TÉCNICO LISBOA

Breast Cancer Multimodality Scalable Interactions

Hugo Emanuel Pereira Menezes e Lencastre

Thesis to obtain the Master of Science Degree in

Information Systems and Computer Engineering

Supervisor: Professor Jacinto Carlos Marques Peixoto do Nascimento

Examination Committee

Chairperson: Professor Alberto Manuel Rodrigues da Silva
Supervisor: Professor Jacinto Carlos Marques Peixoto do Nascimento
Member of the Committee: Professor Augusto Emanuel Abreu Esteves

January 2021

Acknowledgments

I would like to thank my family, specially to my parents, sister and grandmother for their friendship, encouragement and caring over all these years. Every push that I received help me to get to this moment, and sometimes that push was really important.

I would like to thank my girlfriend that supported me, not only she was an inspiration and never gave up on me, but she also helped me reviewing several times this thesis and also my previous works. Without her, I would never be able to finish this chapter of my life.

I would also like to thank my supervisors, Prof. Jacinto Nascimento and Ph.D. student Francisco Maria Calisto for their insight, support, and sharing of knowledge that has made this thesis possible, with their work I was able to make something that I can be proud of.

To my friends, specially Nádia and João for their support and help with this work. Thank you, you have been also a light in my way.

And last but not least, a thank you to Radiology Assistant website for the images and content kindly given and to the specialist in breast cancer that participated in this thesis, your knowledge and comments were truly helpful to the execution of this work.

To each and every one of you – Thank you.

Abstract

Computer-Aided Diagnosis (CADx) systems are essential when diagnosing patients with cancer. Medical Imaging Multimodality Breast Cancer Diagnosis User Interface (MIMBCD-UI) is a Computer-aided Detection (CADe) system that allows to open, view and manipulate medical images in order to diagnose patients with breast cancer. In this work, we aim to improve this system, thus allowing a faster medical image manipulation, by creating automated processes. With Human-Computer Interaction (HCI) techniques, such as Focus Groups, Affinity Diagrams, Interviews, Questionnaires and Scales, we developed functionalities based on the specialists' opinions. Three functionalities were created focusing on reducing steps in the medical image manipulation, without reducing its quality and while making the analysis effortless and faster. It was proven that these functionalities improved the usability, by increasing its value from 86.935 to 91.(1); the workload, by decreasing its value from 29.1(4) to 15.037; and the time of a diagnosis process by reducing the number of clicks by half, when compared with the previous iteration. All the Design Goals and Research Questions were achieved and proven with the results obtained from the tests. With a full base system, the upcoming developments will start by refine our functionalities or the creation of the functionalities that are desired. The ultimate goal is to have this system merging with iterations that are being developed at this instant, Artificial Intelligence (AI) and eXplainable Artificial Intelligence (XAI), which will allow the system to become a complete CADx that could be applied in real scenarios and help to save lives.

Keywords

Computer-Aided Diagnosis; Design Thinking; Human-Computer Interaction; Health Informatics; User-Centered Design; User Interface Design; Usability testing.

Resumo

Sistemas Computer-Aided Diagnosis (CADx) são essenciais para o diagnóstico do cancro da mama. Medical Imaging Multimodality Breast Cancer Diagnosis User Interface (MIMBCD-UI) é um sistema Computer-aided Detection (CADe) que permite abrir, visualizar e manipular imagens médicas de modo a diagnosticar pacientes com cancro da mama. O objetivo deste trabalho foi melhorar o sistema básico de diagnóstico, permitindo assim uma manipulação mais rápida de imagens médicas, através da criação de processos automatizados. Com o uso de técnicas de Human-Computer Interaction (HCI), como Focus Groups, Affinity Diagrams, entrevistas, questionários e escalas, desenvolvemos ferramentas baseadas na opinião de especialistas. Três ferramentas foram criadas tendo como foco a redução do número de passos na manipulação de imagens médicas, sem diminuir a sua qualidade, tornando a sua análise fácil e rápida. Provámos que estas ferramentas melhoraram a usabilidade, melhorando o valor de 86.935 para 91.(1); a carga de trabalho necessária, diminuindo o valor de 29.1(4) para 15.037; e o tempo de processo num diagnóstico, reduzindo o número de clicks para metade, quando comparando com a iteração anterior. As Design Goals e Research Questions foram alcançadas e comprovadas com os resultados obtidos nos testes. Com um sistema de base completo, segue-se a correção e melhoramento das ferramentas agora criadas ou a criação de novas ferramentas. O objetivo será a junção deste trabalho com outras iterações já em desenvolvimento, como a nossa Artificial Intelligence (AI) e eXplainable Artificial Intelligence (XAI), permitindo a atualização para um sistema CADx podendo ser aplicado em casos reais ajudando a salvar vidas.

Palavras Chave

Computer-Aided Diagnosis; Design Thinking; Human-Computer Interaction; Design Centrado no Utilizador; Design de Interface Gráfica; Testes de usabilidade.

Contents

List of Figures

List of Tables

Listings

Acronyms

MIMBCD-UI Medical Imaging Multimodality Breast Cancer Diagnosis User Interface

IHME Institute for Health Metrics and Evaluation

GBD Global Burden of Disease

BI-RADS Breast Imaging-Reporting and Data System

UI User Interface

UTA User Testing and Analysis

IPO Instituto Português de Oncologia de Lisboa Francisco Gentil

HFF Hospital Professor Doutor Fernando Fonseca

SAMS Serviços de Assistência Médico-Social do Sindicato dos Bancários do Sul e Ilhas

SUS System Usability Scale

NASA-TLX NASA Task Load Index

RR Radiologist Room

MG Mammography

CC Cranial Caudal

MLO Mediolateral Oblique

US Ultrasound

CT Computed Tomography

DICOM Digital Imaging and Communications in Medicine

MRI Magnetic Resonance Imaging

XAI eXplainable Artificial Intelligence

AI Artificial Intelligence

HCI Human-Computer Interaction

TDLU Terminal Ductal Lobular Unit

VR Virtual Reality

CADx Computer-Aided Diagnosis

CADe Computer-aided Detection

PACS Picture Archiving and Communication System

JSON JavaScript Object Notation

1

Introduction

Contents

1.1 Motivation and Context

Medical Imaging Multimodality Breast Cancer Diagnosis User Interface (MIMBCD-UI) [13] is a project under development designed to solve problems in the breast cancer area, such as the excessive time necessary to complete a diagnosis, the wrongfully managed resources and the lack of backup in the decision making process. Thus, this project developed a system able to manipulate medical images, along with an Artificial Intelligence (AI) [20, 21] that can interpret them giving a diagnosis, and with an eXplainable Artificial Intelligence (XAI) [22] feature capable to explain it.

Previous iterations identified a lack of development in Computer-Aided Diagnosis (CADx) programs in the detection of breast cancer, using breast images. Thus, since the beginning of this project, eight iterations were created and developed to resolve some of the issues discovered. Each one of these were called an **User Testing and Analysis (UTA)** [13], a process on which a set of tools are created and tested.

In this thesis we developed the Scalable Interactions work, corresponding to the ninth UTA of the MIMBCD-UI project, where a new prototype was developed, tested and compared to older iterations, while using Human-Computer Interaction (HCI) techniques. All this process will be explained throughout this document, however, since our prototype is to be applied to the medical area, an understanding of the subject, *i.e.* the domain, is necessary. Therefore, during this first chapter, we will make a brief introduction to the present situation of cancer worldwide and explain the necessity to create a system aiming to minimize problems regarding diagnoses; the challenges found during our work and the contributions that it gave to the breast cancer health care system and the HCI area; and a brief description of what to expect in the next chapters.

According to the Institute for Health Metrics and Evaluation (IHME) and the Global Burden of Disease (GBD)[1], in 2017, around 56 million people died worldwide, and from those, 9.56 million died from cancer, being this the second leading cause of death. Of all the types of cancer, breast cancer is the fifth with a higher mortality rate, since around 611,625 people perished from this disease. This is the most likely cancer to occur in women, with a chance of around 21%, whereas men, have a chance of less then 1%. Its mortality rate depends mainly on the stage of the cancer and the health system quality of the country where the patient is treated. Even so, it can be reduced if diagnosed and treated in time.

In terms of the stage of the cancer, breast cancer follows a well described scale, called Breast Imaging-Reporting and Data System (BI-RADS), that classifies the development of the disease. This scale is divided into seven categories from zero to six, Figure 1.1, being the categories four and five the most dangerous [1]. Concerning this work, there was no focus on category zero or six, because zero represents the need for more medical images or previous images without which the patient can not get a diagnosis and six represents an already *known diagnosis* made through biopsy [1].

[1] https://ourworldindata.org/ accessed March 2020

Final Assessment Categories			
Category		Management	Likelihood of cancer
0	Need additional imaging or prior examinations	Recall for additional imaging and/or await prior examinations	n/a
1	Negative	Routine screening	Essentially 0%
2	Benign	Routine screening	Essentially 0%
3	Probably Benign	Short interval-follow-up (6 month) or continued	>0 % but ≤ 2%
4	Suspicious	Tissue diagnosis	4a. low suspicion for malignancy (>2% to ≤ 10%) 4b. moderate suspicion for malignancy (>10% to ≤ 50%) 4c. high suspicion for malignancy (>50% to <95%)
5	Highly suggestive of malignancy	Tissue diagnosis	≥95%
6	Known biopsy-proven	Surgical excision when clinical appropriate	n/a

Figure 1.1: BI-RADS scale classification available for breast lesions, from the need to have more information (category 0) to the known result of a biopsy (category 6) [1].

Each category describes how suspicious the lesion is, when present, and how to manage it: the first category means that lesions were not found in the breast image, so its management is a continuous routine screening, being in Portugal every 2 years [23]; in the second category, a lesion is detected but its characteristics show that it is benign, therefore, its management is also a continuous routine screening; in the third category, a lesion is also present and it has a high probability of being benign, however, it has $\leq$ 2% probability of being malignant, thereby it is necessary to do a followup every 6 months or even more periodically; the fourth category reveals that the lesion detected is probably malignant, yet this stage can be divided into three sub categories, from 4.a to 4.c, where all have the same management, tissue analysis also called biopsy, but there are different probabilities that the lesion could be malignant (4.a has a probability between 2-10%, 4.b, between 10-50%, and 4.c, between 50-95%); the fifth category has the highest probability of suspicion that the lesion is malignant, by $\geq$ 95%, and its management is also tissue analysis [1].

To understand why an early diagnosis is important, it is necessary to know what is considered an early or late stage of the breast cancer. Categories *0*, *1* and *6* are excluded from this classification since, *0* and *1* categories means that there is not sufficient information, or a lesion is non existent, and category *6* means that a diagnosis is already known. Categories *2* and *3* are considered early stages, and comprehend calcifications and small masses allocated in the breast and in the armpit. On the other hand, categories *4* and *5* are considered late stages, and those include larger masses that can spread beyond the breast [24].

Statistically, if diagnosed in the first five years the survival rate, for early stages, is between 80 and 90% [25]. However, according to data from 2014, for late stages, the survival rate decreases between 10 and 40% [25]. In 2009, *e.g.* in Algeria, a developing country in Africa, the survival rate was around 59.8% [26], whereas in Portugal, a developed country in Europe, the survival rate was around 83.4% [26].

Our work aimed for the creation of a novel CADx system, based on the ones currently used in hospitals we have an agreement with, which started to be developed in UTA4 [13] of the MIMBCD-UI [13] project. The system was made based on the physicians' necessities, allowing to view and manipulate several types of medical images, such as **Mammography (MG)**, that we called **Modalities**.

This is a new system with new manipulation techniques that, not only makes the manipulation of medical images a faster process, since the user does not need to perform so many operations, but also allows to compare breast asymmetry faster and view the lesion evolution through time. The hospitals that we have an agreement with, are the Hospital Professor Doutor Fernando Fonseca (HFF), Serviços de Assistência Médico-Social do Sindicato dos Bancários do Sul e Ilhas (SAMS), Hospital do Barreiro and the Instituto Português de Oncologia de Lisboa Francisco Gentil (IPO) in Lisbon.

1.2 Challenges

Our goal is not to make another system like many others in the medical area, but to make something that is inspired in the users' needs while helping to resolve the problems found. Therefore, our approach was done in a HCI point of view. Through the course of this thesis, several challenges appeared in different parts of the process and were related both to the HCI area and the computational area.

In every HCI project an overall planning is necessary which, in our case, had to be focused on the user, inexpensive to revisit and changeable at any step, making this, our first challenge. Having that in mind, it was understandable that a strong research background was required before making any progress on a prototype. After reviewing several strategy methods, we chose the **Design Thinking** [27] method which is divided into 5 stages: identify a problem; define and understand it; create ideas to solve the problem; create a prototype based on those ideas; test the prototype [27]. After a complete cycle of the 5 stages, the strategy process can have more iterations [27]. This method is going to be extensively explained in Chapter 4.

Other challenge is related with the functionalities, more specifically with the gather of ideas to create them. Normally, this process is done with a direct interaction between the researchers and the users in their work environment, however, due to the pandemic (COVID-19), this was not possible, which created a HCI challenge. Since our access to the hospital facilities was restricted, we were forced to move the entire process to an online format, that created other challenges.

An online format of this type of projects is a HCI challenge, since we had to reprogram and minimize the new problems that were brought up. When performing the online meetings, the total number of physicians and, also, the managing of those meetings had to be reconsidered. We, also, had to establish roles that allowed or restricted who had the floor during meetings to maintain the order.

Given the procedure chosen, the idea stage was done using HCI techniques, such as **Focus Groups** with **Affinity Diagrams** [7], which will be explained in Chapter 4. However, the latter should imply in-person meeting, so, to overcome this challenge, a tool called *Trello* was used, allowing the same type of process but in an online scenario [7].

The idea behind the diagnosis and the physical space where it is done, the **Radiologist Room (RR)**, was a major computational challenge given that it turned out in a physical limitation to the project. Since we wanted to designing a prototype that is supposed to be used in those conditions, it was necessary to take into account the given constraints and design around them. The first limitation noticed was imposed by the lightning because this is a dark room, therefore, we had to use a limited color pallet to design the whole system. The other characteristic that we noticed was the low amount of hardware available for diagnoses, so it was necessary to develop simple commands.

The *user-test* did not occur in the expected location and conditions, which created its first HCI problem, since the test environment could not be the same for all participants, making the comparison between physicians harder. The second HCI challenge, was regarding both data from the patients and the physicians' demographic data, which had to be anonymized. The physicians' data although private, has important information to the study, so, understanding how and who used our system, allowed the creation of personas that, in some way, represent the medical community.

1.3 Contributions

This thesis contributes to the creation of a novel CADx program to the breast cancer health care system, by providing functionalities that make some **tasks automated** and by allowing the **breast screening asymmetry**, *Coordinated View*, Section 5.5, and the **lesion evolution comparison through time**, *Temporal View*, Section 5.5. Also, the *Recorded View* functionality, that allows to **store the system state**, Section 5.5, was created. In addition, due to the way this program is being developed, it could be adapted to other medical fields related to cancer, given its capability of reading Digital Imaging and Communications in Medicine (DICOM) files.

On the other hand, since this project corresponds to the ninth iteration of the MIMBCD-UI project, the program emerging from this one will be used as a base for future iterations. Currently, there are two other iterations in development, where one is aiming for the creation of an AI [20,21] assistant, using the UTA4 [13] base system; and the other is creating mechanisms to explain the AI result, an area called XAI, where the main goal is to increase the result decision trust [22]. The first iteration will be able to read the exam, prior to the physician, and give a BI-RADS classification by writing, whereas the second iteration, will highlight the lesions found by the AI, in order to explain the decision given.

Scalable Iterations is focused on using several methods that aim to develop functionalities in the users perspective. With methods such as **interviews**, **Focus Groups**, **questionnaires** and others, we started this project with a blank perspective on how to develop a system, and finished it with a system highly capable of being implemented in Hospitals as it is.

1.4 Outline

This document has six more chapters, where the entire process taken is explained. In the **Related Work**, Chapter 2, the **Medical Domain** will be explained in detail and we will also explore and compare other systems with ours. Given that our system has a strong HCI component, a comparison to other systems will be made in and out of this medical area, and explained what we learned from them and which techniques we used in our work.

In the **Design Goals**, Chapter 3, we will explain the measurable goals that we wanted to achieve and our research questions will, also, be introduced.

The **Work Evaluation**, Chapter 4, will have a detailed explanation about how we planned the tests and how we executed them. Furthermore, we will explain the metrics, scales used and how the questionnaires were created and used.

In the **Implementation**, Chapter 5, we will explain in detail the architecture followed, focusing in the technologies that we used to make this system work. In addition, in this chapter, the novel functionalities will be explained in detail and their impact in the diagnosis.

During the **Results and Discussion**, Chapter 6, we will explain, not only, the results that were obtained during the tests, but also the difficulties that we experienced during those tests.

Finally, in **Conclusion and Future Work**, Chapter 7, we summarize the project goals, what was accomplished during this thesis, what was postponed to future work to be improved and what new functionalities are also requested to be done.

2

Related Work

Contents

In this chapter, we will approach the **Medical Domain** in detail, focusing on breast morphology and the different types of lesions that can occur. Also, other systems, Computer-aided Detection (CADe) and Computer-Aided Diagnosis (CADx), will be explained and compared to ours. And, finally, we will take a look on a more HCI point of view, and understand the HCI techniques other projects used, not only when identifying the problem, but also how tests were planned and evaluated. First and foremost, it is important to understand some definitions that we will use in this thesis.

2.1 Definitions

As explained before, we used the concept of *Modalities* referring to every possible type of medical image that can be stored in DICOM [2] file types. In this document we will also use the expression *Multimodality*, when referring to the possibility of having several of these modalities opened side by side in the same overall viewport, allowing different perceptions of a lesion.

Human-Computer Interaction (HCI) is a multidisciplinary field of study, focused on the creation and design of computer technology, with a special attention to the users' needs and their interactions with a computer [28]. This field of study has developed techniques allowing the identification and construction of a full test, and also explaining or grading, if the system meets the users' needs [28].

Concerning this work, we characterized a *physician* as the medical doctor that has the base knowledge for each specialization in this particular area, such as oncologists, surgeons, gynecologists and radiologists, Figure 2.1. Although all off these specialists are trained to examine medical images, only radiologists can perform a diagnosis[1].

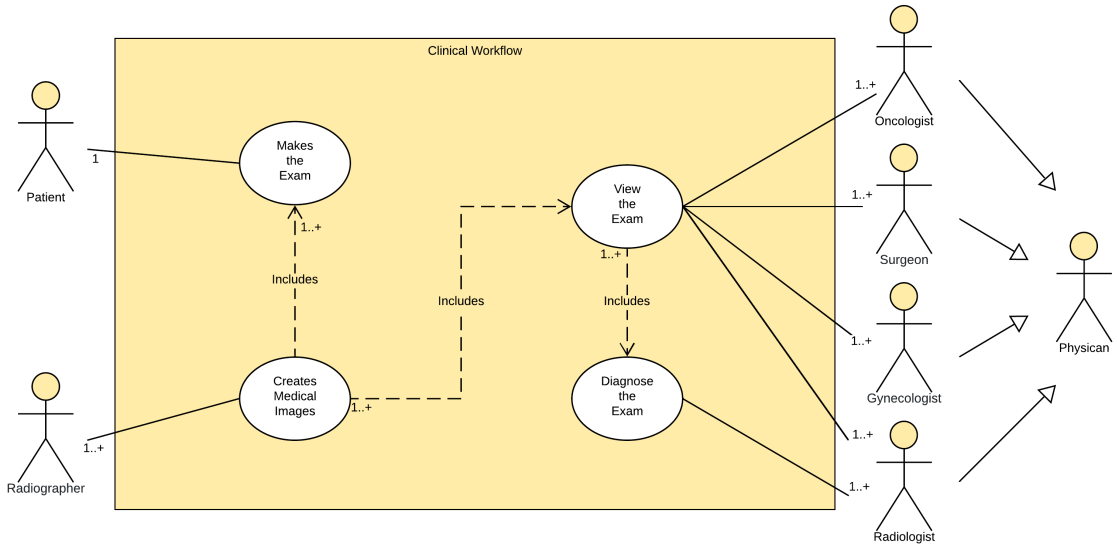


Figure 2.1: A High-Level approach of the clinical workflow that exemplifies each actor present. In this High-Level approach a *Patient* makes a clinical breast exam, that is processed by the *Radiographer* into DICOM [2] files and then uploaded to the DICOM [2] server. Though a series of *Physicians* can view the medical exams, only the *Radiologist* can make the final diagnosis and produce the medical report that every other specialization can consult and decide on what the treatment should be.

[1] American College of Radiology, Link : https://www.acr.org/(...)/About-Radiology accessed March 2020

Thus, in order to have a program that will be used by all, we can not focus our attention only in the **Radiology** specialization. Hereafter, when we mention physicians in general terms, we are referring to all specializations, but when referring to a physician that makes a diagnosis, we mean specifically a *Radiologist*.

2.2 Methodology

In this section, it will be explained several techniques that can be used in a HCI work by providing examples of its use in other systems, and which ones we will use in our system and why. We will explain the techniques that we have available in our project, from the most simple to perform to the most difficult. These techniques offer some benefits in terms of development cost, product quality and user satisfaction [29].

2.2.1 HCI techniques

Questionnaires are the simplest technique to use and to obtain quantitative and qualitative results through a wide range of population, though, creating one can be challenging and it needs to be well planned. A questionnaire can have 3 types of questions: closed questions, open questions and semi-open questions. Closed questions can be categorized as questions that have answers already written, where is not possible to add others. Open questions, invite the participant to write their own answer, without character limitation. Semi-open questions are described like closed questions but, in this case, it is possible to add an *Other* option, with a limitation of 255 characters, if the participant chooses to do it. The results analysis depends on the type of questions made: closed questions can be straightforward whereas open and semi-open questions have to be interpreted by the researcher in order to draw conclusions. For our work, these questionnaires were essential since they provided crucial information such as what already exists in currently used hospital systems, what is desired to implement and who desires it. We made three questionnaires focusing on these points.

Scales are the second easiest technique to execute. These are questionnaires already created and validated by the scientific community, that have a straightforward application. The hardest thing in this case, is to choose the correct scale and why, and to draw conclusions from that scale. For our work, we chose two scales, one that explores the users' experience and usability and other that explores the amount of work that is necessary to have to produce a satisfactory result.

