Health and Technology (2024) 14:671–682
https://doi.org/10.1007/s12553-024-00859-2

ORIGINAL PAPER

Enabling reliable usability assessment and comparative analysis of medical software: a comprehensive framework for multimodal biomedical imaging platforms

Elena Denisova[1,2,3] · Eleonora Tiribilli[1,2,3] · Alessio Luschi[1] · Piergiorgio Francia[1] · Leonardo Manetti[2,3] · Leonardo Bocchi[1,3] · Ernesto Iadanza[4]

Received: 2 February 2024 / Accepted: 31 March 2024 / Published online: 1 May 2024

Abstract

Purpose A literature review reveals that, at the moment, all usability tests for Software as a Medical Device (SaMD) are designed in compliance with international standards but it also reveals a lack of formalization in the implementation and administration of such usability tests, which prevents the comparison of results from different tests for the same class of SaMD. This study aims to provide a reproducible usability testing framework for SaMD to establish a standardized protocol which can ensure repeatability and comparisons of similar SaMD for the visualization of medical images and data.

Methods The devised protocol aligns with international standards and literature recommendations for usability and human factors engineering. It encompasses participant selection, testing environments, equipment setup for various testing methods (HDMI vs. wireless), and hardware interfaces (keyboard/mouse vs. touchscreen), as well as the roles of the required testers. The protocol consists of two distinct sections: exploratory tasks and specific scenarios, to assess software functions and real-life tasks, respectively. Effectiveness and efficiency are evaluated using video analysis and a custom Stopwatch software, while user satisfaction is measured through post-test questionnaires.

Results The usability testing protocol was applied to a Multimodal Biomedical Imaging Platform All-in-One software developed by Imaginalis S.r.l. (Sesto Fiorentino, Italy) for validation. The results of the usability testing protocol applied to the case-study software demonstrate good values of software's effectiveness and efficiency, along with user satisfaction supporting the prior heuristic evaluation. The outcomes confirm the robustness, applicability, and reproducibility of the usability testing protocol, aligning with best practices.

Conclusions The proposed usability testing framework enables reliable usability assessment and comparative analysis of medical software. Furthermore, the obtained results can serve as a reference for assessing other biomedical imaging platforms under development or ready for release.

Keywords Usability · Medical devices · Medical software · Software as a medical device · Imaging platform

1 Introduction

In healthcare systems, new technologies and medical devices (MDs) have become more significant during the latter part of the last century [1]. The increasing use of MDs contributed to the improvement of health care and its quality, leading to new requirements regarding their characteristics and safety. In this sense, the usability of MDs gained interest, making the adoption of standards such as IEC 62366-1:2015 [2] essential for new products released in the market. As a consequence, usability is now part of the whole risk management process and is systematically considered and evaluated in the design,

✉ Ernesto Iadanza
ernesto.iadanza@unisi.it

1 Department of Information Engineering, University of Florence, Florence 50139, Italy

2 Imaginalis S.r.l., Sesto Fiorentino (FI) 50019, Italy

3 Eidolab, Florence 50139, Italy

4 Department of Medical Biotechnologies, University of Siena, Siena 53100, Italy

construction and implementation of MDs. The content of standards and scientific literature dealing with usability shows that the concept of usability is widespread [3].

Among standards addressing usability, UNI EN ISO 26800:2011 describes ergonomic principles for interfaces to improve safety, performance and usability by analysing the target audience, environment, goals, and expected results [4]. EN ISO 9241-11:2018 provides a framework for understanding the concept of usability and applying it to situations where people use interactive systems, other types of systems (including built environments), products (including industrial and consumer products), and services (including technical and personal services). The standard describes usability as "the extent to which a system, product, or service can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" [5]. Finally, EN ISO 25066:2019 addresses various usability testing approaches such as inspection and heuristics review of the interface, tests and post-market surveys [6]. The term usability is, therefore, very broad, and additional features to the list of ISO parameters are included in the scientific literature, such as ease of use, learnability, flexibility, attitude, and memorability [7], considered very important in promoting positive outcomes for both healthcare professionals and patients [3].

Given the complexity of MD usability issues across sectors, disciplines such as Human Factors Engineering (HFE) actually study the topic. HFE addresses the user interface in addition to the design of tools, machines, and systems, by taking into account human capabilities, limitations, and characteristics [8]. The goal is to ensure safe, comfortable, and effective use. Ergonomics, usability engineering, and user-centred design are considered synonyms [9]. The elements that influence user experience, especially in the use of medical software, need to be studied extensively because the way people perceive and use informatics tools can affect how well they understand the data and the outcomes of the analysis [10, 11]. Because of this, it is becoming more crucial to research the usability of medical software because a poorly made or designed interface might be challenging to use and lead to mistakes when using it [12]. Problems affecting medical device interfaces can also be the cause of recalls due to software interface errors. In this sense, interface usability testing is a highly-effective methodology for identifying usage errors and barriers, as well as a practical method for improving the efficacy and efficiency of products, services and systems [13].

Usability testing can be performed during the development stage to reduce errors and optimize the design. Furthermore, testing can be repeated after the product has been distributed, allowing for a prompt identification of problems, and highlighting critical issues [14]. Conducting usability evaluations can face numerous obstacles, such as the lack of adequate and validated guidelines [15] and clear indications of the environment, parameters, data type, and team characteristics. Moreover, the costs associated with evaluating the usability of MDs can be a major limitation. The description of the current context related to usability testing shows that there is an urgent need to define a systematic method to design a cost-effective usability validation procedure.

Zhang et al. [16] modified the existing heuristic evaluation method for software usability evaluation, applying it to medical devices, and using it to evaluate the patient safety of the device by identifying and evaluating usability problems. Shin and Lee [17] proposed a method to design and implement a time-cost effective test procedure for a comprehensive usability validation test by selecting the scenario with the lowest time–cost, starting from an activity diagram based on uFMEA (Use Failure Mode and Effects Analysis). Quality Function Deployment (QFD) is a structured approach to defining customer needs or requirements and translating them into specific plans to produce products to meet those needs. Despite being more suited to different kinds of analysis [18], it may also be used as a tool to develop usability evaluation models. However, using the QFD methodology has limitations, being not a procedure that searches for the optimal solution, but rather a technique designed to match designers' and users' needs in designing a product. Moreover, the QFD methodology is insufficient to understand the correlations among the physical design factors of a product, which is indeed a crucial outcome of usability tests [19].

The literature review reveals that, at the moment, all usability tests for Software as a Medical Device (SaMD) are designed in compliance with international standards or with few changes to the protocol, as described by Zhang et al. [16]. The review also reveals a lack of formalization in the implementation and administration of such usability tests, preventing the comparison of results from different tests for the same class of SaMD. The scope of this study is to provide a reproducible usability testing protocol for medical software to ensure repeatability and comparisons of similar SaMD for visualising medical images and data.

