 or at 1 deviations, except for task seven where the upper quartile is at 2 deviations and at task three where the upper quartile is at 5 deviations. The tasks seem to cause little problems overall, with exception of task 3. A few outliers can be observed in task 1, where majority of the participants did not struggle at all, but 3 participants seemed to have more severe problems with finding the right pathway.

Looking at the satisfaction in use, measured by the SUS, the results are suggesting a high level of average satisfaction perceived by the participants (M = 89.62;SD = 4.55). This is well above the 68% average score of satisfaction indicated by Bangor et al. (2008). No significant differences were identified in terms of satisfaction between expert radiologists and

medical students. The median SUS score of radiologists is slightly higher at 90 compared to the median score of students at 87.5 (see Figure 12).

Figure 12

Comparison of satisfaction between students and radiologists

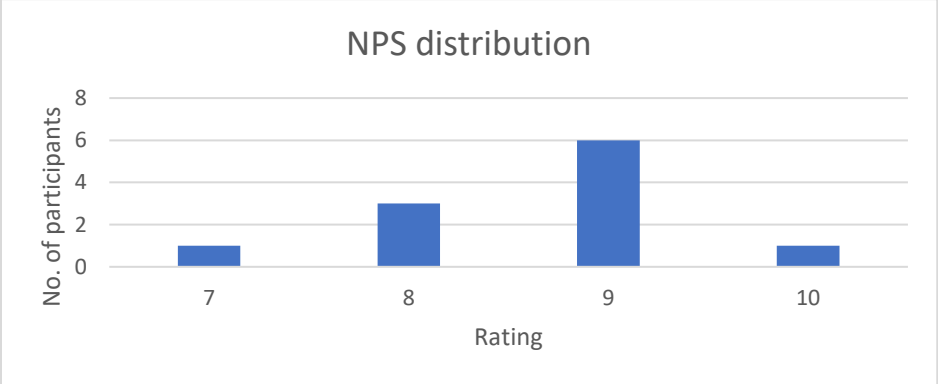


4.2.2 Net Promoter Score

Participants were very positive in terms of intention to use and promote. The mean score on the 10-point Likert scale was quite high (m = 8.85). The overall NPS score was +100, indicating no detractors among the participants. Three participants rated the likelihood to recommend 10 out of 10, and another six participants rated the likelihood to recommend 9 out of 10. Therefore, nine out of 13 participants can be classified as promoters. Three participants gave an 8 out of 10 rating and one participant gave a 7 out of 10 rating. Thereby, the remaining four participants can be classified as "passive" in terms of promotion of usage (see Figure 13).

Figure 13

Distribution of the NPS scores



3.2.3 Usability Issues Experienced by the Users

The incident coding and the subsequent usability problem analysis revealed the existence of 10 UPs in the current design of the prototype (see Table 8).

Table 8

Usability issues which were identified when analyzing the recordings of the interaction

Number	Description	Visibility
01	Drag & Drop is not featured in the prototype	85% (11/13)
02	Settings menu of the viewer tools is confusing	85% (11/13)
03	Concept of annotation tasks is not clear to users	38% (5/13)
04	Buttons in the segmentation menu are confused for annotation tools	15% (2/13)
05	Functionality of the segmentation tools is not intuitive	15% (2/13)
06	“Confirm” button for initiating the AI is confusing	15% (2/13)
07	Tooltips are missing	15% (2/13)
08	“AI should be placed in the viewer toolbar	15% (2/13)
09	Not clear that “automatic segmentation” refers to AI	8% (1/13)
10	Hotkeys and mouse bindings are not implemented in the prototype	8% (1/13)

UPs number 01 and 02 were experienced by 11 out of 13 participants and stand out compared to the other identified issues in terms of visibility. UP 01 (“*Drag & Drop is not featured in the prototype*”) was caused by the lack of interactivity of the prototype, as it refers to the implementation of drag&drop for loading data into the system. UP 02 (“*Settings menu of the viewer tools is confusing”)* was caused by two separate menus in the viewer toolbar which participants were not able to distinguish. One menu for the basic settings of the tool was

opened when clicking on the icon directly and a menu for advanced settings was opened when clicking on the arrows next to the icon (see Figure 14).

Figure 14

Tools in the viewer with adjacent arrows to open advanced settings for each tool

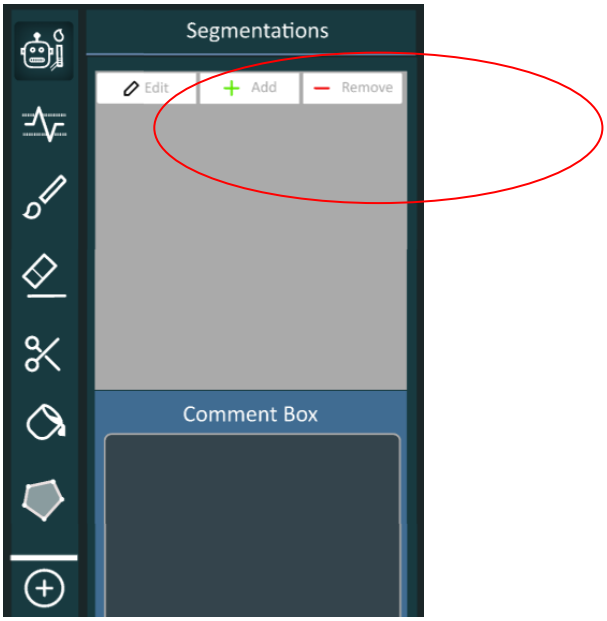
Participants often missed that the arrows next to the icon open another menu for the advanced setting and they tended to click on the icon instead. Thereby, they missed out on important settings for adjusting the view.

UP03 (“*Concept of annotation tasks is not clear”)* was encountered by five participants. The concept of tasks and the data import related to tasks did not seem clear to some participants. It seemed confusing at first for some of the participants, and they did not immediately import the tasks which were assigned to them.

UPs 04 and 05 were found in four out of 13 participants. In UP 04 (“*Buttons in the segmentation menu are confused for segmentation tools”)* the participants confused the annotation tools with the tools for managing the entire segmentation (see Figure 15).

Figure 15

UP 04 showed how participants confuse the buttons for managing the segmenations with the annotation tools



Participants believed that the buttons edit, add, or remove could be used as tools to annotate medical images, but these buttons were used for managing multiple segmentations. For instance, by clicking on “add”, a new segmentation could be added. Instead, the participants needed to click on the tools on the left-hand side in the toolbar. This did not seem clear enough.

UP 05 (“*Functionality of the segmentation tools is not intuitive”*) referred to the icons of the toolbar. The functionality of each icon did not seem entirely clear to four participants. For instance, the functionality of the scissors or the paint bucket were not recognized by everybody. Missing tooltips due to the limited interactivity of the prototype (UP07, 2/13 encountered) could also be a reason for the lack of understanding of the functionality of the tools in the annotation workspace.

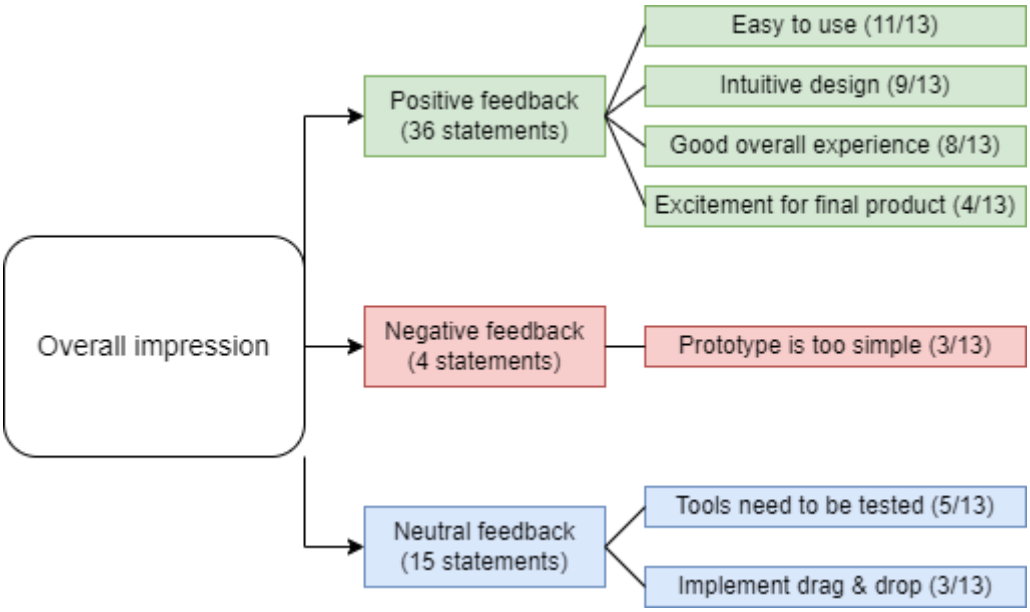
4.2.4 Subjective Experience of the Participants: Interview data and Coding

A summary of the codes and the findings of the semi-structured interview is presented here. Only the most prominent codes in this section which were encountered by at least three participants were highlited. For the full results of the interview analysis with all codes and a selection of quotations for each code, see Appendix R.