Affinity Diagrams is an activity that consist in creating *notes* with ideas and re-arrange them in groups in order to discuss those topics[2]. This technique can be done simultaneously with **Focus Groups**, thus creating the notes while discussing more topics.

Interviews are the first technique that we present that needs to have a person to person contact, where the researcher ask the participant several questions about a subject. There are several types of interviews, such as, Telephone interview, Video interview, *etc*. This is a complicated technique to

[2]Interaction Design Foundation Affinity Diagrams, Link: https://www.interaction-design.org/(...)/affinity-diagrams accessed March 2020

use, because it needs to be well planned, with the right amount of questions for the type of information that the researchers want to obtain, and where even the location needs to be wisely chosen to avoid as many distractions as possible. We have planned for this work to have personal interviews, however due to the pandemic (COVID-19) we discarded this possibility and change all the process to video interviews. Those were done during the tests thus allowing the physician to comment our system and explain theirs during that process.

Focus Groups are essentially interviews that are done in groups but, contrarily to the previous, this promotes a mutual discussion about one or more topics. Focus Groups are interviews with more than two participants, where is included at least one participant and one researcher/interviewer[3]. This HCI technique is complex, given that it is necessary to have a moderator to maintain order and let all participants express their opinions about the subject in question. In this work, we used this technique as our primarily source of ideas, where at the beginning we discussed the domain and proceeded to our system and gather of ideas to future projects.

2.2.2 Related Projects

During this subsection, we explore systems that have a similar domain of application or a very strong HCI approach, that could help us guide our path or take ideas that could be important to explore.

Hatscher, B. *et al.* developed a prototype that translates touchless hand gestures into functions of a special-purpose software for Magnetic Resonance Imaging (MRI)-guided interventions [30]. In this project, the authors started by identifying a cumbersome control system in the prevalent method in clinical practice. This problem was analyzed after the application of an online questionnaire to eleven experienced radiologist. After that, a system was created to allow the user to control the software with four types of gestures: *cursor move*; *mouse click*; *changing slice position*; and *no operation*. A set of tasks were performed to evaluate the system where, at the end, each participant answered to two scales: NASA-TLX [17, 18] and SUS [14, 15]. In conclusion, when comparing the old system against the new one, the time metric results were similar, however, in terms of performance, the new system was worse but more acceptable in the users' eyes [30].

Our approach was similar to the one used in this project. In their case, identifying, defining and creating ideas, is done in one stage, at the execution of the online questionnaire. In our case, we set out different stages with different evaluations techniques for each stage, which will be explained later in Chapter 4. From this project we also took the idea of using scales, that we considered to be a good way of understanding the impact of our work in daily activities.

Li, L. *et al.* developed an interactive online patient decision aid, called *ANSWER-2*, that reduces patient decision conflict and improves their medication-related knowledge and self-management capacity [31]. This web application can be accessed with any device that has internet access and a browser, and was

[3] Interaction Design Foundation Focus Group, Link: https://www.interaction-design.org/(...)/focus-groups, accessed March 2020

developed following international patient decision aid standards with the help of patients, physicians and computer scientists [31].

First, the patients availability was filtered with several required characteristics. From the 117 patients interested in the study, 50 finished it and, from those, 2 did not attend the final interview. The 50 eligible patients in the study were asked to complete a series of outcome measures, with and without the web application. A month later, 48 patients had a semi-structured interview by telephone that had a duration of around 45 minutes with a researcher trained in qualitative research methods. All interviews were recorded and transcribed.

In this study, two scales, Decisional Conflict Scale (DCS) [32] and Partners in Health Scale (PIHS) [33], and one questionnaire, Medication Education Impact Questionnaire (MeiQ) [34], were performed. These scales and questionnaire were used to evaluate the system by understanding how the patient felt while using it, regarding the choices that they made.

This article was chosen given its strong HCI aspect, with the use of two scales and at least one questionnaire, not forgetting the interviews that were also made. We used these techniques as well, to ensure that our work meets the expectations of its users and, ultimately, comparing the results collected with the previous interaction [13] (UTA4). We did not make an after test interview, but a talk-aloud technique during the system testing that, in some cases, became an interview, depending on how much interactive the test was. From this work, we also took the idea of making a transcript of the **Focus Groups**, where some of the ideas were created or discussed [35–38].

Stuijfzand, B. *et al.* aimed to measure the cognitive load of medical students, when interpreting volumetric images such as Computed Tomography (CT) or MRI, by applying HCI techniques and an eye track system [39].

The study was divided into two different studies. The first, used HCI techniques to assess the time per evaluation of the images: how many slices of images were interpreted, how many angle view changes were made and if the right slices were the most viewed during the task. This study does not measure the cognitive load, but uses a variable that is conceptualized as indirect to measure it.

The second study used an eye tracking system that provided more in-depth information about what is being examined by the student.

With these two studies, the researcher understood that students who took longer to locate relevant areas of interest, experienced a higher task-complexity with a higher cognitive load associated [39].

A scale was given to measure the cognitive load, the self-reported one item mental effort scale and pupil-dilation [40–43].

As in our project, the authors used several HCI techniques, such as the *time of completion* or the *number of errors*, by analyzing how many times a feature is used. We call it **Metrics**, and these will be explained in Chapter 4. Although during this project an eye track system was thought, we have put aside this technique, since it is hard to operate and gather information with and it needs to have a physical device in the users' computers. However, because we had to do everything online due to the pandemic, this turned out to be impossible to make.

2.3 Medical Domain

Given the importance of this work and its application, it was required to have a strong knowledge of the domain, the **Breast Medical Domain**. In this section we will explain the differences in breast structures and types of lesions, and why CADx systems are essential to diagnosis and, consequently, in treating the disease.

Primarily, it is important to refer that we considered a **Lesion** to be a **Calcification** or a **Mass** that can appear in the breast or in the armpit, and be detected through breast palpation, if the dimension is considerable, or through medical imaging. Moreover, several images presented bellow were only possible to obtain through different **modalities**. Hereinafter, we will present all these cases.

2.3.1 Brest Density

Breast are not equal regarding density [3]. There are four levels of breast density [3], Figure 2.2: Level 1, describes a low density breast tissue, $<$25% density, where a lesion is easily detected since it is similar to identifying white spots or regions in black screens; Level 2, characterize breasts with a density between $\geq$25% and $<$50%, where small lesions can be harder to detect but larger ones are easily identified; Level 3, corresponds to a density between $\geq$50% and $\leq$75%, in this case, it can be hard to identify masses and harder to detect calcifications; finally, Level 4, means that the breast is extremely dense, with a density $>$75%, and all lesions are hard to identify without the proper tools [3].

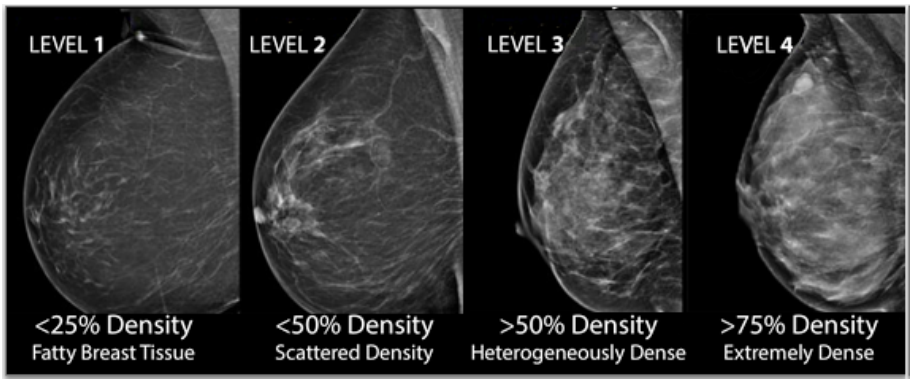


Figure 2.2: The 4 Levels of density in breasts [3]. From left to right, we have the least dense breast, Level 1, to the higher dense level, Level 4, in 25% intervals.

2.3.2 Breast Asymmetry

When referring to a mammography, breast asymmetry represents a morphological range for an unilateral fibroglandular-density finding seen on one or more mammographic projections[4]. In simple terms, if a breast is symmetric, both breast are similar in density and structure (*e.g.*, the tissue formations are similar in both breast).

[4]Radiopaedia Breast Asymmetry, Link: https://radiopaedia.org/(...)/asymmetry-mammography accessed November 2020

A projection refers to different perspectives to view the same exam, by changing the angle. In a Mammography (MG), there are several types of projections, being the most common the Cranial Caudal (CC) and Mediolateral Oblique (MLO), Figure 2.3.

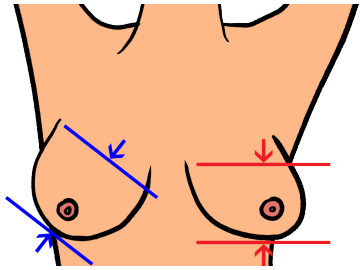


Figure 2.3: The four images taken when doing a MG, each breast will do a CC and a MLO projections. The blue lines represent the MLO projection view, a tilt capture, and the red lines represent the CC projection view, with a horizontal capture. [4]

Regarding to the different types of asymmetry found in a breast, it is possible to find four different types:

- **Asymmetry** refers to an area of fibroglandulair tissue that is only visible in one projection, CC or MLO, and can be characterized by a superimposition of normal breast tissue [3].
- **Focal asymmetry** is also an area of fibrogladulair tissue, although, this is visible in both projections, CC and MLO [3].
- **Global asymmetry** is an asymmetry that is over, at least, one quarter of the breast, and it is usually a normal variant [3].
- **Developing asymmetry** is a new, larger and more noticeable than all other asymmetries [3].

It is important to explain that having an asymmetry does not mean having a lesion. However, the physician will probably ask for more projections or modalities, to evaluate that particular zone to discard or confirm the presence of a lesion [3, 35, 36].

2.3.3 Calcifications

Calcifications are characterized as small calcium deposits [44] that can be formed inside a Terminal Ductal Lobular Unit (TDLU) [5, 44], Figure 2.4. This TDLU is a morphological and functional unit of the breast gland and given the location where this calcifications are created, it can indicate a benign or malignant lesion [45].

If the calcification is formed on the *terminal duct*, right part of Figure 2.5, it will indicate that this calcification is malign, which is called an Intraductal Calcification and corresponds to a stage 4 or 5 of the BI-RADS classification [5, 46]. On the other hand, if it rises reaching the *acini*, left part of the Figure 2.5, and if it is isolated, it means that the lesion is benign [5, 46]. However, if it is not isolated, it will depend on its size. In the presented case, being a small cluster, it will be a benign lesion, also called **Milk** calcification, corresponding to a stage 2 or 3 of the BI-RADS classification [5, 46].

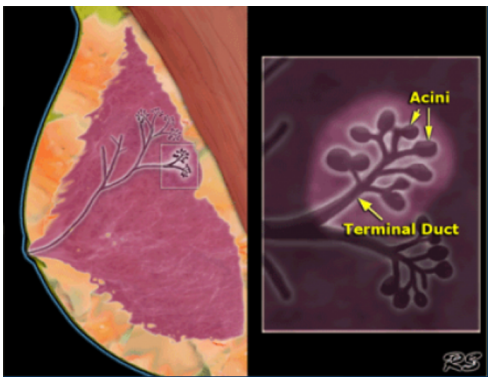


Figure 2.4: The basic functional unit in the breast, also called the TDLU. [5]

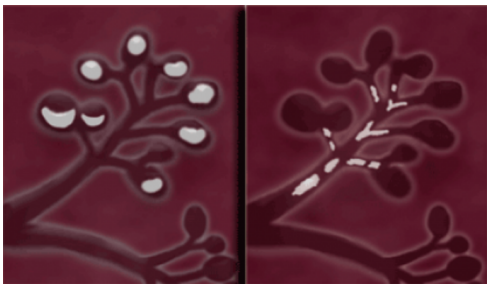


Figure 2.5: Calcifications, are small calcium deposits founded in the *Terminal Duct* and/or the *Acini*. [5]

Calcifications can have different morphology types and different allocations in a breast [5]. Thus, there are a total of fourteen different morphology types, Figure 2.6, that are divided into three categories: Benign; Intermediate Concern; and Malignant [5]. And there are five different allocations of those calcifications, Figure 2.7: **Diffuse**, **Regional**, **Clustered**, **Segmental** or **Linear** [5].

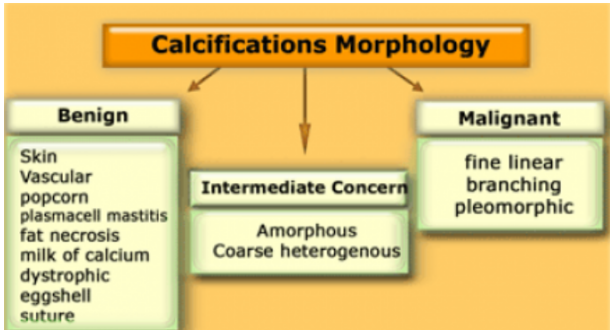


Figure 2.6: Calcification Morphology [5]

Regarding the different calcification formations presented in the Figure 2.7, the top left shows **Diffuse** or **Scattered** formations, known for having isolated calcifications scattered throughout the breast, which corresponds to a benign lesion with a BI-RADS 2 classification [5]; the top right shows a **Regional** formation, where the calcifications are scattered, but with a larger volume (>2cc) and closer together, these are formed in the *acini* and are benign calcifications with a BI-RADS 2 or 3 classification [5]; in the bottom left are present two different formations, **Clustered**, composed by at least 5 calcifications that occur in a small volume of tissue (<2cc) corresponding to a BI-RADS 4 classification [5], and **Linear**, representing an Intraductal Calcification, a malignant formation with a BI-RADS 5 classification [5]; the

bottom right shows **Segmental** formations, also malignant formations with a BI-RADS 5 classification, characterized by calcifications that were formed in the TDLU branches and lobe [5].

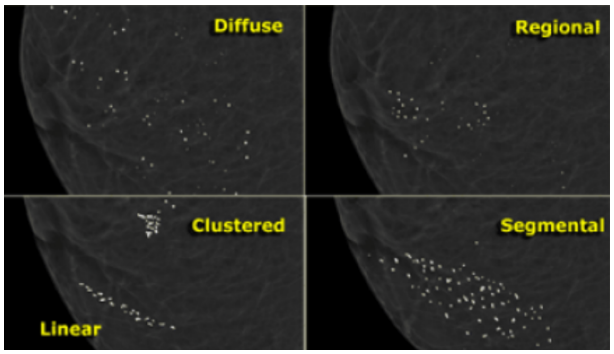


Figure 2.7: The 4 types of calcification groups found in breasts [5]

2.3.4 Masses

Masses are lesions that take a 3D space in a breast. Until these masses are visible, at least, two projections (CC, MLO or others), they are called asymmetry [3]. However, if it is located in more than one projection, it will be considered a mass and will be characterized.

Masses are composed by three different characteristics: **Shape**; **Margins**; **Density** [3]. The different combinations of those characteristics will identify if the mass is a benign or malignant mass [3].

We will start by analyzing the **Shape** aspect, that can be one of three, Figure 2.8: **Round**; **Oval**; **Irregular** [3].

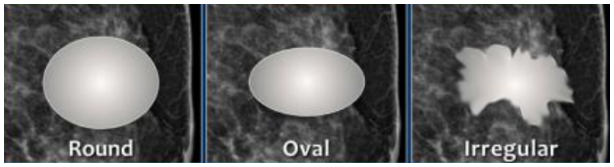


Figure 2.8: The three types of mass shapes possible to detect in lesions [3]

Although the shape is an important characteristic, this alone will not identify the severity of the lesion found, so it is necessary to analyze the other categories. Secondly, there is the **Margin** characteristic, which is divided into five sub-combinations, Figure 2.9: **Circumscribed**; **Obscured**; **Microlobulated**; **Indistict**; **Spiculated** [3].

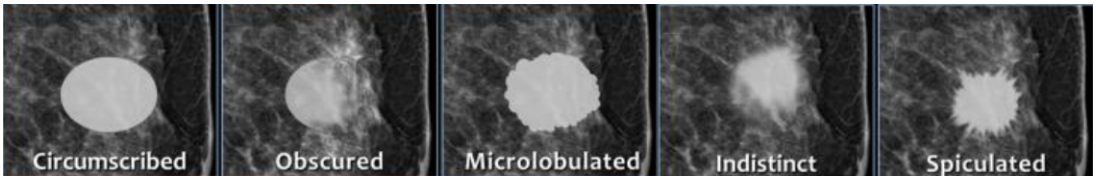


Figure 2.9: The five types of mass margins possible to detect in lesions [3]

With the margin information, it is already possible to draw conclusions. All masses that are circumscribed are benign findings with a BI-RADS 2 classification. Masses that have some of their parts

Obscured will be given a temporally BI-RADS 0, given that this classification means that more images are necessary to reach a conclusion. Masses with margins **Microlobulated** or **Indistict** will be considered a suspicious finding, where more imaging will be necessary BI-RADS 0 or, alternatively, will be classified with BI-RADS 3 or 4. Finally, the margin **Spiculated** indicates a very suspicious finding, that corresponds to a BI-RADS 4 classification [3].

Density category is divided it into three sub-categories, Figure 2.10: **Low density**; **Equal density**; **High density** [3].

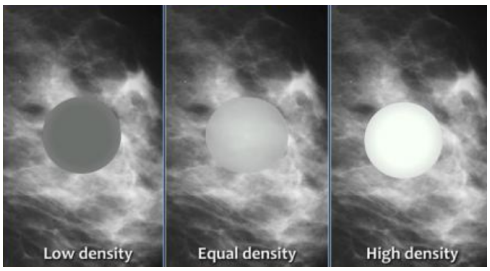


Figure 2.10: The three types of mass densities that can be present in a breast [3]

When a mass has **High density**, it is normally associated with a malignant mass, whereas, **Low density**, is rarely malignant [3]. Depending on those possible **Shapes**, **Margins** and **Densities**, a physician will make a diagnosis with a respective BI-RADS classification [3, 5, 46].

2.3.5 Clinical Workflow

A patient that needs to be examined in the breast cancer area, needs to enter in a clinical workflow, from the acquisition of the medical exam, to the final diagnosis and potential treatment. To explain this workflow, we will divide it in two parts, a more High-Level perspective, the **Overall**, and a more focused approach regarding the acquisition of these medical images, the **Clinician Procedures**.

2.3.5.A Overall

The High-Level workflow happens mainly in the Radiologist Room (RR) and can be divided into three sub-categories: **Image Acquisition**; **Diagnosis**; and **Reporting**. These three categories are represented in Figure 2.11.

The **Image Acquisition** stage refers to the time spent on the image machine itself and the patients' data processing, such as demographic data and clinical records. Regarding this stage, we will focus only in the image acquisition, such as MG, Ultrasound (US), MRI or others. The **diagnosis** stage refers to the time spent by the physician on performing a diagnosis. In this stage, the physician uses images from the medical exams made in the previous stage. The **reporting** stage refers to the time spent by the physician creating the final document, with the information gathered in the diagnosis stage.

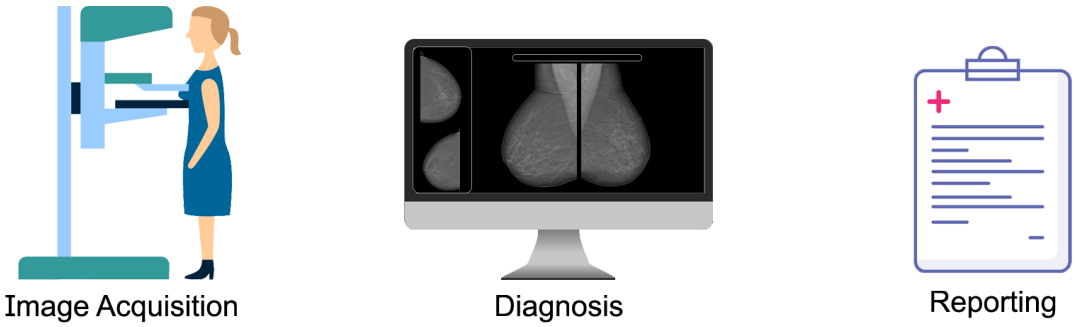


Figure 2.11: Clinical workflow with its three stages: Image Acquisition, Diagnosis and Reporting. [6]

2.3.5.B Clinicians Procedures

The **image acquisition** stage, depends on the hospital specific protocols regarding medical imaging requisitions. These protocols take into account not only which procedure is the most suitable for the diagnosis and the amount of radiation involved in it, but also if the patient can afford the procedure or if there are alternatives that can give the same type of results. Thus, there are four different procedures, represented in Figure 2.12, that can take place, taking into consideration all those characteristics and, also, the time necessary to give the results:

- The physician takes an MG. If the breast is not dense and no lesions are found, the process is concluded. If the breast is dense, an US modality is taken. If both, MG and US are not conclusive, an MRI is required.
- The physician takes an MG and detects a High Risk of cancer from the lesions detected, then both US and MRI are required.
- The physician takes both MG and US, if the exam is not conclusive, then an MRI is required.
- All three modalities are required, MG, US, and MRI from the beginning.

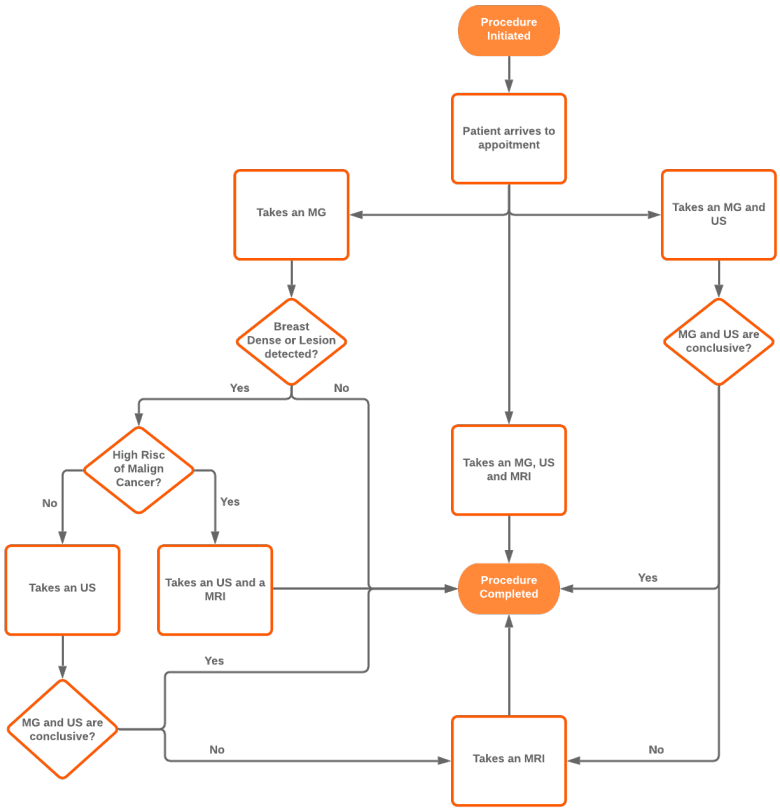


Figure 2.12: The clinical procedure, depending on the hospital, some will take more images with a higher cost and a higher expose to radiation to the patient, but reaching a conclusion faster. On the other hand, others will take more time to take every image, if all are necessary, but if they are not, the procedures ends quicker, with less expenses and less radiation to the patient.

