MIMBCD-UI User Testing & Analysis Guide Towards Assertiveness-based Interactive Agents for Breast Cancer Diagnosis

Technical Report · October 2021

DOI: 10.13140/RG.2.2.22989.92645/1

João Fernandes
Instituto Superior Técnico

MIMBCD-UI
User Testing & Analysis Guide
Towards Assertiveness-based Interactive Agents for Breast Cancer Diagnosis

João Fernandes
joao.g.m.fernandes@tecnico.ulisboa.pt
Computer Science and Engineering Department
Instituto Superior Técnico
University of Lisbon
Portugal
European Union

27/10/2021

MIMBCD-UI:	mimbcd-ui.github.io
Wiki:	github.com/MIMBCD-UI/meta/wiki
User Research:	github.com/MIMBCD-UI/meta/wiki/User-Research
Repository:	github.com/MIMBCD-UI
Private:	github.com/MIMBCD-UI/meta-private

1 Introduction

Artificial Intelligence (AI) has the potential of expanding several clinical domain possibilities [37]. Although these systems have a high level of accuracy, they do not normally take into account the years of professional experience of the clinician. In order to achieve a more persuasive and reliable agent, we need to analyze and collect data regarding clinician's behavior [33]. Communication is essential to increase the reliability of an agent [32] providing a diagnosis to a clinician. One way to achieve that is by aligning the level of assertiveness [31] of the assistant with the years of experience of the clinician. Besides that, providing an explanation of how the AI model achieved a certain output increases trust on the system, solving the problem knows as the "black-box" problem [14].

This work proposes to study how much the level of assertiveness of the assistant impacts the decision-making of clinicians, depending on their category level. The aim is to compare the impact of assertiveness-based communications mediated by an AI assistance and to study each professional category preference. It will also be studying the impact of a proactive and reactive assistant, and how providing an explanation for a clinical diagnosis increases the integration of intelligent agents in a clinical environment [20, 26]. This work will be integrated in a real clinical project [5, 6, 12].

It was decided to divide the user testing and analysis into three different iterations [18, 16, 17, 19]. The first iteration (Section 3.1) consists on the implementation of an AI assistant. On the second iteration (Section 3.2), we talk to clinicians to get feedback, suggestion, and ideas about the outcome of the first iteration. Finally, in the third iteration (Section 3.3), we conduct formal tests with clinicians to evaluate the hypothesis of this work.

The first iteration (Section 3.1) starts with a stage of interviews where it is recorded the routines and *workflow* of the clinicians. The second stage of this iteration consists of a focus group with affinity diagrams that enables to organize the information from the previous stage. In the third stage, it is designed a low-fi prototype in two scenarios: active assistance and passive assistance. Then, in the fourth stage, through iterative development with four clinicians, it is designed an initial user interface in which there are two scenarios (Assertive or Non-Assertive) and two agent behaviors (Proactive or Reactive). In the final stage, the assistant is implemented and integrated with the previous versions of the prototypes [23, 28].

The second iteration (Section 3.3) involves a focus group with two clinicians where there are a couple of questions to be answered and to be organized using affinity diagrams. With these information, there were some design decision to be taken that will be described in this document. Finally, the assistant was completely implemented in two scenarios: (1) Assertive; and (2) Non-Assertive; and two agent behaviors: (a) Proactive; and (b) Reactive. The assertive assistant uses more assertions and statements, while the non-assertive assistant uses more questions and suggestions. On the other hand, the proactive assistant appears with the explanation when the clinician opens the patient, while the reactive agent appears hidden when opening the patient.

In the third iteration (Section 3.2), we will conduct some tests to discuss, conclude, and report the results of our experiments from clinicians' experience regarding the assertiveness levels, agent behaviors and explainability techniques of the assistant. This iteration will be divided in two stages with one month apart. At each stage, each clinician will diagnose three patients with different Breast Imaging Reporting and Data System (BI-RADS). From the three patients in the second stage, one or two will be the same from the first stage but with inverted level of assertiveness and behaviour.

To conclude, several clinicians will interact with all scenarios and agent behaviors, making also their diagnosis of three patients at two different phases, temporally separated by one month. It will be studied the assistant impact on performance, workload, usability, trust and overall experience, depending on the clinician professional experience category (Interns, Juniors, Middles, Seniors). We expect that performance, workload, usability, trust, and overall experience are increasing variables as a function of understanding behavioral differences between the various professional experience categories of clinicians.

2 Description

This document describes the plan for the user tests that will be done to evaluate the impact of an assertiveness-based communication mediated by intelligent agents to support the decision-making process of clinicians on breast cancer diagnosis. The goal of the user tests is to measure the clinician's performance, state their preference and evaluate User eXperience (UX) improvements with assistants of different levels of assertiveness communication and agent behaviors, depending on the clinician's category of professional experience. It will also be studied the impact of the assistant behavior on the clinician's decision-making. The user tests will have two different phases, separated one month from each other, where each clinician will evaluate three patients. Their task is to interact with the assistant and make their BI-RADS diagnosis.

3 Methodology

Here, we describe the methods of our user testing and analysis guide, including information concerning the user tasks, data, and study procedures. The primary goal of the user testing and analysis is to identify the strengths, weaknesses, applicable explanation needs, and design implications of the explanatory forms. In fact, our goal is to incorporate into the future of intelligent agents the communication properties, considering behavioral differences of clinicians during their decision-making process. Next, we are detailing each interaction and respective stage of it across our study.