1. Overall impression of the prototype: The interviews revealed that the general impression of the prototoype was positive. 36 quotations were coded as positive statements, four quotations were coded as negative statements, and 15 statements were coded as neutral statements/suggestions for improvement (see Figure 16).

Figure 16

Number of statements and most prominent codes which were categorized in "overall impression" of the prototype



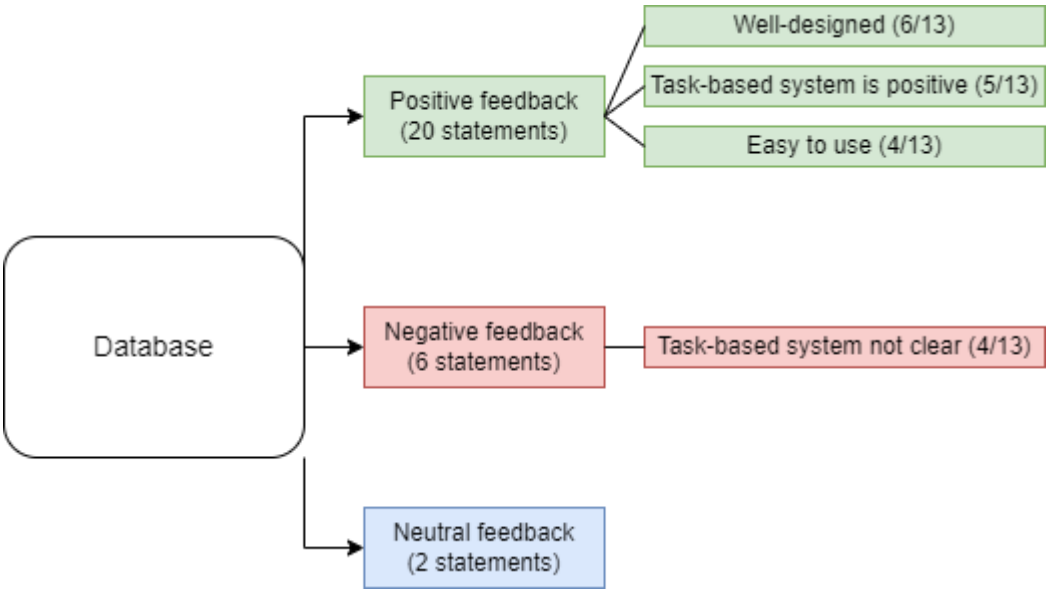
The code *easy to use* was found in 11 out of 13 interviews *("It was very simple. Even as a non-radiologist it was very easy to use it." ¶ 17 in P3; "I liked most its simplicity. It is straightforward so i think this is really important that you don't have to look that long for a feature." ¶ 12 in P6).* Moreover, the code ***intuitive design*** was present in nine interviews (*"It is so much better than the previous one I used in your experiment. It is intuitive and it is similar to other systems we use I guess so you can use knowledge of how you do it in other systems which makes it intuitive." ¶ 14 in P9).* Eight out of 13 participants had a ***good overall experience*** (*"No I am very pleasantly surprised with this new prototype. Great improvements I think."* ¶ 52 in P9), and four participants showed ***excitement for the final product*** *("Nice that you are making this, it could be really useful."* ¶51 in P7). Negative feedback in the overall impression category regarded the level of realism and the simplicity of the prototype. Three participants made statements that the ***prototype is too simple*** *("I am wondering how it is to do more complicated stuff, where hidden features would be or if it was just as simple as this. If that could be implemented."* ¶ 14 in P12*).* Various neutral statements and suggestions for improvement were made, for instance five participants mentioned that ***tools should be tested for a better impression of the prototype*** (*"But the exact handling like the windowing and removing and adding things cannot be tested at this moment. This is also very important how I would experience this program."* ¶ 14 in P10). Three participants suggested to ***implement drag &***

drop in the next design iteration *("And I would really like to drag-and-drop, if possible. " ¶ 19 in P3).*

2. Database: The database received more positive feedback than negative feedback. 20 statements were coded as positive feedback, six statements were categorized as negative feedback, and two neutral statements/suggestions for improvement were made (see Figure 17).

Figure 17

Interview statements and most prominent codes concerning the database

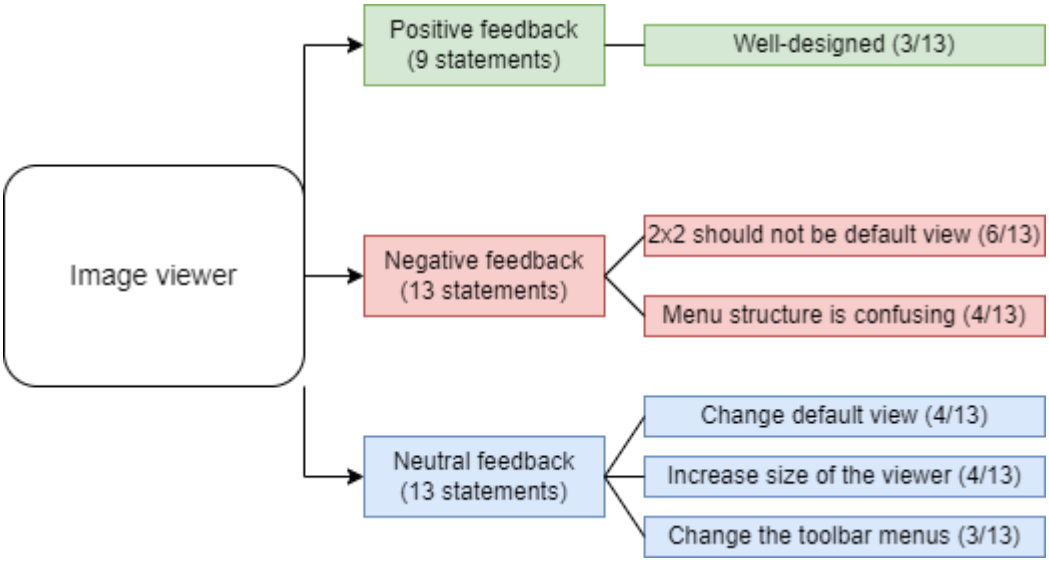


Addressing the postivive feedback, six participants made statements that the database was ***well-designed*** *("That was good. Looks nice." ¶ 16 in P5).* Five participants stated that the ***task-based system is good*** (*"I really like the task-based systems and that is also something that I don't know from the programs which I am currently using." 5:20 ¶ 18 in P5).* Furthermore, five participants found the database ***easy to use*** *("Very easy, really intuitive." 15:38 ¶ 26 in P10).* On the other hand, four participants critizised the task-based system and stated that the ***task-based concept is not clear enough*** (*"When I was assigning the task that was not necessary as straightforward as segmenting. I didn't really understand the concept of tasks. What is a task that I need to assign? I was wondering: What is a task, can I assign different task to the same set of images? It was quite confusing to me." 2:7 ¶ 12 in P2).*

3. Image viewer: The viewer received more negative feedback than the other sections, with nine positive statements, 13 negative statements, and 13 neutral statements/suggestions for improvement (see Figure 18).

