MIMBCD-UI
User Testing Guide

Francisco Maria Calisto
francisco.calisto@tecnico.ulisboa.pt

03/05/2018

Prototype: prototype-breast-cancer
Version: v1.2.0-beta
Milestone: 1.2.0-beta
Release: v1.2.0-beta

DICOM: dicom-server
Commit: 33f4f92f7b016e69e63f13592eb47fa18a4dc8c3

Deployment Environment: Prod.
Deployment Server: Prod.
Link: breastscreening.io/dashboard

Main Server: Production
Port: 8088
Private IP: 10.0.1.23
Public IP: 193.136.138.62
Private Domain: cromo.isrnet

DICOM Server: Test
Port: 8042
Private IP: 10.0.1.23
Public IP: 193.136.138.62
Private Domain: cromo.isrnet
From: 8142

1 Introduction

This document aims to describe the protocol performing a set of tests in the scope of v1.2.0-beta version from the prototype-breast-cancer repository of the MIMBCD-UI project using traditional devices (mouse and keyboard). The goal of the test is to understand the user, performance, efficiency and efficacy metrics. With the session, the sessions will be recorded via video on the computer and using a record, heat-map and triggered event tools. It is guaranteed the confidentiality of the recordings, which will be used only for academic purposes.

Dividing the activity session into four distinct phases per each two activities representing two different scenarios (Single-Modality vs Multi-Modality). Each scenario will have three patients. In both two scenarios, by supporting our traditional devices, the interaction is made with mouse and keyboard. The first phase, is the demographic questionnaire, where we characterise the Radiologist profile. The second phase is the act of classifying those patients. Radiologists will classify each patient by using the BIRADS [1]. On the third phase, we will do a small questionnaire at the end of each scenario using NASA-TLX [15]. Finally, the forth phase we will have a final survey regarding the Usability of each scenario. The well-known scale called System Usability Scale (SUS) [14]. For the user tests we used a two distinct prototype repositories prototype-single-modality and prototype-multi-modality, both are "almost" mirrors of the prototype-breast-cancer with minor changes.

2 Material

For the material and apparatus, it is essential to capture the session apprehending the user interactions. In our case, we will record this interaction by using the QuickTime Player Version 10.4 (928.5.1) to obtain all interactions. We will pair this video tool with a user watch tool called Hotjar. This tool serves the purpose of using several logs of the interaction and gives us visualisation over it. Both instruments will help us to capture where are users interacting. By looking at the test participant's reactions, we find a lot of information regarding the prototype design.

The tools that we choose for the material and apparatus of this User Testing Guide are low-cost and easy to use. Our equipment is a cost-effective and, by using our laboratory materials, bringing it to the radiology room, we enable to capture not only what the user is doing on the screen, but on the body language supported by the interviews and observation.

The material used in the test sessions for the user interface consists of:

- MacBook Pro: it will allow the user to interact with the keyboard and a wireless mouse;
- Wireless Mouse: it will allow the user to interact with a mouse and will complement the keyboard;

2.1 Technical Details

To produce this traditional environment, and since we can simulate with a laptop, the mouse and keyboard interaction, we are using a Microsoft Mobile Mouse 4000 together with the MacBook Pro (Retina, 13-inch, Early 2015) with a standard integrated keyboard on the laptop.

2.2 Software

To track our user interactions across our system, we are using Hotjar. This tool is an analytic package allowing us to follow our users remotely. It also provides two critical pieces of functionality, among others, that can aid in remote user testing. First of all, the heatmaps allow us to see where users are clicking, tapping and scrolling on our system. Second, it records a video playback of the entire user session. The tool shows evidence of being useful for our studies while we successfully used it in the past.

To record the task activities and the interview, we used QuickTime [16]. The QuickTime (Apple Computer) tool is available for MacBook Pro to movie, audio and screen recording. Despite of have an overall of features, we just used it for our user's screen recording. It provides this functionalities at minimum requirements and compatible to our apparatus.