The developed protocol was applied to a Multimodal Biomedical Imaging Platform All-in-One, designed by Imaginalis S.r.l. (Sesto Fiorentino, Italy) with different users tested in a real-life scenario. To ensure the reproducibility of the testing protocol, custom software was designed and developed for recording and calculating test results (see Section 2.5).

2 Materials and methods

The Graphic User Interface (GUI) of the tested software allows users to view medical data, images, and other relevant information. The designed protocol admits only two

modalities for mirroring the screen during the test: cabled (via High-Definition Multimedia Interface - HDMI connection) or wireless. These modalities are available on all possible devices on which the tested SaMD can be installed regardless of its type: workstations, laptops, and tablets. These modalities are mandatory and cannot be changed to reduce variability and ensure the reproducibility of the testing protocol. Therefore, the machine on which the SaMD is installed must have an Operative System (OS) that supports display mirroring and at least one HDMI port or a stable WiFi connection.

According to the IEC 62366-1:2015 [2], the goal of usability testing is reducing risks. However, for this specific scope, the fastest access to all basic functionalities and overall user satisfaction was identified as the main goal beyond the evaluation of the risks. Moreover, the usability test may also reveal missing useful functionalities in the current release of the software [20] which were not detected with analytical approaches.

2.1 The multimodal biomedical imaging platform

The platform under testing is intended for pre-, post-, and intra-operative usage in the human and veterinary fields. Despite the imaging platform supports the visualization of any DICOM (Digital Imaging and COmmunications in Medicine) image, it was mainly developed to satisfy the needs of orthopaedics, and, therefore, is mainly focused on Computed Tomography (CT), fluoroscopy, and radiography. However, at the moment of preparing the first iteration of the usability testing procedure, the CT modality was the one with the highest level of readiness, leading us to limit the tasks in the administered tests, as described in this work, to CT acquisitions only. The proposed software provides a three-dimensional (3D) representation of the CT volumetric data, including Multiplanar Reformation (MPR) and 3D volumetric reconstruction views (see Fig. 1). The basic controls (zoom, pan, rotation, and scroll) enable the series navigation. The MPR functionalities include image enhancement, colour inversion, slab thickness control, annotation and measurement, histogram window control, and intensity presets. 3D functionalities cover different volume crop and rendering modes, auto-play, transfer functions control, and preset saving. Moreover, advanced functionalities such as compare, surgical planning, and multislice viewing are supported.

The set of proposed functionalities is based on the interview of a large group of orthopaedics, radiologists, and veterinaries. Additionally, the imaging software systems already present in the market were analysed. The GUI was developed according to Gestalt visual design principles. Moreover, the law that predicts the time taken to acquire a target (Fitts) and the human choice-reaction time (Hick-Human) were taken into account.

After completing the first proposal of the GUI design, the method of heuristic evaluation proposed by Zhang was applied. The result of the evaluation was analysed and the

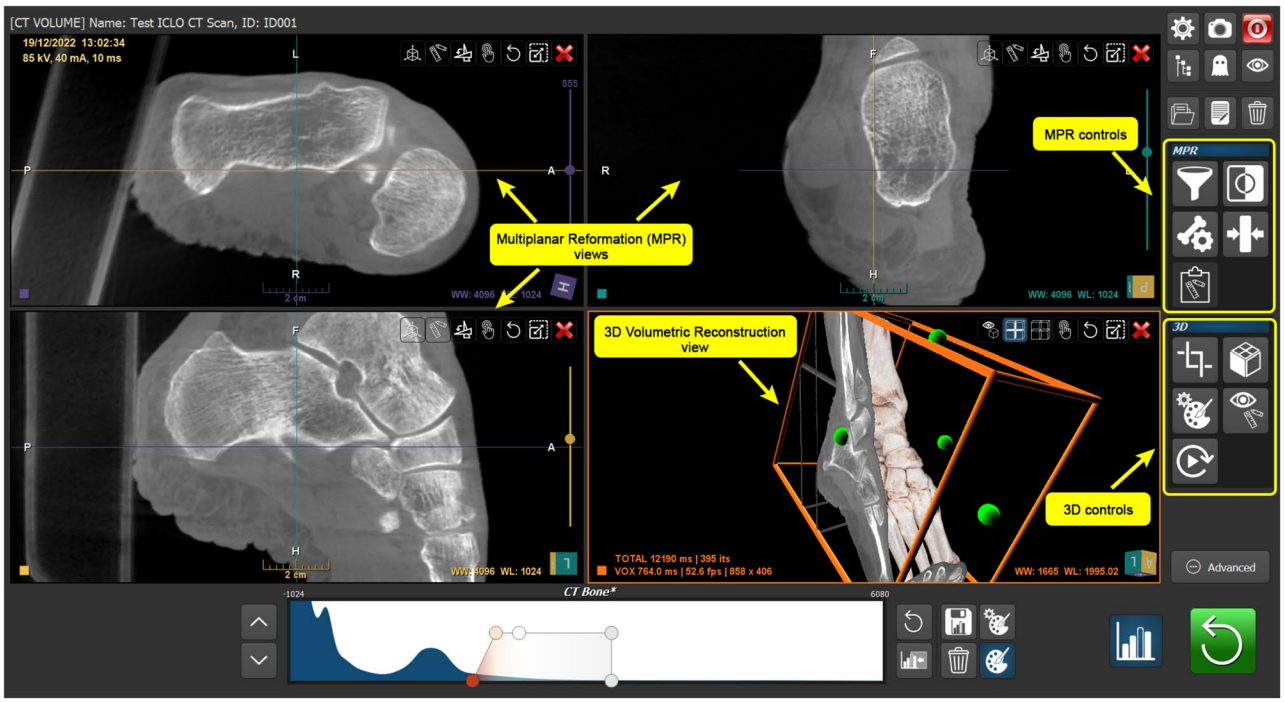


Fig. 1 Multimodal Biomedical Imaging Platform All-in-One

requested modifications and fixes were implemented. The heuristic evaluation process was repeated until only minor issues remained. At that point, the development was suspended and the current version of the software was used for usability testing.

2.2 Participants

The complexity of the software and the environment in which it is meant to be utilized determine the optimal number of participants. According to research [21–23], between three and twenty people can yield trustworthy results, with five to ten being a good start. Generally speaking, a higher number of testers is needed for more difficult, risky initiatives, but fewer participants are needed to test more creative ideas. Following these considerations, twelve volunteers were recruited to compose the testing population.

2.3 Environment

According to the above-mentioned requirements, only two modalities are allowed for mirroring the screen during the administration of the usability test: HDMI cable and wireless. In both modalities, the testing occurred in person.

2.3.1 HDMI testing

The environment for the HDMI test consisted of two adjacent rooms: one room designated as the test room and a second one as the observation room (see Fig. 2).