2.4 Background Work

As previously mentioned, we chose a design process called **Design Thinking** [27], however, it is important to explain that there are other design processes. In previous work, more specifically in the UTA4 [13] of the MIMBCD-UI project [13], the **Usability Methods** [47, 48] process was chosen. Thus, by having a point of comparison to this work, it is important to understand why we did not follow the same process.

Usability Methods process can be divided into several categories [47]: (i) **Model-based**; (ii) **Inspection-based**; (iii) **User-based**; and (iv) **Scenario-based**. However, is for the researcher to define which one to implement. In the previous iteration, the authors chose the **User-Based** as a primary method and the **Inspection-Based** as their secondary method.

User-Based method, gathers inputs from users when interacting with the User Interface (UI). Questionnaires are the most widely used methods in this category, which measures the user subjective preferences after trying the system.

Inspection-Based method, is used in the early stages of the system, when creating and developing the UI, by evaluating prototypes or parts of the system that are not tested by users. This method is typically used by the evaluator that inspects the UI. The UI usability is inspected depending on a series

of heuristics of usability [49]. Although this method is cost and time-efficient, it requires multiple usability experts to maximize its effective measurement [50].

Model-Based method can provide questions regarding how users would perform a specific task concerning computational models of human behavior and cognitive processes [51]. This method was not used because the authors were not focusing on understand how the participant would use the system but in testing what was already made.

Scenario-Based is a method where scenarios are created in order to get users to do tasks, measure how they are done and if the test is successful. Often it needs to be complemented with other usability measures [47]. This method was not used in initial states of the project, given that no actual tasks were possible to be executed, but only after the first High-Fidelity prototype was developed.

Since MIMBCD-UI project, each iteration changed the design process in order to respond to the necessities of that prototype stage. This is not different in our project, we had the necessity of improving a system by creating good functionalities in a small period of time, with a reduced cost and with a very strong HCI component.

Given our necessities, we chose a design process that was focus on a pre-test scenario, with several techniques for gathering information about the physicians' needs, and how those were seen in their point of view. Nevertheless, it was also required a design process that could allow us to review our steps and re-process some techniques if necessary, which can be done with the loop process that **Design Thinking** [27] creates. In Chapter 4, we will explain this process in detail.

2.5 Computer-Aided Diagnosis

Our project was created under the assumption that CADx systems are undeveloped in this area. However, we were just focusing on a diagnosis point of view, without considering the aid that the AI could bring to the diagnosis, and, thus, becoming a real CADx system. Computer-Aided Diagnosis (CADx), means that the computer helps the user when executing a task, though, we can not consider our project a CADx system since it is still limited and only helps making some of the tasks automated. Ultimately, with our project, we will not have a CADx system, but a base for a future one, where future UTAs, with the introduction of the AI [20], and its XAI [22] will give rise to a full CADx system.

Thus, what makes a system a CADx system? To answer this question, we need to see it from the perspective of the computer while doing tasks or scenarios, in order to reduce or compliment the users' work. One way is to have an AI program that examines a medical image before hand and gives a result, or even sort the cases for the physician according to the severity of its discoveries. The other way, is when the AI program lets the physician analyze the medical image and gives a second opinion after those images are reviewed. In our general project both hypothesis are still on the table, and a final choice has not been made yet. Though, both hypothesis can, in some way, coexist when separating them in two phases: first, the system can sort the cases by priority without giving a result; and second, the system waits for the physician to ask for the second opinion, in order to confirm or not the physicians' idea. With this approach, a final CADx system will be concluded.

2.6 Overview Clinical Area

Given the importance of this type of programs, it was necessary to find out all about the present systems used, and, within the literature, what is being done in this field and what is their focus.

Regarding the present systems, we know that they only have available the simplest tools [52], which will be explained in Section 5.4, and these are the same simple manipulation tools already available in image manipulation programs: moving the image; zoom; changing color/contrast; and identify and measure lesions [52].

Therefore, in order to have a system that could compete with those in hospitals, we had to be able to give something new, aside from those tools. Several projects analyzed a Virtual Reality (VR) [53] approach; a touchless approach [54], that detects and interprets hand movements; and touch controls that uses touchscreens for a more direct contact [55–57]. The conclusions drawn by the latter refer that a more familiar apparatus, keyboard and mouse, is better and more convenient for the diagnosis when comparing to the touchscreen option [55, 56]. The others have an inconvenience given the lack of resources in the equipment available in the RR, at the present moment. However, these three approaches, only brought new input in controlling systems, not giving any new functionalities.

Other programs did not focus on the diagnosis, but in the portability of the exams [58], allowing the exams to be opened outside the RR. Given that our system works in a browser system, if the Orthanc server is running in the public domain with medical authentication, we can access through our system in any place that has internet access. If the server is running in an intranet domain, a Virtual Private Network (VPN) can be established for physicians.

We need to analyze programs that focus on the normal apparatus, with a server that can be accessed at any moment and that are meant to be used in a diagnosis [59].

The most similar program to ours is PaxeraUltima [60], which has the same tools used currently in hospitals systems, the same apparatus and it enables the patients' files analysis remotely. At the moment this program is a complete CADx system. It has two types of systems with the aided capacity. The **EraBot** that allows the physician during the exam to talk to another physician in order to have a second opinion and an AI system called **PaxeraAI** that receives information about the cases, in particular, images, lesions annotations, diagnosis and questionnaires about that patient. With this information, the most important cases can be prioritized giving physicians the support that they need.

The PaxeraUltima [60] program, also uses DICOM [2] files, therefore, all the different modalities are acceptable, but it is not an open-source program.

3

Design Goals

Contents

3.1 Objectives

During the development of our project, we established several objectives that we thought to be necessary to achieve, in order to deliver a finished product with utility to the user. This system, had to present functionalities capable of providing solutions effortlessly and without any drawback at any time. Furthermore, we wanted to improve the base system capabilities, so we also changed some of the little aspects of the system (*e.g.*, buttons icons; more detailed image list information; rewrite the help text; rewrite code functionality to reduce bugs; *etc*).

3.2 Design Goals

To understand if our system reached the objectives proposed, we chose three Design Goals, in other words, attributes that we wanted to have in our prototype. The first Design Goal was **Usability**, which means that we aimed for a ease of use prototype, which can be evaluated by the System Usability Scale (SUS) [14, 15]. In order to reach this goal, it was required to have an equal or higher score than previous iterations [13, 61], 86.9 out of 100.

The second Design Goal was **Efficiency**, meaning that the maximum work is done with the less input possible. This can be proved by having less non-critical errors and, at the same time, less click actions to make a task. For this, we used the **Count use of a tool** metric to measure how many clicks are necessary to make a task, and the **Errors** metric to measure the non-critical errors that are done during that task. We set as the limit to have no more than six non-critical errors, two per each functionality, since it was the average obtained in the previous work [13], and less clicks than in previous iterations [13].

The last Design Goal chosen was **Productivity**, to prove that our systems' productivity was, at least, equal or higher than in previous iterations [13]. To measure this, the user had to do similar tasks with less effort and time. The time was measured with the Count use of a tool metrics and the effort, which corresponds to the workload of the system, through the NASA Task Load Index (NASA-TLX) [16–18] scale analysis. To reach this goal, it was necessary to demonstrate the systems' capabilities in reducing the time with the functionalities that we developed and by having a workload total score lower than 30.92 [13, 62].

3.3 Design Methods

Design methods are techniques used to help design a system, that we used to understand better the domain and the works' necessities. We chose the Design Thinking [27] method for this thesis, given the necessity of understanding well the domain beforehand and focusing on the problem and its theoretical solution, before developing any functionalities. Thus reducing the production costs and giving a final result with a higher rate of acceptance by the users.

We implemented techniques such as: **interviews**, **Focus Groups**, **Affinity Diagrams** and **low-fidelity prototypes**. These were used during the 5 stages of the **Design Thinking** [27] process and

allowed us to meet our goals.

Usually, the first stages of the process are not done during the thesis, given the time limitations associated. Thus, two of our three functionalities started at the third stage of the process, but the **Recorded View** functionality made through the entire process, Section 5.5. In the first stage we used past **interviews** to identify a problem that was overlooked, the storing of the system state. The second stage, understanding that problem, was discussed between the researchers and, off-the-record, with a Senior Physician, that liked the idea and told us what was available in current systems. The third stage encompassed all the functionalities developed, Section 5.5, where **Focus Groups** with several physicians were done, and where **Affinity Diagrams** were created for gathering of ideas. It was during this stage that the other functionalities were chosen to be developed [35–38, 52]. Over the course of this work, we were able to make only one **Low-Fidelity** prototype, for the **Temporal View** functionality, that was used in the **Focus Groups** opening the way to the third research question. During the fourth stage, we used **interviews** and **Think-aloud** to obtain more qualitative data from present and, sometimes, future work.

3.4 Research Questions

During this work, we developed three Research Questions, that allowed us to understand some particularities of the system that was developed.

RQ.1 When and who will use the functionalities?

- H1.1 The functionalities were rejected;
- H1.2 The functionalities were used, only in cases of doubt;
- H1.3 The functionalities were used, but only by inexperienced physicians;
- H1.4 The functionalities were used by all physicians.

RQ.2 What is the impact of the functionalities, in the clinic workflow?

- H2.1 The usability of the system increased;
- H2.2 The workload impact was affected in a positive way;
- H2.3 Diagnostic time per patient was reduced.

RQ.3 What is the best method to represent the lesion evolution?

- H3.1 Old annotations on top of the more recent image, using a time bar;
- H3.2 With a time bar in the left viewport, and a more recent image in the right viewport;
- H3.3 With a time bar in the right viewport, and a more recent image in the left viewport;
- H3.4 A time bar in each viewport;
- H3.5 The lesion evolution tool was rejected.

4

Work Evaluation

Contents

4.1 Design Thinking

The **Design Thinking** [27] method was created because *"rather than asking designers to make an already developed idea more attractive to consumers, companies are asking them to create ideas that better meet consumers' needs and desire"* (Brown, 2008, p. 2) [27]. Thereby, before the development of the tools, physicians were asked about which ideas were better fitted for them, using interviews and questionnaires. **Design Thinking** [27] can be divided into five sub-categories [63]: **Defining the Problem**; **Needfinding and Benchmarking**; **Bodystorm**; **Prototype**; **Test**.

4.1.1 Defining the Problem

This is the stage were a general idea of a product, is thought of, however, it is still unknown which path to take in order to start the development process. Subsequently, it is done the analysis from the point of view of potential users with questionnaires, interviews, Focus Groups and other methods, in order to identify problems to be taken into account [63].

4.1.2 Needfinding and Benchmarking

This stage focus is to understand why a previously specified problem is happening and what could be its solution. There are several methods to be used for the discussion and creation of solutions [63]. One of these are the **Affinity Diagrams**, which are based in the creation of notes or cards with different ideas or ways that could resolve the issue.

4.1.3 Bodystorm

The bodystorm stage has an audience already defined to whom the researchers are going to discuss the ideas created [63]. This stage is also called the ideate stage [63].

4.1.4 Prototype

As expected, the prototype stage is when the development of the functionalities that were decided to be developed, are done. In this stage every piece of information from previous stages is gathered in order to determine if all the necessities were checked for when creating the feature.

4.1.5 Test

This is the last stage of Design Thinking [63]. In this stage, the prototype is tested to give the possibility to the user of giving feedback, by collecting metrics, questionnaires and scales [63]. In the next section we will explain the test in more detail.

4.2 Prototype Development

To create and evaluate the functionalities, Section 5.5, that we presented in this document, we made a test with the objective of allowing us to make an optimal path of the functionalities discovery and, also, to understand how efficient and productive our system was going to be, when compared to previous iterations [13].

In the first stage of the **Design Thinking** [27] process, the **Defining the problem**, we analyzed videos from previous iterations [13] and identified the necessity of **saving a state** during the manipulation of medical images. With this concern, we created the idea of developing the **Recorded View** functionality, Section 5.5. The other functionalities were ideas gathered in informal interviews between researchers and physicians in previous iterations [13], but confirmed in the third stage of this process with the **Interaction Tools** questionnaire.

In the second stage, the **Needfinding and Benchmarking**, we had a discussion between the researchers about the best and simplest way to implement the **Recorded View** functionality, Section 5.5. We did not use a Design Method to make this discussion, but several meetings to understand all points of view.

In the third stage, **Bodystorm**, we gathered all the ideas collected and discussed them with the physicians in four different **Focus Groups**, transcripts available online [35–38]. Here, we were able to clarify our doubts about the domain and the functionalities that we thought would be chosen for development, Section 5.5. During these **Focus Groups** we used **Affinity Diagrams** [7], where our questions were answered and where we wrote the ideas, Figure 4.1.

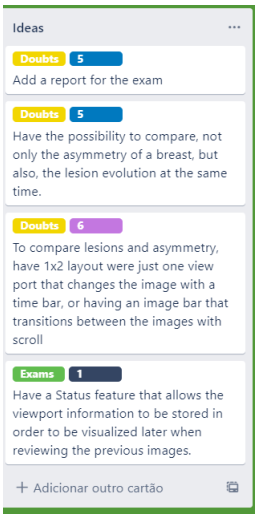


Figure 4.1: These notes have some of the ideas gathered during the brainstorm stage that was made, being a small sample of a much larger document [7]. The numbers shown in the labels correspond to the question number of that text label: in this example, the ”Doubt” label is related to ”What happens when we have doubts about a patient”, and the ”Exam” is about the functionalities that are used in the exams.

With the information gathered from previous stages, we developed the functionalities chosen, Section 5.5, in the **Prototype** stage, which were approved by the physicians as a good path to follow. First, we gave two questionnaires, Appendix B [52], one Demographic to get to know our users and their

experience [64], and other to know more about the tools that are already available in their work systems and what they would like to implement. After analyzing the answers, we decided which functionalities had priority to be developed, Section 5.5.

In the **Test** stage we made individual appointments with each physician and gave them a set of four studies, so they could test each functionality, Section 5.5. During these tests, some of the physicians showed more willingness to talk about their work and what they thought of the functionalities, so we made a non scripted interview to try to gather even more information. With the others, the design method **Talk-aloud** was used to understand what they were thinking regarding the functionalities. After each test, each physician answered to an after test questionnaire, Appendix B, and two scales, SUS [14, 15] and NASA-TLX [16–18], concluding the prototype test.

4.3 Evaluation

Our evaluation criteria was divided into three categories: **metrics**; **questionnaires**; and **scales**. In this section we will explain what we did with each one of them.

4.3.1 Metrics

In the **Test** stage we had the objective to obtain several metrics that would help us to understand better the functionalities developed, Section 5.5, by responding to our **Design Goals** and **Research Questions**.

The measures are as follows:

- Time on task
- Number of errors
- Count use of a tool

4.3.1.A Time on task

The time on task [65] was a very important metric, however, given that we changed our test and had small interviews between them, its relevance dropped. This issue was resolved by having a time per click and count the number of clicks with the metric **Count use of a tool**, instead of the measurement of time itself [11, 66].

4.3.1.B Number of errors

We considered two types of errors, **Critical Errors** and **Non-Critical Errors**. Both will be demonstrated as follows. **Critical Errors** occur when the scenario can not be completed, or several errors appear from that error. An example of a **Critical Error** could be a situation where the user can not continue the diagnosis alone, or when an image can not be loaded. **Non-Critical Errors** are errors that can be

recovered or, if not detected, do not result in processing problems or unexpected results. Although **Non-Critical Errors** can go unnoticed by the participant, when they are detected, it can be associated with frustration of the participant. These errors may be procedural, in which the participant does not complete a scenario in the most optimal way (*e.g.*, when trying to move an image, it changes luminosity). **Non-Critical Errors** can always be recovered during the process of completing the task.

4.3.1.C Count use of a tool

In order to understand if our functionalities are being useful to physicians, in each task performed, we counted how many clicks and what operations were done. Both standard tools and our new functionalities were counted, Section 5.5. With this information, we were able to know how and when our new functionalities were used in the diagnosis.

4.3.2 Questionnaires

We made three questionnaires for this thesis, where two of them were given before the test: the **Demographic**, and the **Interaction Tools**; and one given after the tasks were completed: **Interaction Tools Pos-Tasks**. Both the **Interaction Tools** and the **Interaction Tools Pos-Tasks** are available in the Appendix B [67]. The results of these questionnaires are presented in Section 6.5.1 and 6.5.2.

4.3.3 Scales

In this project, we chose two scales to measure the evolution of the system. The first is the System Usability Scale (SUS) [14, 15], that measures the usability of the system. This was very important for our work because we need to have a system well-received by the physician community. If the usability of the system was bad, physicians would not use it. Secondly, the NASA Task Load Index (NASA-TLX) [16–18] measures the workload of the system, which means that it measured how a task was performed and how difficult was to perform it. With this measurement, we expect to reduce the workload of the system. Both will be explained below.

4.3.3.A System Usability Scale

SUS [14, 15] is a reliable tool for measuring the usability of the system. It consists of ten statements with five possible answers, from *Strongly Disagree* to *Strongly Agree*. It is easy to use and can be applied to a small number of participants with reliable results. With this scale, we got to draw usability conclusions, helping to create a system that is better received by users.

4.3.3.B NASA Task Load Index

NASA-TLX [16–18] scale *"measures the workload estimated from one or more operators while they are performing a task or immediately afterward"* (Hart, 2006 , p. 904) [17]. It consists of six questions with 20 possible answers (1 corresponds to Low and 20 corresponds to High). This scale can also be applied

to a small number of participants. By using this we measured all sub-scales and understood if with these new functionalities, the task of diagnosing a patient, is easier or harder to accomplish.

5

Implementation

Contents

The goal for this project was to develop new functionalities to a system already proven effective in past research [13]. Thus, it is essential to explain how this will work and how the system is constructed, from the technologies chosen, to the system communication and interaction with the user. This system is supposed to be used in a Radiologist Room (RR), therefore, we had to be aware of the radiologists' limitations in terms of hardware and their setup for user tests. However, it was necessary to offer the possibility of having all functionalities in the most limited hardware. Also, given that this system has such important and personal data, it was imperative to have a good security system in order to be accepted in a hospital environment.

5.1 Language

When creating a system that already has a pre-established location where to be implemented, it is necessary to know what are the conditions of this particular location. In our case, we knew that a RR has computers with internet access and that installing a software could lead to maintenance problems. Therefore, we had to create a system usable in browsers, with encrypted connections, only accessible by a medical account with card authentication and with little maintenance in the user side.

For this browser approach, the simplest and more compatible programming language is **JavaScript**, which is able to open, manipulate and save medical images, by using one external library, the **CornerstoneJS**[1], that will allow the manipulation of DICOM [2] files. However, to open and store those files, it is necessary to use a Picture Archiving and Communication System (PACS) server, that also provides security.

5.2 System Architecture

Following the clinical workflow, after a medical imaging exam is completed, it is uploaded to a server thus making it available to the physician. This server needs to have unique characteristics, given that it is necessary for it to be compatible with a type of file able to store modalities in high definition, called DICOM [2]. It is, also, necessary to provide security measures to protect the data that is attached to the exams and enable it to be accessed from any location with an authentication process. Given these necessities, we chose the **Orthanc Server**[2] [68], a Picture Archiving and Communication System (PACS) server given that it is specialized in medical image storing.

The physician will then proceed with the diagnosis by manipulating and analysing the exam, using for that a system with specific tools that allow image manipulation. Since we chose to have a browser approach, it was required a tool compatible with **JavaScript**, more specifically a library. We chose the **CornerstoneJS**[1] library. This is an open-source library that allows to manipulate DICOM [2] files and with some modifications to create new functionalities.

[1] Link: https://github.com/cornerstonejs accessed March 2020
[2] Link: https://www.orthanc-server.com/ accessed March 2020

The final stage of the clinical workflow is the reporting stage, though this is not a focus point in our project. In the present hospital system, this stage is done with another program where the physician speaks directly to the program while it generates the report.

5.3 DICOM Server

As previously mentioned, our server needs to be able to store and manage medical images and have internet accessibility, facilitating the communication between hospitals and clinics. A Picture Archiving and Communication System (PACS) server enables image storing, facilitates communication and is remotely accessible. A DICOM server is a PACS server that is specialized in medical image storing.

Given this requirement, we chose the **Orthanc Server**[3] [68], a DICOM server, lightweight, accessible by **JavaScript**, and, most importantly, open-sourced and able to run in hospitals servers, or in an intranet. Orthanc server[3] [68] does not need software installation, no third-party dependencies and is cross-platform, so any type of server can run it without any problem. With this type of approach, the necessity of physical state exams (*e.g.*, discs) is no longer required, and so, in future medical appointments, the patient does not need to have their exams. Also, because the demographic information of the patient is already in the system, there is less amount of work to be done, which reduces data mistakes. With our new functionalities and with this server, if a previous exam needs to be compared with a new one, it will be possible to do it at the moment, because the information is centralized and easily available in the diagnostic program.

Moreover, given that our system is written in JavaScript like the server, it is able to export the data in a way that is almost direct to read, so a JavaScript Object Notation (JSON)[3] file is the obvious choice, since the **Orthanc Server**[3] [68], allows the data to be transmitted in JSON[4]. Also, a feature that this server has, is the possibility of downloading exams in a **PNG** file format, in order to give them to patients if they ask. Finally, the **Orthanc server**[3] [68] will ease some actions of scripting from people that are not comfortable to use them, by providing a web application of server management, that can be used by any physician.

