# MIDA
# User Testing & Analysis Guide
# Clinical Opinion About Intelligent Agents

Francisco Maria Calisto
francisco.calisto@tecnico.ulisboa.pt
Instituto Superior Técnico
University of Lisbon
Portugal

21/11/2019

| | |
|---|---|
| **BreastScreening:** | breastscreening.github.io |
| **Meta:** | github.com/BreastScreening/meta |
| **Datasets:** | github.com/BreastScreening/meta/wiki/Datasets |
| | |
| **MIDA:** | mida-project.github.io |
| **Meta:** | github.com/mida-project/meta |
| **Datasets:** | github.com/mida-project/meta/wiki/Datasets |
| | |
| **MIMBCD-UI:** | mimbcd-ui.github.io |
| **Meta:** | github.com/MIMBCD-UI/meta |
| **Datasets:** | github.com/MIMBCD-UI/meta/wiki/Datasets |

# 1 Introduction

During the breast cancer screening, missing cancers may not be identified until they are more advanced and less agreeable to treatment [19]. Artificial Intelligence (AI) in the medical workflows may help with this challenge [26]. Studies have demonstrated the ability of AI to meet the human's performance on various clinical tasks [39, 40]. As a lack of medical professionals threatens the adequacy and availability of clinical services worldwide [27, 38], the scalability of AI could improve to higher care.

The role of Human-AI Interaction (HAII) in healthcare delivery the appropriate settings in which it can be applied, and its impact on the quality of care have yet to be evaluated [41]. There have been several attempts at addressing the effects of HAII across multiple workflows and different levels of clinical expertise [16, 43]. However, the use case of breast cancer diagnosis to address the effects from varied representations of AI-based supported by intelligent agents is still scarce. This explains why it is an open topic research, and the motivation behind the proposed research of this User Testing and Analysis (UTA) guide.

# 2 Description

We took impressions from another domain study [21, 22], where we applied several Human-Computer Interaction (HCI) techniques to extract important user needs and translate those needs into system requirements [1, 4, 10, 36]. In this study, we follow the same logic to assess the user needs concerning the adoption of AI in a clinical environment. Additionally, our research work involves the development of two categories of diagnostic systems that are different in terms of requirements. First of all, an annotating system [2, 5, 11] is providing clinicians the ability to label [29] medical images and, consequently, generate various *datasets* [6] that will be consumed by the AI models. Second, the AI models will be trained by consuming these *datasets* and provide recommendation to clinicians as a second reader or as an autonomous patient diagnostic [9, 12, 30].

Our research work deals with several challenges [8, 23, 25, 28], where the goal is to address and surpass these challenges through an HCI approach. We follow an human-centered perspective to understand the user needs and improve our novel systems through interaction. Nevertheless, it is important to provide contextualization [7] of the work done until now and the future directions that will be introduced in this document.

Although the current work under this research has already studied the preliminary acceptance and trust of AI systems [13], more work should be done. To strengthen the research evidence for such claims, further work must measure a more detailed technology adoption in the context of medical imaging diagnosis. In this work, a continuing quest to ensure clinicians' acceptance of AI is an ongoing clinical challenge [24, 31, 32, 33, 34]. However, this challenge has occupied researchers to such an extent that AI adoption in the clinical domain is now considered an opportunity.

In this UTA, we aim to demographically assess the main characteristics and user profiles of the medical imaging community. Additionally, we will address the community acceptance to the AI topic so that we can understand the potential adoption of AI in the clinical workflow. As a demographic and domain study, this UTA is the 8th (UTA8) reporting guide of our research work and is an iteration [35] from the previous (UTA7) reporting guide [3]. The previous iteration (UTA7), titled as "Assistant Introduction: User Testing Guide For A Comparison Between Multi-Modality and AI-Assisted Systems", guided us through the introduction of an AI-assisted system in the clinical workflow. However, we did not properly study the demographic characteristics that influence the adoption of AI in medical practice. Hence, we will study the existent technology acceptance theories, adapting the Unified Theory of Acceptance and Use of Technology (UTAUT) to develop a model to evaluate the acceptance of AI in the clinical workflow.