3 Description

To verify our work, we identified measurable and explicit targets. By having several goals, including that a value percentage of the users should be able to operate the tasks without the need of help. On the same rate value, the user should be able to start and complete the medical diagnosis tasks over the system with little errors or mitigating those errors. Measuring the expected number of errors with a relation between our laboratory pilot tests. On the laboratory pilot tests we aim to test our prototypes with Researchers. The Researchers are in the context of the system and know well the functionalities so that we need to expect a percentage value over their results compared to Clinicians and not the same benefits. Last but not least, both users (Researchers and Clinicians) should be able to understand in a similar time amount the meaning of all visible controls. By the similar amount of time, it is expected to have a variance of the percentage value between Researchers and Clinicians of the same value percentage of the early goals described in this paragraph.

We tested each objective in early laboratory and field tests so that we could take the appropriate corrective actions. Also, we expect to run early field tests with Researchers and Clinicians to highlight issues that we overlooked and ignored during the prototyping phase. To support interaction use by the Clinicians, we will try to emphasise several key factors on our user tests. The tasks must be simple, low intrusive, support for natural interaction and the system must always give visibility and the task current-state.

3.1 Devices

Traditional interaction remains the most common way to interact with user interfaces in a clinical environment. Unfortunately, most of this interaction is made by low profile equipment that makes users produce more errors and take more time interacting with those user interfaces.

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

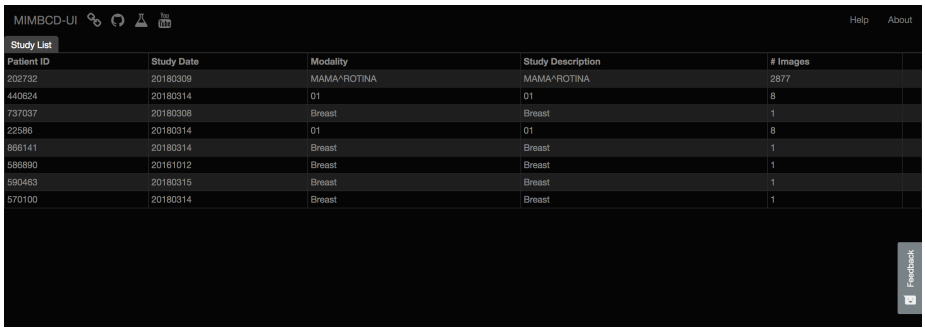

Figure 1: List of patients.

As we can see in Figure 2, it shows the first task in our User Interface (UI), where the patient's breasts are on a small left column. The options are in a short row near of the viewport and described below. We also have the tabs where the user can change the patient. The centre viewport shows the **DICOM** image, and it can be configured to display a number up to four **DICOM** images at the same time. The viewport has some text information on it (yellow) with the details of the metadata.

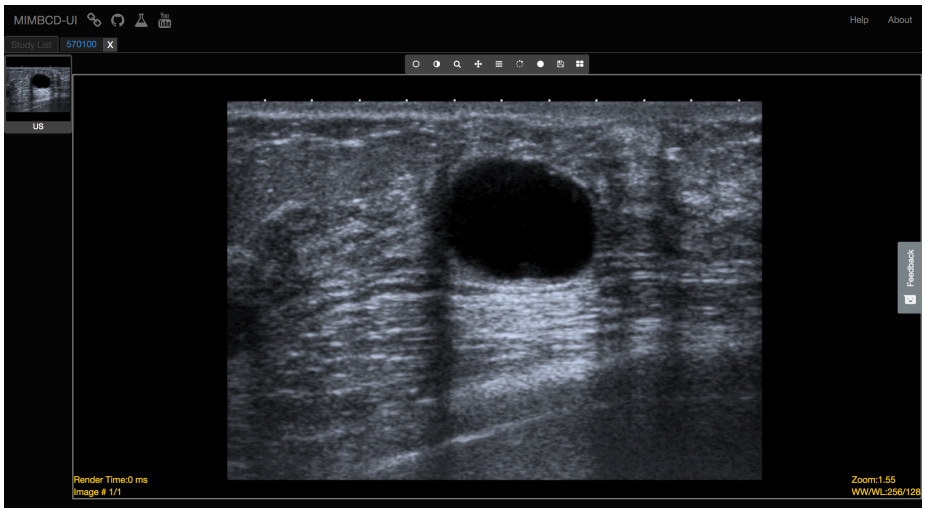

Figure 2: Viewer of the **DICOM** images.