Figure 18

Feedback on the image viewer and the most frequent codes which were identified

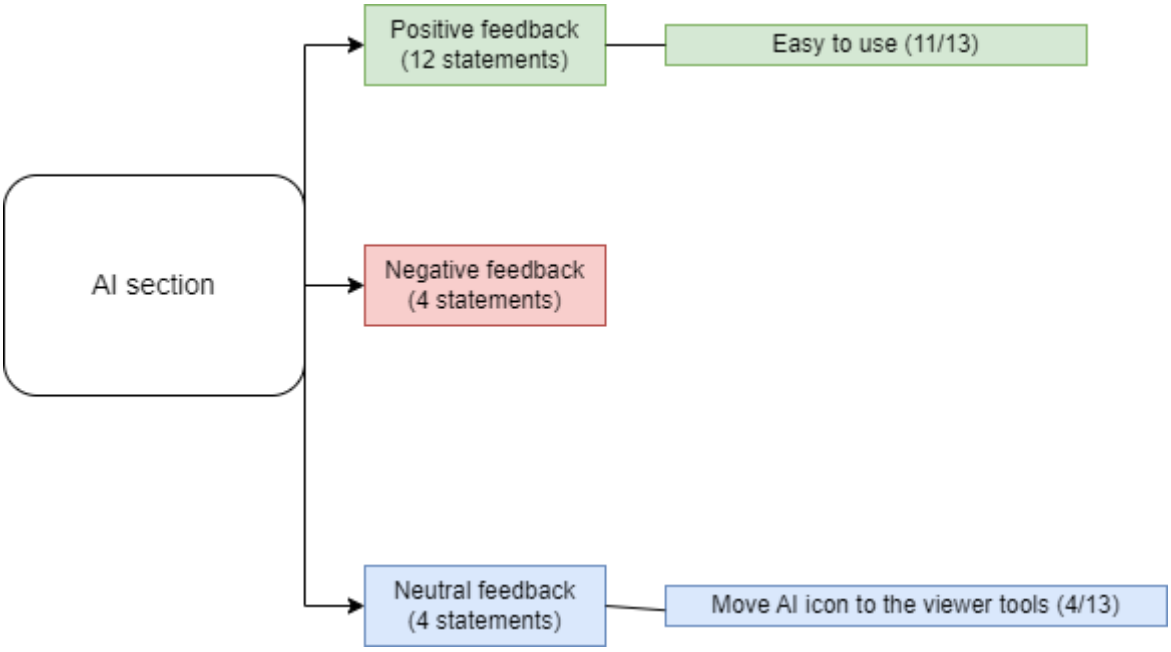


The most prominent positive code was that the viewer was ***well-designed***, which was mentioned in three interviews (*"The viewer looked nice." ¶ 24 in P6).* The most criticism regarded the default view which was set to 2x2. Six participants mentioned that ***2x2 should not be the default view*** (*"I don't really need a 2x2 view here. Just like a big axial and make a small coronal and sagittal on the side, that would do it." ¶ 23 in P2).* Four participants voiced that the ***menu structure of the viewer tools is confusing*** (*"For me it was also confusing that you have the grid logo and the arrows on the right. These are two functionalities in one button. You would expect that if you press one button, then you would get all the information behind it. But now if you press the grid, you refer to the arrows, so that was a bit confusing to me." ¶ 15 in P1).* As neutral statements/suggestions for improvement, four participants suggested to ***change the default view when opening the viewer*** *("I think I can change the layout, right? Usually when I review the case I want the axial screen the biggest and the other screens smaller."* ¶ 32 in P10). Three out of 13 participants mentioned to ***increase the size of the viewer*** (*"I would say take as much space as possible for the scan. Because I don't want to stare at the small scan. I want to have a huge image with lots of detail." ¶ 47 in P9).* Furthermore, three participants suggested to ***change the grid layout menu*** (*"It should maybe be all in the same drop-down menu, the different options for the grid." ¶ 18 in P2).*

4. AI integration: The AI section received mostly positive feedback. Overall, 12 statements were coded as positive feedback, four statements were coded as negative feedback, and eight neutral statements/suggestions for improvement were counted (see Figure 19).

Figure 19

Statements and most frequent codes regarding the AI integration

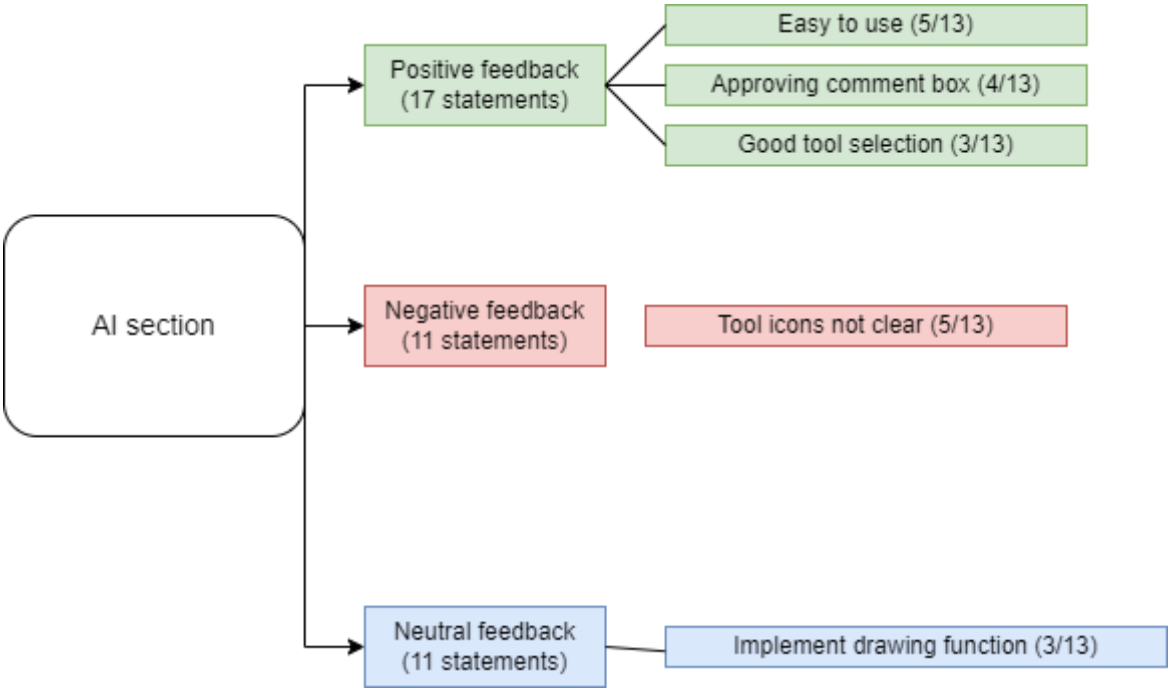


11 out of 13 participants found the AI ***easy to use*** *("But loading it and getting it to work on the case seemed easy and intuitive. It will be nice to see how easy that is. In some programs even if you only have to adjust small things it can be really annoying." ¶ 40 in P9).* The most prominent suggestion for improvement was to ***place the AI icon in the viewer toolbar***, which was suggested by four out of 13 participants (*"My first thought was to go to the toolbar. I get that there is another tab for that, but if it is just selecting something from a list it could also be in the position of the wrench. Other than that it works fine." ¶ 22 in P12).*

5. Annotation workspace: The annotation workspace also received mostly positive feedback, with 17 quotations marked as positive feedback, 11 quotations marked as negative feedback, and 11 neutral statements/suggestions for improvement (see Figure 20).