# 3 Methodology

A substantial level of activity has witnessed the use of a wide range of exploratory techniques. In fact, these exploratory techniques are examining many different systems in countless different contexts, to the extent that even the most cursory examination will reveal a variety of user perspectives, contexts, analysis, theories, and research methods [44]. Such situation has in turn led to an element of confusion, as it is often current to be forced of picking specific characteristics across a wide variety of models and theories. In response to this confusion, and in order to harmonize the literature associated with acceptance of new systems, Venkatesh et al. [42] developed a unified model – created and studied by these authors as UTAUT – that brings together alternative views on user and innovation acceptance.

As a reporting guide, the document proposes the application of a model based on the UTAUT. In this work, we are using this model as the constructs to study the determinants for adoption of AI systems in medical imaging diagnosis. The idea is to test the model via Confirmatory Factor Analysis (CFA) and Structural Equation Modeling (SEM) while using clinicians' responses (expected $n > 300$ clinicians) to a formulated UTAUT questionnaire.

Future results will show how an increased understanding of a vital role of safety, security, privacy, and trust in usage intention of intelligent agents in these medical fields. It is expected that such results will show to this research how improvements of the workflow performance for a clinical AI system is a strong predictor of adoption, while medical professional experience (*i.e.*, Interns, Juniors, Middles and Seniors) and medical specialities (*e.g.*, Radiologists, Surgeons, Dermatologists, Neurologists, etc) are essential moderators of behavioral intention. The future empirical findings will provide valuable theoretical contributions to HCI and AI researchers concerning the design and implementation of intelligent agents by explaining the reasons behind adoption and usage of AI systems in the clinical workflow.

# 4 Roles

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

## 4.1 Facilitator

- Provides overview of the study to participants;
- Responds to participant's requests for information;

## 4.2 Ethics

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

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

# 5 Apparatus

Questionnaires and data collection will be performed through the Google Forms platform. The group of questionnaires will be sent via e-mail and social networking platforms for an expected $n > 300$ number of participants. For this study, we used e-mail to spread our questionnaires, but also used Facebook groups, Reddit and LinkedIn. To support us on the LinkedIn social network, we will use LinkedHelper2, a powerful LinkedIn automation software. The LinkedHelper2 tool is providing us an automatically invite targeted level of contacts with a personal note. With that functionality, we can spread easily our questionnaires thanks to the auto-responder messaging system of this tool.

# 6 Evaluation

The study followed the required ethical standards. Specifically, the study complies with the provisions of the General Data Protection Regulation [Regulation (EU) 2016/279 of the European Parliament and of the Council of 27 April 2016], and follows the recommendations of the Declaration of Helsinki for research. There are no risks and benefits associated with the users' participation. User's participation is completely voluntary.

Quantitative and qualitative studies will be conducted with clinical experts and healthcare professionals. Analysis will be performed by descriptive categorization statistics of statements. In this document, results (Section 11) are being briefly described, but the evaluation process is further reported.

The raw data will be analyzed as follows. In accordance with the recommendations of the two-stage procedure [37], CFA will be used to test validity and reliability of the model [15]. SEM will be used as a preferable technique to regression as it allows simultaneous analysis of all relationships through multiple regression, while also allowing for both observed and latent variables to be analyzed at the same time, and providing overall fit statistics [17].

# 7 Tasks

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.

List of stand alone tasks:

**Task 1.1:** Read carefully the study description;

**Task 1.2:** Fill the consent form section and accept voluntarily to participate in the study;

**Task 2.1:** Fill the user characterization section and proceed;

**Task 2.2:** Fill the medical experience section and proceed;

**Task 3.1:** Fill the experience with AI systems section and proceed;

**Task 3.2:** Fill the AI systems section and proceed;

**Task 4.1:** Fill the conclusions section and submit;

# 8 Metrics

Herein, we outline the theoretical analysis, the methodology employed to create the metrics, and the benefits that can be obtained. In order to define the metrics, we need to generate a consensus from many perspectives as to what was important to measure and how the measures should be calculated. On an ongoing basis, it is envisioned that these metrics would evolve and become much more comprehensive and complex; however, it is critical that the early-stage metrics be meaningful and feasibly generated from data that were clear, concise, and accessible. In order to begin this process of measurement, reporting, and analysis with as much consensus as possible, our research team convened to work on identifying the first metrics. One interesting debate focuses around whether safety, security, privacy, and trust indicators (Section 3) should be just that, an indicator, or an all-inclusive calculation, similar to the technology acceptance and adoption items.