Manual annotation is adopted by us thanks to Freehand ROI and Probe annotation features, both from CornerstoneJS. According to the CornerstoneJS Library, the user can create an annotation by setting up consecutive landmarks around a Region of Interest (ROI). The markers finish a lesion annotation when it interconnects the historical. Additional features available in our User Interface (UI) includes on-demand increment of the number of landmarks, and throw transformations of the shape of an annotation.

3.2 User Interactions

The systems have several buttons (Figure 3) that allows the user to interact or access to a set of user interface features. Each item of the following list represents each metaphoric icon of Figure 3.

Figure 3: Toolbar of the System available features.

The buttons are (from left to right of Figure 3) as follows:

- WW/WC
- Invert
- Zoom
- Pan
- Stack Scroll
- Freehand
- Probe
- Save
- Window Controller

3.3 Usability Evaluation Technique

Usability and functionality are the significant elements which can significantly affect the performance of a medical system. While some prior studies [3] have investigated the functionality of healthcare systems, the usability issue has mostly been overlooked in the existing Health Informatics (HI) literature regarding Human-Computer Interaction (HCI).

The following Table 1 is presenting six evaluation questions to have in mind during evaluation. The purpose of this questions is to facilitate systematic user studies in a clinical environment and support the identification of usability problems. The proposed issues involve various aspects of workload combined with either need for satisfaction or division of attention.

Number	Issues of Content Key Questions
1	How do you perceive this activity?
2	Could it be done in a more intuitive way?
3	What are the consequences?
4	Why did you do as you did with this activity?
5	Is this activity relevant for you?
6	Could you suggest another way to do this activity?

Table 1: Usability Evaluation Questions

The influence of perceived activity [8] is an important variable for our empirical analysis. In fact, the trust of the user increases when the user perceived that the system is usable and that there will be a consequent increase of the clinician trust in our system. This arguments explains the first question, the *How do you perceive this activity?* question. For the second question, the *Could it be done in a more intuitive way?* question, we aim to conclude if there is some solution for a more intuitive way of perceive the activity. Third, we intend to filter possible consequences of the clinician workflow by asking *What are the consequences?* directly to the clinician. The fourth question, underlines the reasons why the clinician did that way, with the question *Why did you do as you did with this activity?* we can understand the process of achieving the activity goal and the clinician's interpretation of it. On the fifth question, where we ask *Is this activity relevant for you?*, we aim to understand the potential relevance of our system to the clinician. Last but not least, the six question is present to give the clinician opportunity to suggest improvements, reflecting the reasons why we ask the *Could you suggest another way to do this activity?* question.

To conclude this section, by doing this questions, we aim to support our user studies by giving our users, the clinicians, the opportunity of improving our empirical analysis regarding user's *open answers*. However, the results should be treated with caution. Several bias exists since we are doing here an ambiguous approach.

4 Methodology

The hereby prototype used is the v1.2.0-beta version of our prototype-breast-screening. The purpose of this prototype is to integrate a web-based tool by using medical imaging for the breast screening diagnosis. This web-based tool was created with a front-end and bak-end architecture utilising common programming languages, libraries, frameworks and tools including JavaScript (JS) [7], NodeJS [18], HammerJS, CornerstoneJS [9] and Orthanc [10]. The central component of this prototype is a web-based PACS [4] pairwise with ubicous web technologies and based on the **Open Source (OS)** CornerstoneJS library [6, 9].

4.1 Environments

This section describes the user environment over interaction, the so called **Radiology Room (RR)** (Figure 4). This guide is based on soft-copy diagnosis using computer workstations in their current reading room environment. It will be here where we take impressions regarding the efficacy of radiologists, and their recommendations based on their experience for improvements on the soft-copy reading environment. Several studies demonstrated that radiologist fatigue levels and performance are related to environmental factors such as monitor brightness and ambient room light in addition to workstation enhancements. Supported by this guide, our research aims to conduct an investigation for the several environmental variables. We expect to analyse the noise, temperature, seating ergonomics and so on.