Figure 20

Interview statements and most frequent codes concerning the annotation workspace

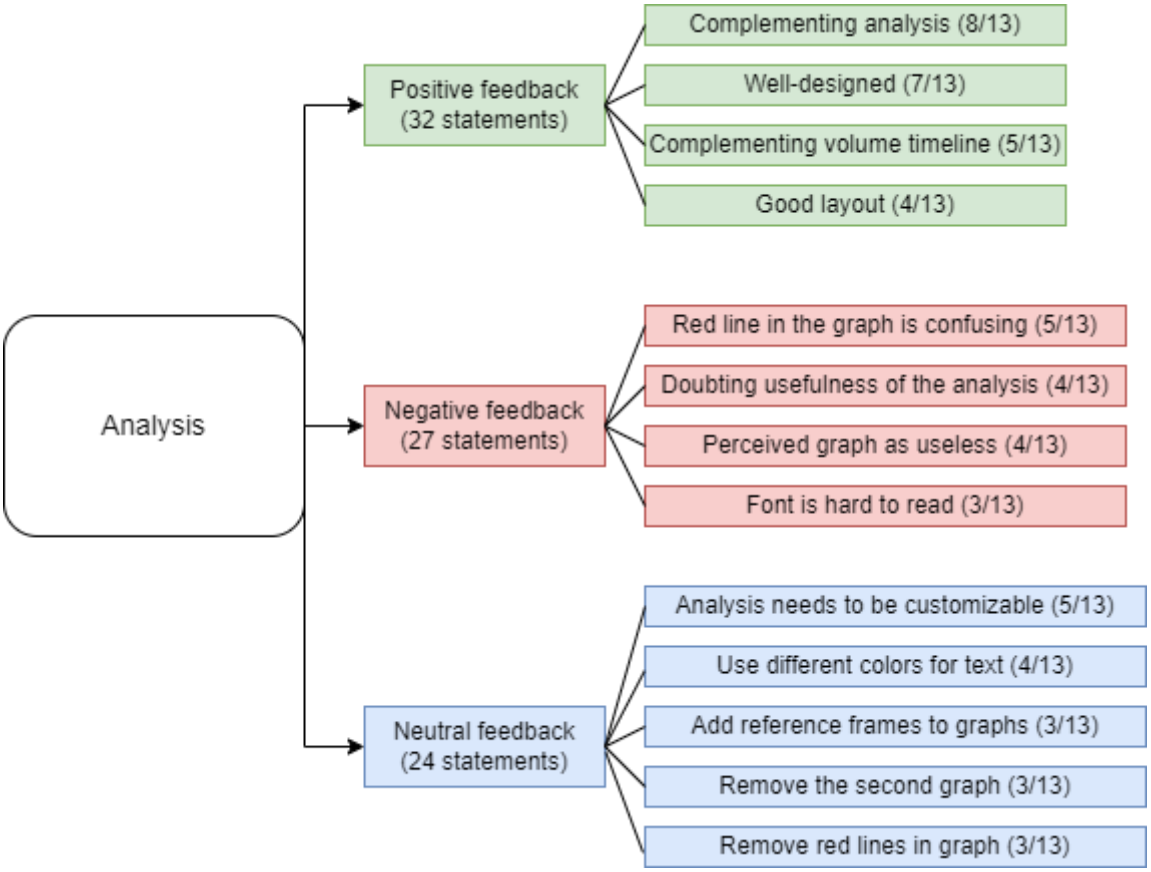


Five out of 13 participants labelled the annotation workspace ***easy to use*** *("I think the functionality and the symbols make sense." ¶ 26 in P12).* Four participants found that the ***comment box is a good idea*** *("Very good, especially the comment box, as I said. So not everybody has to open an excel sheet and type comments separately. I would say this is very smart." ¶ 26 in P5).* Three participants complemented the ***good tool selection*** *("I think it really has the right selection of tools as a basic selection" ¶ 41 in P7).* Negative feedback on the annotation workspace regarded the design of the icons, as five participants found the ***tool icons not intuitive*** *("Maybe the icons are not that intuitive, for example when I needed to remove segmentation there was a button that stated "remove", but I needed to click on the eraser." ¶ 11 in P1).* As a suggestion for improvement, three participants emphasized on implementing ***a drawing function*** as an additional tool for annotation (*"You have like a brush and fill between dots but you don't really have a pen, right?" ¶ 52 in P13).*

6. Analysis section*:* The analysis section generated by far the most quotations and it seemed like participants were eager to discuss this section specifically. The feedback was mixed, and it seems like the participants had controversial opinions about this new section. 32 statements were coded as positive feedback, 27 statements were categorized as negative feedback, and 24 neutral statements/suggestions for improvement were made (see Figure 21).

Figure 21

Feedback on the analysis section with the most prominent codes



A total of eight participants ***complemented the analysis section*** because it is something new and it supports users with further means for the analysis of the output (*"I think it would be absolutely useful. I think this is really missing in clinical life and this is important for RECIST evaluations for example. I think the reports of the future will also have something like this in it. For the patient or for the clinician, this is super important for the future." ¶ 4 in P5).* Furthermore, seven participants mentioned that the analysis section was ***well-designed*** *("I think it looks nice but it depends on what you want to use the program for and who are the intended end users." ¶ 4 in P11).* Five participants mentioned that the ***tumor volume timeline is useful*** (*"but the timeline is always useful. It always gives some idea of what it is happening and the exact percentages which is nice considering the RECIST." ¶ 4 in P8).* Four participants suggested that the ***layout of the analysis section is good*** (*"I think the layout is really convenient because you read from left to right. You select a patient, you see which CT scans there are and then you see the analysis of it. So I think it is a good order to present the view." 1:19 ¶ 2 in P1).* Negative feedback on the analysis section regarded the design of the graphs. Five participants mentioned that the ***red lines in the graph are perceived as confusing*** (*"The first graph is*

confusing to me, I don't like the jumping red line. This is very strange, in a graph like that I don't think the volume would go up and down and up and down but the way it is presented is confusing." ¶ 4 in P2). Four participants ***doubted the usefulness of an analysis section*** in a platform for medical image annotation (*"A radiologist might be interested, I am more of a researcher. If I want any information about the scans then the segmentations were already checked by someone, it might not be necessary" ¶ 4 in P12).* Another four participants ***perceived the histogram analysis as useless*** which was a historgram to compare the severeness of the present case to other cases (*"Tumor volume in comparison with other cases… You mean patients? Maybe this is some standard analysis which I am not familiar with, but I am not sure how useful this is. Are the other patients from the same clinical trial or not?" 8:17 ¶ 4 in P8).* Moreover, three participants found the ***font used in the analysis section hard to read*** (*"There were also quite small letters in red which is not that easy to pick up with your eyes and I had to look very closely to read it. Maybe I would increase it and check if this is the best contrast possible in your screen." ¶ 19 in P10).* The most prominent neutral statement was that the analysis section needs to be customizable in the future, based on the different AI algorithms or studies which are employed. Five participants mentioned the ***need for customizability in the future*** (*"Yeah, exactly. In general I would say that it is useful, but it really depends on what kind of information you need for a project and if the analysis section is tweakable." ¶ 17 in P11).* Furthermore, four participants suggested to ***change the font and the color*** to make it bigger and to change the color of the font to be brighter (*"the red numbers are difficult to read for me and also if you are colorblind the numbers in red and green are really hard to see. I would prefer it in white or black, but not in these colors. I think the yellow letters in the middle in the second graph are also difficult to read. I think if it fills your whole screen it is easier to see because the numbers are bigger." 4:22 ¶ 4 in P4).* About three participants suggested to ***add a reference frame to the graphs in the analysis (****"I would not use this one. It depends on where you are working, and what is the average case. Is it in your hospital, in this program? In some hospitals the volume can be very low, and here we see a lot of cases with severe load so it depends on what do you think what the average point is. I prefer making the other graphics bigger and skip this one." ¶ 6 in P4").* Three participants advocated for ***removing the red line in the tumor volume timeline*** *("Just the red lines are a little misleading. Maybe just remove them." ¶ 6 in P7").* Another three participants suggested to simply ***remove the second graph in the analysis***, as it does not add much value *("If I use the percentages and the change over time, I think for me the volume and the percentage change are most important and I want to see it large and clear. How the other cases are doesn't really matter, and the most important for*

me is individual patients. Of course I understand that you also have to consider the wishes of pulmonologists." ¶ 18 in P10).

4.3 Discussion

The results of the user test suggested that the new prototype has good overall usability and it also received mostly positive feedback in the interview sessions. The new UI seems to resolve some of the issues which have been identified in the first usability assessment. Task completion rates and deviations are better overall, and less usability problems have been observed in the second usability test. Moreover, the satisfaction with the new UI seems to be much higher (see Table 9).