In this research, we utilized the unprecedented opportunity presented by the need of technology acceptance at an international scale to better understand the factors affecting the adoption and use of AI systems in clinical environments. We are particularly interested in investigating the role of safety, security, privacy, and trust indicators in the adoption context of AI assistants that supported clinicians on medical imaging diagnosis. We also want to understand the effect of moderator variables (gender, age, education, and clinical experience) to relate with the achieved metrics of technology acceptance.

# 9 Goals

The research goals of this work are: (i) to investigate the effects of AI on technology adoption; (ii) to increase our understanding of differences in the determinants of technology use; and (iii) to improve the explanatory power and predictive accuracy of parsimony questionary based on known UTAUT constructs for broader application in HCI research. For testing the hypothesis the questionnaire will comprise several questions (items) for responses on a Likert-type scale, ranging from 1-"Strongly disagree", 2-"Disagree", 3-"Undecided", 4-"Agree" and 5-"Strongly agree". To ensure the content validity of the questionnaire used to assess each construct, all items regarding the measurement of constructs were adapted from previous studies and carefully reworded to fit the context of AI systems which can ge generalized for diagnostic systems.

# 10 Challenges

In addition to the challenges already highlighted in the presented document, we must accomplish the participation issues. The difference in knowledge and expertise levels between the participants will inhibit communication and participation of participants in different ways. Moreover, the factor that posed challenges to participants are involving them to a nominal adoption of consequences in the perceptions and practice, related ethical and self conflicts in presence of results. Challenges are presented through this document and for practitioners to improve both study and research.

As in any large-scale measurement and evaluation effort, designing and validating the measures will be one of the most important and difficult challenges to overcome. This document should be a stimulus to re-examine how we approach existing challenges and study some aspects of human behavior, such as clinicians' relationship with AI assistance and its role during medical imaging diagnosis, for instance, in breast cancer disease. Against the lack of AI acceptance, this document provided the first detailed research study on the adoption of AI assistants, designed to mitigate the medical error on world-wide clinical institutions. While we expect that some of our findings will not generalize beyond diagnostic systems, others provide early insight into the increasingly important role of safety, security, privacy, and trust in AI adoption and usage.

## 11 Results

For analysis of the collected data, absolute and percentage values of the distribution will be obtained, and the results will be presented as descriptive statistics. CFA will be conducted in Python using MaximumLikelihood Estimation (MLE) [14]. This approach will be followed by path analysis of the structural relationships that are also expected to be conducted in Python with SEM libraries [20]. Moderation analysis will also be undertaken in Python [18]. All results and statistical analysis will be published and made public as curated *datasets* in our MIMBCD-UI organization on GitHub. A list of available *datasets* is already public and online (github.com/MIMBCD-UI/meta/wiki/Datasets).

## 12 Acknowledgements

A special thanks for the support and revisions provided by Hugo Lencastre and Nádia Mourão. We would like to thank Doctor Clara Aleluia, Doctor Gisela Andrade, Dr. Willian Schmitt, Dr. Ana Sofia Germano and Dr. Pedro Marques from the HFF for the generous support and medical expertise. Also, an immense thank for Doctor Cristina Ribeiro da Fonseca. My appreciation goes also to Bruno Cardoso and Bruno Dias for help and above all for the good companionship. Thanks to Professor Daniel Gonçalves, Professor Daniel Simões Lopes and Daniel Mendes for the technical inputs and network. Last but not least, thank to my advisors Professor Jacinto C. Nascimento and Professor Nuno Jardim Nunes. We also want to provide a special acknowledgment to Professor Ramtin Zargari Marandi who, among others, gave us important information and comments regarding the presented report. This 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, BL89/2017-IST-ID grant. We would like to convey Hospital Fernando Fonseca (HFF) for the collaboration.

## 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.

**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.

**UI** User Interface.

**US** UltraSound.

**UTA** User Testing and Analysis.

**UTAUT** Unified Theory of Acceptance and Use of Technology.

# References


[1] 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), October 2017. A Medical Imaging Tool for a Multimodality use of Breast Cancer Diagnosis on a User Interface.

[2] 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.