The chosen rooms were adjacent to allow an HDMI cable to be passed through. That enabled the duplication of the test machine screen on the monitor of the observer (see Section 2.4). As a precaution, the test machine was not connected to the Internet, while superfluous operative system processes were suspended to avoid compromising software productivity (see Appendix A).

2.3.2 Wireless testing

For wireless testing, no external cable was needed for mirroring, as a Google Meet session was set up on the same machine where the tested software was installed to share the screen. In this scenario, the superfluous operative system processes were not suspended, to evaluate their impact on the performance of the SaMD during the test.

2.4 Test conductors

The testing procedures require a minimum set of two people for conducting the test:

- **Moderator**: in charge of managing the progress of the test; responsible not only for administering the tasks, but also for observing the user's facial expressions, resolving any problems, and answering the possible questions arising during the session;
- **Observer**: responsible for reporting the user's performance of the tasks, tracking down the time taken to perform each task, and leaving comments on eventual issues and user's difficulties. In specific cases, when reporting the task performance while taking notes may result complicated, the presence of more than one Observer can be useful.

Optionally, a third person (namely the **Recorder**) can be involved. The Recorder observes and analyses the footage coming from the camera which frames the user from the entire scene's perspective during the test. In the absence of this third person, the footage recorded with the external camera can still be analysed after the conclusion of the test.

2.5 Equipment

A list of tools needed for carrying out a usability test for both HDMI and wireless modalities for touchscreen and mouse/

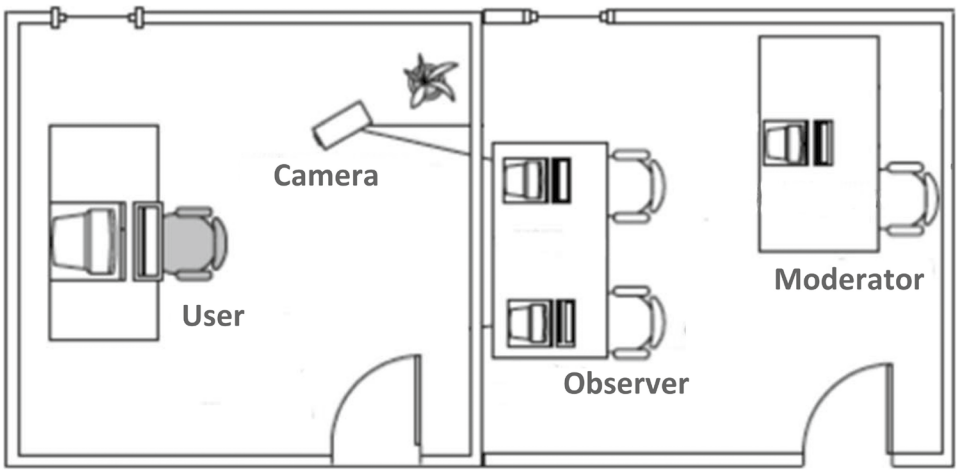


Fig. 2 Test Environment Scheme

keyboard configuration was defined in Appendix A. The settings for the different modalities/configurations are quite similar and are described in Tables A.1 and A.2.

Custom *Stopwatch* software was developed to ease the observer's tasks. The software is developed in C/C++ and the source code has been made publicly available at https://github.com/eletiri93/Stopwatch. The software enables the observer to record the amount of time spent on each task during the usability test while also noting any noteworthy user behavior: the GUI shows the current task, a stopwatch, and a space for taking notes. On the right side of the screen, a table summarises the recorded times and notes. After the completion of each task, the observer is able to export a CSV (Comma Separated Values) file containing the recorded times and notes.

2.6 Exploratory tasks and specific scenarios

Based on the experience acquired during the heuristics evaluation, the list of 55 exploratory tasks (see Appendix F) was produced. The objective was to ensure that participants did not become overly fatigued, while covering all of the most crucial capabilities within realistic time constraints. The tasks had to be clear, short, and as independent as possible (e.g., the failure of one task should not compromise the success of the following tasks). Moreover, four specific scenarios were developed with the help of an external radiologist consultant (see Appendix G). The exploratory tasks and specific scenarios can be modified to allow the testing protocol to be tailored to the specific application and the provided functionalities of the tested software.

2.7 Test evaluation

Tests were evaluated in terms of effectiveness, efficiency, and user satisfaction. The first two parameters were evaluated with the aid of a purposely developed *Stopwatch* software, notes taken by the Observer, and a camera footage. User satisfaction was evaluated by using a post-test questionnaire administered at the end of each test session.

More specifically, the effectiveness evaluates the participant's capacity to finish each suggested task, independent of the amount of time required. It is evaluated through the following score system:

- Score: 1. Failure. The user fails to complete the task, despite some suggestions
- Score: 2. Partial success. The task is partially completed or completed after suggestions
- Score: 3. Complete success. The task is completed without any difficulties or suggestions

On the other hand, the efficiency measures how fast each participant completes each assignment and is evaluated by timing the performance of each task. The above-mentioned *Stopwatch* software was used to record timestamps. The recorded value corresponds to the time interval between the end of the reading of the task by the moderator, and the moment when the user asserts the completion of the task.

Finally, scores given to each of the statements proposed in the post-test questionnaire were analysed to assess user satisfaction. The agreement scale was used, where **5** represented the *fully agree* and **1** represented *strongly disagree* options. All statements were designed to have consistent meanings (e.g., "The software is intuitive to use", "I had no problem using the basic features").

For comparison purposes, it is important to evaluate only the functionalities common to all thecompared SaMD. This enables a direct comparison of effectiveness and efficiency, task by task, and scenario by scenario. It is also crucial for assessing the user satisfaction evaluation: if the tasks and scenarios differ between the assessed SaMD, user satisfaction may be higher for the simpler test suite and lower for the specific and innovative features tested. However, the methods for comparing similar SaMD using the proposed testing protocol are beyond the scope of this paper.

2.8 Statistical analysis

Efficiency, effectiveness, and user satisfaction variables' distribution were tested using the Shapiro-Wilk test. The variables were statistically described as mean ± standard deviation (SD) for normally distributed quantitative data; median and interquartile range (IQR) for non-normally distributed data; frequency count and percentage for qualitative data. The difference between the groups of participants and the hardware modalities was evaluated using the independent *t-test* for normally distributed quantitative variables and the Mann-Whitney test for non-normally distributed quantitative variables. The relationship between efficiency, effectiveness, and user satisfaction was tested with the Spearman test. When a significant difference was detected, Cohen's d (for normal distribution) or Cliff's delta (for non-normal distribution) was calculated as a measure of the difference. The significance level for all tests was set to 0.05 ($p < 0.05$).

2.9 Pilot testing

Three weeks before the test administrations, the *Pilot Test* in touchscreen mode was carried out to validate the proposed test method, as well as the above-described environment, equipment, and tasks. Four persons were involved in the pilot test administration: a moderator, two observers, and a

user. The observers prepared all the necessary environment and equipment, while the moderator made sure that all the procedures were followed correctly. The pilot test was very useful, as it uncovered some task-related issues, such as the duration of some of them or the used lexicon.