5.4 Image Manipulation and Visualization

Given our idea of a browser based system, our functionalities had to be fast and light, in order to avoid overruning the capacity of the system, that can be low. **CornerstoneJS**[1] library is still in development, although, at this present state, it already has all the characteristics necessary to implement our system.

It was explained before, how the image manipulation is done, when using the **CornerstoneJS**[1] library. Nevertheless, it is important to explain that this is also the library that loads the image into the viewport and allows its visualization. One of our new functionalities, the **Temporal View**, enable us to understand the lesions evolution in the image by compare two similar images with different time periods

[3]Link: https://www.json.org/json-en.html accessed March 2020

with the possibility of having a time-bar for quick swap of images. This time bar was done using **Vanilla JS**[4], since design libraries could have several dependencies and some incompatibilities with our system. A Pseudo Code explanation of how the program handles the image is in Listing A.1 of the Appendix A.

Firstly, the **Orthanc server**[3] [68] needs to be initiated by a console command or using the Orthanc UI. Only then, the **Radiographer**, can upload medical images, previously converted into DICOM [2] files, into the server, allowing them to be accessed in the system.

The system is now ready to be initiated, and the physician can authenticate with their medical account, while the scripts present in the **CornerstoneJS**[1] library are loaded. If a successful login is made, the system will try to seek the files in the online connection to the server, although, if an error is encountered, another search can be done immediately. With the files available, the system will present a list, that has the patients' study information.

At this moment, the diagnosis stage begins. The physician selects one patient to perform a diagnosis, the patient images are loaded and displayed in two areas. One, is the list of exams available on the left side of the screen and the other is the viewport, where the first images are loaded automatically. At this point, the physician can make two types of choices, opening other image by dragging it to the viewport or click on it, or select tools to manipulate the image. The basic tools available are the following:

1. Layout;
2. Image Position (Move/Pan);
3. Zoom;
4. Change Luminosity (WW/WC);
5. Invert;
6. Stack Scroll;
7. Length Measurement;
8. Annotations (Freehand).

With these tools, the system is similar to the ones currently used in hospitals. The Layout allows physicians to have one or more images or modalities in a screen, which is important to the visualization of breast asymmetry. Each image will appear in certain areas that we called *viewports*. The Move or Pan tool, allows to move an image in a certain space. The Zoom tool, allows to see lesions more closely.

There are also tools that change the characteristics of the image itself, such as Contrast or Luminosity, which enables to hide breast tissue and expose lesions; or the Invert tool, that switches black and white colors making the lesions pop-up when the whole breast is visible. Finally, the most advanced tools available are the Stack Scroll which allows to scroll through the several images present in a MRI; the Measurement tool to measure the lesion and position; and the Annotation tool that allows to delineate the border of the lesion, which is very important for the understanding of its evolution [69–71].

5.5 Functionalities

In the MSc project [72], four functionalities were proposed and, from those, the **3D Module view** [73] was not developed given the time frame available and the results obtained from our questionnaires [52].

[4]Link: http://vanilla-js.com/ accessed March 2020

Thus, three functionalities were created: **Recorded View**, **Coordinated View** and **Temporal View** which we will explain in detail in this section, along with some of the opinions given by the physicians during the tests.

5.5.1 Recorded View

During previous studies, we have noticed that physicians repeated the same actions several times. This occurred because, after the manipulation of a medical image, if the physician wanted to change to a previous image for comparison, the already done actions were not stored. Therefore, when loading the previous image, the modifications done on the first one would be lost. This was also mentioned by a Senior physician, during a Focus Group, where we asked "Is Storing information important to you?" [35–38].

> ***"It is essential, because we need to go back to previous images to review and understand them, and having this capability is very important to us because it eases our work and understanding"***.
>
> Senior Physician

To address this issue, we developed the **Recorded View** functionality, Figure 5.1. Thus, the luminosity tool, **WW|WC**, the **Zoom** and **Pan** tools, which are important to store, will now have their last value stored and not all the values from each action made. Regarding, the **Invert** tool, given that with just one click it activates or deactivates, we chose not to include in this feature.

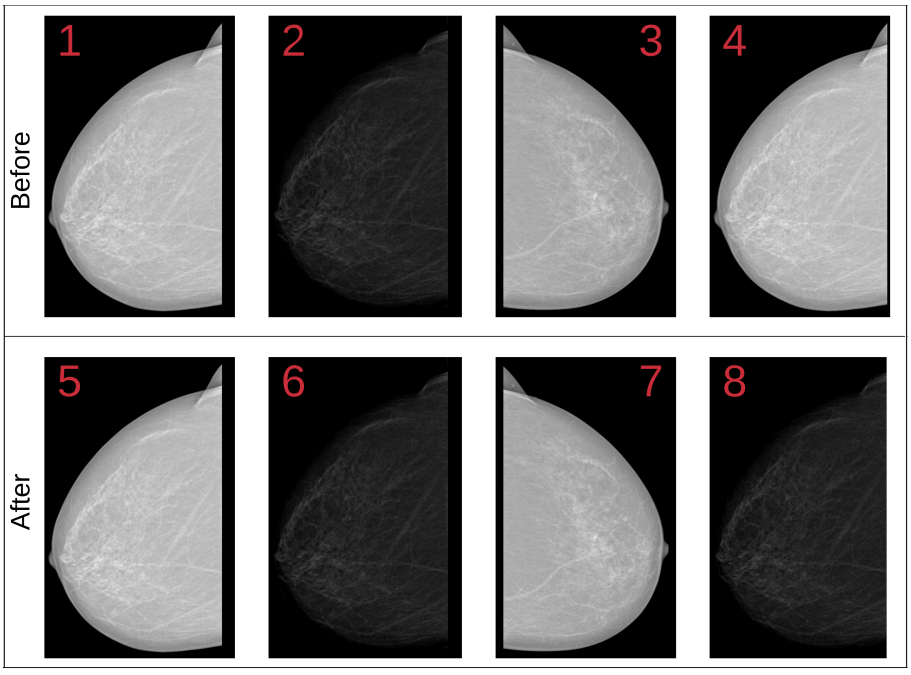


Figure 5.1: The Recorded View tool effect. The images in the line **Before**, represent how the system worked, whereas the ones in the line **After**, represent how the process is now conducted. [8]

Given this need of a saved state, we chose to create a variable for each individual image, with the following structure, which will be used in order to process it when it is loaded:

```
var Image = {
    WW|WC: [Value1,Value2],
    Zoom: [Value3],
    Pan: [Value4,Value5],
};
```

WW|WC tool has two dimensions, *windowWidth* and *windowCenter*, that make the image brighter or darker, giving visibility to the lesions that could be hidden in the breast tissue. These are stored in **Value1** and **Value2**, respectively.

Zoom tool has one dimension, the *scale*, that, when used, is frequent to makes changes in the **Pan** also. We chose to work with these two tools separately, however, they are always interconnected. We stored the **scale** dimension as **Value3** in the structure. This tool allows the user to focus on particular parts of the image in order to identify small lesions or tissue that could be hard to see.

Pan tool has two dimensions, translation **x** and **y**, that we mapped as **Value4** and **Value5** respectively. Contrary to the **Zoom** tool, when the values of **Pan** change, it does not mean that the **scale** dimension will change. In order to avoid conflict with duplicated values, every time **Pan** or **Zoom** are performed, an update on these three dimensions (**scale**; **x**; **y**) will be done at the same time. The **Pan** tool allows the physician to shift the image to a specific point were is more convenient to be analyzed.

5.5.2 Coordinated View

The **Coordinated View** functionality allows the user to load both sides of a breast to both viewports at the same time. This feature makes this process automatic, giving the possibility to verify breast asymmetry more quickly. With this functionality, it is also possible the simultaneous manipulation of all images loaded in all viewports, Figure 5.2, whereas the current hospital system does not provide this function, Figure 5.3.

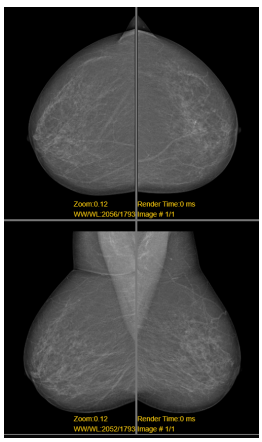


Figure 5.2: Coordinated View final result. [9].

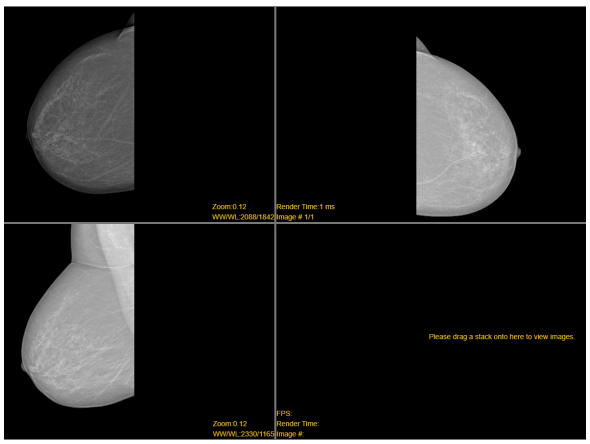


Figure 5.3: The present system without the Coordinated View, in multimodality mode, that shows the problems mentioned previously [9]

Since this functionality can open and modify images, it was divided in two aspects. When opening the image, there are two options, clicking on the image or dragging it to the wanted viewport. The clicking action opens the chosen image and its opposite (the same modality or projection but different laterality), whereas the dragging action only opens a particular image in the viewport chosen, giving, this way, the possibility of having two simple but different behaviours. When a case study is open by the system, it always reads the most recent image in the list, and if this image is an US, it will open a 1 x 1 viewport, if it is a MG, it will open a 1 x 2 viewport to allow breast asymmetry comparison.

> ***"The idea is always comparing with the other projection laterality (breast asymmetry), if one click does that, the better."***
>
> Intern Physician

Regarding the image modification, it can be set on or off. When it is on, every image opened will have the same behavior, in other words, if we change the **WW|WC** or use the **Invert** tool in one image, it will also happen on the others. The remaining tools will have an individual effect in the chosen image. When set off, each image needs to be modified individually.

> ***"It is helpful when you can modify several modalities/projections at the same time, without losing the option of modifying only one image."***.
>
> Senior Physician

This feature is very useful to the physician since it can reduce at least one operation per image. As mentioned before, we used **CornerstoneJS**[1] library to perform this operation.

5.5.3 Temporal View

The **Temporal View** functionality enables us to compare the temporal growth of a lesion. First, we developed the idea of comparing the lesions trough annotations, Figure 5.4, by adding to the new image the previous annotations, using an annotation hover.

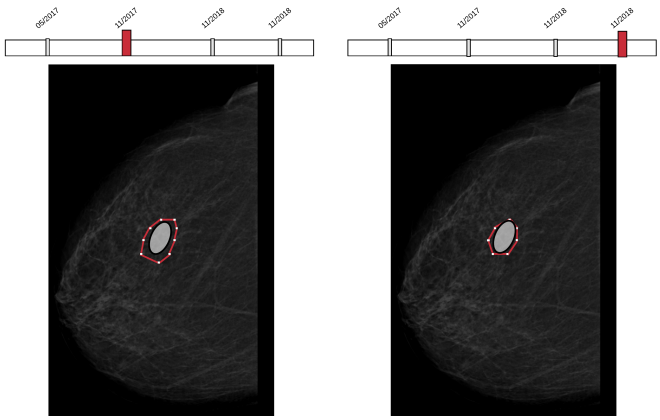


Figure 5.4: Temporal View, low-fidelity prototype, created before the Focus Group that established how the functionality had to be made. [10]

This was proven wrong in our low-fidelity prototypes when, in the Focus Groups, a Junior physician said [35–38]:

> ***"We compare the evolution by having the same breast projection through several years, first starting with the MLO projection and after, the CC projection,(...), in those we see the breast asymmetry and the lesions"***.
>
> Junior Physician

This made us discard the original approach and develop a functionality that does the a similar action as the **Coordinated View** but in a temporal aspect, by putting two similar images from different time periods side by side, Figure 5.5.

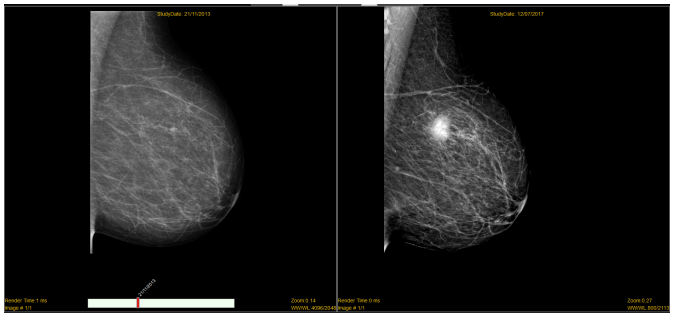


Figure 5.5: Temporal View effect, the high-fidelity prototype, after understanding the need that existed when analyzing the evolution of the breast lesion. Here is presented the oldest image on the left, with no lesion, and the newest image with a visible lesion, on the right [10].

5.6 Enhancement Understanding

With the pandemic (COVID-19), there were several changes that had to be done in several stages of the **Design Thinking** [27] process, being the prototype and test stages also adapted to this reality. We implemented in the system a tracking tool that registers the patient data and what functionalities were used in the test. This **Metric** was called **Count use of a tool** and allowed us to have a higher sense of what was being done by the physician without being there to see it.

As shown in text A.2 of the Appendix A, a real representation of the *Count Use of a Tool* metric, starts with the identification of the exam that is being analyzed and the DICOM [2] images ID's that are available in that study, with information about the modality, projection and laterality. For research proposes, we also toke notes of the time at the beginning and end of the test. We present the state of the test, by mentioning what is the layout present in the moment, what DICOM [2] images are opened in that layout and if the automated functionality, **Coordinated View**, is enabled or not. Then, after saying what images are open, each action or manipulation of the functionalities or tools used in the images, will be described on the front of each image. At the end, a summary of the tools and functionalities used is made, where each will have a counter to know how many times it was used.

6

Results and Discussion

Contents

6.1 Equations

Statistics are fundamental when planning, collecting and organizing information, analysing data and drawing conclusions in a research project. Thus, with this in mind, in this section we will briefly introduce some functions and tests for the statistical analysis of our work.

Standard Deviation, represented as σ, measures the amount of variations or dispersion in data samples [74], Equation 6.1.

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}} \tag{6.1}$$

Where:

N = total number of observations in all samples combined;

x = data sample;

μ = mean of the data;

Tukey Fences [75] is a technique used to discover values that do not belong in a data sample, since they do not follow the norm. These values are called outliers and are outside of the interval, equation 6.6, discovered by using the following equations 6.2, 6.3, 6.4, 6.5.

$$Q1 = First\ Quartile, \quad Q3 = Third\ Quartile, \tag{6.2}$$

$$IQR = Q3 - Q1, \tag{6.3}$$

$$Below\ Outlier < Q1 - 1.5 \times IQR, \tag{6.4}$$

$$Above\ Outlier > Q3 + 1.5 \times IQR, \tag{6.5}$$

$$[\ Below\ Outlier\ \ ;\ \ Above\ Outlier\ ] \tag{6.6}$$

Where:

$First\ Quartile\ (Q1)$ = the middle value between the smallest value and the median of the data set;

$Third\ Quartile\ (Q3)$ = the middle value between the median and the highest value of the data set;

$Interquartile\ Range\ (IQR)$ = measure of variability based on dividing a data set into quartiles;

$Below\ Outlier$ = minimum value accepted in a data set;

$Above\ Outlier$ = maximum value accepted in a data set;

The **Kruskal Wallis** [76] test, which is used in data sets that are non-normalized, is a non-parametric method for testing whether samples originate from the same distribution. The choice of this method will be explained in the sections ahead. The test Equations are the following 6.7, 6.8 and 6.9.

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2} \tag{6.7}$$

$$\bar{r}_i = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i} \tag{6.8}$$

$$\bar{r} = \frac{1}{2}(N+1) \tag{6.9}$$

Where:

N = total value of observations from all groups;

n_i = the value of observation the group i;

r_{ij} = the rank of observations j from group i;

When doing the **Kruskal Wallis** [76] test, depending on the result obtained, it could be necessary to make a *Post Hoc* test to extract more information of the data collected. For this *Post Hoc* test, we chose a **Dunn's** [19] test, Equations 6.10, 6.11 and 6.12, used often to determine which groups are significantly different.

$$z = \frac{|(\bar{R}_i - \bar{R}_j)|}{s.e.} \tag{6.10}$$

$$\bar{R}_i = \frac{R_i}{n_i} \tag{6.11}$$

$$s.e. = \sqrt{\frac{n(n+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \tag{6.12}$$

Where:

n = the total sample size;

n_i = the size for that group i;

R_i = the rank of the group;

To understand the results that are given by the **Kruskal Wallis** [76] test and why we used the **Dunn's** [19] test afterwards, we need to explain the $chi - square$ formula, equation 6.13.

$$X^2 = \sum \frac{(O-E)^2}{E} \tag{6.13}$$

Where:

O = each Observed value;

E = each Expected value;

6.2 Data

For our work, a **test** was developed for a population of 11 physicians, 5 specialized in breast cancer detection and 6 still doing their specialization course. However, we only took into consideration the data from 10 physicians, given that one did not completed every step of the test. These users are divided into 4 groups: Intern; Junior; Middle; and Senior.

- Intern - Physician without specialization;
- Junior - Physician with specialization and less than 5 years of experience;
- Middle - Physician with specialization and 5 to 10 years of experience;
- Senior - Physician with specialization and more than 10 years of experience;

For our tests, we had 6 Interns, 2 Juniors, 0 Middles and 2 Seniors. All of them have a medical degree with a specialization, or in the making, in radiology, but only 1, a Senior physician, has also both specialties of Senology and Mastology.

Unfortunately, we were not able to have the same representation for each group in these tests, given our inability to contact or have available more physicians.

As it was mentioned in Section 4, we gave a set of four patient cases so that our users could test our functionalities. During those tests, we measured several metrics, two scales, one related with the systems' usability and the other with the systems' workload and three questionnaires, two before the test and one after.

Since the beginning of this project, we wanted to compare and surpass the results of UTA4 [13]. This fourth iteration was the foundation in which we built our functionalities and improved the basic system, thus making it simpler and more user-friendly. In the next sections, we will compare the data and results that we obtained with the ones from that User Testing and Analysis (UTA).

6.3 Measures

6.3.1 Time and Count use of a tool

The time taken to finalize the test was a crucial aspect to take into consideration and is presented in Figure 6.1, however, to compare these times with the ones taken in the previous iteration, a different path was followed. This path focused on the difference of clicks instead of the time per say since there is a difference between the tasks in each iteration.

We know that, in UTA4 [13], opening a 1 x 2 viewport with both images, 4 clicks are necessary, whereas with our new functionalities, the same is done with 1 click [66]. Thus, each action can take up to 2 seconds to be executed, so we are taking half the time to do the same task. In the previous iteration, if we have already the 1 x 2 viewport chosen, the switch of projections takes 2 clicks, whereas with our functionalities, take again half the clicks to do it [13, 66].

			Time		
	Physician	Watched	Begin	End	Total
	2	Yes	29,52	41,06	11,54
	3	Yes	15,44	23,3	7,86
	5	Yes	4,41	12,5	8,09
	6	Yes	7,38	18,22	10,84
	8	Yes	30,52	45,11	14,59
	11	Yes	42,55	53,45	10,9
	15	Yes	32,44	38,24	5,8
	33	Yes	4,48	14,04	9,56
	37	Yes	24,05	50	25,95
	44	Yes	24,07	37,46	13,39
Total	10				**118,52**

Median	10,87
Average	11,852
Mode	

Figure 6.1: The table represents the time that took to make the test. It is important to notice that during the test an open interview was done and this could impact the total time to make the test [11]

The **Recorded View** is one of the most time-saving functionality, given that it enables the maintenance of a state when an image is being manipulated. Regarding the **Temporal View** feature, it made possible to compare 2 images with 2 clicks, but in the older version, it was necessary at least 5 clicks, being one of them a scroll through a list that has similar images [66]. This way, we proved that each feature had an impact in the system regarding the time, with several improvements in different points [66].

6.3.2 Errors

In this section we explain what errors occurred during the tests, for which we re-watched the videos in order to take notes of all errors that could be detected. With this information, we concluded that each physician had the possibility of encountering an average of 2.2 errors, 1.9 non-critical and 0.3 critical, Figure 6.2. Although, in our point of view, this is not a representation of the reality [12].

Usually, when this kind of tests are done, the work environment, the hardware and the software is always the same, so there is the same base to every participant, however, given the pandemic, all our tests were made remotely. Each physician used the hardware/software they had at their disposal, so there were non-supported browsers and devices, some of them used mouses and others did not, and some had 200MB/s of internet connection and others less then 10MB/s which made the download of high-quality images much more complicated.

			Errors									
			Critical		Non-Critical					Total		
	Physician	Watched	Browser	Device	Zoom	Grab	Two finger	Load Error	Annotation	Critical	Non-critical	Totais
	2	Yes	0	0	1	0	0	0	0	0	1	**1**
	3	Yes	0	0	1	0	0	1	1	0	3	**3**
	5	Yes	0	0	1	0	0	0	0	0	1	**1**
	6	Yes	0	0	0	0	0	0	0	0	0	**0**
	8	Yes	1	0	1	0	0	3	0	1	4	**5**
	11	Yes	0	1	0	1	2	1	0	1	4	**5**
	15	Yes	0	0	1	0	0	0	0	0	1	**1**
	33	Yes	0	0	0	0	0	0	0	0	0	**0**
	37	Yes	1	0	3	0	0	2	0	1	5	**6**
	44	Yes	0	0	0	0	0	0	0	0	0	**0**
Total	10		**2**	**1**	**8**	**1**	**2**	**7**	**1**	**3**	**19**	**22**

Median		0	0	1	0	0	0	0	0	1	1
Average		0,2	0,1	0,8	0,1	0,2	0,7	0,1	0,3	1,9	2,2
Mode		0	0	1	0	0	0	0	0	1	1

Figure 6.2: The table represents all errors that occurred in the tests, per physician, and the basic statistic of errors [12].

6.3.2.A Critical

During the tests we had some critical errors, not because of the functionalities implemented but due to the use of non-supported browsers or devices. A total of 3 critical errors were found, with an average of 0.3 [12]. However, we have to consider that this result could be explained by the small population sample used and the fact that the tests were not done in a controlled environment.