3.1 Iteration 1

The first iteration started with interviews, where four radiologists and four members of the *BreastScreening-AI* development participated. The interviews took around 30 minutes each session, where participants addressed key components of their institution's *workflow*.

The focus group consisted of four researchers and four radiologists using affinity diagrams to arrange *workflow* practices and ideas in greater depth. The affinity diagrams allowed to identify a variety of functionalities, such as the need for literature assistance during the agent diagnostic. Besides that, it was also understood what should be the level of assertiveness of the assistant (Assertive *vs* Non-Assertive) depending on the clinician's level (*i.e.*, Intern, Junior, Middle, or Senior). One Senior referred that a more quick and effective communication is better, giving the idea that a non-assertive and proactive assistant would be preferred for a more experienced clinician. On the contrary, a Junior argued that having the last chance and interacting with an assistant that uses affirmative sentences is better, giving the idea less experienced clinicians would prefer an assertive and reactive assistant.

After the result of the interviews and the focus group, it was designed a Low-Fi prototype of *BreastScreening-AI* along with four clinicians. At this stage, there were two scenarios: (1) Active assistance; and (2) Passive assistance. All clinicians agreed that the second scenario (Passive assistance) was the best and most appropriated. After that, through iterative development with four clinicians, it was designed an initial user interface, where there are two scenarios (Assertive or Non-Assertive) and two agent behaviors (Proactive and Reactive). Besides, the clinician can also accept or reject the assistant BI-RADS diagnosis recommendation and see explanation of why the assistant reached that final diagnosis.

Finally, we implemented the proposed assistant and integrated this assistance scenarios in the BreastScreening framework. However, it is important to refine some design details again with clinicians before the formal user tests (Section 3.3) in a real-world clinical setting. Hence, we structured a second iteration (Section 3.2) for this purpose.

3.2 Iteration 2

Building off from the first iteration (Section 3.1), which point to requirements of the system, we conducted several other focus groups, but this time to formulate the next design of our solution. At this second iteration, the focus group consists of a discussion with the clinician where we try to get the answers from the questions presented below to gather more ideas, while showing the prototype and clustering the information with affinity diagrams. Our questions go from information appearance to clinical needs concerning the design and order of recommendations for a better clinician's understanding of the assistance results. After all the focus group meetings, we formulated the final version of the assistant that will be designed and implemented [38] under this study.

The second iteration consists of a focus group with 2 clinicians, where we seek answers to the following questions:

- Should the information about co-variables appear with the assistant's recommendation? Or just when the clinician clicks in the button "Explain"?
- Are the number of lesions a relevant co-variable for the clinician? If yes, what is its priority and importance for the clinician? And if there should be any color distribution for the number of lesions?
- Is it useful for the clinician that each calcification counts as a lesion?
- Regarding masses, besides showing the mass category, we also show the severity of the lesion in a certain region percentage of the mass. Is this information useful for the clinician?
- Regarding the family and personal history, does the clinician think it is necessary to specify something? Or is it enough to refer that there is family or personal history?
- Should a diagnosis of BI-RADS 4 be classified as 4A, 4B or 4C?
- What would be most useful for the clinician to the color to represent? The accuracy of the result of the model? The severity of the co-variable? Or relate the color with the level of assertiveness of the assistant? And how should that color be distributed?
- Would it be useful to have a caption for the meaning of the colors? If so, should the caption always appear or appear just when the mouse is over a co-variable with a color?
- Should the assistant's explanation contain different information between mammography and ultrasound?
- What are clinicians expecting from a second opinion, and when do they use it?

3.3 Iteration 3

At this iteration, several clinicians will use the final version of the assistant. Our goal is to achieve a number of clinicians between 10 and 30. They will have three patients to diagnose, out of a list of thirty patients.

At a first stage, the clinician will evaluate 3 patients with 3 different BI-RADS diagnosis. One with low BI-RADS, 1, one with medium, 2 or 3, and one with high, 4 or 5. The assistant can be assertive or non-assertive. It can also have a proactive or reactive behaviour. What is important to have is a 50/50 relation between assertive and non-assertive assistants, as well as proactive and reactive assistants. This works either for all clinicians or each professional category (Interns, Juniors, Middles and Seniors).

At a second stage, one month after the first stage, clinicians will diagnose 50% of the same patients that they have diagnosed in the first stage. This means some clinicians will diagnose again 1 or 2 patients from the ones they've diagnosed in the first stage. Beside that, the patients that diagnose again will invert the assistant level of assertiveness and behaviour. For instance, if in the first stage the assistant was assertive and had a proactive behaviour, in the second stage it will be non-assertive and have a reactive behaviour. The conditions that were present in the first stage still remain in the second stage.

4 Roles

The roles involved in our study are as follows. An individual may play multiple roles, as well as the study may not require all roles.