2.10 Testing procedure

After the administration of the pilot test and its further analysis, the final testing procedure was set up. Before the participant arrived, the moderator and the observers verified the instrumentation, making sure that nothing was missing or abnormally working by using a dedicated checklist (see Appendix B). A single testing session performed by one user took about 1.5 h and consisted of the following steps:

1. Introduction to the test, including the software description and the desirable goals (about 5 mins). See Appendix C
2. Signing of the recording agreement (about 2 mins). See Appendix E
3. Compilation of the pre-test questionnaire (about 5 mins). See Appendix D
4. Execution of exploratory tasks (45-50 mins). See Appendix F
5. Specific scenarios (10-15 mins). See Appendix G
6. Compilation of the post-test questionnaire, including user's feedback (10-15 mins). See Appendix H

Figure 3 illustrates the entire workflow of the testing protocol as described in this section.

3 Results

Tests were conducted from October 12, 2022, to November 24, 2022. Figure 4 illustrates the environment and equipment configuration.

According to the observations, done in Section 2.2, twelve users were recruited for the testing. Six of the participants had a clinical background, and six had an engineering background (see the detailed distribution of the sample of

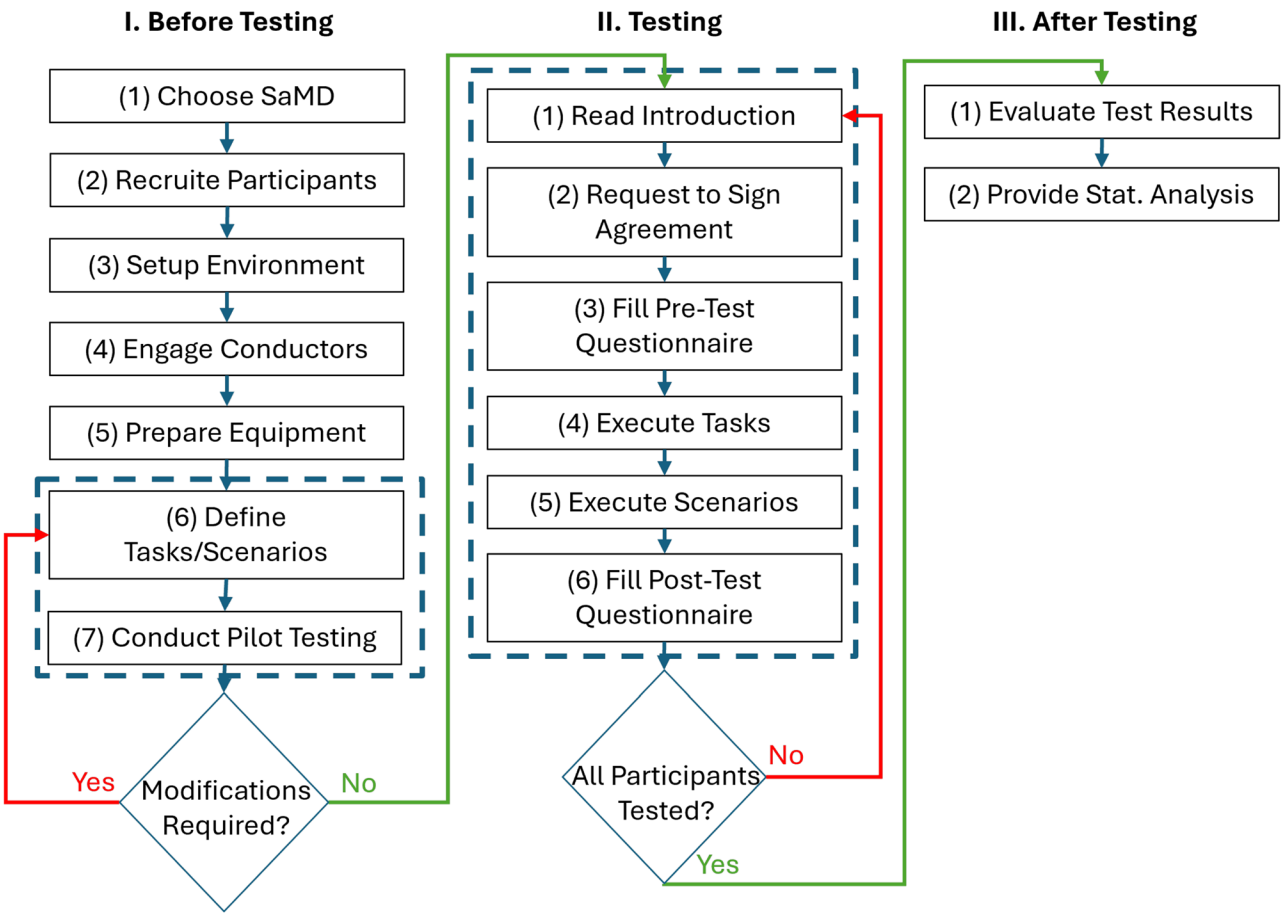
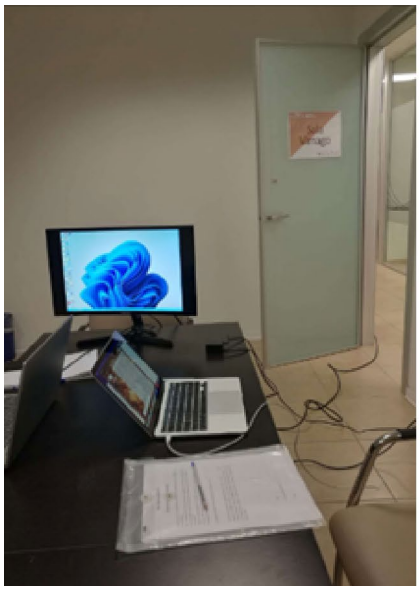
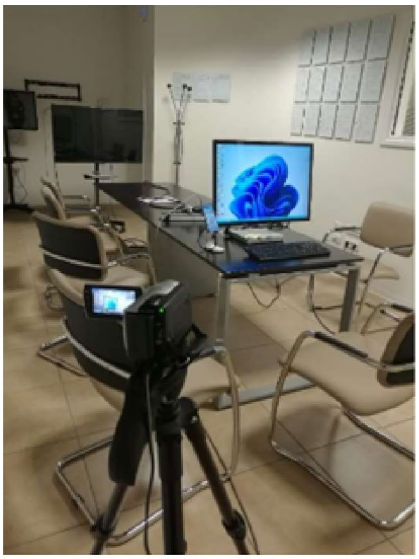


Fig. 3 The flow chart of the proposed testing protocol illustrates the sequence of steps to perform before testing (I), during testing (II), and after testing (III). Dashed-border boxes represent the steps that could be repeated more than once



(a) The observation room: laptop for the moderator; laptop for the observer; printed task list and supportive documentation; external monitor, connected via an HDMI cable to the test workstation.