Table 9

Comparison of the results of both user tests in phases 1 and 3

Measurement	Old UI	New UI
Task completion rate (all tasks)	82% (92/112)	97% (114/117)
Mean deviations (all tasks)	2.52	1.07
Identified usability problems	16	10
Mean SUS score	50.21	89.62

Nevertheless, the comparison of the user tests has to be treated with caution as there are several limitations that need to be considered. The level of realism of the prototype was limited. Some functionalities could only be simulated up to a certain point, and therefore not every aspect of the proposed design could be tested thoroughly. This usability test assessed the layout and the steps which need to be taken for specific workflows, but other aspects of use such as real annotation of medical images and the handling of the different tools could not be simulated. Moreover, both usability tests were conducted with a different set of tasks, at a different point in time, with a different set of participants, and under different circumstances. Random effects and individual differences cannot be accounted for. In order to produce a more reliable comparison between both UIs which also accounts for learning effects, A/B testing with repeated measures should be conducted, as it is done in Schmettow et al. (2017). Therefore, the results of the two tests do not necessarily prove that the new design is superior to the previous implementation. However, as the different measurements all point towards the same direction and since the interview feedback was also positive for the most part, it could be justified as a trend and a preliminary assessment of the UIs.

Overall, it seemed remarkable how positively the participants reacted towards the newly proposed UI. Although the prototype was not fully functional, it received a lot of

compliments and a lot of positive feedback, already. The usability test revealed that the proposed design still has flaws, but regardless the participants seemed very content with the new prototype. This could be the case because the proposed design fills a gap and the participants perceived the application as useful, when further developed. The proposed approach for data sharing and task assignment in medical image annotation could make the work for annotators less burdensome and more efficient. The availability of medical image data and data sharing still constitute bottlenecks in the successful development of AI models (Panayides et al., 2020; Willemink et al., 2020), and new approaches for more efficient data sharing seemed to be valued by the experts who participated in the user test. Moreover, the implementation of AI models and the proposed analysis of AI-generated results could make AI algorithms in medical imaging easier to use and more accessible for clinical experts without technical knowledge in AI. The design of UIs for the interpretation of AI-generated results is challenging, but it could help to bridge the imbalance between end-users and AI developers (Chen, Gomez, Huang, & Unberath, 2022). Chen et al. (2022) reported a lack of formative user research to inform the design and development of transparent AI models in medical imaging and argue that these shortcomings put contemporary research on transparent AI at risk of being incomprehensible to users and recommend formative user research as a first step to understand user needs and requirements for transparent AI models which are usable for a clinicians. The present research introduces first ideas on how to apply conventional HCD methods to the development of a new UI for the integration of AI algorithms by putting the user in the center of development. Various HFE methods for the development and the testing of the new prototype were applied, and the positive feedback which was generated in the final assessment emphasize that these methods are feasible for evidence-based concept generation and development of AI-featured products.

4.4 Suggestions for the Further Development and Improvements for the Prototype

On top of the built prototype, the analysis of the observed usability issues and the subjective experience also provided key suggestions about how to further improve the design. Below improvements are reported which could be implemented based on the results and the suggestions which have been reported above.

4.4.1 Overall improvements: Drag & Drop, Customizability, and Tool Implementation

A major improvement for the next design iteration is the implementation of drag & drop to move elements or data from one position to another. 11 out of 13 participants tried to drag & drop (UP01) in order to move medical image data or tasks in the UI. This feature was not

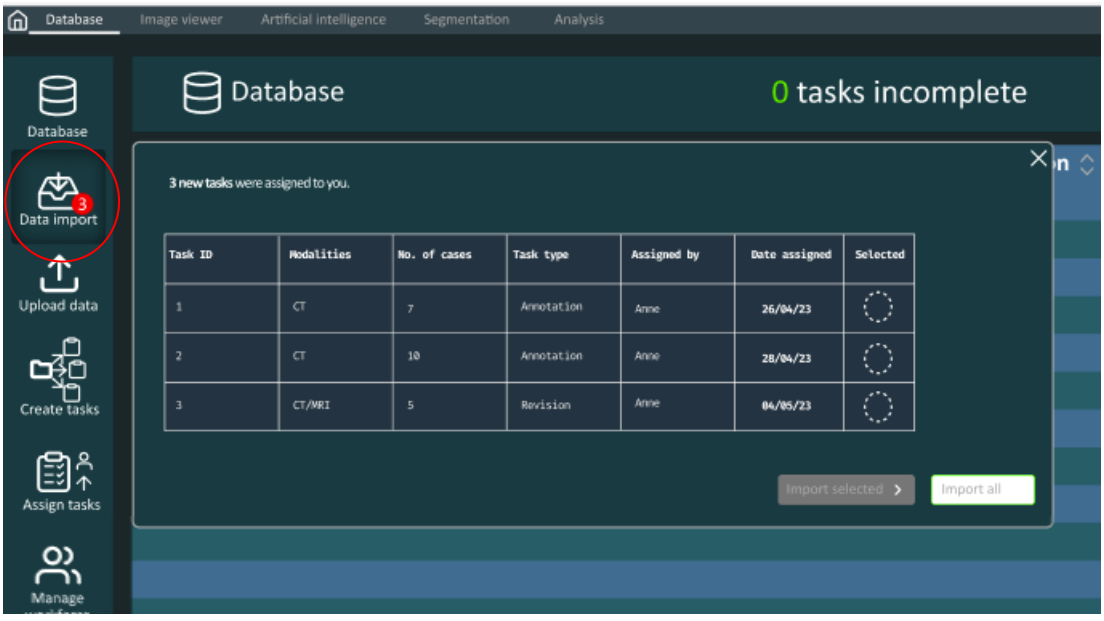
integrated in the current design iteration due to the limited interactivity of the prototype, but it should definitely be a feature in the upcoming design iteration. Moreover, the customizability of the UI could be another improvement. Hotkeys and key bindings should be available in the next design iteration and they should be adjustable based on the users' preferences. Customizing the workflow could facilitate efficiency and satisfaction when expert users are making regular use of the program. Moreover, viewing tools and annotation tools should be available for testing in the next design iteration. Five participants mentioned in the interview that the handling of the tools is essential for a better impression and a more realistic assessment of the quality of the program. Therefore, it should be possible for the users to interact with tools and models in real time in the next design iteration to facilitate better insights.

4.4.2 Refining the database: Changing Labels and Introducing New Users to Task

The database only caused one issue in the user test and the feedback from the participants during the interview session was positive. However, an issue that was observed during the usability test and also voiced by the participants during the interview session regarded the task-based data import. The main problem with this feature was that some participants did not fully grasp the idea behind a task-based annotation system. This issue could be tackled by changing the label from "my tasks" to "data import" to make it more clear that this feature is about loading medical image data into the system (see Figure 22).

Figure 22

Changing the name from "my tasks" to "data import"



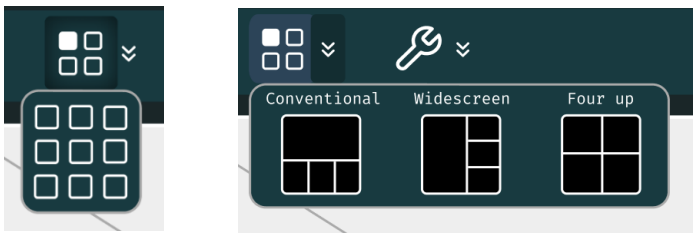
Moreover, an introduction and a description of the task-based functionalities should be provided when using the program for the first time. The task-based approach to data sharing constitutes a novelty and users will not be familiar with it. Therefore, users need to be accustomed with the functioning of the system before they can use it effectively. Instructions on the task-based annotation system could be shown when starting the program for the first time or a help section could be implemented so that users become aware of the functionality before they start utilizing the software.

4.4.3 Updating the Image Viewer: Changing the Menu Structure, the Default View, and the Layout

Two UPs were observed in the image viewer section during the usability test, and moreover, the viewer has been subject to plenty of criticism in the interviews. For this reason we advise to give the image viewer a makeover. The most prominent issue which has been observed in 11 out of 13 participants during the usability test and also has been voiced by four participants during the interview is that the menu structure of the tools in the viewer is confusing (see Figure 23).