Due to the lack of compatibility of devices, like tablets, some functionalities could not be tested, however, this problem was solved with a quick test adjustment in order to have a functional test. Regarding the browser choice, although our project was design for three browser types (Chrome, Firefox, and Edge), one physician used the Safari browser, which created problems when loading and opening images. This physician did not had other browser, so it was not possible to resolve this issue during the test. In this case, we had to show the physician the expected behavior in a supported browser which has rightly indicated the idea that was in place, and thus making it possible to comment the system and use the new functionalities [12].

During the test development we noticed the existence of one critical error, that did not happen during the user test itself. When using the mark tool to make a border of the lesion, if at some point we changed the tool or the viewport, the data would corrupt and could crash the system. This problem needs to be addressed in future UTAs and redesign to prevent this type of behavior.

6.3.2.B Non-Critical

As expected, non-critical errors happened during our tests, with the basic tools and our new functionalities. The total number of errors that occurred during all tests was 22, where 19 were non-critical errors, with an average of 1.9 [12].

The most common non-critical error was related to the **Zoom** tool, a total of 8 errors, were almost every physician that used our system made it. **Zoom** action is done by clicking in the right mouse key

or clicking in the zoom button and using the left mouse key while pulling the mouse towards us or in the opposite direction. The system and the tool by default is registering the pulling towards us as a **Zoom In** and in the opposite direction a **Zoom Out** and every physician made the first use upside down.

The second most common non-critical error was related to loading an image, a total of 7 of these errors occurred. However, these errors happened mostly in tests with critical errors, related to the browser or the device, and one test where no error, critical or non-critical, was made. In one of these cases, a critical error had and impact so severe that images were loaded incorrectly or not at all [12]. Other type of errors had little impact, 2 errors, that only happened when using a tablet and when a physician tried to make an annotation and used the incorrect tool.

6.4 Scales

In this section, the focus will be to present and explain the results that were obtained from testing our functionalities. Each scale was only presented to the user at the end of the study and not after testing each functionality, given that our aim is to use the system as a whole and not individual functionalities. However, this could influence the final results, since each error is counted by the physicians perspective to the entire system.

6.4.1 System Usability Scale - basic statistics

As mentioned before in Section 4.3.3.A, the SUS [14, 15] measures the usability of the system by presenting 10, positive and negative, statements. Here, the user chose one of five points between *strongly disagrees* to *strongly agrees*, which is also known as a 5-points likert-scale.

To analyse this scale and its results, it is necessary to treat separately the positives and negatives statements. Regarding the positive statements, the final result for each statement is the subtraction of one value to the answer given by the physician. In negative statements, the final result for each statement is equal to 5 subtracted by the answer given by the physician. Thus, for each statement, the score is between 0 and 4. Adding then, the final values of the 10 statements, the minimum score is 0 while the maximum score is 40, per user. Furthermore, these scores need to be converted from a scale from 0-40 to a scale from 0-100, for this, the value previously obtained has to be multiplied by 2.5 [61].

Thereby, a final score lower than 67, is considered to be in the *Poor* category, and below 51, in the *Awful* category. A score equal to 68 is in the *OK* category and between 69 to 80.3, in the *Good* category, although it means that there are some improvements that can be done to the system. Finally, a score above 80.3 means that it is in the *Excellent* category, a very user-friendly system.

The results obtained from our 10 physicians are as follows [61]:

Median = 91.25;

Mean = 88.75;

σ = 10.25;

As previously mentioned, to detect outliers, we used the **Tukey Fences** [75] test, Equations 6.2, 6.3, 6.4, 6.5, 6.6. Thus, if a data point is outside the obtained range, it is considered an outlier. When applying these equations to our data, we have the following results:

$$Q1 = 84.375, \quad Q3 = 95 \ , \tag{6.14}$$

$$IQR = Q3 - Q1 = 95 - 84.375 = 10.625 \ , \tag{6.15}$$

$$BelowOutlier < Q1 - 1.5 \times IQR < 84.375 - 1.5 \times 10.625 < 68.4375 \ , \tag{6.16}$$

$$AboveOutlier > Q3 + 1.5 \times IQR > 95 + 1.5 \times 10.625 > 110.9375 \ , \tag{6.17}$$

$$[\ 68.4375 \ ; \ 110.9375\ ] \tag{6.18}$$

Given that the maximum value can not be higher than 100, we will consider that the *Above Outlier* will never be reached, however, any data point below 68.4, will be considered an outlier [61]. Regarding our data, there was found an outlier with a SUS [14, 15] score of 67.5 which enters in the *Poor* category, although very close to the *OK* category. Nevertheless it is an outlier, and will be removed from the statistic. Still, the opinion gathered during the test was taken into consideration.

Updating our results, the removal of this outlier leaves us with the following [61]:

Median = 92.5;

Mean = 91.(1);

σ = 7.648;

The majority of our data points are in the higher score ranges, in the categories *Good* and *Excellent* [61].

In the previous iteration [13], a test was also performed focusing on the usability and workload of the basic system. The difference from that test to ours was the task performed, since they asked the physicians to make a diagnosis, whereas in ours the focus was on testing the functionalities in real cases.

These are the results of that iteration [61]:

Median = 87,5;

Mean = 86.935;

σ = 9,811;

When applying the **Tukey Fences** [75] test equations to find out if there is an outlier:

$$Q1 = 80, \quad Q3 = 95 \ , \tag{6.19}$$

$$IQR = Q3 - Q1 = 95 - 84.375 = 15 \ , \tag{6.20}$$

$$BelowOutlier < Q1 - 1.5 \times IQR < 80 - 1.5 \times 15 < 57.5 \ , \tag{6.21}$$

$$AboveOutlier > Q3 + 1.5 \times IQR > 95 + 1.5 \times 15 > 117.5 \ , \tag{6.22}$$

$$[\ 57.5 \ ; \ 117.5\ ] \tag{6.23}$$

With these results we can conclude that these tests did not had any outliers regarding SUS [14, 15], maintaining the results of Mean, Median and σ previously given [61]. Figure 6.3 represents a bar graph with our work (orange columns) and UTA4 (blue columns) data points, in five step intervals, where the height of the bar indicates how many physicians obtained a score in that five step interval. The columns with a red border are alerting to the outliers in both data samples.

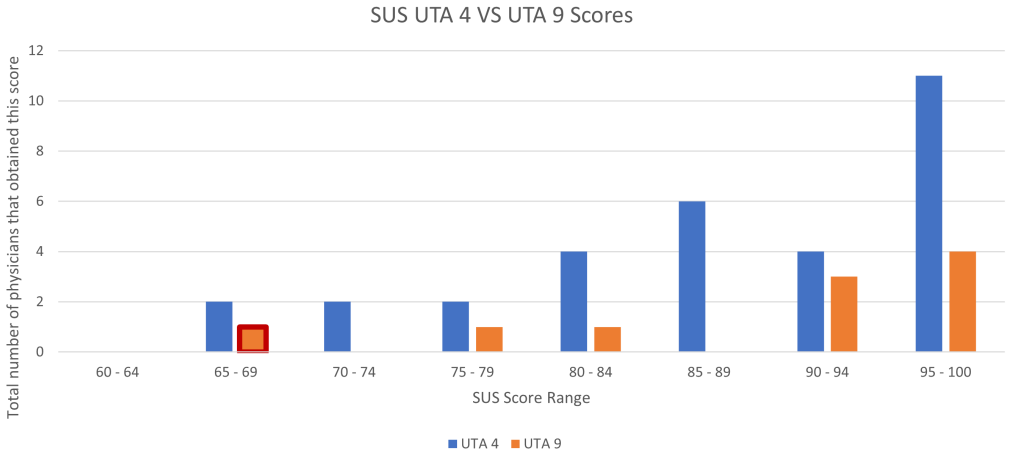


Figure 6.3: Comparison between UTA4 [13] and UTA9. In this graph we can see both data points using the same scale: in the vertical axis, the total number of users that obtained a score in that score range, and, at the horizontal axis, the SUS [14, 15] score ranges with 5 score distance. The previous iteration [13] is represented in blue columns and our work score is represented in orange columns. The red border column means that the data point is an outlier. Both graphs are not-normal distributions and demonstrate the satisfactory results presented in both UTAs.

Comparing both iterations, they both have a SUS [14, 15] score in the *Excellent* category, which shows that in terms of usability we are taking into consideration every idea and correction that was given. This work was able to improve the SUS result in every case, with better mean, median and σ, which means that our data is more compact and closer to the higher categories [61].

6.4.2 NASA Task Load Index - basic statistics

In the Section 4.3.3.B, we presented a scale that aims to measure the workload involved in completing a task, which, in this case, was the manipulation of several images with different functionalities. As in the previous scale, we did not applied this per functionality, but to the entire system.

This scale had 6 questions for the user to answer by using a scale from 0 to 100, where 0 means *Very Low* and 100 means *Very High*. In the case of the performance question, the scale is inverted, thus, 0 corresponds to *Perfect* and 100 corresponds to *Failure*. In these cases, there are 20 marks of answer possibilities, where each mark has an interval of 5 points. Regarding its analysis, we want the opposite of the System Usability Scale [14, 15], so, the closest to 0 the better, meaning the less workload the physician has, the better the system is.

The mathematical process is complicated, but it starts by multiplying each answer given by 5. Then, the questions are compared between each other, following the method in the Table A.1 of the Appendix A. If the value given to one question is higher than the other, that question has one point and the

other zero points, however, if the value given is equal to both, the question that is being compared to the second one receives the point. There are, in total, 15 comparisons where each question adds points [62]. After this, each initial value already multiplied by 5, will be multiplied by the points previously added and summed the total score by each physician. These final values are then divided by 15, giving a result between 0 and 100 for each user [62].

Regarding the interpretation of our results, there are a few categories that we need to follow:

- Low : 0 - 9;
- Medium : 10 - 29;
- Somewhat High : 30 - 49;
- High : 50 - 79;
- Very High : 80 - 100;

From our test and analysis, we had the following results [62]:
Median = 11.1(6)7;
Mean = 17.7(3);
σ = 15.075;

As in the previous scale, there is the need of finding out if there are any outliers, which is done, again, by applying the **Tukey Fences** [75] test. The results are the following:

$$Q1 = 6.25, \quad Q3 = 21.(6)7 \ , \tag{6.24}$$

$$IQR = Q3 - Q1 = 21.(6)7 - 6.25 = 15.41(6)7 \ , \tag{6.25}$$

$$BelowOutlier < Q1 - 1,5 \times IQR < 6.25 - 1.5 \times 15.41(6)7 < -16.875 \ , \tag{6.26}$$

$$AboveOutlier > Q3 + 1.5 \times IQR > 21.(6)7 + 1.5 \times 15.41(6)7 > 44.791(6)7 \ , \tag{6.27}$$

$$[\ -16.875 \ ; \ 44.791(6)7\ ] \tag{6.28}$$

With this test, any value that is outside the interval [-16.875 ; 44.791(6)7], will be considered an outlier. Since any value below 0 is impossible to happen in the test, we will only consider the *Above Outlier*. We identified an outlier with the score 46 [62], that enters in the category *Somewhat High*.

By removing this outlier from our data sample results, we have the following results [62]:
Median = 11;
Mean = 15.037;
σ = 13.186;
In previous iterations [13], the NASA-TLX [16–18] scale obtained the following results [62]:
Mean = 30.925;
Median = 22.(3);

σ = 21.178;

The outliers, were again calculated using the same **Tukey Fences** [75] test:

$$Q1 = 14, \quad Q3 = 41.(6)7 \ , \tag{6.29}$$

$$IQR = Q3 - Q1 = 41.(6)7 - 14 = 27.(6)7 \ , \tag{6.30}$$

$$BelowOutlier < Q1 - 1.5 \times IQR < 14 - 1.5 \times 27.(6)7 < -27.5 \ , \tag{6.31}$$

$$AboveOutlier > Q3 + 1.5 \times IQR > 41.(6)7 + 1.5 \times 27.(6)7 > 83.1(6)7 \ , \tag{6.32}$$

$$[\ -27.5 \ ; \ 83.1(6)7\ ] \tag{6.33}$$

From these results, one outlier was identified with a score of 84.(3) [62], which enters in the *Very High* category, visible in the Figure 6.4.

By removing this outlier from the statistics of that iteration, the results update are the following [62]:

Mean = 29.1(4);

Median = 21.5;

σ = 19.035;

In the bar graph presented in Figure 6.4, it is possible to observe that our data is condensed between the first five ranges, that go from the *Low* to the *Medium* category, and two values far ahead, one in the *Somewhat High* and one outlier. Regarding the previous iteration [13], the results were scattered along the several ranges, with none in the better range [0;5]. Both outliers are presented with a red border, being the orange bar from our work and the blue from the previous iteration.

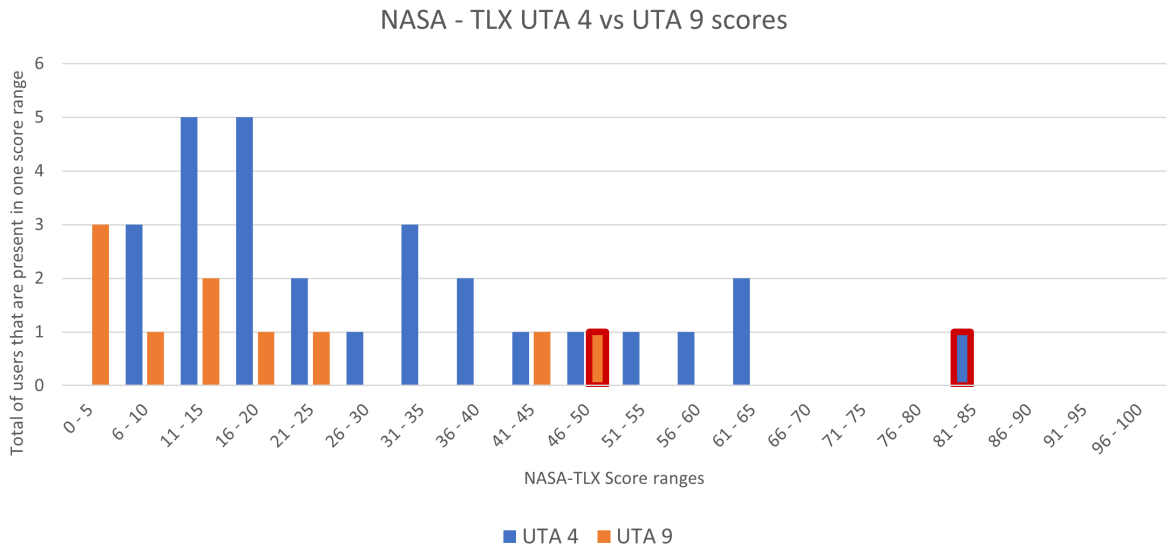


Figure 6.4: Results of the application of the NASA-TLX [16–18] scale in UTA4 [13] and UTA9. In the vertical axis, we have the total number of users that obtained a score in that score range, and in the horizontal axis we have the score ranges by 5 score steps. In this graph, we can observe that we have, again, a non-normal distribution with two outliers, the columns with a red border.

These results show a positive evolution since the previous iteration, where several data points were in the *Somewhat High* or *High* category of workload, meaning that for some tasks, the physician had to make an effort to realize it, whereas in our work, the data points are in the first score ranges, meaning that the level of workload was *Low* or *Medium*.

Although as is possible to observe, our outlier and the last physician with a valid score, both have scores that are already in the range of *Somewhat High*, [30;49], which shows that is necessary yet to refine some of the functionalities and even to have an interview with both physicians in order to understand their difficulties and what is the best path to improve the system.

6.4.3 Advanced statistics

As said in the previous subsections, our data does not follow a normal distribution, given that we have little results, which are not converging. However, we have several spikes in our data, and used statistics tests in order to understand them better.

We had several mechanisms that could be used to make sense of this data, but we chose the **Kruskal Wallis** [76] test, that explains if the medians of two or more groups are different or not, equations 6.7,6.8 and 6.9.

This test has two Hypotheses, the $H0$ referring that the population medians are equal, and the $H1$ referring that there is not enough evidence to suggest that the medians are unequal. This test also shows if there is a significant difference between groups, however, it does not explain which ones are different. In this case, we need to do a *Post Hoc* test that we will explain next.

For each SUS [14, 15] and NASA-TLX [16–18], we have the following results, when running the **Kruskal Wallis** [76] test, for the same data sample size [77]:

$$c = 10 \quad , \quad chi - square = 16.919 \quad ; \tag{6.34}$$

Where:

c = sum of physicians;

$chi - square$ = value of the data sample for 1 degree of freedom and α = 0.05 for that population size;

From these values, if the statistic H of the **Kruskal Wallis** [76] test is higher then the $chi - square$, the null hypothesis is rejected, $H0$, if not, the $H1$ is rejected. If the null hypothesis is rejected a *Post Hoc* test is necessary to explain the differences between groups.

These are the results for **Kruskal Wallis** [76] test [77]:

$$\text{SUS UTA9} \quad H = 1.407 \ , \quad p - value = 0.495 \ , \tag{6.35}$$

$$\text{SUS UTA4} \quad H = 3.213 \ , \quad p - value = 0.201 \ , \tag{6.36}$$

$$\text{NASA-TLX UTA9} \quad H = 3.652 \ , \quad p - value = 0.161 \ , \tag{6.37}$$

$$\text{NASA-TLX UTA4} \quad H = 0.659 \ , \quad p - value = 0.719 \ , \tag{6.38}$$

In order to accept $H0$, the chi-square value needs to be bellow the statistic value, which is not the case. Therefore, we conclude that we do not have the ability of saying that the groups have equal medians, so they are not similar. However, this does not tell us much, and we need to perform a *Post Hoc* test to know how unequal they are. For this we chose the **Dunn's** [19] test, Equations 6.10, 6.11, 6.12, that can be used to pinpoint which specific medians are significant in non-parametric data sets.

This *Post Hoc* test has a Null Hypothesis, $H0$, meaning that there is no difference between groups, whereas Hypothesis $H1$, shows that difference. **Dunn's** [19] test has different ways to be calculated with adjustments, though none of the adjustments seemed good to be used, so we did the test with no adjustments.

The test results are presented in Table A.3 of Appendix A [77]. The negative values of -1, are present when the same group is compared; a value equal to 0 means that the groups are similar; and the closer to 1, the most dissimilar they are. With our results, we understood that our groups are not similar, although some are more similar than others (*e.g.*, Senior group and Intern group in the NASA-TLX in our work, 0.0560, and the same groups in System Usability Scale in previous iteration, 0.096), proving the hypothesis $H1$. This result supports the **Kruskal Wallis** [76] test, where we were not able to say that the groups were similar.

6.5 Questionnaires

During this thesis, three questionnaires were made, but only two of them were about the system itself. In the first questionnaire, done before the test, the focus was on the state of currently used systems and what new functionalities the users were interested in being implemented, whereas the second questionnaire, made after the test, was regarding the functionalities tested and what was the users' opinion.

6.5.1 Interaction Tools Questionnaire

First and foremost, it was necessary to understand what type of tools are already used in current systems, so this first questionnaire started by presenting images that represented different tools, where the users had to pick whichever was applied to their systems. The bar graph presented in Figure 6.5 shows the answers given: all the users had the **Zoom** tool and could **Measure the lesion size**; 9 out of 10 had **Luminosity** and the **Invert** tool; and 7 or less had the **Move** tool, **Local Zoom** and **Lesion annotations** [52]. This questionnaire is in the Appendix B of this document.

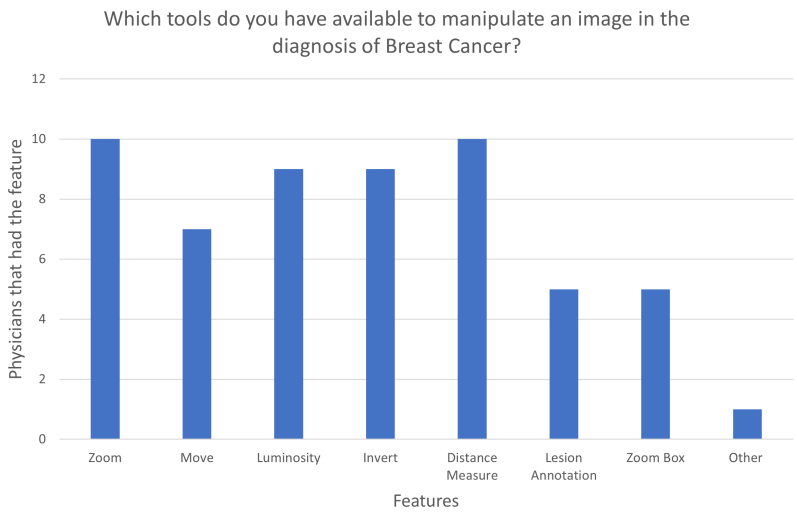


Figure 6.5: Present situation in systems used in Hospitals, with a count per tool of what is present in each institute were our physicians work.

Secondly, it was necessary to understand what physicians wanted to be implemented in the systems, so the questionnaire presented several options that were discussed in previous tests. Regarding the possibility to compare the medical images simultaneously in order to analyze the lesion evolution, two choices were presented: the lesion evolution with a line border and the comparison between breasts from different time periods, both accepted by all users. With their feedback regarding these ideas and the Focus Group, the **Temporal View** functionality was created [35–38, 52]. 9 out of 10 physicians had also the necessity of having some kind of automation, when opening or manipulating an image, so that the opposite image could change as well, so the **Coordinated View** functionality was created [52].

Another option that had a high number of acceptance, 8 out of 10, was the possibility of having a **MRI 3D space awareness**, *i.e.*, knowing the position of the slice in the human body. The least wanted options, were a **3D representation of the lesion**, **Re-adjustment of the breast** to the center of the screen and **Automatically zoom In/Out** [52].

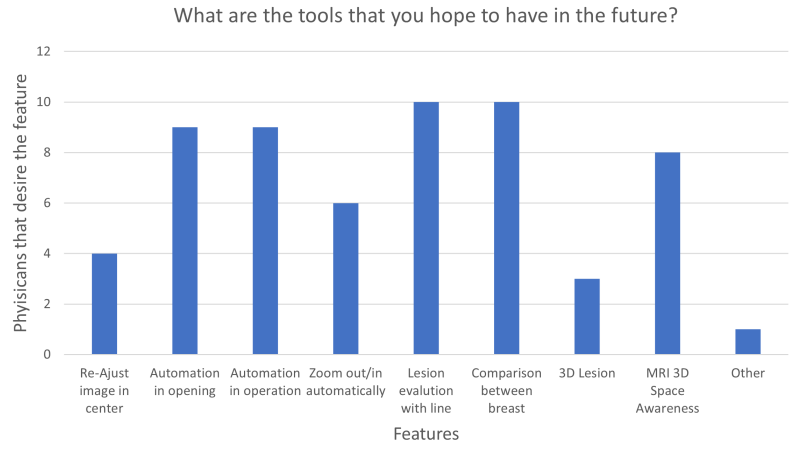


Figure 6.6: The score of the functionalities that are desired by physicians to be implemented in the system.