(b) The test room for touchscreen modality: touchscreen display placed on the support with height regulation; workstation with tested software installed, connected through an HDMI cable to an external display; smartphone, placed on support with angular regulation.



(c) The test room for mouse/keyboard modality: traditional display, mouse and keyboard placed on a large desk; comfortable chair; workstation with tested software installed, connected through an HDMI cable with an external display; smartphone, placed on support with angular regulation; camera placed on a tripod.

Fig. 4 Test Environment & Equipment Setup

users in Fig. 5). Seven tests were conducted via HDMI connection and five using Google Meet. Ten tests were administrated in touchscreen modality, while the remaining two were administrated using a mouse/keyboard.

The statistical analysis of the participants did not reveal any significant differences between the groups of engineers and clinicians. Only two explorative tasks showed a significant difference in the efficiency (task 6 with Cliff's $\delta = -0.6944$, and task 33 with Cliff's $\delta = -1$). One task showed a significant difference in the efficiency between the two modalities (HDMI vs. wireless - test 7 with Cliff's $\delta = -0.8857$), and one showed a significant difference, again on the efficiency only, between mouse/keyboard vs. touchscreen (task 19 with Cliff's $\delta = -1$). Consequently, the results of all participants were combined for evaluation. Data collected during the pilot test was not included in the analysis, as the final version of the protocol was edited afterwards due to the subsequent considerations.

Figure 6 shows the statistical description of the effectiveness as frequency count and percentage for each explorative task, while Fig. 7 shows the statistical description of the efficiency for each task in terms of median and IQR. Figure 8 presents effectiveness and efficiency for the specific scenarios.

Figure 9 reports the statistical description of user satisfaction as frequency count and percentage for each assessing question.

The statistical analysis of efficiency and effectiveness indicated a statistically significant negative monotonic relationship ($\rho = -0.658495$, see Figs. 10 and 11), while no statistically significant relationship with satisfaction was found. Two outliers were detected.

At the end of testing, user feedback was collected to improve the existing functionalities in future software updates.

4 Discussion

A number of considerations were made at the end of the first usability testing iteration.

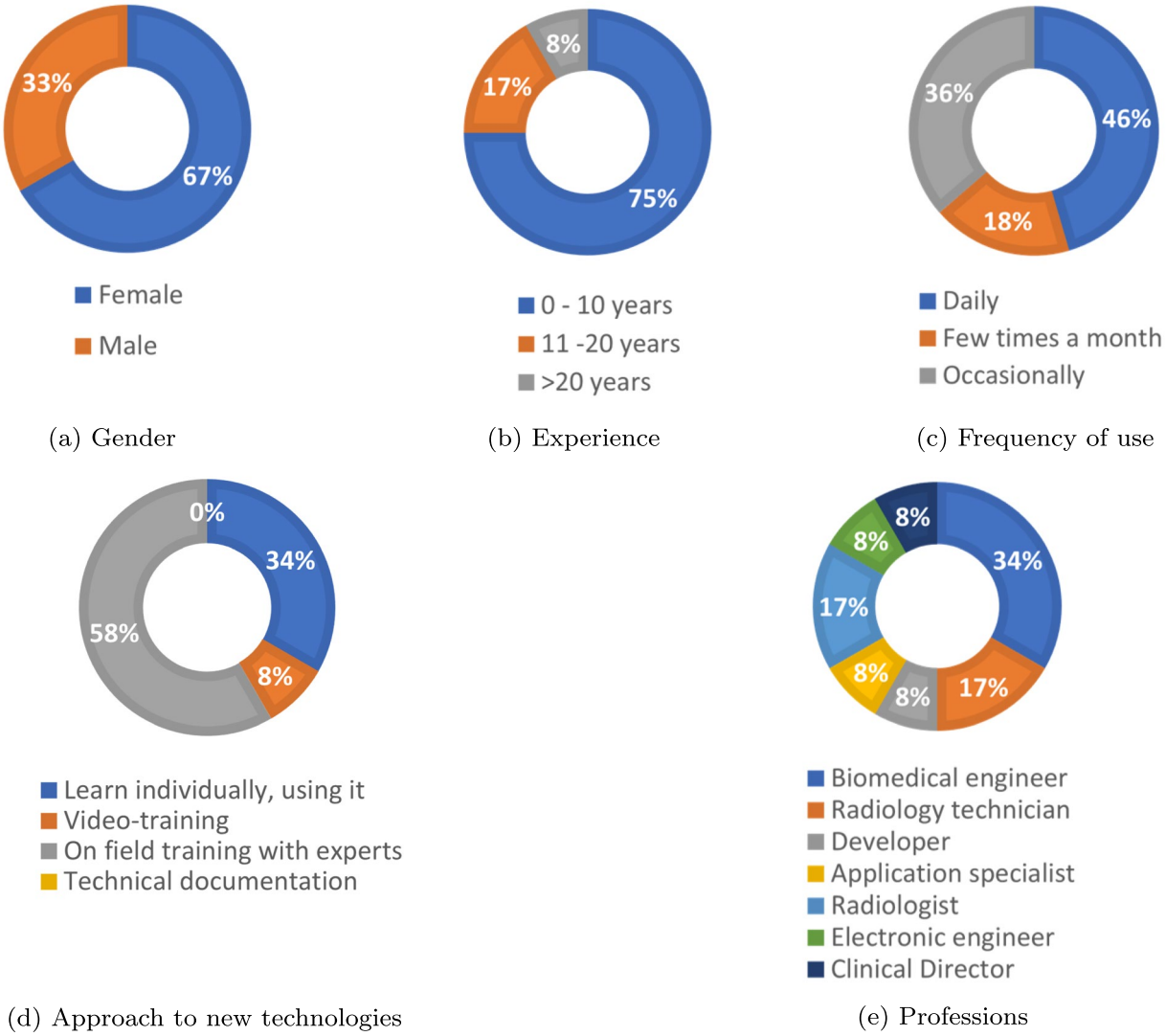


Fig. 5 Pretest questionnaire results. Distribution of test samples in terms of gender (**a**), experience (**b**), frequency of use of medical software (**c**), approach to new technologies (**d**), and professions (**e**)