Figure 23

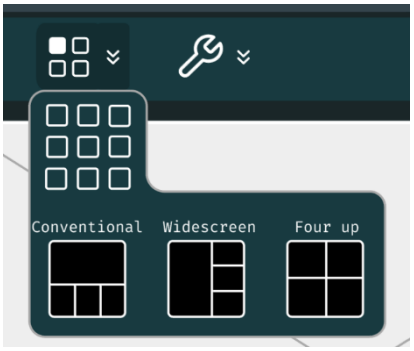
The arrows next to the icon open a separate menu which was difficult to find



To fix this issue, the menu structure could be changed, and the two menus in the viewer could be combined into a singular menu with advanced options (see Figure 24).

Figure 24

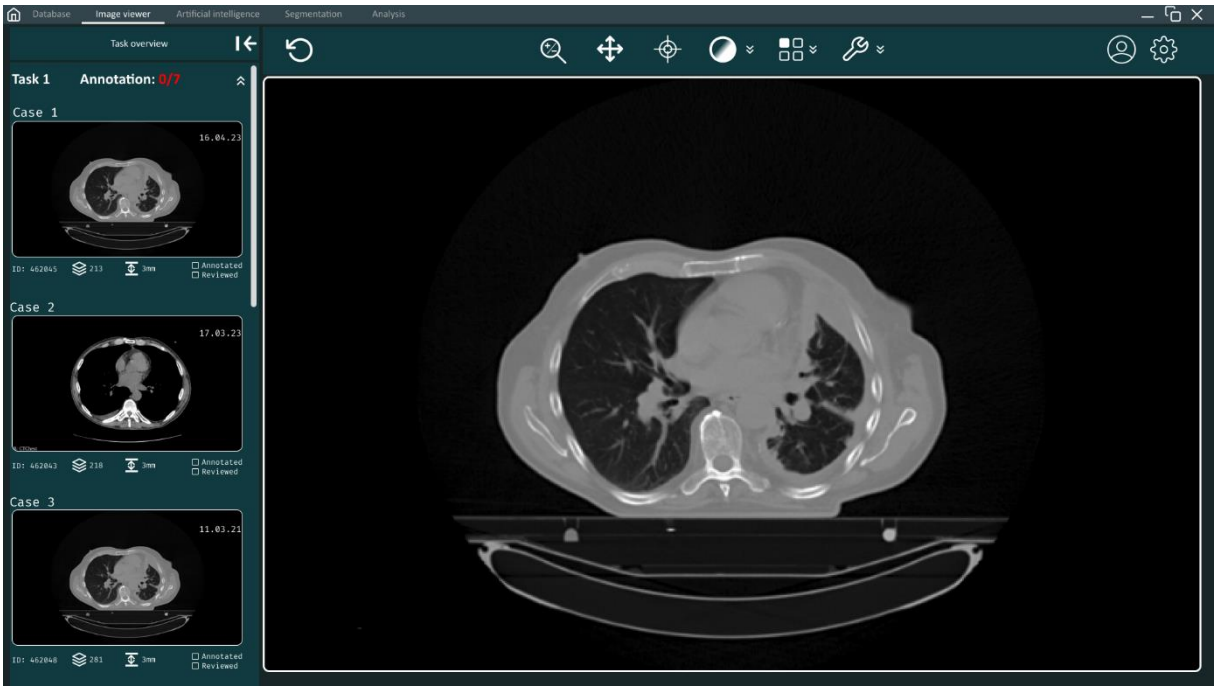
Alternative layout for the menu to avoid confusion abot the menu structure



Another design which was critized by the users was the default layout of the viewer. We used a 2x2 view as default, but six out of 13 participants argued that 2x2 should not be the default view. Instead, the axial should be the default view, and accordingly, we suggest to use the axial view as a default starter view when opening the image viewer. Moreover, two participants argued that the viewer should cover even more space as the viewer is the most important feature for seeing details on the medical images. Thereby, we adjusted the viewer to give it more space next to changing the default view (see Figure 25).

Figure 25

Larger image viewer with a singular axial view as default view

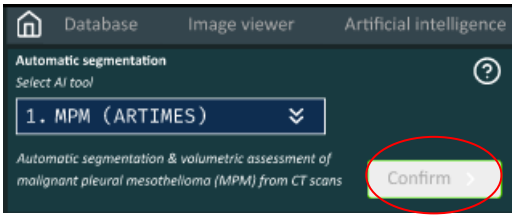


3.4.4 Changing the Labelling of a Button and the Position of the AI Icon

The implementation of the AI in the prototype did not seem to cause many problems during the usability test, and also in the interviews the participants seemed to be satisfied with the proposed AI integration. Nevertheless, a few minor UPs related to AI were observed during the usability test. For instance, UP06 revealed that two participants did not grasp that automatic segmentation needs to be initiated with the “confirm” button in the segmentation workspace (see Figure 26).

Figure 26

Participants needed to click on "confirm" in order to intitate the AI

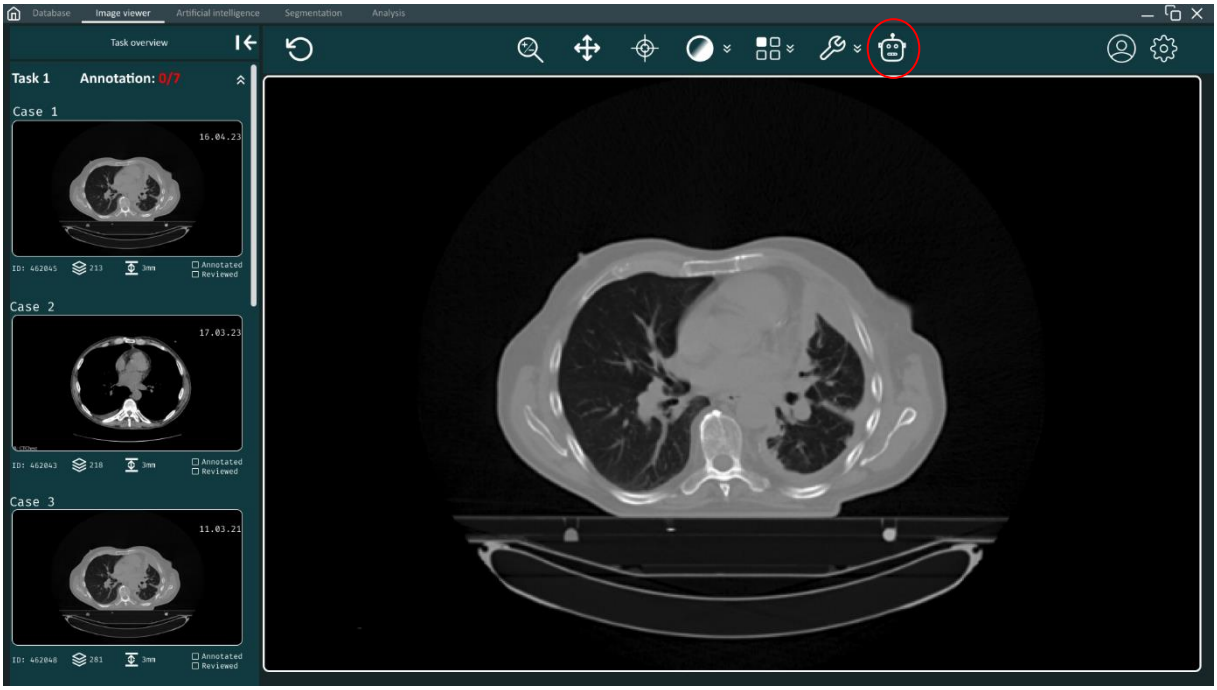


To make it more clear that an action needs to be taken in order to run the automatic segmentation algorithm the label of the button could be changed. For instance, it could be called "*initiate AI*" instead.

Another minor UP which was observed during the usability test and also voiced by two participants regarded the location of the AI icon. It was decided to implement the icon in the annotation workspace, because the AI algorithm is performing automatic segmentation and the functionality corresponds with the manual segmentations which can be performed in the annotation workspace. However, the two participants argued that they would expect the AI functionalities to be grouped with the viewing functionalities in the toolbar. For the further development of the UI changing the location would indeed make sense, because other AI models with other purposes than automatic segmentation may be added in the future, and in that case it would not make sense to group the AI with the annotation tools. The new location offers more flexibility, also for AI models which are not performing automatic segmentation. A more accessible location would therefore be next to the viewing tools (see Figure 27).