Lastly, we asked physicians to give a priority to this choice, where the most voted functionality was the lesion comparison, our **Temporal View**; then, the automation both in opening and manipulating images, our **Coordinated View**; and the **MRI 3D space awareness** [52]. **Recorded View** feature was not discussed in these questionnaires, but chosen during the **Focus Groups** by asking physicians how important it was for them to be able to revisit images and analyse them again with the image manipulation already done, quote 5.5.1. Since this is a basic tool, we chose to prioritize it instead of the **MRI 3D space awareness**.

6.5.2 Interaction Tools Pos-Tasks Questionnaire

After the execution of the test, a questionnaire was given to physicians so they could rate, from a 1 (*Dislike*) to 5 (*Like*) likert-scale, and point out their likes and dislikes about the functionalities tested: **Recorded View**; **Coordinated View**; and **Temporal View**. Here, all the three functionalities obtained a score of 4 or 5 and, in all cases, the maximum score was given by, at least, 70% of the users [78]. This questionnaire is also in the Appendix B of this document.

With this questionnaire we obtained several comments for each functionality tested. Regarding the **Recorded View**, the vast majority wanted to have the possibility of reset the image to its original state regarding the luminosity, because sometimes it was necessary to analyze again the images without the changes already done. In the **Coordinated View**, the comments were mostly positive, with a single change, a reset but now focused only on the **Zoom** and **Pan/Move** tools [78].

> ***"It makes sense to have different resets, for the zoom and luminosity."*** (...)
> ***"It would be perfect to have the possibility to make the reset in just one image or in both, depending of the presented situation."***
>
> Intern Physician

Finally, in the **Temporal View**, physicians showed interest in having a second time bar in the other image which could help compare older exams side by side. Also, they would like to have the possibility of having a choice of side in a settings area, and have this functionality work not only in 1 x 2 viewport but also in a 2 x 2 viewport [78].

6.6 Design Goals

In Section 3.2 was presented three **Design Goals** that would be answered by the results given in this chapter. Those three goals are the following: **Usability**; **Efficiency**; and **Productivity**. For the first goal, the System Usability Scale [14, 15] was used, where we aimed to obtain a value above 86.9, that would mean an improvement from the fourth iteration [13]. Since our results showed a total score of 91.(1), it means that our system is in the *Excellent* category and above our target.

The **Efficiency** goal was explored with two approaches, first reducing the times by using a **Time** and **Count Use of a tool** metric, and second, the total number of **non-critical errors** which had to be less

than 6 for each physician. For the first approach, we aimed to reduce the time of interactions, where any type of improvement would be sufficient to reached it. In Section 6.3.1, we demonstrate that with our new functionalities this was possible, since the number of clicks necessary are reduced to half of what existed, consequently, reducing the overall diagnosis time. Regarding the second approach, we demonstrate in Section 6.3.2, that our mean value is equal to 1.9 in **non-critical errors**, a better result that was proposed as a goal. With these two approaches proven, we can conclude that another goal was met.

Lastly, the **Productivity** goal was explored with two approaches as well, being one the **Time** and **Count Use of a tool** metrics, and the second the data from the NASA-TLX [16–18] to evaluate the workload of the system. For this goal the set a score lower than 29.1(4), the fourth iteration [13] score. In Section 6.4.2 we presented the data that we collected from our tests, where we achieved a score of 14.593, indicating a *Medium* workload, and proving, once again, our goal.

6.7 Research Questions

Here we explore how we evaluated the research questions mentioned in Section 3.4, basing our conclusions on the data gathered during the **Design Process**.

The first research question, **RQ.1**, "When and who will use the functionalities", will be answered with the results obtained in the questionnaires, **Interaction Tools** and **Interaction Tools Pos-Tasks**, and based on the comments given by the participants. With these questionnaires we understood which functionalities the users wanted to see implemented in the system and what was their opinion after the test. The results from the post-task questionnaire showed us that our functionalities were well received, since all of them registered a 4 or 5 classification (in a scale form 1 to 5). However, it is still impossible to determine if our functionalities would indeed be used, if they were available. We can prove that with this information, our scale results and with the opinions given by the physicians, it is probable that the hypothesis **H1.1** would be rejected. Hypothesis **H1.2** is also discarded because all functionalities developed can be used in any situation, since they are not specific to any type of difficulty. We can also conclude that, regardless of the level of the physicians' expertise, the functionalities could be used, discarding the hypothesis **H1.3**. Finally, the hypothesis **H1.4** is the one accepted for this research question, since all the functionalities were, as previously said, well received.

The hypothesis from the second research question, **RQ.2**, "What is the impact of the functionalities, in the clinic workflow?", can be validated with the responses given in both scales and metrics, **Time** and **Count use of a tool**. Regarding the hypothesis **H2.1**, the results from the System Usability Scale [14, 15] showed a condensed data in the higher ranges, categories *Good* and *Excellent*, proving that the usability has increased from the previous work [13] thus accepting this hypothesis. The hypothesis **H2.2**, was meant to understand how the workload was affected with the introduction of the new functionalities. The NASA-TLX [16–18] showed that we obtained better results than in previous UTAs [13], with the majority of the data points in the first five ranges, *Low* or *Medium* workload, which represents a good evolution. With these results we proved the reduction of the workload present in the system, also accepting this

hypothesis. The hypothesis **H2.3**, focused on the time of a task by reducing the clicks necessary to make an action which is demonstrated by the automation that the functionalities provide, proving that the reduction of some steps can reduce the total time.

The third and last research question, **RQ.3**, "What is the best method to represent the lesion evolution?", explores the best representation for the **Temporal View** feature. We presented in the MCs project [72] a low-fidelity prototype, where we took a screenshot of the system and added annotations on top of the lesion with a timeline that could change the annotation in order to compare the lesion, however, this low-fidelity prototype showed us that this was not the right way to proceed. In the **Focus Groups** and **interviews**, physicians told us that they prefer to see the images side by side from different dates. With this type of configuration, even breast asymmetry over-time could be seen, rejecting hypothesis **H3.1**. The time bar hypothesis, **H3.2**, **H3.3** and **H3.4**, are opposite to each other, so accepting one will reject the others. Physicians gave their opinions while doing the test, and some of them were the following quotes:

> ***"I would prefer to have the most recent image in the left side, I have this in my daily system "***.
>
> Intern Physician

> ***"It is essential to have the most recent at the right side"***.
>
> Senior Physician

> ***"It is not common to compare two past images of breasts but could be a necessity to have that possibility."***.
>
> Senior Physician

We can see, with these quotes, that physicians have different ways to approach the problem, however, in the end, all of them agreed that having the possibility of two time bars, one in each side, could resolve the problem. Thus, the hypothesis **H3.2** and **H3.3** were rejected and the hypothesis **H3.4** was accepted. Also, hypothesis **H3.5**, is rejected by the acceptance of the **H1.4**.

7

Conclusion and Future Work

Contents

7.1 Conclusion

Breast cancer is one of the most usual cancers regarding women, with a mortality rate of nearly 13%, although with little incidence in men. This thesis purpose was to **enhance tools** that aim to help physicians when making a patient diagnosis, by simplifying the process and, consequently, making it faster.

Previous work from the **Medical Imaging Multimodality Breast Cancer Diagnosis User Interface (MIMBCD-UI) project** [13], not only built a base system for this thesis, but proceeded to make some groundbreaking progress in the AI [20] field by enabling the medical data analysis and providing results. This thesis is a small but significant part of the development of this major project, though, instead of focusing on the AI part, we identified the **physicians' needs** regarding basic tools that are not present in systems currently used in hospitals.

Following a **Design Thinking** [27] process, a **HCI technique**, the first stage was understanding what already existed in prior iterations by analyzing their videos, where problems were found that could have an impact in breast cancer diagnosis. Then, the problems were defined, and ways to address them were studied in the second and third stages. Here, we had **Focus Groups** [7, 35–38] where we discussed the problems and created the ideas to resolve them. From the **Focus Groups** and **questionnaires**, we chose three functionalities to develop: **Recorded View**, **Coordinated View** and **Temporal View**. The fourth stage was the prototyping, where these functionalities were developed. Even in this process other changes were necessary besides the main functionalities that were designed to be implemented.

The fifth and final stage was the testing, were all functionalities and small interface changes were used, analyzed and criticize by our experts. Given our results and the feedback from the participants, we can conclude that we made the desired and necessary functionalities focused on the final users. Each test was made with the intent of demonstrating the capabilities of the system in real-life scenarios, showing us that our mental process of executing some tasks were correct and that we managed to anticipate some problems that other programs have. In these tests, we had a better performance, with a higher efficiency and productivity, when comparing to the prior iteration [13]. Regarding the scales measured, we had excellent results in SUS [14, 15] with an average score of 91.(1) out of 100, which enters in the *Excellent* category, and good results in the NASA-TLX [16–18], where we had an average of 14.593, a result that is considered *Medium*. When analyzing the click patterns with the **Count use of a tool** metric, it is almost certain that the time per patient was proved to be reduced. The **Errors** metric can give us a sense of the systems' performance, given that having fewer errors leads to a better path to the final result. Since in the course of each tests, an interview was made, it is impossible to measure this kind of ideal path to a result, however, we were able to identify different types of errors that were done, correcting them for future work. After concluding the **Design Thinking** [27] process, with all the data gathered, we were able to prove all the **Design Goals** and **Research Questions**.

With this thesis, an iteration of the MIMBCD-UI [13] project, we made a huge step forward in giving users' functionalities that could, eventually, help in daily cases. Although this project is not actually in use in hospitals, it does apply present technology, making the system far more advanced when compared to some of the current systems available.

7.2 Future Work

The development of this thesis allowed the creation of a more robust system which will be the base for future iterations of the MIMBCD-UI [13] project. The combination of this project, the seventh iteration [20] and the eXplainability [22] of the AI, will form a CADx system, enabling lesion identification that could not be seen, along with a BI-RADS classification result.

7.2.1 New functionalities

In the course of the **Design Thinking** process, Chapter 4, we discovered desirable functionalities that were not developed. Being the most wanted, the **MRI 3D space awareness**, Figure 7.1, that allows to perceive a 2D lesion slice of the MRI modality in a 3D way, by identifying the location of that slice in the body.

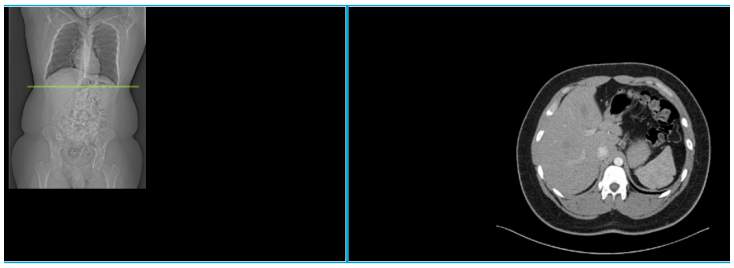


Figure 7.1: An example of a **MRI 3D Space Awareness**, called by the **CornerstoneJS** team as a reference line tool, where in the left is shown a Z axis representation and in the right the slice of the MRI with the X and Y axis represented.

7.2.2 Existent tools and system

There were also problems found in the system that were not thought of during the developed process as a necessity. In the overall system, we can focus on two aspects: the *patients' page*, which contains all patients' available to be diagnosed; and the *diagnostic page*, which contains the images and data for a particular patient.

In the *patients' page*, a set of patients (Rows) is loaded so it is possible for the physician to choose the case that is to be analyzed. In each study, there is several columns that represent information about that study, Figure 7.2. These columns can be improved so that they could hold more useful information for the physician. However, in order to determine which information would be more useful in a real case scenario, it would be necessary to do more tests with the physicians.

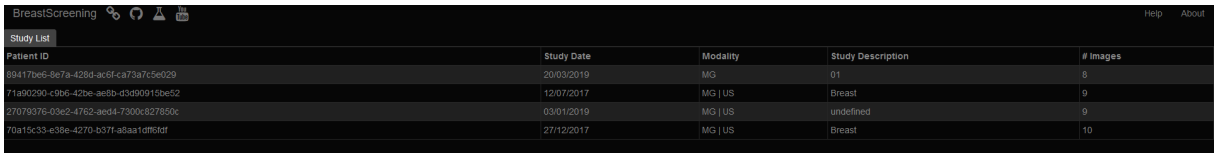


Figure 7.2: An actual representation of the information available to the physician before opening the image.

Regarding the *diagnostic page*, several points were addressed in order to improve the quality of the system, starting with the information presented in the breasts list, Figure 7.3, where, in this work, was added the projection information for each MG modality, being also necessary the addition of the laterality information of the MG. In this list some physicians had difficulties to pinpoint some of the exams from a specific date, so it was suggested to have a container to group the images by date.

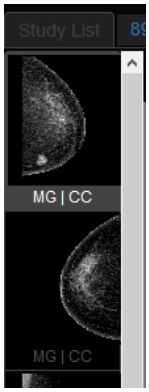


Figure 7.3: An actual representation of the list of images that are at the disposal of the physician to choose and evaluate, this has the basic information and are not organized.

It is also important, though not critical, to mitigate one other problem the current system has regarding the **Zoom** tool. Although we did not receive any comment in this regard, we noticed that every physician used this tool upside down in the first time, in other words, when they wanted to make **Zoom In** they end up making **Zoom Out**. The suggestion of improving this tool was made by one physician, by incorporating it into the automated tool and focusing on a quadrant, Figure 7.4 and 7.5, with different behaviours depending of the breast that are loaded.

> ***"Having an automated zoom will be interesting."***
>
> ***"It is normal to have the desire to be able to compare both, new and old, breasts and opposite breast in terms of zoom by quadrants."***
>
> Intern Physician

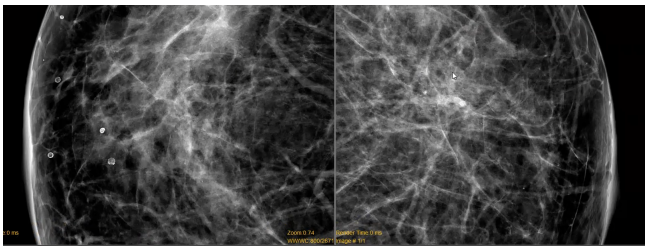


Figure 7.4: An example of automated zoom in opposite images, while in the left we decrease the X axis value, in the right image we increase it in order to show the same quadrant.

Regarding the developed functionalities and starting with the **Record View** functionality, although we made everything that was asked in the interviews, we discovered that, in some cases, there is the necessity of being able to reset all the transformations done to an image. This reset could be total (*e.g., Luminosity, Zoom and Pan*) or in individual aspects.

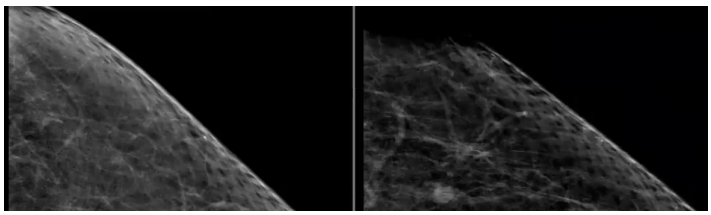


Figure 7.5: An example of automated zoom in similar images, maintaining the same values for the translation, in some ways, coping one value to the other.

For the **Temporal View**, there is the necessity of developing a personal settings file that the physician needs to pre-fill, so that the most recent image available opens on their favorite side for analysis. Several physicians also think that the created time bar needs to be in every viewport, in order to change the most recent image with older ones, if necessary. The third and final aspect was the possibility of having this functionality working with 2 x 2 viewport, to enable the analysis of both projections, CC and MLO, in older images and the most recent ones, at the same time.

With this thesis, improvements suggested and with the fusion with UTA7 [20] and UTA10 [22], we will have a full, but bare, CADx system, that can have a serious impact in the breast cancer diagnosis, making it safer, faster and, consequently, helping saving more lives.

Bibliography

[1] "BIRADS For Mammography," https://radiologyassistant.nl/breast/bi-rads-for-mammography-and-ultrasound-2013#introduction-standard-reporting, 2014.

[2] P. Mildenberger, M. Eichelberg, and E. Martin, "Introduction to the dicom standard," *European radiology*, vol. 12, no. 4, pp. 920–927, 2002.

[3] "BIRADS For Mammography," https://radiologyassistant.nl/breast/bi-rads-for-mammography-and-ultrasound-2013, 2014.

[4] N. Mourão and H. Lencastre, "Mg cc and mg mlo projections," 2020. [Online]. Available: http://rgdoi.net/10.13140/RG.2.2.19033.34409

[5] "Breast Calcifications Differential Diagnosis," https://radiologyassistant.nl/breast/breast-calcifications-differential-diagnosis, 2008.

[6] Macrovector_official, "Business vector created by macrovector_official: Freepik.com," 2020. [Online]. Available: https://www.freepik.com/free-photos-vectors/business

[7] H. Lencastre, "UTA9: Affinity Diagrams dataset," 2020. [Online]. Available: https://mimbcd-ui.github.io/dataset-uta9-ad/

[8] H. Lencastre, F. M. Calisto, and J. C. Nascimento, "Recorded view feature," 2020. [Online]. Available: http://rgdoi.net/10.13140/RG.2.2.22788.07048

[9] H. Lencastre, F. M. Calisto, and J. C. Nascimento, "Coordinated view feature," 2020. [Online]. Available: http://rgdoi.net/10.13140/RG.2.2.21110.34888

[10] H. Lencastre, F. M. Calisto, and J. C. Nascimento, "Temporal view feature," 2020. [Online]. Available: http://rgdoi.net/10.13140/RG.2.2.11882.88000

[11] H. Lencastre, "UTA9: Time dataset," 2020. [Online]. Available: https://mimbcd-ui.github.io/dataset-uta9-time/

[12] H. Lencastre, "UTA9: Errors dataset," 2020. [Online]. Available: https://mimbcd-ui.github.io/dataset-uta9-errors/

[13] F. M. 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. [Online]. Available: http://rgdoi.net/10.13140/RG.2.2.15187.02084

[14] F. M. Calisto and J. C. Nascimento, “Medical imaging multimodality breast cancer diagnosis user interface: Sus survey template file,” 2018. [Online]. Available: 10.13140/RG.2.2.24758.86088

[15] J. Brooke *et al.*, “Sus-a quick and dirty usability scale,” *Usability evaluation in industry*, vol. 189, no. 194, pp. 4–7, 1996.

[16] F. M. Calisto and J. C. Nascimento, “Medical imaging multimodality breast cancer diagnosis user interface: Nasa-tlx survey template file,” 2018. [Online]. Available: http://rgdoi.net/10.13140/RG.2.2.26978.79044

[17] S. G. Hart, “Nasa-task load index (nasa-tlx); 20 years later,” *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 50, no. 9, pp. 904–908, 2006. [Online]. Available: https://doi.org/10.1177/154193120605000909

[18] A. Ramkumar, P. J. Stappers, W. J. Niessen, S. Adebahr, T. Schimek-Jasch, U. Nestle, and Y. Song, “Using goms and nasa-tlx to evaluate human–computer interaction process in interactive segmentation,” *International Journal of Human–Computer Interaction*, vol. 33, no. 2, pp. 123–134, 2017.

[19] O. J. Dunn, “Multiple comparisons among means,” *Journal of the American statistical association*, vol. 56, no. 293, pp. 52–64, 1961.

[20] F. M. Calisto, “Assistant introduction: User testing guide for a comparison between multi-modality and ai-assisted systems,” Instituto Superior Técnico, Tech. Rep. 7, 04 2019.

[21] F. M. Calisto, P. Miraldo, N. J. Nunes, and J. C. Nascimento, “Breastscreening: A multimodality diagnostic assistant,” 2018. [Online]. Available: http://rgdoi.net/10.13140/RG.2.2.29816.70409

[22] N. Mourão, “Master project: 2d breast cancer diagnosis explainable visualizations,” Master’s thesis, Instituto Superior Técnico, Avenida Rovisco Pais 1, 1049-001 Lisboa - Portugal (EU), 01 2020, introduction of explainable methods on medical decision making. [Online]. Available: http://rgdoi.net/10.13140/RG.2.2.31605.93928/4

[23] L. Marques, “Cancro da mama,” *Revista Portuguesa de Medicina Geral e Familiar*, vol. 19, p. 463–468, 2003. [Online]. Available: https://rpmgf.pt/ojs/index.php/rpmgf/article/view/9976/9714

[24] “Breast cancer,” https://www.wcrf.org/dietandcancer/breast-cancer, 2018.

[25] W. H. Organization *et al.*, *WHO position paper on mammography screening*. World Health Organization, 2014.

[26] C. Allemani, H. K. Weir, H. Carreira, R. Harewood, D. Spika, X.-S. Wang, F. Bannon, J. V. Ahn, C. J. Johnson, A. Bonaventure, R. Marcos-Gragera, C. Stiller, G. Azevedo e Silva, W.-Q. Chen, O. J. Ogunbiyi, B. Rachet, M. J. Soeberg, H. You, T. Matsuda, M. Bielska-Lasota, H. Storm, T. C. Tucker, and M. P. Coleman, "Global surveillance of cancer survival 1995–2009: analysis of individual data for 25.676.887 patients from 279 population-based registries in 67 countries (concord-2)," *The Lancet*, vol. 385, no. 9972, pp. 977 – 1010, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0140673614620389

[27] T. Brown *et al.*, "Design thinking," *Harvard business review*, vol. 86, no. 6, p. 84, 2008.

[28] M. Beaudouin-Lafon, "An overview of human-computer interaction," *Biochimie*, vol. 75, no. 5, pp. 321–329, Jan. 1993. [Online]. Available: https://doi.org/10.1016/0300-9084(93)90165-o

[29] M. Fonseca, P. Campos, and D. Gonçalves, *Introdução ao Design de Interfaces*. FCA, Portugal, Lisbon, 2012, vol. 1.