4.1 Facilitator

- Provides overview of the study to participants;
- Defines tasks and purpose of the user testing to participants;
- Responds to participant's requests for information;

4.2 Data Logger

- Records participant's actions;
- Records participant's comments for later transcripts;

4.3 Ethics

All persons involved with this guide are required to adhere to the following ethical guidelines:

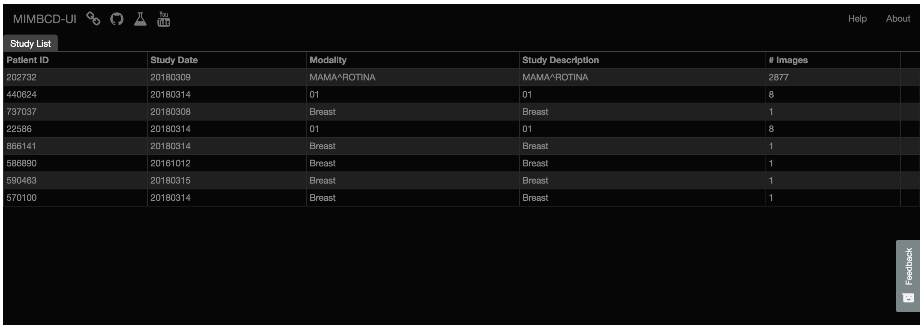
- An individual participant's name should not be used in reference outside the set of questions;
- A description of the participant's answers should not be reported to his or her superior;

5 Apparatus

The session will be held remotely via Zoom. We will use the recording tool included in Zoom to record the session, in order to save all interactions.

5.1 User Interactions

On Figure 1, the user can select the list of patients. The list has a table with several patient information. The first column is the *Patient IDentifier (ID)*; we used it as an identifier of the patient. In that way, we can have anonymized information with no reference to the patient name. The second column is the *Study Date*, the third column is the *Modality* of the used **Digital Imaging and Communications in Medicine (DICOM)** image, the fourth column is the *Study Description* of the used study and the last column is the number of *Images*.



Patient ID	Study Date	Modality	Study Description	# Images
202732	20180309	MAMA^ROTINA	MAMA^ROTINA	2877
440624	20180314	01	01	8
737037	20180308	Breast	Breast	1
22586	20180314	01	01	8
866141	20180314	Breast	Breast	1
586890	20161012	Breast	Breast	1
590463	20180315	Breast	Breast	1
570100	20180314	Breast	Breast	1

Figure 1: List of Patients.

The systems have several buttons (Figure 2) that allows the user to interact or access to a set of user interface features. Each item of the following list represents each metaphoric icon of Figure 2.

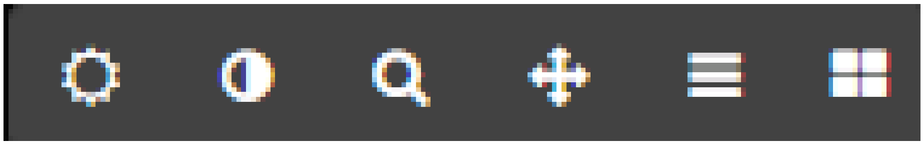


Figure 2: Toolbar of the System available features.

The buttons are (from left to right of Figure 2) as follows:

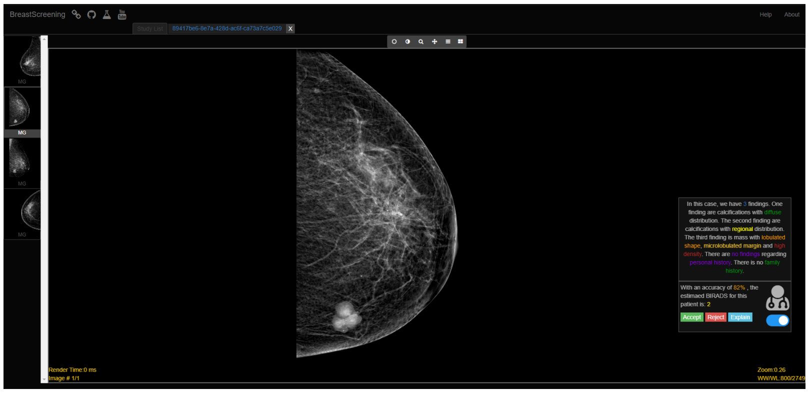
- WW/WC
- Invert
- Zoom
- Pan
- Stack Scroll
- Layout

As we can see in Figure 3, it shows the first task in our User Interface (UI), where the patient's breasts are on a small left column. The options are in a short row near of the viewport and described below. We also have the tabs where the user can change the patient. The center viewport shows the **DICOM** image, and it can be configured to display a number up to four **DICOM** images at the same time. The viewport has some text information on it (yellow) with the details of the metadata [35].

For the assistant, we provide the recommendations of our *bot-like* system with anthropomorphic characteristics [24]. This *bot-like* will give clinicians information regarding the patient's achieved severity of the breast (BI-RADS), and the respective interpretation by text. The interpretation is an analysis of the patient's co-variables and an explanation to clinicians concerning the lesion severities across each image. There is also the option to hide/show the assistant.

Finally, there are 3 buttons:

- Accept (green button) - accept the recommendation of the assistant
- Reject (red button) - reject the recommendation of the assistant that open a popup to input the clinician's opinion regarding BI-RADS diagnostic
- Explain (blue button) - highlight the lesions in the **DICOM** image and, for a reactive agent, appear the AI explanation



In this case, we have 3 findings. One finding are calcifications with diffuse distribution. The second finding are calcifications with regional distribution. The third finding is mass with lobulated shape, microlobulated margin and high density. There are no findings regarding personal history. There is no family history.