Figure 27

Changing the location of the AI icon to the viewer toolbar

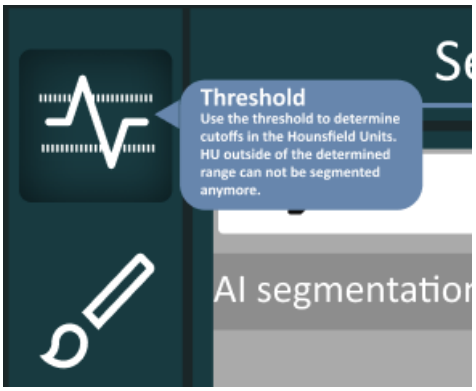


4.4.5 Minor changes in the segmentation workspace regarding the Icon Design and Tooltips

As the verbalized feedback was mostly positive and only one UP was observed in the interaction with the annotation workspace only minor adjustments to this section are advised. The biggest issue was that the functionality of the icons was not clear enough (UP05, 2/13). Five out of 13 participants also voiced this issue in the retrospective interview. Participants confused the functionality of the tools, because they seemed too similar and the difference between the tools was not clear. For instance, the difference between the scissors and the eraser or the difference between the paint brush and the paint bucket was confused often times. This issue will be present especially for new users, once the functionality of the tools is learned the users will be able to differ between the tools more easily. To increase the learnability and to make the functionality of the tools for new users more clear, tooltips could be implemented. When hovering over a tool with the cursor, a text box displays information such as the name and the functionality of the element (see Figure 28).

Figure 28

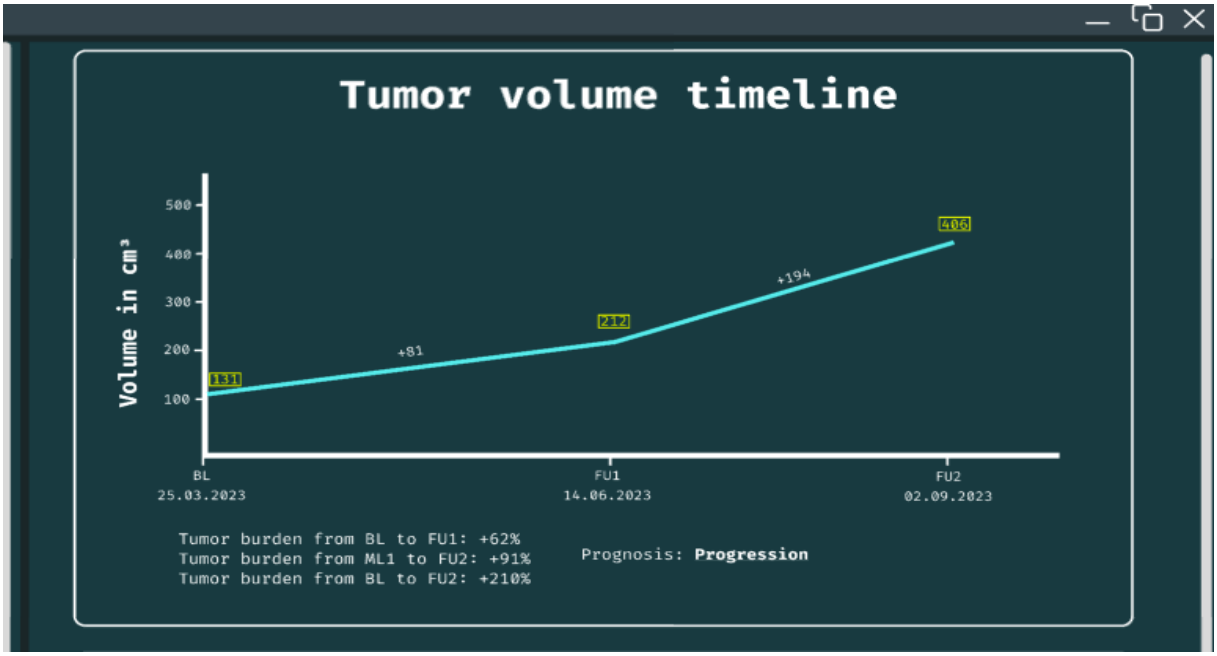
Tooltips should help new users to grasp the functionality of the tools more easily



The analysis section was subject to many different types of feedback and comments in the interviews. It received many compliments, but it was also subject to much criticism. Eight out of 13 participants approved the analysis section and seemed interested in the implementation of analysis features. A lot of criticism regarded the readability of the graphs and the font in the analysis section. Four out of 13 participants claimed that the font is hard to read, and five out of 13 participants said that the graphs are hard to interpret. Red letters were used on a dark background and this seemed hard to read. Issues regarding the readability of graphs and fonts can be fixed easily by changing the color and the fonts in the analysis section. The same accounts for graphs, an updated color design should make the graph easier to read (see Figure 29).

Figure 29

Updated font size, font color, and graph design in the analysis section



More profound issues with the analysis regard the customizability of the section. For instance, five out of 13 participants mentioned that the analysis needs to be customizable depending on the AI model and the study type. This means that the analysis section needs to become much more sophisticated, and developing an automated analysis section for AI-generated results is a complicated task overall. Four participants doubted the overall usefulness of the analysis section and advocated for focusing on the design of the other sections. Overall, the analysis section may be the most difficult to design because it requires a novel conceptual design and it depends on the implemented AI models and the types of data which need to be analyzed. Therefore, the automated analysis will be harder to develop than other sections of the prototype and it could cosntitute a design project on its own which is added to the rest of the program in later iterations. Nevertheless, the participants seemed highly interested in the analysis section and the presentation of AI-generated results to users may be an intriguing research topic for the future.

5. Conclusion

The present work constitutes a first step in the design of a human-centered platform that features AI algorithms in medical imaging. It is an example for the methodological transfer and the adaption of classical HFE methodology to modern use cases. After the initial assessment of the current implementation of the AI algorithm revealed usability issues and low satisfaction with the current implementation of the AI algorithm, HCD methodology was applied to develop a prototype for a novel UI for the integration of AI algorithms in medical imaging. The prototype was designed from scratch by cycling through two iterative design rounds with a subsequent expert review and a user test. HCD methods which were employed involved user research, heuristic design, and usability testing to ensure that the design meets the users' needs as closely as possible and to ensure that the proposed design is working as intended. The final assessment of the prototype revealed that the participants seemed to be satisfied with the novel concepts and the suggested design of the new UI. Final suggestions for further improvement of the prototype were provided to support potential future development processes. The improved prototype and the documentation of the design could be used as a foundation for the further development of a novel software architecture for the implementation of AI in medical imaging.

References

Abras, C., Maloney-Krichmar, D., & Preece, J. (2004). User-centered design. *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications, 37*(4), 445-456.

Alberdi, E., Povyakalo, A., Strigini, L., & Ayton, P. (2004). Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography. *Academic radiology, 11*(8), 909-918.

Alon-Barkat, S., & Busuioc, M. (2023). Human–AI interactions in public sector decision making:"automation bias" and "selective adherence" to algorithmic advice. *Journal of Public Administration Research and Theory, 33*(1), 153-169.

Asan, O., & Choudhury, A. (2021). Research trends in artificial intelligence applications in human factors health care: mapping review. *JMIR human factors, 8*(2), e28236.

Babich, N. (2017). Prototyping 101: The difference between low-fidelity and high-fidelity prototypes and when to use each. Retrieved from https://blog.adobe.com/en/publish/2017/11/29/prototyping-difference-low-fidelity-high-fidelity-prototypes-use

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction, 24*(6), 574-594.

Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., . . . Balkenhol, M. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama, 318*(22), 2199-2210.

Briganti, G., & Le Moine, O. (2020). Artificial intelligence in medicine: today and tomorrow. *Frontiers in medicine, 7*, 27.

Bruseberg, A., & McDonagh-Philp, D. (2001). New product development by eliciting user experience and aspirations. *International Journal of Human-Computer Studies, 55*(4), 435-452.

Carayon, P., Hoonakker, P., Hundt, A. S., Salwei, M., Wiegmann, D., Brown, R. L., . . . Wang, Y. (2020). Application of human factors to improve usability of clinical decision support for diagnostic decision-making: a scenario-based simulation study. *BMJ quality & safety, 29*(4), 329-340.