[30] B. Hatscher, A. Mewes, E. Pannicke, U. Kägebein, F. Wacker, C. Hansen, and B. Hensen, "Touchless scanner control to support MRI-guided interventions," *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, no. 3, pp. 545–553, Sep. 2019. [Online]. Available: https://doi.org/10.1007/s11548-019-02058-1

[31] L. C. Li, C. D. Shaw, D. Lacaille, E. Yacyshyn, C. A. Jones, C. Koehn, A. M. Hoens, J. Geldman, E. C. Sayre, G. G. Macdonald, J. Leese, and N. Bansback, "Effects of a web-based patient decision aid on biologic and small-molecule agents for rheumatoid arthritis: Results from a proof-of-concept study," *Arthritis Care & Research*, vol. 70, no. 3, pp. 343–352, Feb. 2018. [Online]. Available: https://doi.org/10.1002/acr.23287

[32] M. M. Garvelink, L. Boland, K. Klein, D. V. Nguyen, M. Menear, H. L. Bekker, K. B. Eden, A. LeBlanc, A. M. O'Connor, D. Stacey, and F. Légaré, "Decisional conflict scale use over 20 years: The anniversary review," *Medical Decision Making*, vol. 39, no. 4, pp. 301–314, May 2019. [Online]. Available: https://doi.org/10.1177/0272989x19851345

[33] M. W. Battersby, A. Ask, M. M. Reece, M. J. Markwick, and J. P. Collins, "The partners in health scale: The development and psychometric properties of a generic assessment scale for chronic condition self-management," *Australian Journal of Primary Health*, vol. 9, no. 3, p. 41, 2003. [Online]. Available: https://doi.org/10.1071/py03022

[34] R. H. Osborne, G. R. Elsworth, and K. Whitfield, "The health education impact questionnaire (heiQ): An outcomes and evaluation measure for patient education and self-management interventions for people with chronic conditions," *Patient Education and Counseling*, vol. 66, no. 2, pp. 192–201, May 2007. [Online]. Available: https://doi.org/10.1016/j.pec.2006.12.002

[35] N. Mourão, H. Lencastre, and F. M. Calisto, "Mimbcd-ui uta8 - focus group - transcript," 06 2020. [Online]. Available: http://rgdoi.net/10.13140/RG.2.2.26477.20963

[36] N. Mourão, H. Lencastre, and F. M. Calisto, “Mimbcd-ui uta9 - focus group - transcript,” 06 2020. [Online]. Available: http://rgdoi.net/10.13140/RG.2.2.23750.91207

[37] N. Mourão, H. Lencastre, and F. M. Calisto, “Mimbcd-ui uta10 - focus group - transcript,” 06 2020. [Online]. Available: http://rgdoi.net/10.13140/RG.2.2.20395.46881

[38] N. Mourão, H. Lencastre, and F. M. Calisto, “Mimbcd-ui uta11 - focus group - transcript,” 06 2020. [Online]. Available: http://rgdoi.net/10.13140/RG.2.2.12006.86085

[39] B. G. Stuijfzand, M. F. van der Schaaf, F. C. Kirschner, C. J. Ravesloot, A. van der Gijp, and K. L. Vincken, “Medical students' cognitive load in volumetric image interpretation: Insights from human-computer interaction and eye movements,” *Computers in Human Behavior*, vol. 62, pp. 394–403, Sep. 2016. [Online]. Available: https://doi.org/10.1016/j.chb.2016.04.015

[40] B. P. Bailey and S. T. Iqbal, “Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management,” *ACM Transactions on Computer-Human Interaction*, vol. 14, no. 4, pp. 1–28, Jan. 2008. [Online]. Available: https://doi.org/10.1145/1314683.1314689

[41] F. G. W. C. Paas, “Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach.” *Journal of Educational Psychology*, vol. 84, no. 4, pp. 429–434, Dec. 1992. [Online]. Available: https://doi.org/10.1037/0022-0663.84.4.429

[42] O. Palinko, A. L. Kun, A. Shyrokov, and P. Heeman, “Estimating cognitive load using remote eye tracking in a driving simulator,” in *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications - ETRA '10*. ACM Press, 2010. [Online]. Available: https://doi.org/10.1145/1743666.1743701

[43] R. Zheng and A. Cook, “Solving complex problems: A convergent approach to cognitive load measurement,” *British Journal of Educational Technology*, vol. 43, no. 2, pp. 233–246, Mar. 2011. [Online]. Available: https://doi.org/10.1111/j.1467-8535.2010.01169.x

[44] S. Gosling, R. Scott, C. Greenwood, P. Bouzy, J. Nallala, I. D. Lyburn, N. Stone, and K. Rogers, “Calcification microstructure reflects breast tissue microenvironment,” *Journal of Mammary Gland Biology and Neoplasia*, vol. 24, no. 4, pp. 333–342, 2019. [Online]. Available: https://doi.org/10.1007/s10911-019-09441-3

[45] D. H. Levine, P. P. Ghoroghchian, J. Freudenberg, G. Zhang, M. J. Therien, M. I. Greene, D. A. Hammer, and R. Murali, “Polymersomes: A new multi-functional tool for cancer diagnosis and therapy,” *Methods*, vol. 46, no. 1, pp. 25 – 32, 2008, single molecule detection and therapeutic technologies. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1046202308000820

[46] C. Balleyguier, S. Ayadi, K. Van Nguyen, D. Vanel, C. Dromain, and R. Sigal, “Birads™ classification in mammography,” *European journal of radiology*, vol. 61, no. 2, pp. 192–194, 2007.

[47] S. M. Stuij, “Usability evaluation of the kinect in aiding surgeon computer interaction,” Master’s thesis, University of Groningen, 2013. [Online]. Available: https://fse.studenttheses.ub.rug.nl/11090/

[48] P. Å.krlj, C. Bohak, J. Guna, and M. Marolt, “Usability evaluation of input devices for navigation and interaction in 3d visualisation,” *International SERIES on Information Systems and Management in Creative eMedia (CreMedia)*, no. 2014/1, pp. 19–23, 2015. [Online]. Available: http://www.ambientmediaassociation.org/Journal/index.php/series/article/view/222

[49] J. Nielsen, “Finding usability problems through heuristic evaluation,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI ’92. New York, NY, USA: ACM, 1992, pp. 373–380. [Online]. Available: http://doi.acm.org/10.1145/142750.142834

[50] W. Jorritsma, F. Cnossen, and P. Van Ooijen, “Merits of usability testing for pacs selection,” *International journal of medical informatics*, vol. 83, 10 2013.

[51] F. Paternò, *Model-based design and evaluation of interactive applications*. Springer, 01 2001, vol. 10.

[52] H. Lencastre, “UTA9: Questionnaire Before Test dataset,” 2020. [Online]. Available: https://mimbcd-ui.github.io/dataset-uta9-qbt/

[53] M. Sousa, D. Mendes, S. Paulo, N. Matela, J. Jorge, and D. S. Lopes, “VRRRRoom,” in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, May 2017. [Online]. Available: https://doi.org/10.1145/3025453.3025566

[54] L. Yu, P. Svetachov, P. Isenberg, M. H. Everts, and T. Isenberg, “Fi3d: Direct-touch interaction for the exploration of 3d scientific visualization spaces,” *IEEE transactions on visualization and computer graphics*, vol. 16, no. 6, pp. 1613–1622, 2010.

[55] F. M. Calisto, A. Ferreira, J. C. Nascimento, and D. Gonçalves, “Towards touch-based medical image diagnosis annotation,” in *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces*, ser. ISS ’17. New York, NY, USA: ACM, 2017, pp. 390–395. [Online]. Available: http://doi.acm.org/10.1145/3132272.3134111

[56] F. M. Calisto, “It-medex closing workshop: Towards touch-based medical image diagnosis annotation,” 2019. [Online]. Available: http://rgdoi.net/10.13140/RG.2.2.30479.43682

[57] C. Lundström, T. Rydell, C. Forsell, A. Persson, and A. Ynnerman, “Multi-touch table system for medical visualization: Application to orthopedic surgery planning,” *IEEE transactions on visualization and computer graphics*, vol. 17, pp. 1775–84, 12 2011.

[58] C. Ardito, P. Buono, M. F. Costabile, and R. Lanzilotti, “Two different interfaces to visualize patient histories on a pda,” in *Proceedings of the 8th Conference on Human-computer Interaction with Mobile Devices and Services*, ser. MobileHCI ’06. New York, NY, USA: ACM, 2006, pp. 37–40. [Online]. Available: http://doi.acm.org/10.1145/1152215.1152223

[59] S. Grange, T. Fong, and C. Baur, "M/oris: a medical/operating room interaction system," in *Proceedings of the 6th international conference on Multimodal interfaces*. ACM, 2004, pp. 159–166.

[60] "EraBot," https://paxerahealth.com/products/pacs-system/web-based-pacs/, 2019.

[61] H. Lencastre, "UTA9: SUS dataset," 2020. [Online]. Available: https://mimbcd-ui.github.io/dataset-uta9-sus/

[62] H. Lencastre, "UTA9: NASA-TLX dataset," 2020. [Online]. Available: https://mimbcd-ui.github.io/dataset-uta9-nasa-tlx/

[63] C. Meinel, L. Leifer, and H. Plattner, *Design Thinking : Understand, Improve, Apply*. Springer, Berlin, Heidelberg, 2011, vol. 86.

[64] H. Lencastre, "UTA9: Demographic dataset," 2020. [Online]. Available: https://mimbcd-ui.github.io/dataset-uta9-demographic/

[65] C. Delgado and L. Wolf, "Time on task: Perceived and measured time in online courses for students and faculty," *Journal of Nursing Education and Practice*, vol. 7, no. 5, 2017.

[66] H. Lencastre, "UTA9: Count use of a tool dataset," 2020. [Online]. Available: https://mimbcd-ui.github.io/dataset-uta9-logs/

[67] P. Merrell, E. Schkufza, Z. Li, M. Agrawala, and V. Koltun, "Interactive furniture layout using interior design guidelines," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 87:1–87:10, Jul. 2011. [Online]. Available: http://doi.acm.org/10.1145/2010324.1964982

[68] S. Jodogne, "The Orthanc ecosystem for medical imaging," *Journal of Digital Imaging*, vol. 31, no. 3, pp. 341–352, Jun 2018. [Online]. Available: https://doi.org/10.1007/s10278-018-0082-y

[69] F. M. Calisto, "Medical imaging multimodality annotating framework," 2020. [Online]. Available: http://rgdoi.net/10.13140/RG.2.2.16086.88649

[70] F. M. Calisto, "Breast cancer medical imaging multimodality lesion contours annotating method," Instituto Superior Técnico, Tech. Rep., 2020. [Online]. Available: http://rgdoi.net/10.13140/RG.2.2.14792.55049

[71] F. M. Calisto, N. Nunes, and J. C. Nascimento, "Breastscreening: On the use of multi-modality in medical imaging diagnosis," in *Proceedings of the International Conference on Advanced Visual Interfaces*, ser. AVI '20. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3399715.3399744

[72] H. Lencastre, "Master project: Breast cancer multimodality scalable interactions," 2020. [Online]. Available: http://rgdoi.net/10.13140/RG.2.2.35800.24329/4

[73] H. Lencastre, F. M. Calisto, and J. C. Nascimento, "3d module view feature," 2020. [Online]. Available: http://rgdoi.net/10.13140/RG.2.2.28345.52322

[74] J. M. Bland and D. G. Altman, “Statistics notes: Measurement error,” *BMJ*, vol. 312, no. 7047, pp. 1654–1654, Jun. 1996. [Online]. Available: https://doi.org/10.1136/bmj.312.7047.1654

[75] D. C. Hoaglin, “John w. tukey and data analysis,” *Statistical Science*, vol. 18, no. 3, pp. 311–318, Aug. 2003. [Online]. Available: https://doi.org/10.1214/ss/1076102418

[76] W. H. Kruskal and W. A. Wallis, “Use of ranks in one-criterion variance analysis,” *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, Dec. 1952. [Online]. Available: https://doi.org/10.1080/01621459.1952.10483441

[77] H. Lencastre, “UTA9: Advance Statistics dataset,” 2020. [Online]. Available: https://mimbcd-ui.github.io/dataset-uta9-advstatic/

[78] H. Lencastre, “UTA9: Questionnaire After Test dataset,” 2020. [Online]. Available: https://mimbcd-ui.github.io/dataset-uta9-qat/

Code and Tables

Listing A.1: Pseudo Code of the Scalable Interactions Black Box

```
while (Change):
    if (UI change):
        if (Simple UI Change):
            #Proceed alteration as before
        if (Drag image change):
            #Change DICOM image in View-Port
        if (Click image change):
            if(One View || Two Views):
                #Open the image clicked and the opposite
            else:
                #Open the image clicked and the opposite
                #Open the other image modality and the opposite
        if (Temporal Comparison View):
            if (Not Active Before):
                #Open 2 views and reset them
                if(Drag image Successful):
                    #Open the image dragged and the similar from another time
                        ↪ period, with time line
                else:
                    #Temporal Comparison View disable
                    #Open the image clicked and the opposite
            else:
                if(Drag image Successful):
                    #Open the image dragged and the similar from another time
                        ↪ period, with time line
                else:
                    #Temporal Comparison View disable
                    #Open the image clicked and the opposite
    if (Image Manipulation):
        if (Simple Image Manipulation):
            if (Coordinated View):
                for each (View-Port Open):
                    #Proceed image manipulation
            else:
                #Proceed image manipulation
```

Table A.1: Relationship table that is use to compare the different questions, from the result of the questionnaire the questions will be compared question against other question, as show in the table and the question with higher value has one point.

1	Mental Demand	Physical Demand
2	Temporal Demand	Performance
3	Effort	Frustration
4	Mental Demand	Temporal Demand
5	Effort	Physical Demand
6	Performance	Frustration
7	Effort	Mental Demand
8	Temporal Demand	Frustration
9	Physical Demand	Performance
10	Mental Demand	Performance
11	Temporal Demand	Effort
12	Frustration	Physical Demand
13	Frustration	Mental Demand
14	Physical Demand	Performance
15	Temporal Demand	Effort

Table A.2: Real use case of the Count Use of a Tool in a diagnosis

Count Use of a Tool Real Case

==========Study internalId==========
dca23c

==========Study==========
70a15c33-e38e-4270-b37f-a8aa1dff6fdf

==========Dicom Studies Present==========
b373727c-0855b96f-aa38b69f-4020129f-fb0252ed/file 15/12/2017 Modality: MG|CC
c8f025d5-7ab3c45e-894e80b6-91f5a4d4-d4172068/file 15/12/2017 Modality: MG |CC
60f3159e-4c6b91a7-a4595acb-11ccf003-e6815979/file 15/12/2017 Modality: MG |MLO |L
acff29fe-46581788-9daacf05-aba25f1f-cd89c588/file 15/12/2017 Modality: MG |MLO |R
76b200c6-836606ad-443fd3f1-2b12168d-eacb6d54/file 14/09/2016 Modality: MG |CC |L
818ebeb2-e726c0c5-ebd76f6d-132c30f1-314ba544/file 14/09/2016 Modality: MG |CC |R
984bdfd5-1c8a2283-606108ed-2d0ceb74-20bd6f17/file 14/09/2016 Modality: MG |MLO |L
ceb2e228-bf8e0b46-580a7646-eaa3eb03-9c029113/file 14/09/2016 Modality: MG |MLO |R
d0cbdd06-2657fd97-1c802bf5-ad8241f5-14202ce7/file 27/12/2017 Modality: US
1626b730-6f5b1562-61d4b7f2-a3eeaae7-50514193/file 27/12/2017 Modality: US

==========Study Date==========
27/12/2017
==========Time Begin==========
20:01:54
==========Actions List==========
20:01:54 |Layout = 1x2
— Automation is Activated
— Image Open → c8f025d5-7ab3c45e-894e80b6-91f5a4d4-d4172068
— Image Open → b373727c-0855b96f-aa38b69f-4020129f-fb0252ed
20:08:33 |Layout = 1x1
— Automation is Activated
— Image Open → c8f025d5-7ab3c45e-894e80b6-91f5a4d4-d4172068
20:08:59 |Action On Image → c8f025d5-7ab3c45e-894e80b6-91f5a4d4-d4172068 - Zoom
20:09:02 |Action On Image → c8f025d5-7ab3c45e-894e80b6-91f5a4d4-d4172068 - Zoom
20:09:02 |Action On Image → c8f025d5-7ab3c45e-894e80b6-91f5a4d4-d4172068 - Pan
20:09:04 |Action On Image → c8f025d5-7ab3c45e-894e80b6-91f5a4d4-d4172068 - Pan
20:09:10 |Action On ALL Images open - WW/WC
— Image Open → b373727c-0855b96f-aa38b69f-4020129f-fb0252ed
— Image Open → c8f025d5-7ab3c45e-894e80b6-91f5a4d4-d4172068
20:10:07 |Layout = 1x2
— Automation is Activated
— Image Open → c8f025d5-7ab3c45e-894e80b6-91f5a4d4-d4172068
— Image Open → b373727c-0855b96f-aa38b69f-4020129f-fb0252ed
20:10:27 |Layout = 1x2
— Automation is Activated
— Image Open → acff29fe-46581788-9daacf05-aba25f1f-cd89c588
— Image Open → 60f3159e-4c6b91a7-a4595acb-11ccf003-e6815979
20:10:52 |Automation is Deactivated
20:11:01 |Action On Image → acff29fe-46581788-9daacf05-aba25f1f-cd89c588 - WW/WC
20:11:18 |Automation is Activated
20:11:26 |Action On ALL Images open - WW/WC

20:12:48 |Layout = 2x2
— Automation is Activated
— Image Open → c8f025d5-7ab3c45e-894e80b6-91f5a4d4-d4172068
— Image Open → b373727c-0855b96f-aa38b69f-4020129f-fb0252ed
— Image Open → acff29fe-46581788-9daacf05-aba25f1f-cd89c588
— Image Open → 60f3159e-4c6b91a7-a4595acb-11ccf003-e6815979
20:13:41 |Layout = 2x2
— Automation is Activated
— Image Open → 818ebeb2-e726c0c5-ebd76f6d-132c30f1-314ba544
— Image Open → 76b200c6-836606ad-443fd3f1-2b12168d-eacb6d54
— Image Open → ceb2e228-bf8e0b46-580a7646-eaa3eb03-9c029113
— Image Open → 984bdfd5-1c8a2283-606108ed-2d0ceb74-20bd6f17
20:14:23 |Layout = 1x2
— Automation is Activated
— Image Open → 818ebeb2-e726c0c5-ebd76f6d-132c30f1-314ba544
— Image Open → 76b200c6-836606ad-443fd3f1-2b12168d-eacb6d54
20:14:28 |Layout = 1x2
— Automation is Activated
— Image Open → 818ebeb2-e726c0c5-ebd76f6d-132c30f1-314ba544
— Image Open → c8f025d5-7ab3c45e-894e80b6-91f5a4d4-d4172068
20:15:02 |Action On ALL Images open - WW/WC
20:15:10 |Automation is Deactivated
20:15:13 |Action On Image → c8f025d5-7ab3c45e-894e80b6-91f5a4d4-d4172068 - WW/WC
20:15:32 |Action On Image → c8f025d5-7ab3c45e-894e80b6-91f5a4d4-d4172068 - Pan
20:15:36 |Automation is Activated
20:15:40 |Action On Image → 818ebeb2-e726c0c5-ebd76f6d-132c30f1-314ba544 - Zoom
20:15:43 |Action On Image → 818ebeb2-e726c0c5-ebd76f6d-132c30f1-314ba544 - Pan
20:15:48 |Action On Image → 818ebeb2-e726c0c5-ebd76f6d-132c30f1-314ba544 - Zoom
20:15:53 |Action On Image → c8f025d5-7ab3c45e-894e80b6-91f5a4d4-d4172068 - Zoom
20:15:56 |Action On Image → c8f025d5-7ab3c45e-894e80b6-91f5a4d4-d4172068 - Pan
20:16:52 |Temporal Comparison started
20:16:52 |Layout = 1x2
— Automation is Activated
— Loading images to comparison
— Image Open → b373727c-0855b96f-aa38b69f-4020129f-fb0252ed
— Image Open → 76b200c6-836606ad-443fd3f1-2b12168d-eacb6d54

==========Time End==========
20:19:16

==========End Counter==========

WW/WC - 5
Invert - 0
Zoom - 5
Pan - 5
Stack Scroll - 0
Freehand ROI draw - 0
Probe - 0
Automated - 4 |At the end the tool was Activated
Temporal Comparison - 1
Save - 0

Table A.3: Dunn's [19] test Table results.

SUS UTA9	Intern	Junior	Senior
Intern	-1	0.376406	0.323839
Junior		-1	0.933586
Senior			-1
SUS UTA4			
Intern	-1	0.291589	0.095512
Junior		-1	0.617075
Senior			-1
NASA-TLX UTA9			
Intern	-1	0.632816	0.055996
Junior		-1	0.241887
Senior			-1
NASA-TLX UTA4			
Intern	-1	0.839232	0.417077
Junior		-1	0.619230
Senior			-1

B

Documents

Interaction Tools

Welcome to the MIMBCD-UI project.

This project is developing a framework to ease the detection of Breast cancer, not only by reducing the diagnosis time, but also, to provide the user with new techniques to detect lesions. We will develop an AI system with eXplanability techniques to give a second opinion, or to be a filter system that will diagnose a patient case and reorder the priority of the patient by the classification.

In order to respond this questionnaire, is essential to respond our previous questionnaire about Demographic Data, https://forms.gle/e7SzRttWprUtw1By7 .

This questionnaire will be able to be performed in less than 5 minutes.

It is of utmost convenience that you answer with rigor and honesty, as this is the only way to fulfill the objective of this questionnaire.

There are no right or wrong answers in relation to any of the items, the intention being only to ascertain the data provided.