With an accuracy of 82% , the estimaed BIRADS for this patient is: 2

Figure 3: Viewer of the **DICOM** images.

6 Evaluation

In Table 1, it is presented the main Research Questions to address during evaluation. With these questions, we intend to understand the impact of assertiveness in the clinicians' decision-making, clinical workflow and user experience. Besides that, we want to see the impact of the assistant behavior on the clinicians' diagnosis decision.

Number	Research Questions
RQ1.	Should the assistant agent interact with all clinicians in the same way to improve their performance?
RQ2.	Will the perception of assertiveness-based assistance and different behaviours be the same for each professional category?
RQ3.	Adapting the assertiveness levels and assistant behaviour to the professional category will result into User Experience (UX) improvements?
RQ4.	How much influence will assistant behaviour have on the clinicians' decision-making?

Table 1: Research Evaluation Questions

Next, we address the hypotheses associated with each research question:

- **RQ1** Should the assistant agent interact with all clinicians in the same way to improve their performance?
  - **H1.1** Less experienced clinicians are performing better with reactive and more assertive assistants.
  - **H1.2** Higher experienced clinicians are performing better with proactive and low assertive assistants.
- **RQ2** Will the perception of behavior and assertiveness-based assistance be the same for each professional category?
  - **H2.1** Less experienced clinicians will prefer to interact with reactive and more assertive assistants.
  - **H2.2** Higher experienced clinicians will prefer to interact with proactive and less assertive assistants.
- **RQ3** Adapting the assertiveness levels and assistant behavior to the professional category will result into UX improvements?
  - **H3.1** The UX of assertive and reactive agent is better for less experienced clinicians.
  - **H3.2** The UX of non-assertive and proactive agent is better for higher experienced clinicians.
- **RQ4** How much influence will assistant behavior have on the clinicians' decision-making?
  - **H4.1** A proactive behavior will have more influence on the clinicians' decision-making.
  - **H4.2** A reactive behavior will have less influence on the clinicians' decision-making.

For the first question, enumerated as **RQ1.**, we want to see how clinicians with different categories of professional experience interact with different assistant behavior and levels of assertiveness. We will study if less experienced clinicians perform better with an assertive and reactive assistant (**H1.1.**) and if more experienced clinicians perform better with a non-assertive and proactive assistant (**H1.2.**).

The second question, enumerated as **RQ2.**, addresses the perception of clinicians with different categories of professional experience regarding the assistant assertiveness level and behavior. We hypothesize that a less experienced clinician will prefer an assertive and reactive agent (**H2.1.**), while a more experienced clinician will prefer a non-assertive and proactive agent (**H2.2.**).

In the third question, enumerated as **RQ3.**, we study improvements of user experience (UX). We predict that the UX will be better for less experienced clinicians with an assertive and reactive agent (**H3.1.**). While for the higher experienced clinicians, we predict a better UX experience with a non-assertive and proactive agent (**H3.2.**).

Finally, in the fourth question **RQ4.**, we'll see the influence of the assistant behavior on the clinician's decision-making. From the results of the focus groups we had, the hypothesis predict that a proactive behavior (**H4.1.**) will have more influence than a reactive behavior (**H4.2.**).

7 Tasks

In this section, we will describe the tasks that facilitators (Section 7.1) and participants (Section 7.2) have to follow in the third iteration (Section 3.3). It is important to notice that an assistant with reactive behavior has a slightly different set of tasks of an assistant with proactive behavior. The task descriptions below are required to be reviewed by all researchers and facilitators (Section 4) to ensure that the content, format, and presentation are representative of real final questionnaires and study. Their acceptance is to be documented prior to this study.

7.1 Tasks for Facilitators

Before starting the tasks, the facilitator should give a brief introduction of the purpose and objective of this work. They should talk about the "black-box" problem and the need for explainability. Then say that the agents that provide explanations need to gain trust and confidence of the clinicians, which can be gain through better communication (*i.e.*, assertiveness) and behaviour (i.e. proactiveness/reactiveness) of the agent. After this, proceed to the tasks.

The facilitator is responsible for providing the tasks and material to the participants. They must record the session with the authorization of the participant, as well as answer to questions the participants may have. It is also necessary to record the time that each participant to diagnostic each patient and write it on the list of scenarios.

List of standalone tasks for facilitators:

Task 7.1.1.1: Provide the consent form to clinicians;

Task 7.1.1.2: Provide the form for the user characterization;

Task 7.1.2.1: Open the list of patients;

Task 7.1.2.2: Open the list of scenarios;

Task 7.1.2.3: Open the DICOM server and deployed prototypes available at meta-private wiki;

Task 7.1.3.1: Provide post-task questionnaires;

7.2 Tasks for Clinicians

For the clinicians tasks, it is important to notice that there is a difference between a task with a proactive assistant and a reactive assistant. In the proactive assistant, the clinician makes one normal diagnostic, with the assistant on the screen. In the reactive assistant, the clinician makes two diagnostics: (1) before seeing the assistant recommendation; (2) after seeing the assistant recommendation. The clinician should be warned about this before going to **Task 7.2.2.1**.