Figure 4: Radiology Room

4.2 Case Studies

The functionality of the prototype will be best demonstrated by a series of case studies. By describing the expected workflow and capabilities of the research study at the **RR** specific environment and changes of the workflow by using our system prototype. The study implies the evaluation of medical imaging features of several breast lesions. The primary goal of this case studies analysis is to generate a receiver operating characteristic to evaluate the performance and validation of our system. Let us consider a list of hypothetical use cases for the research investigation that evaluates the interaction and usability performance of the prototype. Therefore, the following list will show the preliminary case studies.

List of case studies to analyse our solution prototype:

- Multi-modality Imaging Features of a Breast Cancer Diagnosis;
- Single-modality Imaging Features of a Breast Cancer Diagnosis;
- Intuitive Controls Analysis;
- Performance Measurement Values Acquisition;
- Usability Measurement Values Acquisition;
- Radiologist Validation;

We expect to demonstrate several uses through a series of case studies, including implementation of our research prototype for both multi-modality and single-modality view studies and other imaging research, and creation of a novel tool for the purpose. By creating a set of questions, we will try to achieve and feed this case studies. It is automatically associated with all cases. The radiologist may interact with our system manipulating the medical imaging and report to us difficulties and improvements. The number of questions is not restricted to the present document, since the interview will be open and suggestive. There is no limit to the number of questions that can be asked per case but it should fit the amount of expected time per each.

The data will be collect from the study video and observations into a spreadsheet for further analysis. We will report the results of this tests and conclusions. This guide and respective use cases will be iteratively improved.

5 Procedures

Participants will take part in the tests at our formed institution protocols (e.g. Hospital Fernando Fonseca (HFF)) with the v1.2.0-beta version of our prototype-breast-screening. The interaction with the system will be used in a typical **RR** environment. Note takers and data logger(s) will monitor the sessions for observation in the **RR**, connected by screen recording feed. The test sessions will be recorded and further analysed.

5.1 Briefing

A presentation of the systems and it's use and capabilities will be made. Participants will be presented to the available interactions and will be explained how to interact with the prototype, underlining the limitations. The facilitator will brief the participants on the prototype application and instruct the participant that they are evaluating the application, rather than the facilitator evaluating the participant. Participants will sign an informed consent that acknowledges: the participation is voluntary, that participation can cease at any time, and that the session will be videotaped but their privacy of identification will be granted. The facilitator will ask the participant if they have any question.

5.2 Post-Task Questionnaire

Our metrics will refer the user performance measured against specific performance goals necessary to satisfy several requirements of our system. For our **Post-Task Questionnaire** we will use **SUS** to measure the usability of our system each time a *scenario* is completed. From a set of tasks (see **Tasks** section) we aim to cover our two scenarios, **Single-Modality** and **Multi-Modality**. Therefore, the **SUS** will allow the facilitator to quickly and easily assess the usability of a given scenario. This scale has several attributes [2] that make it a good choice for our clinical usability participants. Those attributes are as follows.

List of the scale attributes:

- The survey is technology agnostic, making it flexible enough;
- The survey is relatively quick and easy to use;
- The survey provides a single score on a scale that is easily understood;
- The survey is nonproprietary, making it a cost effective tool;

The facilitator will explain that the amount of time taken to complete the *tasks* will be measured and that exploratory behaviour outside the *task* flow should not occur until after task completion. At the beginning of each task, the participant will listen the task description from the facilitator and begin the task. *Time-on-task* measurements begins when the participant starts the *task*, measured until the end of each *task*.

5.3 Training Session

The participant will receive and overview the usability test procedure. However, the user will not receive information how to annotate and interact in all degrees of freedom. With the aim of disabling users to get their work done before the test tasks. It will take advantage of a "surprise" acknowledgement.

5.4 Execution of Tasks

The *tasks* were derived from test scenarios developed from **Case Studies**. Due to the range and extent of functionality provided by our prototype, and the short time from which each participant will be available, the *tasks* are the most common and relatively complex of available functions. The *tasks* are the identical for all participants of a given user role in the study.

The *tasks* will be performed by several classes of radiology experience. Professionals from Radiology Seniors, Juniors and Interns will be performing these *tasks*. On the **RR** the Radiologist is characterised [5, 12] as a physician who examines and interpret Medical Imaging (MI) [11], such as X-Rays, CT Scans or MRIs.