This questionnaire is confidential and voluntary. The treatment of this, in turn, is carried out in a global way, not being subject to an individualized analysis, which means that your anonymity is respected.
*Obrigatório

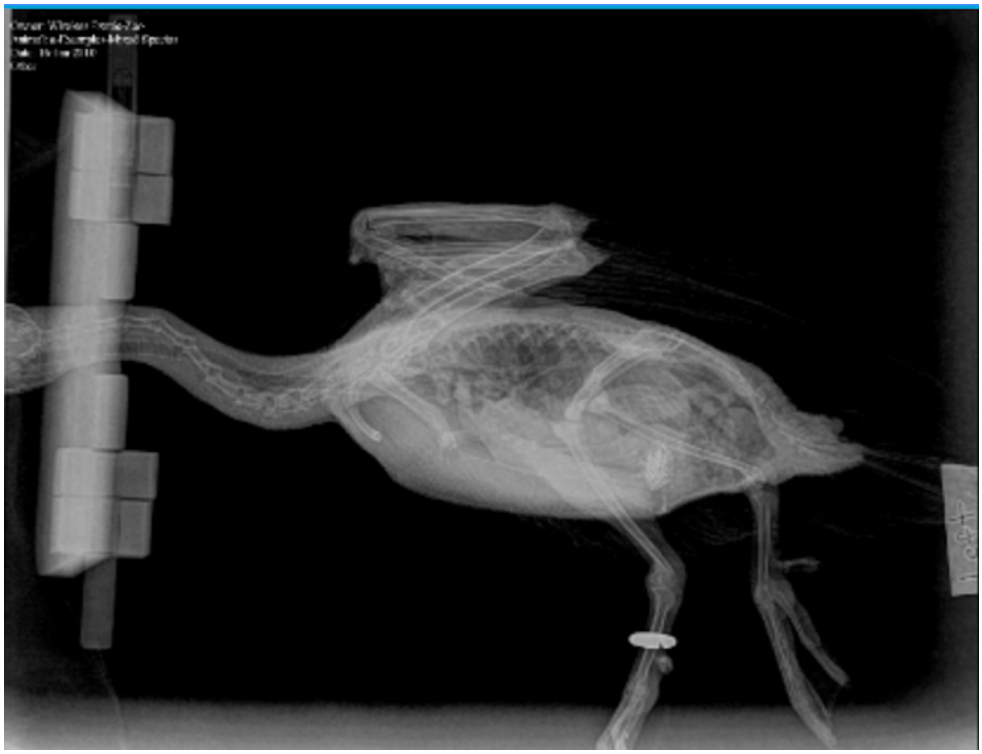
Demographic Data

1. ID *

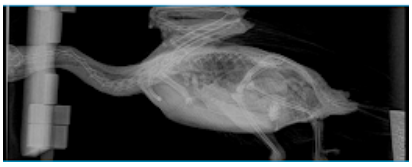
   Do not change this information. Please continue.

Present Tools Available

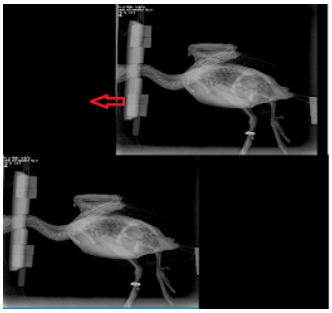
2. Which tools do you have available to manipulate an image in the diagnosis of Breast Cancer? *



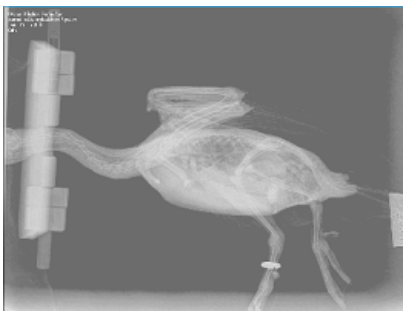
Marcar tudo o que for aplicável.



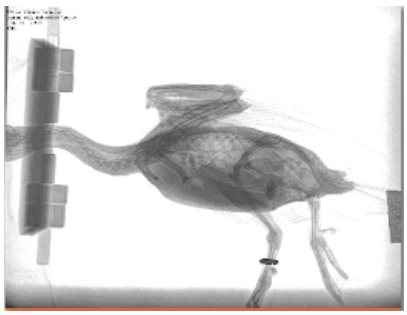
☐ Zoom - https://tools.cornerstonejs.org/examples/tools/zoom.html



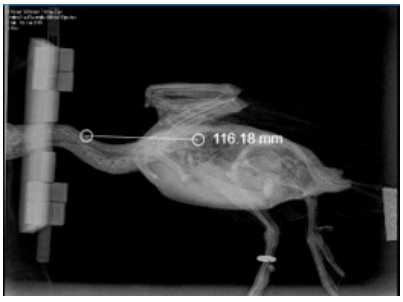
☐ Move - https://tools.cornerstonejs.org/examples/tools/pan.html



☐ Luminosity



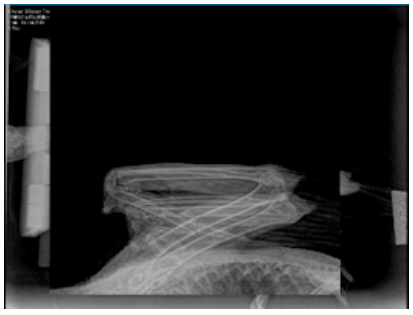
☐ Invert



☐ Distance measure - https://tools.cornerstonejs.org/examples/tools/length.html

☐ Lesion Annotation (Probe or Freehand) - https://tools.cornerstonejs.org/examples/tools/text-marker.html

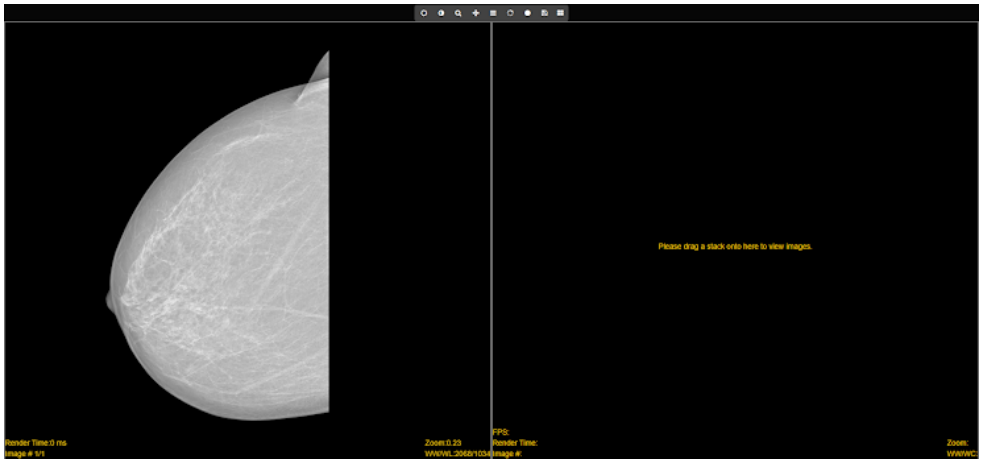
Outra: ☐ _____



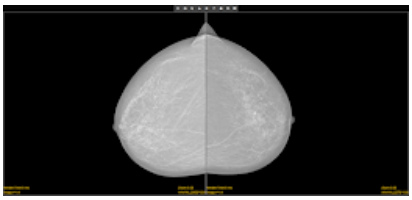
☐ Zoom in in that point for a instance of time- https://tools.cornerstonejs.org/examples/tools/magnify.html

Future Tools Proposal

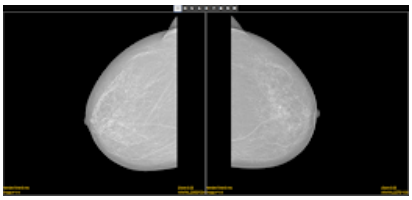
3. What are the tools that you hope to have in the future? *



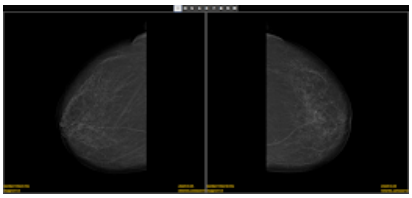
Marcar tudo o que for aplicável.



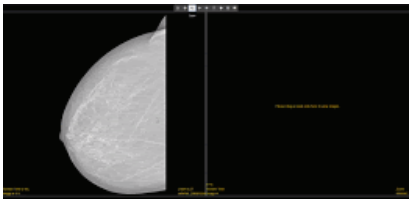
☐ Re-ajust of the image in the center



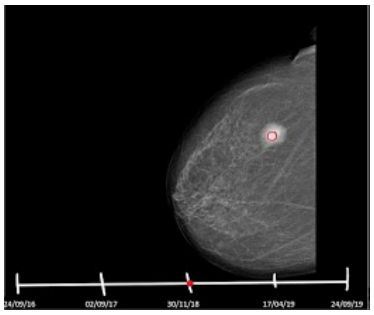
☐ When open a Modality View, the oposite view opens in the oposite ViewPort



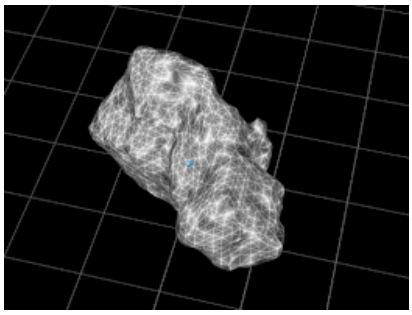
☐ When a operation is done at one modality View, all open at the same time, perform the same manipulation



☐ Zoom out/in automatically to enable a easy annotation of the lesion

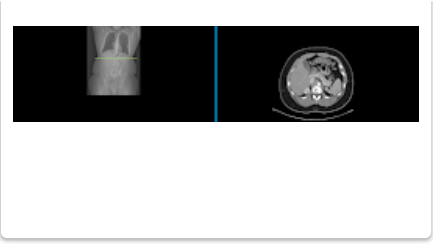


☐ Show the evolution of a lesion over the present lesion



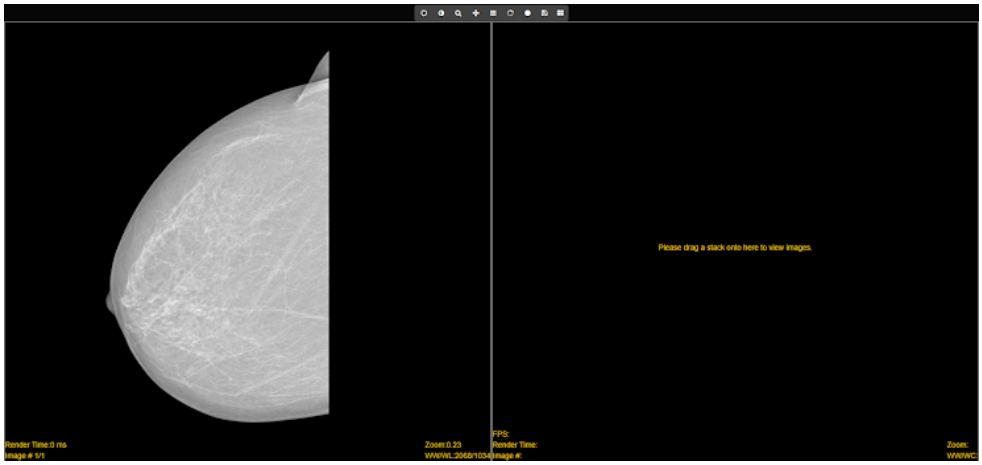
☐ When a 3D view is enable, create a representation of the lesion without the necessity to see several 2D images

Outra: ☐ _____



When using a RMI, knowing where that slice of image is relative to the body.

4. What order of priority do you give to each future feature presented? *



Marcar tudo o que for aplicável.

	Not Needed	Not Very Import	2	3	4	Very Important
Re-ajust of the image in the center	☐	☐	☐	☐	☐	☐
When open a Modality View, the oposite view opens in the oposite ViewPort	☐	☐	☐	☐	☐	☐
When a operation is done at one modality View, all open at the same time, perform the same manipulation	☐	☐	☐	☐	☐	☐
Zoom out/in automatically to enable a easy annotation of the lesion	☐	☐	☐	☐	☐	☐
Show the evolution of a lesion over the present lesion	☐	☐	☐	☐	☐	☐
When a 3D view is enable, create a representation of the lesion without the necessity to see several 2D images	☐	☐	☐	☐	☐	☐
When using a RMI, knowing where that slice of image is relative to the body.	☐	☐	☐	☐	☐	☐

Interaction Tools Pos-Tasks

Welcome to the BreastScreening project.

In this project, we aim at developing a CADx system to ease the detection of Breast cancer, not only by reducing the diagnosis time, but also, to provide the user with new techniques to detect lesions. We will develop an AI system with eXplanability techniques to give a second opinion, or to be a filter system that will diagnose a patient case and reorder the priority of the patient by the classification.

In order to respond this questionnaire, is essential to respond our previous questionnaire about Demographic Data, https://forms.gle/e7SzRttWprUtw1By7 and to perform all Scenarios from the Phase 3.

This questionnaire as the goal to understand if the tools created are important to the daily diagnosis, and what could be improved in those, with this questionnaire we will know if the right path was taken.

This questionnaire will be able to be performed in less than 5 minutes. It is of utmost convenience that you answer with rigor and honesty, as this is the only way to fulfill the objective of this questionnaire.

There are no right or wrong answers in relation to any of the items, the intention being only to ascertain the data provided. This questionnaire is confidential and voluntary. The treatment of this, in turn, is carried out in a global way, not being subject to an individualized analysis, which means that your anonymity is respected.
*Obrigatório

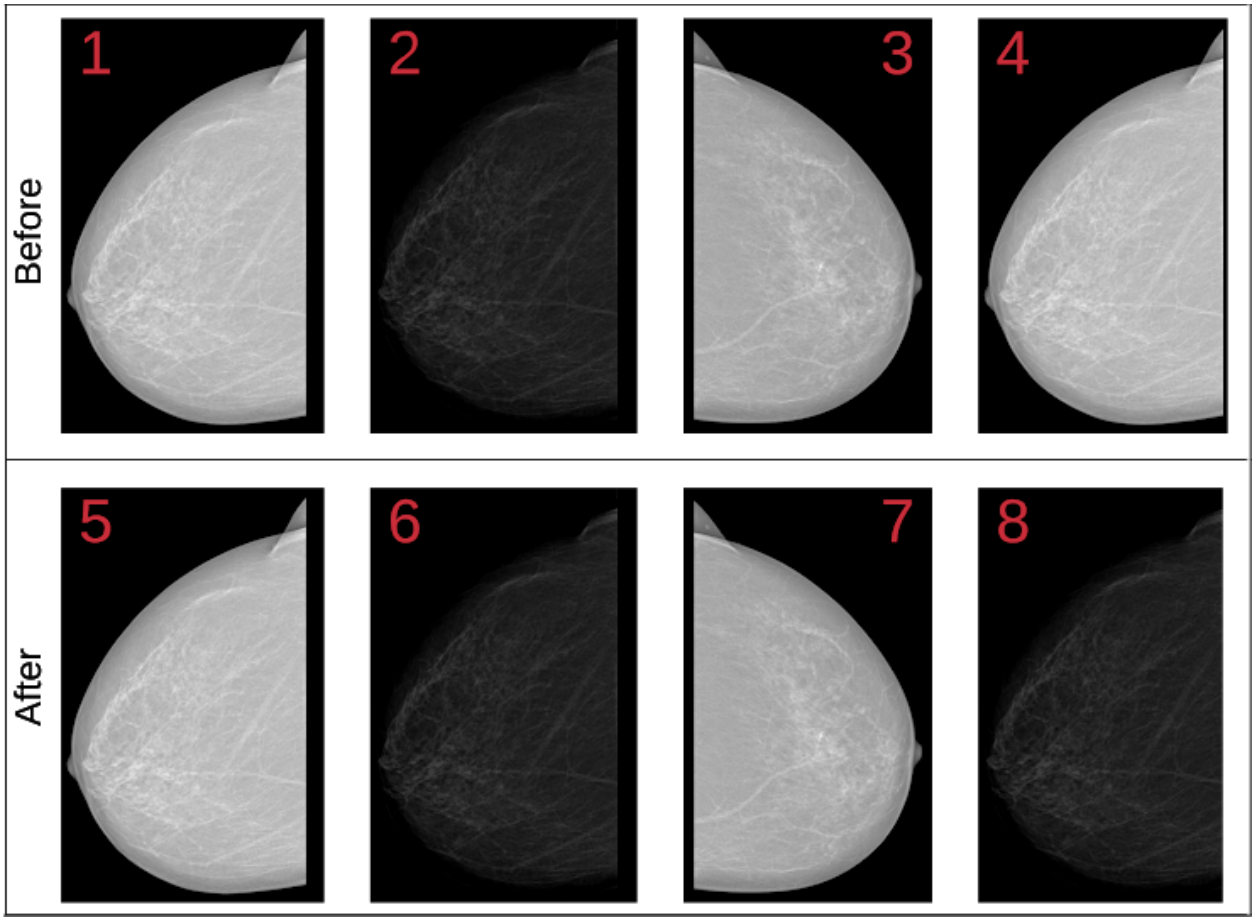
Demographic Data

1. ID? *

   Do not change this information. Please continue.

Recorded State

2. How much do you liked the feature? *



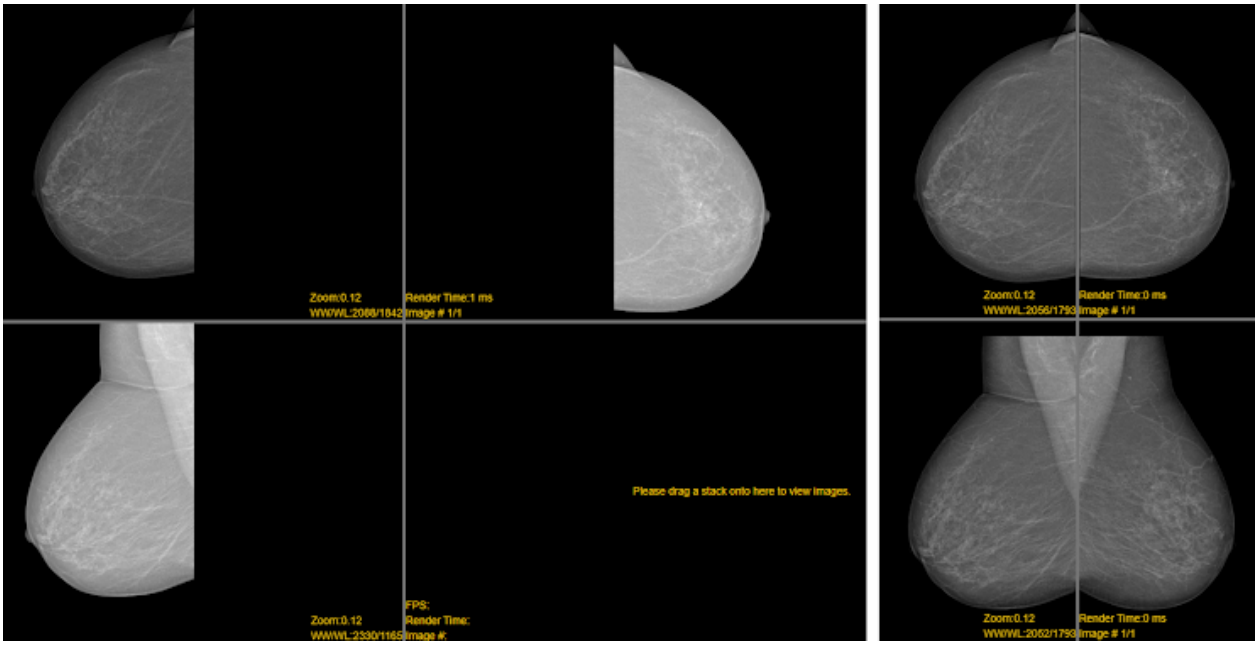
Marcar apenas uma oval.

	1	2	3	4	5	
Dislike	◯	◯	◯	◯	◯	Like

3. What did you like /dislike in the feature and what changes you want to see in this feature?

Coordinated actions and automation

4. How much do you liked the feature? *



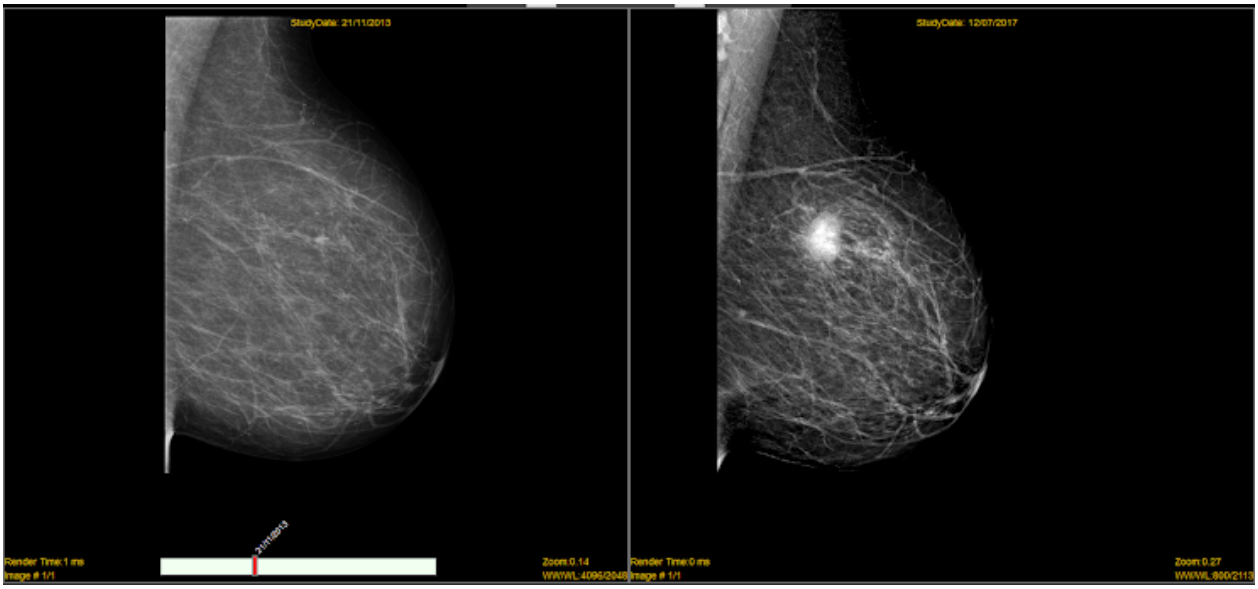
Marcar apenas uma oval.

	1	2	3	4	5	
Dislike	◯	◯	◯	◯	◯	Like

5. What did you like /dislike in the feature and what changes you want to see in this feature?

Temporal comparison on the evolution of the lesion

6. How much do you liked the feature? *



Marcar apenas uma oval.

	1	2	3	4	5	
Dislike	◯	◯	◯	◯	◯	Like

7. What did you like /dislike in the feature and what changes you want to see in this feature?