List of standalone tasks for clinicians:

Task 7.2.1.1: Fill the consent form and accept voluntarily to participate in the study;

Task 7.2.1.2: Fill the form for the user characterization and proceed to the next steps of the study;

Task 7.2.2.1: Open the deployed prototype depending on the scenario order presented at list of scenarios;

Task 7.2.2.2: Open the patient (Figure 1) by following the order available at the list of patients;

Task 7.2.2.2.1: Interact with the interface by selecting the provided tools (Figure 2) available in our framework;

Task 7.2.2.2.2: Explore the patient;

Task 7.2.2.2.3: Classify the patient;

Task 7.2.2.4: Go back to **Task 7.2.2.2** if there are more patients or scenarios;

Task 7.2.3.1: Fill post-task questionnaires;

8 Metrics

Our user test metrics will address system accuracy, assertiveness perception and satisfaction, recall and precision, assistant influence on decision-making and diagnostic experience.

8.1 Patient Classification

For the patient classification, we will use the well known scale for classifying the breast cancer disease called BI-RADS [2]. The BI-RADS scale is a scheme for putting the findings from breast into a small number of well-defined categories [29].

The BI-RADS assessment categories are:

- 0 - Incomplete;
- 1 - Negative;
- 2 - Benign Findings;
- 3 - Probably Benign;
- 4 - Suspicious Abnormality;
- 5 - Highly Suspicious of Malignancy;
- 6 - Known Biopsy Proven Malignancy;

For each participant, we will ask the respective examination and respective BI-RADS value. We intend to study inter-variability and intra-variability i.e. how we vary assertiveness between patients and in the same patient, at different times. An example of inter-variability would be: Patient A diagnosed by Clinician X where the assistant is proactive and assertive and Patient B diagnosed by Clinician X where the assistant is reactive and non-assertive, all in the first set of tests. An example of intra-variability would be: Patient A diagnosed by Clinician X where the assistant is proactive and assertive and, one month later, the same Patient A is diagnosed by Clinician X where the assistant is reactive and non-assertive .We intend to study how these variables influence the clinicians' decision-making as well as the rates of false positives and false negatives.

8.2 Workload

To measure the workload, we used the NASA Task Load Index (NASA-TLX) [34] scale. The scale is a subjective workload assessment tool that will allow us to perform subjective workload assessments on our participants. For the purpose, it will be used a repository [7, 10] to cover this need of content.

By incorporating a multidimensional rating procedure, NASA-TLX derives an overall workload score based on a weighted average of ratings on six sub-scales:

- Mental Demand
- Physical Demand
- Temporal Demand
- Performance
- Effort
- Frustration

8.3 Usability

To measure the usability, we used the System Usability Scale (SUS) [30]. The SUS provides a "quick and dirty", reliable tool for measuring the usability. It consists of a 10 item questionnaire with ten response options for respondents; from *Strongly Agree* to *Strongly Disagree*. Originally created by John Brooke in 1986, it allows you to evaluate a wide variety of products and services, including hardware, software, mobile devices, websites and applications. For the purpose, it will be used a repository [8, 11] to cover this need of content.

When using SUS, participants are asked to score the following 10 items with one of ten responses that range from **Strongly Agree** to **Strongly Disagree**:

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

8.4 Trust

The Dimensions Of Trust Scale (DOTS) [9, 13] was introduced on a recent work [3, 4], introducing the concept of measuring trust across *AI* systems. For that, it will be used a repository [9] supporting our user tests. DOTS is a scale to measure the trustworthiness of our *AI-Assisted* system. Therefore, we created a three items list of questions on a 20-point scale of Likert-style [22].

The following list, represents the questions adapted from this model [25] addressing each of the three items, *i.e.*, *understanding*, *capability* and *benevolence* [3]:

1. I understand what the system is thinking. (**Understanding**)
2. The system seems capable. (**Capability**)
3. The system seems benevolent. (**Benevolence**)

8.5 Qualitative Evaluation

Qualitative and subjective evaluations regarding ease of use, satisfaction and design suggestions will be collected. This collection will be done via *open-ended questions* [1, 27], and during debriefing at the conclusion of the session.

The *open-ended questions* will utilize free-form responses and feedback, when possible. Whenever possible, it's best to ask *open-ended questions* so we can find out more than we can anticipate. We will test our questions by trying to answer them with short answers, and rewrite those to find out more about *how* and *what*. In some cases, we won't be able to accommodate free-form or write-in answers, though, and then it is necessary to limit the possibilities.

8.6 Comparison between prototypes

Clinicians will compare assertive assistant with non-assertive assistant as well as proactive assistant with reactive assistant. For this, they will rate perceived reliability, trustworthiness, capability and overall preference on a 7-point Likert Scale. In fact, there will be two 7-point Likert Scale: one for assertiveness and one for behaviour. These scales will range from 1 ("totally assertive assistant" or "totally proactive assistant"), 2 ("much more assertive than non-assertive" or "much more proactive than reactive"), 3 ("slightly more assertive than non-assertive" or "slightly more proactive than reactive"), 4 ("neutral" for both), etc. to 7 ("totally non-assertive assistant" or "totally reactive assistant") [36].