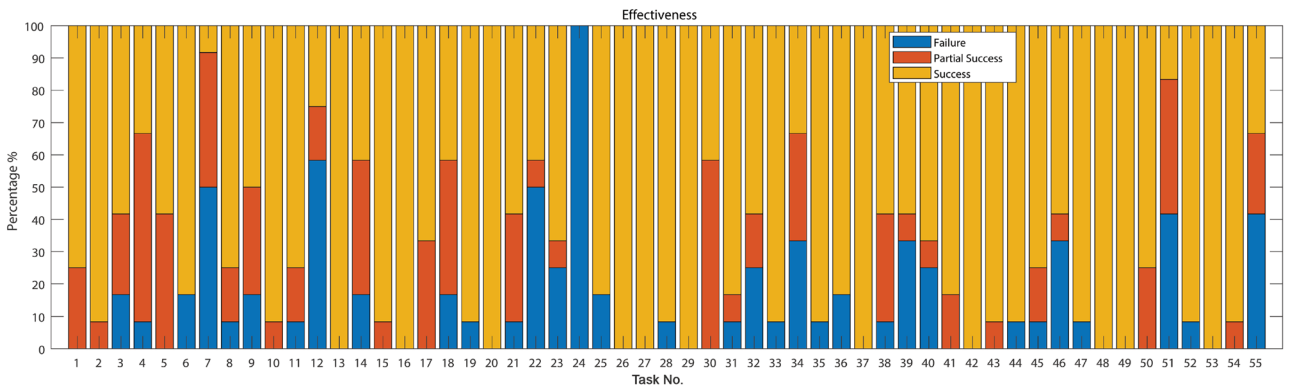


Fig. 6 Effectiveness for explorative tasks, evaluated as frequency count and percentage obtained for each task

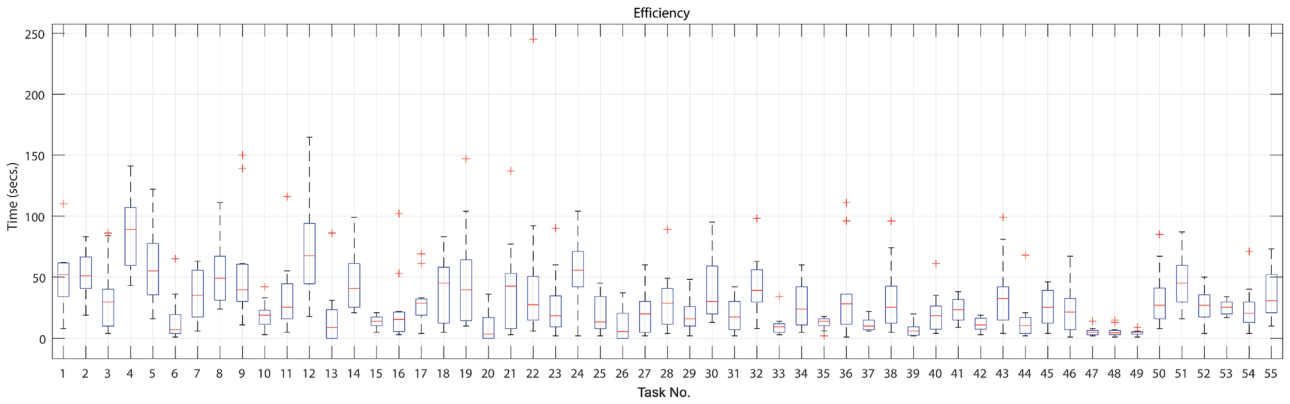


Fig. 7 Efficiency evaluation for explorative tasks. In the graph, the box represents the IQR, spanning from the 25th to the 75th percentile. Higher IQRs suggest significant variability in the time taken to complete the task. Smaller IQRs with medians (red line inside the box) close to the bottom indicate tasks completed very quickly. Outliers are denoted as (+)

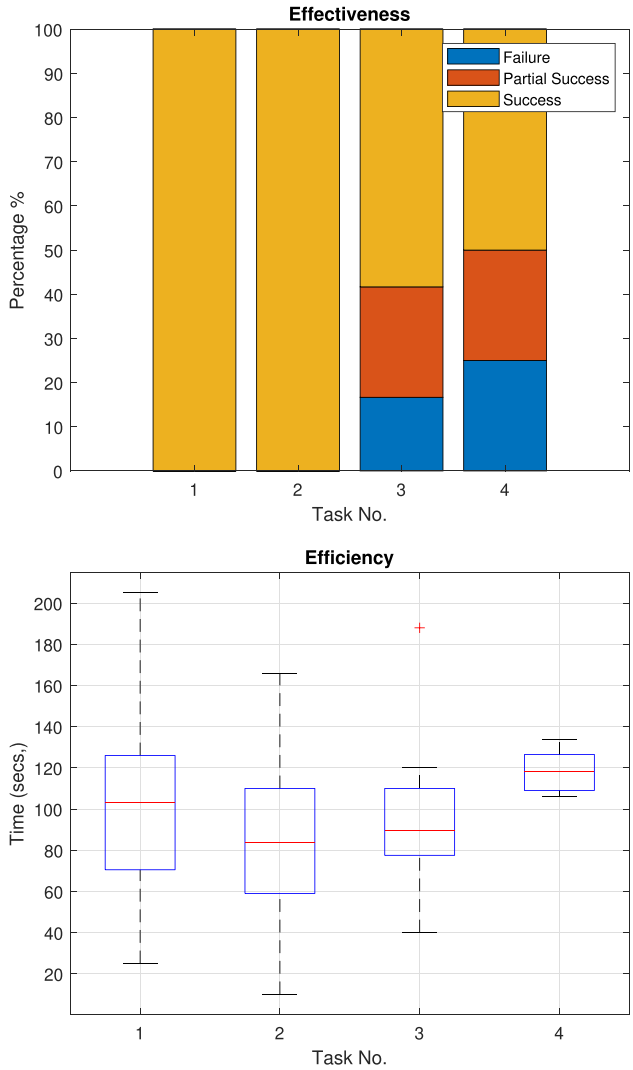


Fig. 8 Effectiveness and efficiency evaluation for specific scenarios. Effectiveness: frequency count and percentage. Efficiency: median, IQR, range, and outliers

As demonstrated in Section 3, the analysis of the two environmental testing modes (HDMI and wireless) revealed no significant variations. This suggests that any of the two modes can be chosen for testing, based on hardware specifications and the preferences of the test conductors. Furthermore, the results highlight that the suspension of superfluous operating system processes on the test machine has no impact on the performance of the software and the test suite, and it is therefore discretionary.

Remote testing may allow a larger group of participants while obtaining similar results as those performed in-person [24, 25], but it shows some limitations and drawbacks. The main challenge of remote testing consists of the participants being in charge of the whole environment and equipment set-up, which may lead to an unwanted and unexpected set of problems. Another issue which might be faced are the possible different time zones for the participants and the test conductors. Due to these reasons, the developed usability protocol takes into consideration only in-person testing, minimizing the sources of variability that remote testing could introduce, thus improving its reproducibility and consistency.

It is worth noting, that the test description and the terminology used can greatly influence the success (or failure) of the test results. It is indeed very important to carefully formulate the tests and the tasks, asking, when possible, for the help of an external specialised end user for implementing the appropriate terminology and lexicon.

The results show that none of the participants from the selected testing group chooses to learn a new technology by reading the technical documentation.