5.5 Post-Activity Questionnaire

After completing all *tasks* and scenarios, participants will be asked to complete a questionnaire to classify the prototype according to various parameters regarding the workload. To measure this, we will use the well known scale called NASA-TLX [15]. It consists of a set of six rating scales to evaluate the workload of the participant in a *task* or a set of *tasks*. NASA-TLX is used in **Human-Computer Interaction (HCI)** research to identify users' performance, mental demand, emotion, etc. We will use this scale questionnaire to identify participants' workload during the various stages of the workflow.

6 Tasks

During our usability tests, we need to ask participants to provide a subjective assessment of their experience using our system. There are several widely used questionnaires giving us different prons-and-cons. However, in most cases, a single question instrument [17] is the right method for a quantitative usability testing. By taking less time and effort to answer, participants are pursuing to this phase after task while it is minimally disruptive.

We aim to test the performance of the diagnostic between Single-Modality and Multi-Modality views. Therefore, we will try to understand if more precisely the radiologist encounters the most accurate severity (BIRADS) of the breast lesions [13]. For this purpose, Figure 5 illustrates the idea. We have three patients; each patient has three images in the respective modalities: (i) MG; (ii) US; and (iii) MRI. In a first activity (Single-Modality) the physicians will observe a single image/patient, whose own Single-Modality follows the order that we placed in Figure 5. Then, in a second activity (Multi-Modality), they observe ALL images with all modalities (Multi-Modality) per each patient.

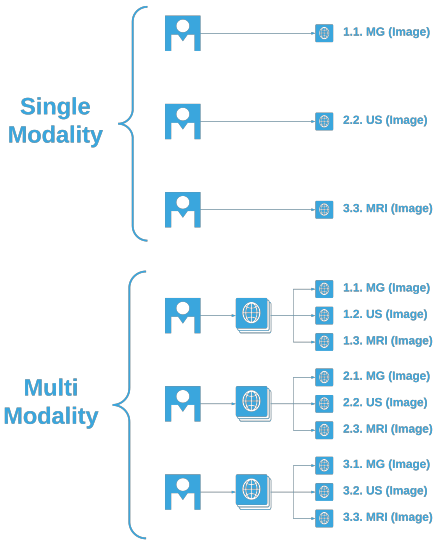

Figure 5: Single vs Multi

In our **User Testing Guide** a set of tasks is necessary and carefully crafted. Our test studies involve asking participants to perform a set of tasks. By looking at what our user need to do with our system, our tasks are realistic as possible. We are not describing the exact steps participants need to take. We achieve that by avoiding the precise language used as labels in our system. The tasks are emotionally neutrals. And we did several pilot tests to prevent misleading situations saving us from wasting resources by accidentally use a lousy task or from getting bad data. The tasks are as follows.

List of stand alone tasks:

Task 1.1: Classify *Patient 1* on a **Single-Modality**;

Task 1.2: Classify *Patient 2* on a **Single-Modality**;

Task 1.3: Classify *Patient 3* on a **Single-Modality**;

Task 2.1: Classify *Patient 1* on a **Multi-Modality**;

Task 2.2: Classify *Patient 2* on a **Multi-Modality**;

Task 2.3: Classify *Patient 3* on a **Multi-Modality**;

7 Measurements

Our measurements refers to user performance measured against specific performance goals necessary to satisfy requirements. *Task* completion success rates, adherence to dialog scripts, error rates and subjective evaluations will be used. *Time-to-completion* of *tasks* will also be collected. The measures are as follows.

The tests are intended to achieve the following measures:

- BIRADS Classification;
- Time measurement;
- Number of clicks;
- Number of errors;
- Efficiency;
- Difficulty;
- Experience;

To prioritise recommendations, a method for problem difficulty and degree severity classification will be used in the analysis of the collected data during evaluation process. The approach treats problem severity has a combination of several factors. Those factors are measuring the impact of the problem and the frequency of users experiencing issues during the evaluation.