The questions we seek to answer are the following:

- Which assistant was more reliable?
- Which assistant was more capable?
- Which assistant did you prefer overall?

8.7 Time Completion

We will measure the Time on Task (ToT) [15, 21]. Specifically, by recording the time each clinician spends doing the proposed task, excluding qualitative and subjective evaluations.

8.8 Critical Errors

Critical Errors are deviations at completion from the targets of the scenario. Obtaining or otherwise reporting of the wrong data value due to participant workflow is a Critical Error. Participants may or may not be aware that the task goal is incorrect or incomplete.

An example of a Critical Error, could be a situation where the participant cannot open a patient. From this error, we cannot even proceed to the next tasks and complete the user test. Despite the independent completion of the scenario is the goal, we need to guarantee the execution of the test, however, when these errors occur, the facilitator must act.

Critical Errors can also be assigned when the participant initiates, or attempts to initiate, an action that will result in the goal state becoming unobtainable. In general, Critical Errors are unresolved errors preventing completion of the task or errors that produce an incorrect outcome.

8.9 Non-Critical Errors

Non-Critical Errors, are errors that are recovered from and by the participant. Or, if not detected, do not result in processing problems or unexpected results. Although Non-Critical Errors can be undetected by the participant, when they are detected they are generally frustrating to the participant.

The non-critical errors may be procedural, in which the participant does not complete a scenario in the most optimal means (*e.g.*, excessive steps and keystrokes). These errors may also be errors of confusion (*e.g.*, initially selecting the wrong function, using a UI control incorrectly, such as attempting to edit an unsuitable field). Non-Critical Errors can always be recovered from during the process of completing the scenario. Exploratory behavior, such as opening the wrong menu while searching for a function, will be coded as a non-critical error.

9 Goals

In this section, we will describe the goals that we intend to achieve. We will assess performance and perception-related metrics, as well as qualitative and error-related metrics.

9.1 Completion Rate

Completion Rate is the percentage of test participants who successfully complete the task without critical errors. A critical error is defined as an error that results in an incorrect or incomplete outcome. In other words, the completion rate represents the percentage of participants who, when they are finished with the specified task, have an ”output” that is correct.

*A **Completion Rate** of **90%** is the goal for each task in this usability test.*

Note: If a participant requires assistance in order to achieve a correct output then the task will be scored as a critical error and the overall completion rate for the task will be affected.

9.2 Error-Free Rate

Error-Free Rate is the percentage of test participants who complete the task without any errors (critical or non-critical errors). A non-critical error is an error that would not have an impact on the final output of the task but would result in the task being completed less efficiently.

*An **Error-Free Rate** of **80%** is the goal for each task in these tests.*

9.3 Subjective Measures

Subjective opinions about specific tasks, time to perform each task, features, and functionality will be surveyed. At the end of the test, participants will rate their satisfaction and measure the workload, usability and trust of the overall system by answering open-ended questions and three simple questionnaires (NASA-TLX, SUS and DOTS). Combined with the interview/debriefing session, these data are used to assess attitudes of the participants.

9.4 Hypothesis Approval

The main goal is to confirm the hypothesis presented in the research questions. We intend to evaluate:

- Performance - measuring ratio of false positives and false negatives using BI-RADS as well as time completion of each task.
- Preference and UX - comparing both levels of assertiveness and behaviours as well as measuring workload, usability and trust (NASA-TLX, SUS, DOTS, respectively
- Influence on decision-making - studying intervariability and intravariability

The goal is to study and confirm the influence of these assistant properties. In the end, we aim to answer how these assistant properties are influencing the different categories of clinicians when diagnosing breast cancer.

10 Challenges

Besides the challenges already mentioned in this document, we may have other challenges with participants. The different experience/knowledge of the participants will constrain the communication with the participant on the test, as well as scheduling meetings with participants due to their busy work. Moreover, due to the COVID-19 pandemic, our tests will be held in remote sessions to avoid propagation of the pandemic.

11 Results

At the end of these tests, a master thesis will be elaborated over the achievements of this guide. It will consist of a report and a presentation of the results; evaluation of the metrics against the pre-approved goals, subjective evaluations, and suggestions for future work. We will publish and link all information concerning the results, used prototypes, as well as datasets in our `sa-uta11-results` repository (github.com/MIMBCD-UI/sa-uta11-results) with the respective documentation and wiki.

For more information, please follow our scientific contributions at:

researchgate.net/project/Medical-Imaging-Multimodality-Breast-Cancer-Diagnosis-User-Interface

12 Statements

In this section, we provide some statements of this document. Our statements inform the reader with information concerning a document disclaimer, the credits from direct and indirect document contributors, as well as acknowledgements for the recognition of all involved researchers at this stage. Bellow, the document is providing the necessary information to understand the study context under this research.

12.1 Disclaimer

The purpose of this document is to provide information concerning the steps to follow during our user testing and analysis. This document summarizes the process to guide facilitators of the work into the user testing process and future analysis of the results. The document does not reflect the views of the author(s) nor is an evident report of scientific achievements. This document was not peer reviewed by external reviewers. Instead, the research team self reviewed the document within several team elements and iterations.