Overall, the results for user satisfaction showed that the tested software was deemed intuitive to use and that the majority of users believed that it could improve their work. Users did not encounter any difficulties using basic features

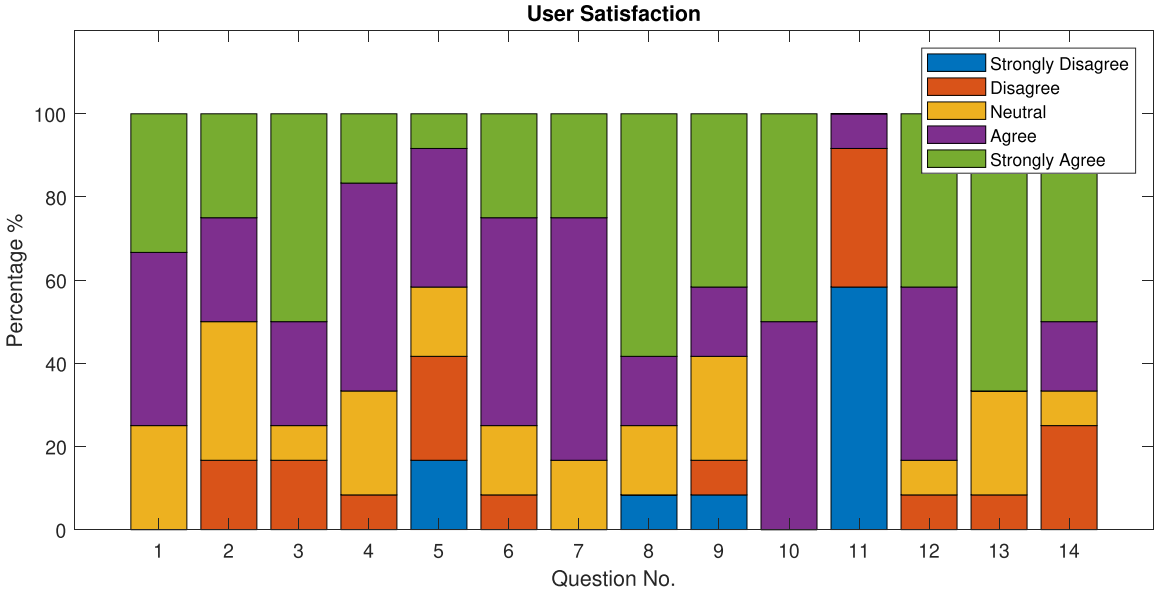


Fig. 9 Post-test questionnaire. Responses in a range from 1 to 5, were evaluated as frequency count and percentage obtained for each question

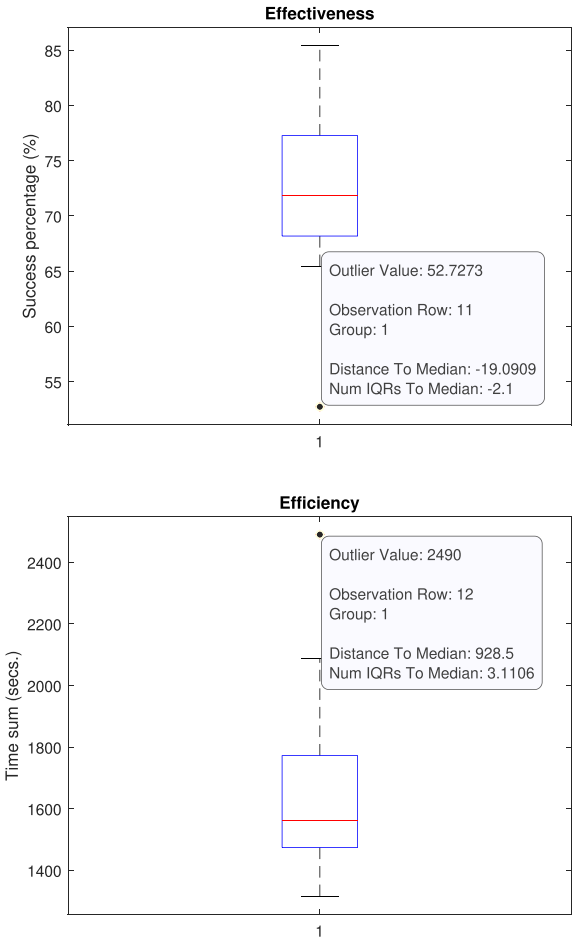


Fig. 10 Top to bottom: effectiveness in terms of success percentage; efficiency in terms of overall time spent. Two outliers were detected

in the mouse/keyboard or in the touchscreen configuration. However, the vast majority of them reported the need for initial support in the software usage. The touchscreen was preferred by the majority of involved users.

The results also highlight that statistically significant differences emerged only on the efficiency of four exploratory tasks: engineers were faster in performing tasks 6 and 33, task 7 was completed faster with HDMI compared to the

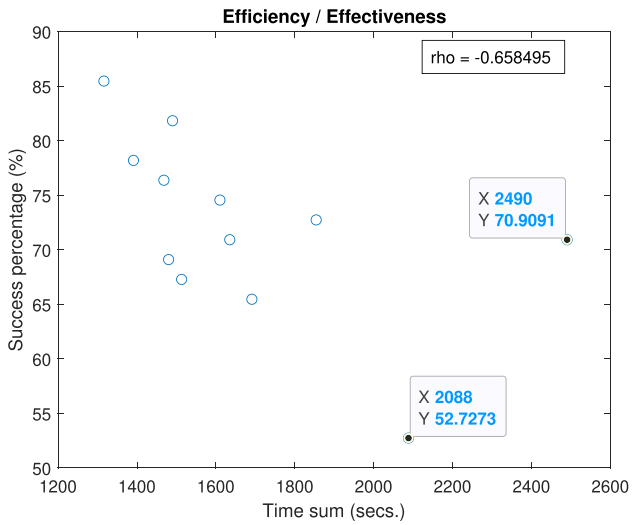


Fig. 11 Correlation between effectiveness and efficiency. Each point on the scatter plot represents the relationship between effectiveness (y-coordinate) and efficiency (x-coordinate), providing a visual representation of the distribution and spread of data and confirming the presence of two outliers

wireless setting, and task 19 was completed more quickly with mouse than with touchscreen. This aspect underlines that, together with the HDMI modality and the mouse configuration, the academic background/profession is the only personal feature among the analysed five users (Fig. 5) that can slightly influence the performance of the test in terms of time. Finding no significant differences in the effectiveness points out that the successful completion of each task is not related to any personal background or administration modality. Therefore, the effectiveness outcomes of the tested software reveal that some specific tasks (e.g., task no. 7, 12, 22, 24) actually shed light on possible weak points of the GUI. Moreover, the statistical analysis of efficiency and effectiveness indicated a statistically significant negative monotonic relationship, which suggests that as more time passes during task execution, the probability of successfully completing the task decreases.