[3] Francisco Maria Calisto. Assistant introduction: User testing guide for a comparison between multi-modality and ai-assisted systems. Technical report, Instituto Superior Técnico, 2019.

[4] Francisco Maria Calisto. It-medex closing workshop: Towards touch-based medical image diagnosis annotation, 2019.

[5] Francisco Maria Calisto. Breast cancer medical imaging multimodality lesion contours annotating method. Technical Report 116801, Instituto Superior Técnico, Avenida Rovisco Pais 1, 1049-001 Lisboa - Portugal (EU), October 2020. Method and process using a system to annotate and visualize masses and microcalcifications of breast cancer lesions in a multimodality strategy.

[6] Francisco Maria Calisto. Medical imaging multimodality annotating framework. In *PhD Open Days 2020*, POD '20, pages 1–2. Instituto Superior Técnico, October 2020.

[7] Francisco Maria Calisto. Towards the human-centered design of intelligent agents in medical imaging diagnosis thesis proposal problems and contributions, November 2020.

[8] Francisco Maria Calisto, Hugo Lencastre, Nuno Jardim Nunes, and Jacinto C. Nascimento. Breastscreening: Towards breast cancer clinical decision support systems. In *National Science Summit 2019*, NSS '19, pages 1–2. Fundação para a Ciência e Tecnologia, July 2019.

[9] Francisco Maria Calisto, Hugo Lencastre, Nuno Jardim Nunes, and Jacinto C. Nascimento. Medical imaging diagnosis assistant: Ai-assisted radiomics framework user validation. In *Keep In Touch 2019*, KIT '19, pages 1–2, Avenida Rovisco Pais 1, 1049-001 Lisboa - Portugal (EU), 2019. Instituto Superior Técnico.

[10] Francisco Maria Calisto, Pedro Miraldo, Nuno Jardim Nunes, and Jacinto C. Nascimento. Breastscreening: A multimodality diagnostic assistant.

In *LARSyS 2018 Annual Meeting*, LARSyS AM '18, pages 1–2. Interactive Technologies Institute, June 2018.

[11] 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.

[12] Francisco Maria Calisto, Nuno Jardim Nunes, Jacinto C. Nascimento, and Pedro Miraldo. Breastscreening: A multimodality diagnostic assistant. In *National Science Summit 2018*, NSS '18, pages 1–2. Fundação para a Ciência e Tecnologia, June 2018.

[13] 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.

[14] Heining Cham, Evgeniya Reshetnyak, Barry Rosenfeld, and William Breitbart. Full information maximum likelihood estimation for latent variable interactions with incomplete indicators. *Multivariate behavioral research*, 52(1):12–30, 2017.

[15] Marcus Crede and Peter Harms. Questionable research practices when using confirmatory factor analysis. *Journal of Managerial Psychology*, 2019.

[16] Krzysztof J. Geras, Ritse M. Mann, and Linda Moy. Artificial intelligence for mammography and digital breast tomosynthesis: Current concepts and future perspectives. *Radiology*, 293(2):246–259, 2019. PMID: 31549948.

[17] Joseph F Hair Jr, Marko Sarstedt, Christian M Ringle, and Siegfried P Gudergan. *Advanced issues in partial least squares structural equation modeling*. saGe publications, 2017.

[18] Andrew F Hayes and Nicholas J Rockwood. Regression-based statistical mediation and moderation analysis in clinical research: Observations, recommendations, and implementation. *Behaviour research and therapy*, 98:39–57, 2017.

[19] Nehmat Houssami and Kylie Hunter. The epidemiology, radiology and biological characteristics of interval breast cancers in population mammography screening. *npj Breast Cancer*, 3(1):12, April 2017.

[20] Anna A Igolkina and Georgy Meshcheryakov. semopy: A python package for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, pages 1–12, 2020.

[21] Hugo Lencastre and Francisco Maria Calisto. Feedbot user testing guide daily task. Technical report, Instituto Superior Técnico, April 2019.

[22] Hugo Lencastre and Francisco Maria Calisto. Feedbot user testing guide user characterization. Technical report, Instituto Superior Técnico, March 2019.

[23] Hugo Lencastre, Francisco Maria Calisto, and Jacinto C. Nascimento. 3d module view feature, January 2020.

[24] Hugo Lencastre, Francisco Maria Calisto, and Jacinto C. Nascimento. Coordinated view feature, January 2020.