12.2 Credits

For purposes of determining the work credits from direct and indirect contributors, the following information is describing and recognizing all involved to reduce authorship disputes and facilitating collaboration. The corresponding author and first contributor, namely João Fernandes, is responsible for the prototype development, methodology, formal analysis, and writing the original draft of this document. The second contributor is Prof. Jacinto Nascimento, who supervised the first contributor and corresponding author of this document. Moreover, the second contributor is also responsible for research administration, data, and funding acquisition. The third contributor, namely Francisco Maria Calisto, is responsible for conceptualization and validation of the study, investigation of the experiments, co-supervision of the first contributor with the second contributor, and writing review & editing, as well as final data curation. The fourth contributor, is Prof. Nuno Nunes and is the formal arguer of the work developed under the master thesis of the first contributor. Additionally, the fourth contributor is also responsible for the provision of computing resources, as well as critical review, commentary, and revision of writing. Finally, the fifth contributor, namely Prof. Carlos Martinho, is partially responsible for the conceptualization and formulation of the idea under this research.

12.3 Funding

This work was supported in various funding sources and collaborations. The work was partially supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013 and Instituto Superior Técnico (IST-ID) through the FCT/UID/EEA/50009/2013 project reference, BL89/2017-IST-ID, and PD/BD/150629/2020 grants. We would like to convey Hospital Fernando Fonseca for the collaboration.

12.4 Acknowledgements

A special thanks for the support provided by Carlos Santiago, Margarida Parrado Morais and Miguel Bastos. An important recognition goes to Dr. João Maria Abrantes for the participation and data curation. We would like to thank Dr. Clara Aleluia, Dr. Gisela Andrade, Dr. Willian Schmitt, Dr. Ana Sofia Germano and Dr. Pedro Marques from the Hospital Fernando Fonseca for the generous support and medical expertise. Also, an immense thank to Dr. Cristina Ribeiro da Fonseca for the participation in this work. Our appreciation goes also to Bruno Oliveira, Luis Ribeiro Gomes, Bruno Cardoso and Bruno Dias for help, and above all for the good companionship. Thanks to Prof. Daniel Gonçalves, Prof. Daniel Simões Lopes and Daniel Mendes for the technical inputs and network. Last but not least, thanks to Prof. Jacinto C. Nascimento and Prof. Nuno Jardim Nunes for advising this work. We also want to provide a special acknowledgment to Prof. Ramtin Zargari Marandi who, among others, gave us important information and comments regarding the presented report.

Acronyms

AI Artificial Intelligence.

BI-RADS Breast Imaging Reporting and Data System.

CC CranioCaudal.

CFA Confirmatory Factor Analysis.

DICOM Digital Imaging and Communications in Medicine.

DOTS Dimensions Of Trust Scale.

HAII Human-AI Interaction.

ID IDentifier.

MG MammoGraphy.

MLE MaximumLikelihood Estimation.

MLO MedioLateral Oblique.

MM Multi-Modality.

MRI Magnetic Resonance Imaging.

NASA-TLX NASA Task Load Index.

SEM Structural Equation Modeling.

SS Single-Modality.

SUS System Usability Scale.

ToT Time on Task.

UI User Interface.

US UltraSound.

UTA User Testing and Analysis.

UTAUT Unified Theory of Acceptance and Use of Technology.

UX User eXperience.

References

[1] Julia Abelson, Kathy Li, Geoff Wilson, Kristin Shields, Colleen Schneider, and Sarah Boesveld. Supporting quality public and patient engagement in health system organizations: development and usability testing of the p ublic and p atient e ngagement e valuation t ool. *Health Expectations*, 19(4):817–827, 2016.

[2] Corinne Balleyguier, Salma Ayadi, Kim Van Nguyen, Daniel Vanel, Clarisse Dromain, and Robert Sigal. Birads™ classification in mammography. *European journal of radiology*, 61(2):192–194, 2007.

[3] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th international conference on intelligent user interfaces*, pages 258–262, 2019.

[4] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

[5] Francisco Calisto. Medical imaging multimodality breast cancer diagnosis user interface. Master's thesis, Instituto Superior Técnico, Avenida Rovisco Pais 1, 1049-001 Lisboa - Portugal (EU), 10 2017. A Medical Imaging Tool for a Multimodality use of Breast Cancer Diagnosis on a User Interface.

[6] Francisco M. Calisto, Alfredo Ferreira, Jacinto C. Nascimento, and Daniel Gonçalves. Towards touch-based medical image diagnosis annotation. In *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces*, ISS '17, page 390–395, New York, NY, USA, 2017. Association for Computing Machinery.

[7] Francisco Maria Calisto. Mimbcd-ui/nasa-tlx: v1.0.0-alpha, September 2018.

[8] Francisco Maria Calisto. Mimbcd-ui/sus: v1.0.0-alpha, September 2018.

[9] Francisco Maria Calisto. Dimensions Of Trust Scale (DOTS) LaTeX Version: v1.0.2-alpha, May 2019. See: https://github.com/mida-project/dots.

[10] Francisco Maria Calisto and Jacinto C. Nascimento. Nasa-tlx survey, 2018.

[11] Francisco Maria Calisto and Jacinto C. Nascimento. Sus survey, 2018.

[12] Francisco Maria Calisto, Nuno Nunes, and Jacinto C. Nascimento. Breastscreening: On the use of multi-modality in medical imaging diagnosis. In *Proceedings of the International Conference on Advanced Visual Interfaces*, AVI '20, New York, NY, USA, 2020. Association for Computing Machinery.