5 Conclusion

This work presents a usability study on medical imaging software, focusing on the Multimodal Biomedical Imaging Platform All-in-One by Imaginalis S.r.l., compliant with the DICOM standard. The article provides a formal protocol for repeatable analysis, allowing for comparison of tests on similar SaMD for the visualization of medical images and data.

The protocol adheres to international standards, covering participant selection, testing environment setup, minimum number of required testers, as well as their roles and specific assignments, equipment requirements for HDMI and wireless modalities, and hardware interfaces (keyboard/mouse, touchscreen). The usability test comprises exploratory tasks and specific scenarios for evaluating software functions in real-life tasks.

Results were evaluated in terms of effectiveness, efficiency, and user satisfaction. The study validates the initial heuristic evaluation of the GUI, confirming the protocol's robustness, applicability, and reproducibility, aligned with best practices.

The exploratory tasks and specific scenarios are the only aspects of the proposed protocol that may need adjustment for different visualization SaMD. This is because different software may have varying functionalities.

Thus, the proposed usability testing framework enables reliable usability assessment and comparative analysis of medical software. Besides, the obtained results can serve as a reference for comparing biomedical imaging platforms under development or ready for release.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12553-024-00859-2.

Acknowledgements The authors would like to thank Laura Gatti for setting up the initial draft of the proposed protocol, and Alice Mati for designing the specific testing scenarios.

Funding Open access funding provided by Università degli Studi di Siena within the CRUI-CARE Agreement.

Availability of data and material Test equipment, checklist, test introduction, pre-test questionnaire, record agreement, exploratory tasks, specific scenarios, and post-test questionnaire are provided as Supplementary Materials .

Code availability Not applicable.

Declarations

Conflict of interest Elena Denisova and Eleonora Tiribilli work for Imaginalis S.r.l. (Sesto Fiorentino, Italy) and are currently PhD candidates at the Department of Information Engineering at the University of Florence (Italy). Leonardo Manetti works for Imaginalis S.r.l. (Sesto Fiorentino, Italy). Leonardo Bocchi coordinates the joint lab "Eidolab", between the Department of Information Engineering of the University of Florence and Imaginalis S.r.l. The other authors have no conflicts of interest.

References

1. Pecchia L, Pallikarakis N, Magjarevic R, Iadanza E. Health technology assessment and biomedical engineering: global trends, gaps and opportunities. Med Eng Phys. 2019;72:19–26.
2. ISO/TC 210. IEC 62366-1:2015. Medical devices — Part 1: Application of usability engineering to medical devices. 1st ed. International Organization for Standardization; 2015.
3. Formicola R, Amici C, Mor M, Bissolotti L, Borboni A. Design of medical devices with usability in mind: a theoretical proposal and experimental case study using the lepre device. Designs. 2023. https://doi.org/10.3390/designs7010009.
4. ISO/TC 159/SC 1. EN ISO 26800:2011. Ergonomics - General approach, principles and concepts. 1st ed. International Organization for Standardization; 2011.
5. ISO/TC 159/SC 4. EN ISO 9241-11:2018. Ergonomics of human-system interaction - Part 11: Usability: Definitions and concepts. 2nd ed. International Organization for Standardization; 2018.
6. ISO/TC 159/SC 4. EN ISO/IEC 25066:2019. Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE). 1st ed. International Organization for Standardization; 2019.
7. Shackel B. Usability – context, framework, definition, design and evaluation. Interact Comput. 2009;21(5–6):339–46. https://doi.org/10.1016/j.intcom.2009.04.007.

8. Iadanza E. Clinical engineering handbook. 2nd ed. 2019. https://doi.org/10.1016/C2016-0-04298-5.
9. Gosbee J. Human factors engineering and patient safety. Qual Saf Health Care. 2003;11:352–4. https://doi.org/10.1136/qhc.11.4.352.
10. Luschi A, Caltagirone L, Mondovecchio C, Miniati R, Iadanza E. Assessing the impact of a CIS/PACS technology for a cardiology department using QFD methodology. IFMBE Proceedings. 2016;57:965–8.
11. Iadanza E, Fabbri R, Luschi A, Melillo P, Simonelli F. A collaborative restful cloud-based tool for management of chromatic pupillometry in a clinical trial. Heal Technol. 2019;10:25–38.
12. Machado Paixão-Cortes VS, Dos Santos da Silva Tanus M, Paixão-Cortes WR, De Souza ON, De Borba Campos M, Silveira MS. Usability as the key factor to the design of a web server for the CREF protein structure predictor: the WCREF. Information. 2018. https://doi.org/10.3390/info9010020.
13. Hass C. A practical guide to usability testing. Springer International Publishing; 2019, Ch. A Practical Guide to Usability Testing, p. 107–24.
14. Dumas JF. A practical guide to usability testing. Intellect Books; 1999.
15. Russ AL, Saleem JJ. Ten factors to consider when developing usability scenarios and tasks for health information technology. J Biomed Inform. 2018;78:123–33.
16. Zhang J, Johnson TR, VL Patel, Paige DL, Kubose T. Using usability heuristics to evaluate patient safety of medical devices. J Biomed Inform. 2003;36(1):23–30 Patient Safety.
17. Shin J, Lee H. Optimal usability test procedure generation for medical devices. Healthcare. 2023. https://doi.org/10.3390/healthcare11030296.
18. Luschi A, Monti M, Iadanza E. Assisted reproductive technology center design with quality function deployment approach. IFMBE Proc. 2015;51:1587–90.
19. Jin B, Ji YG, Choi K, Cho G. Development of a usability evaluation framework with quality function deployment: from customer sensibility to product design. Hum Factors Ergon Manuf Serv Ind. 2009;19:177–94. https://doi.org/10.1002/hfm.20145.
20. Markonis D, Holzer M, Baroz F, De Castaneda RLR, Boyer C, Langs G, Müller H. User-oriented evaluation of a medical image retrieval system for radiologists. Int J Med Informatics. 2015;84(10):774–83.
21. Virzi RA. Refining the test phase of usability evaluation: how many subjects is enough? Hum Factors. 1992;34:457–68.
22. Faulkner L. Beyond the five-user assumption: benefits of increased sample sizes in usability testing. Behav Res Methods Instrum Comput. 2003;35:379–83.
23. Turner C, Lewis J, Nielsen J. Determining usability test sample size, vol. 3. CRC Press; 2006, Ch. Determining Usability Test Sample Size, p. 3076–80.
24. Andreasen MS, Nielsen HV, Schrøder SO, Stage J. What happened to remote usability testing? an empirical study of three methods. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '07. Association for Computing Machinery; 2007. p. 1405–14.
25. Brush AJB, Ames MG, Davis J. A comparison of synchronous remote and local usability studies for an expert interface. In: Dykstra-Erickson E, Manfred T. CHI '04 Extended Abstracts on Human Factors in Computing Systems. New York: Association for Computing Machinery; 2004. pp. 1179–82.