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—addressing ethical challenges. *The New England journal of medicine, 378*(11), 981.

Chen, H., Gomez, C., Huang, C.-M., & Unberath, M. (2022). Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. *npj Digital Medicine, 5*(1), 156.

Currie, G., Hawk, K. E., Rohren, E., Vial, A., & Klein, R. (2019). Machine learning and deep learning in medical imaging: intelligent imaging. *Journal of medical imaging and radiation sciences, 50*(4), 477-487.

Djamasbi, S. (2014). Eye tracking and web experience. *AIS Transactions on Human-Computer Interaction, 6*(2), 37-54.

Esposito, E. (2018). Low-fidelity vs. high-fidelity prototyping. Retrieved from https://www.invisionapp.com/inside-design/low-fi-vs-hi-fi-prototyping/

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *nature, 542*(7639), 115-118.

Felmingham, C. M., Adler, N. R., Ge, Z., Morton, R. L., Janda, M., & Mar, V. J. (2021). The importance of incorporating human factors in the design and implementation of artificial intelligence for skin cancer diagnosis in the real world. *American Journal of Clinical Dermatology, 22*(2), 233-242.

Filice, R. W., & Ratwani, R. M. (2020). The case for user-centered artificial intelligence in radiology. *Radiology: Artificial Intelligence, 2*(3).

García-Peñalvo, F., Vázquez-Ingelmo, A., García-Holgado, A., Sampedro-Gómez, J., Sánchez-Puente, A., Vicente-Palacios, V., . . . Sánchez, P. L. (2021). Application of artificial intelligence algorithms within the medical context for non-specialized users: the CARTIER-IA platform.

Harte, R., Glynn, L., Rodríguez-Molinero, A., Baker, P. M., Scharf, T., Quinlan, L. R., & ÓLaighin, G. (2017). A human-centered design methodology to enhance the usability, human factors, and user experience of connected health systems: a three-phase methodology. *JMIR human factors, 4*(1), e5443.

Hegde, V. (2013). *Role of human factors/usability engineering in medical device design.* Paper presented at the 2013 Proceedings Annual Reliability and Maintainability Symposium (RAMS).

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer, 18*(8), 500-510.

Houde, S., & Hill, C. (1997). What do prototypes prototype? In *Handbook of human-computer interaction* (pp. 367-381): Elsevier.

Hwang, E. J., & Park, C. M. (2020). Clinical implementation of deep learning in thoracic radiology: potential applications and challenges. *Korean Journal of Radiology, 21*(5), 511.

ISO. (2018). ISO 9241-11:2018 Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts.

ISO. (2019). ISO 9241-210:2019 Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems.

Kotter, E., & Ranschaert, E. (2021). Challenges and solutions for introducing artificial intelligence (AI) in daily clinical workflow. In (Vol. 31, pp. 5-7): Springer.

Krueger, R. A. (2014). *Focus groups: A practical guide for applied research*: Sage publications.

Lekadir, K., Osuala, R., Gallin, C., Lazrak, N., Kushibar, K., Tsakou, G., . . . Tsiknakis, M. (2021). FUTURE-AI: Guiding Principles and Consensus Recommendations for Trustworthy Artificial Intelligence in Medical Imaging. *arXiv preprint arXiv:2109.09658*.

Lencioni, R., & Llovet, J. M. (2010). *Modified RECIST (mRECIST) assessment for hepatocellular carcinoma.* Paper presented at the Seminars in liver disease.

Lewis, J. R. (2006). Usability testing. *Handbook of human factors and ergonomics, 12*, e30.

Liew, C. (2018). The future of radiology augmented with artificial intelligence: a strategy for success. *European journal of radiology, 102*, 152-156.

Lindgaard, G., & Dudek, C. (2003). What is this evasive beast we call user satisfaction? *Interacting with computers, 15*(3), 429-452.

Maguire, M. (2001). Methods to support human-centred design. *International Journal of Human-Computer Studies, 55*(4), 587-634.

Mao, J.-Y., Vredenburg, K., Smith, P. W., & Carey, T. (2005). The state of user-centered design practice. *Communications of the ACM, 48*(3), 105-109.

Molich, R., & Nielsen, J. (1990). Improving a human-computer dialogue. *Communications of the ACM, 33*(3), 338-348.

Muratovski, G. (2021). *Research for designers: A guide to methods and practice*: Sage.

Murphy, D. J., & Gill, R. R. (2017). Volumetric assessment in malignant pleural mesothelioma. *Annals of translational medicine, 5*(11).

Nielsen, J. (1993). Iterative user-interface design. *Computer, 26*(11), 32-41.

Nielsen, J. (2020). 10 Usability Heuristics for User Interface Design. Retrieved from https://www.nngroup.com/articles/ten-usability-heuristics/

NKI. (n.d.). Our vision - A cure for every cancer. Retrieved from https://www.nki.nl/about-us/our-vision/

Norman, D. A., & Draper, S. W. (1986). User centered system design: New perspectives on human-computer interaction.

Olsen, A., Smolentzov, L., & Strandvall, T. (2010). Comparing different eye tracking cues when using theretrospective think aloud method in usability testing. *Proceedings of HCI 2010 24*, 45-53.

Omoumi, P., Ducarouge, A., Tournier, A., Harvey, H., Kahn, C. E., Louvet-de Verchère, F., . . . Richiardi, J. (2021). To buy or not to buy—evaluating commercial AI solutions in radiology (the ECLAIR guidelines). *European radiology, 31*(6), 3786-3796.

Panayides, A. S., Amini, A., Filipovic, N. D., Sharma, A., Tsaftaris, S. A., Young, A., . . . Kurc, T. (2020). AI in medical imaging informatics: current challenges and future directions. *IEEE Journal of Biomedical and Health Informatics, 24*(7), 1837-1857.

Pianykh, O. S., Langs, G., Dewey, M., Enzmann, D. R., Herold, C. J., Schoenberg, S. O., & Brink, J. A. (2020). Continuous learning AI in radiology: implementation principles and early applications. *Radiology, 297*(1), 6-14.

Reichheld, F. F. (2003). The one number you need to grow. *Harvard Business Review, 81*(12), 46-55.

Robinson, B., Nowak, A., Robinson, C., & Creaney, J. (2008). Malignant mesothelioma. *Textbook of Lung Cancer*, 206-222.

Scheetz, J., Rothschild, P., McGuinness, M., Hadoux, X., Soyer, H. P., Janda, M., . . . Keel, S. (2021). A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology. *Scientific reports, 11*(1), 1-10.

Schmettow, M., Schnittker, R., & Schraagen, J. M. (2017). An extended protocol for usability validation of medical devices: Research design and reference model. *Journal of biomedical informatics, 69*, 99-114.

Schönberger, D. (2019). Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications. *International Journal of Law and Information Technology, 27*(2), 171-203.

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction, 36*(6), 495-504.

Sujan, M., Furniss, D., Grundy, K., Grundy, H., Nelson, D., Elliott, M., . . . Reynolds, N. (2019). Human factors challenges for the safe use of artificial intelligence in patient care. *BMJ health & care informatics, 26*(1).

Thrall, J. H., Li, X., Li, Q., Cruz, C., Do, S., Dreyer, K., & Brink, J. (2018). Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. *Journal of the American College of Radiology, 15*(3), 504-508.

Tractinsky, N., Katz, A. S., & Ikar, D. (2000). What is beautiful is usable. *Interacting with computers, 13*(2), 127-145.

Van Den Haak, M., De Jong, M., & Jan Schellens, P. (2003). Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. *Behaviour & information technology, 22*(5), 339-351.

Wang, S., Zhou, M., Liu, Z., Liu, Z., Gu, D., Zang, Y., . . . Tian, J. (2017). Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation. *Medical image analysis, 40*, 172-183.

Wickens, C. D., Gordon, S. E., Liu, Y., & Lee, J. (2004). *An introduction to human factors engineering* (Vol. 2): Pearson Prentice Hall Upper Saddle River, NJ.

Willemink, M. J., Koszek, W. A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., . . . Lungren, M. P. (2020). Preparing medical imaging data for machine learning. *Radiology, 295*(1), 4-15.

Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., . . . Sellitto, M. (2021). The ai index 2021 annual report. *arXiv preprint arXiv:2103.06312*.