[25] Hugo Lencastre, Francisco Maria Calisto, and Jacinto C. Nascimento. Recorded view feature, January 2020.

[26] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S. Corrado, Ara Darzi, Mozziyar Etemadi, Florencia Garcia-Vicente, Fiona J. Gilbert, Mark Halling-Brown, Demis Hassabis, Sunny Jansen, Alan Karthikesalingam, Christopher J. Kelly, Dominic King, Joseph R. Ledsam, David Melnick, Hormuz Mostofi, Lily Peng, Joshua Jay Reicher, Bernardino Romera-Paredes, Richard Sidebottom, Mustafa Suleyman, Daniel Tse, Kenneth C. Young, Jeffrey De Fauw, and Shravya Shetty. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89–94, January 2020.

[27] S Moran and H Warren-Forward. The australian breastscreen workforce: a snapshot. *Radiographer*, 59(1):26–30, 2012.

[28] Nádia Mourão, Francisco Maria Calisto, and Jacinto C. Nascimento. Mimbcd-ui: Ai visual explaination - lesions types, January 2020.

[29] Nádia Mourão, Francisco Maria Calisto, and Jacinto C. Nascimento. Mimbcd-ui: Ai visual explanation - label lesion, January 2020.

[30] Nádia Mourão, Francisco Maria Calisto, and Jacinto C. Nascimento. Mimbcd-ui: Ai visual explanation - uml, January 2020.

[31] Nádia Mourão, Hugo Lencastre, and Francisco Maria Calisto. Mimbcd-ui uta10 - focus group - transcript, June 2020.

[32] Nádia Mourão, Hugo Lencastre, and Francisco Maria Calisto. Mimbcd-ui uta11 - focus group - transcript, June 2020.

[33] Nádia Mourão, Hugo Lencastre, and Francisco Maria Calisto. Mimbcd-ui uta8 - focus group - transcript, June 2020.

[34] Nádia Mourão, Hugo Lencastre, and Francisco Maria Calisto. Mimbcd-ui uta9 - focus group - transcript, June 2020.

[35] Nádia Mourão, Hugo Lencastre, Francisco Maria Calisto, and Jacinto C. Nascimento. User testing architecture tree, March 2020.

[36] Bruno Oliveira, Francisco Maria Calisto, José Borbinha, and Lilian Gomes. Adaptive q-sort matrix generation: A simplified approach. Technical report, INESC-ID, 2015.

[37] S Rahi, M Ghani, F Alnaser, and A Ngah. Investigating the role of unified theory of acceptance and use of technology (utaut) in internet banking adoption context. *Management Science Letters*, 8(3):173–186, 2018.

[38] Abi Rimmer. Radiologist shortage leaves patient care at risk, warns royal college. *BMJ: British Medical Journal (Online)*, 359, 2017.

[39] Jiayi Shen, Casper J P Zhang, Bangsheng Jiang, Jiebin Chen, Jian Song, Zherui Liu, Zonglin He, Sum Yi Wong, Po-Han Fang, and Wai-Kit Ming. Artificial intelligence versus clinicians in disease diagnosis: Systematic review. *JMIR Med Inform*, 7(3):e10010, August 2019.

[40] Eric J. Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, January 2019.

[41] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, John Paoli, Susana Puig, Cliff Rosendahl, H. Peter Soyer, Iris Zalaudek, and Harald Kittler. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234, August 2020.

[42] Viswanath Venkatesh, James YL Thong, and Xin Xu. Unified theory of acceptance and use of technology: A synthesis and the road ahead. *Journal of the association for Information Systems*, 17(5):328–376, 2016.

[43] Walter F. Wiggins, M. Travis Caton, Kirti Magudia, Sha-har A. Glomski, Elizabeth George, Michael H. Rosenthal, Glenn C. Gaviola, and Katherine P. Andriole. Preparing radiologists to lead in the era of artificial intelligence: Designing and implementing a focused data science pathway for senior radiology residents. *Radiology: Artificial Intelligence*, 0(ja):e200057, 0.

[44] Michael D Williams, Nripendra P Rana, and Yogesh K Dwivedi. The unified theory of acceptance and use of technology (utaut): a literature review. *Journal of enterprise information management*, 2015.