[13] Francisco Maria Calisto, Nuno Jardim Nunes, and Jacinto C. Nascimento. Medical imaging diagnosis assistant: Dimensions of trust scale (dots) survey template file, 2019.

[14] Francisco Maria Calisto, Carlos Santiago, Nuno Nunes, and Jacinto C. Nascimento. Introduction of human-centric ai assistant to aid radiologists for multimodal breast image classification. *International Journal of Human-Computer Studies*, 150:102607, 2021.

[15] Cheryl Delgado and Linda Wolf. Time on task: perceived and measured time in online courses for students and faculty. *Journal of Nursing Education and Practice*, 7(5):27–32, 2017.

[16] João Fernandes. Extended abstract: Towards assertiveness-based interactive agents for breast cancer diagnosis. Technical report, Instituto Superior Técnico, Avenida Rovisco Pais 1, 1049-001 Lisboa - Portugal (EU), 1 2022.

[17] João Fernandes. Master project: Towards assertiveness-based interactive agents for breast cancer diagnosis. Technical report, Instituto Superior Técnico, Avenida Rovisco Pais 1, 1049-001 Lisboa - Portugal (EU), 1 2022.

[18] João Fernandes. Towards assertiveness-based interactive agents with different behaviours for breast cancer diagnosis. Master's thesis, Instituto Superior Técnico, Avenida Rovisco Pais 1, 1049-001 Lisboa - Portugal (EU), 11 2022.

[19] João Fernandes and Francisco Maria Calisto. Mimbcd-ui uta11 - qualitative data - transcript, 05 2022.

[20] Renato Hermoza, Gabriel Maicas, Jacinto C. Nascimento, and Gustavo Carneiro. Post-hoc overall survival time prediction from brain mri. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1476–1480, April 2021.

[21] Jue Huang, Christine Ulke, Christian Sander, Philippe Jawinski, Janek Spada, Ulrich Hegerl, and Tilman Hensch. Impact of brain arousal and time-on-task on autonomic nervous system activity in the wake-sleep transition. *BMC neuroscience*, 19(1):1–11, 2018.

[22] Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. Likert scale: Explored and explained. *British Journal of Applied Science & Technology*, 7(4):396, 2015.

[23] Hugo Lencastre, Francisco Maria Calisto, and Jacinto Nascimento. Breast cancer multimodality scalable interactions, 01 2021.

[24] Kaifeng Liu and Da Tao. The roles of trust, personalization, loss of privacy, and anthropomorphism in public acceptance of smart healthcare services. *Computers in Human Behavior*, 127:107026, 2022.

[25] Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734, 1995.

[26] Daniela Medley, Carlos Santiago, and Jacinto C. Nascimento. Cycoseg: A cyclic collaborative framework for automated medical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 9 2021.

[27] Rajan Merchant, Rubina Inamdar, Kelly Henderson, Meredith Barrett, Jason G Su, Jesika Riley, David Van Sickle, David Stempel, et al. Digital health intervention for asthma: patient-reported value and usability. *JMIR mHealth and uHealth*, 6(6):e7362, 2018.

[28] Nádia Mourão and Jacinto Nascimento. Master thesis: 2d breast cancer diagnosis explainable visualizations, 01 2021.

[29] S Obenauer, KP Hermann, and E Grabbe. Applications and literature review of the bi-rads classification. *European radiology*, 15(5):1027–1036, 2005.

[30] Konstantina Orfanou, Nikolaos Tselios, and Christos Katsanos. Perceived usability evaluation of learning management systems: Empirical evaluation of the system usability scale. *The International Review of Research in Open and Distributed Learning*, 16(2):227–246, 2015.

[31] António C Pacheco and Carlos Martinho. Alignment of player and non-player character assertiveness levels. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 15, pages 181–187, 2019.

[32] Raul Paradeda, Maria José Ferreira, Raquel Oliveira, Carlos Martinho, and Ana Paiva. The role of assertiveness in a storytelling game with persuasive robotic non-player characters. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, pages 453–465, 2019.

[33] Corina Pelau, Dan-Cristian Dabija, and Irina Ene. What makes an ai device human-like? the role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Computers in Human Behavior*, 122:106855, 2021.

[34] Anjana Ramkumar, Pieter Jan Stappers, Wiro J Niessen, Sonja Adebahr, Tanja Schimek-Jasch, Ursula Nestle, and Yu Song. Using goms and nasa-tlx to evaluate human–computer interaction process in interactive segmentation. *International Journal of Human–Computer Interaction*, 33(2):123–134, 2017.

[35] Young June Sah and Wei Peng. Effects of visual and linguistic anthropomorphic cues on social perception, self-awareness, and information disclosure in a health website. *Computers in Human Behavior*, 45:392–401, 2015.

[36] Mike Schaekermann, Graeme Beaton, Elaheh Sanoubari, Andrew Lim, Kate Larson, and Edith Law. Ambiguity-aware ai assistants for medical data analysis. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14, 2020.

[37] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44, 2019.

[38] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang'Anthony' Chen. Chexplain: Enabling physicians to explore and understand data-driven, ai-enabled medical imaging analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.