Through the questionnaire after the test session, we intend to obtain the answers to the following questions for each *task*:

- Difficulty of lesion classification;
- Difficulty of interaction;
- Difficulty translating;
- Difficulty performing the several features;
- Degree of classification;
- Degree of interaction;

8 Goals

The next sections will describe the goals for prototype-breast-screening expectations. We will try to assess performance-related metrics such as time and correctness of participants completing *tasks* for our expectations. Our expectations are based of the results obtained at the lab as pilot tests.

8.1 Completion Rate

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

*A **Completion Rate** of **100%** is the goal for each task in this usability test.*

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

8.2 Error-Free Rate

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

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

8.3 Time on Task (TOT)

The time to complete a scenario is referred to as "time on task". It is measured from the time the participant begins the scenario to the time which the participant signals completion.

8.4 Subjective Measures

Subjective opinions about specific tasks, time to perform each task, features, and functionality will be surveyed. At the end of the test, participants will rate their satisfaction with the overall system. Combined with the interview/debriefing session, these data are used to assess attitudes of the participants.

9 Reporting Results

A Test Report will be provided at the end of this tests. It will consist of a report and/or a presentation of the results; evaluation of the metrics against the pre-approved goals, subjective evaluations, and specific issues of the system, as well as, recommendations for resolution. The recommendations will be categorically sized by development to aid in implementation strategy. The results will be translated to a spreadsheet (view only). Also, more related information can be found at Test 4: Single-Modality vs Multi-Modality.

References

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

[2] Aaron Bangor, Philip T Kortum, and James T Miller. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction*, 24(6):574–594, 2008.

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

[4] Robert E Cooke Jr, Michael G Gaeta, Dean M Kaufman, and John G Henrici. Picture archiving and communication system, June 3 2003. US Patent 6,574,629.

[5] Ruth Ann Ehrlich and Dawn M Coakes. *Patient Care in Radiography-E-Book: With an Introduction to Medical Imaging*. Elsevier Health Sciences, 2016.

[6] Joseph Feller, Brian Fitzgerald, et al. *Understanding open source software development*. Addison-Wesley London, 2002.

[7] David Flanagan. *JavaScript: the definitive guide*. " O'Reilly Media, Inc.", 2006.

[8] Carlos Flavián, Miguel Guinalíu, and Raquel Gurrea. The role played by perceived usability, satisfaction and consumer trust on website loyalty. *Information & management*, 43(1):1–14, 2006.

[9] Jason Hostetter, Nishanth Khanna, and Jacob C Mandell. Integration of a zero-footprint cloud-based picture archiving and communication system with customizable forms for radiology research and education. *Academic radiology*, 2018.

[10] S. Jodogne, C. Bernard, M. Devillers, E. Lenaerts, and P. Coucke. Orthanc – A lightweight, RESTful DICOM server for healthcare and medical research. In *Biomedical Imaging (ISBI), IEEE 10th International Symposium on*, pages 190–193, San Francisco, CA, USA, April 2013.

[11] Y Kobashi, Y Munetomo, A Baba, S Yamazoe, and T Mogami. Evaluation of the ossification of the cervical poste-rior longitudinal ligament utilizing x-ray, ct and mr imaging. *Orthop Res Traumatol Open J*, 2(1):35–39, 2017.

[12] Diana L Miglioretti, Rebecca Smith-Bindman, Linn Abraham, R James Brenner, Patricia A Carney, Erin J Aiello Bowles, Diana SM Buist, and Joann G Elmore. Radiologist characteristics associated with interpretive

performance of diagnostic mammography. *Journal of the National Cancer Institute*, 99(24):1854–1863, 2007.

[13] American College of Radiology. BI-RADS Committee. *Breast imaging reporting and data system*. American College of Radiology, 1998.

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

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

[16] Melissa R Rowell, Frank M Corl, Pamela T Johnson, and Elliot K Fishman. Internet-based dissemination of educational audiocasts: a primer in podcasting-how to do it. *American Journal of Roentgenology*, 186(6):1792–1796, 2006.

[17] J Sauro. 10 things to know about the single ease question (seq). *Measuring U, 2012*, 2012.

[18] Jim Wilson. *Node. js 8 the Right Way: Practical, Server-side Javascript that Scales*. Pragmatic Bookshelf, 2018.