



Risk-based Comprehensive Usability Evaluation of Software as a Medical Device

Noemi Stuppia¹, Federico Sternini^{1,2}^a, Federica Miola¹, Giorgia Picci³, Claudia Boarini³,
Federico Cabitza⁴^b and Alice Ravizza¹

¹USE-ME-D srl, I3P Politecnico di Torino, Torino, Italy

²PolitoBIOMed Lab, Politecnico di Torino, Torino, Italy

³Dedalus spa, Firenze, Italy

⁴DISCO, Università degli Studi di Milano-Bicocca, Milano, Italy


Keywords: Usability, Software as a Medical Device, User Interface.


Abstract: Introduction: Usability evaluation is a core aspect in risk assessment of medical devices, as it aims to ensure the device interface safety, avoiding that usability problems at interface level are not related to harm. Methods: Our research group applied our risk-based approach, international reference standards and guidelines to the usability evaluation of a large family of SaMD. The methodology used for the evaluation is an elaboration of regulatory prescriptions and is composed of a combination of quantitative and qualitative methods. In particular, the usability evaluation is structured in a two-stage evaluation composed by formative and summative evaluation. The formative stage is propaedeutic for the planning of the summative evaluation. The final assessment included the analysis of quantitative data collected through three questionnaires and a user test. Results and discussion: Risk-based task analysis led to the identification of the most common use error emerged during the user test performance. The three questionnaires led to different results: Heuristic analysis allowed the identification of violations to the heuristic principles as perceived by the users and their severity; SUS questionnaire provided an indicator of general device usability; the interview identified the usability problems of each device with respect to their functionalities. Conclusions: The study allowed the extensive assessment of the devices, the identification of usability issues, and the classification in terms of criticality of each issue. In conclusion the study led to different proposals to solve the issues and design changes.

1 INTRODUCTION

Patient care, two words that carry an array of diverse practices that have a shared scope: to prioritize patient health while limiting any unnecessary or potential harm." To err is human", is a long-lasting thought that in 1999 opened the conversation on the consequence of human error in healthcare and triggered a new approach to improve patient safety through design (Institute of Medicine (US) Committee on Quality of Health Care in America, 2000). The removal of all the root causes of any hazardous situation is unfeasible, but by factoring in the human element within the

design process, the manufacturer can mitigate risks associated with proper use. The risk mitigation approach is a core regulatory requirement for medical device approval by authorities, worldwide. It is specifically addressed in the European Medical Device Regulation EU 2017/745. International standards apply, and IEC 62366-1:2016 provides a systematic approach for the manufacturer to analyze, specify, develop and evaluate the usability of a medical device as it relates to safety (International Electrotechnical Commission, 2020). The standard provides a framework that is suitable for all medical devices. Nevertheless, no indication in the standard is

^a <https://orcid.org/0000-0002-5510-2296>

^b <https://orcid.org/0000-0002-4065-3415>

provided regarding the selection of the most adequate methods for each medical device and research is progressing for the proposal of best usability evaluation process (Kwak et al., 2021; Schmettow et al., 2017). Continuing the work of the research group for the identification of the most adequate strategy for the usability evaluation for each device (D. Ravizza et al., 2019), in this paper, we present how our team chose the regulatory-approved methods for usability assessment and used them for the usability assessment of 10 medical software of different complexity, and the result of this methodology.

2 METHODS

The international standard aims to reduce the risk of medical errors due to poor interface design through the definition of methods of usability evaluation of the interface. Similarly, the standard also applies to the documentation that accompanies a device and to the training of the intended users. Following the standard requirements, we defined an integrated and comprehensive approach defining a two-step usability evaluation phase that includes both methods available at the state of the art and innovative methods proposed within the context of this study and previous studies (Sternini et al., 2021). Each phase has a different purpose; therefore, different techniques are used accordingly (D. Ravizza et al., 2019). In each phase, we defined the chosen techniques and the outcomes that each step should provide.

2.1 Formative

The first phase described in the IEC standard is the formative evaluation, which aims to iterate the design of the user interface to achieve the minimization of usability-based risks.

The first activity was the definition of the software functions and requirements, core to planning all the further testing activities. The software primary operating function, as already defined in the technical file, were paired with one testable requirement. The technical testable requirement was defined as the capability to complete the primary operating function with predetermined usability criteria that are consistent with the intended use (e.g. in case of primary operating function for patient incoming in an emergency ward, the technical testable requirement is the capability of the device to support triage, that is to allow for the efficient association between the patient and the proper colour code). The target quality level was identified in terms of the number of times the

product would meet the testable requirement as well as the number of bugs and unclear user interface features, for example icons.

Subsequently, the formative evaluation was performed following these iterative steps:

- Preliminary analysis: it included as a first step a general, "quick-and-dirty" overview of the core product functionalities and general interface aspect. We completed a cognitive walkthrough (International Electrotechnical Commission, 2016) and brainstorming, in a team composed of usability experts, to identify potential use errors, applicable standards, known errors and complaints, and relevant literature. The main goal of this step is the definition of interface strengths and weaknesses. The latter ones are then mapped into a device risk analysis by using the relevant questions listed in risk management international standard ISO 14971 (International Organization for Standardization, 2019) as a reference. With this introductory knowledge we drafted a task list, which is defined as a sequence of actions that are necessary to achieve the task goal for each operating function and each user profile foreseen in the software.
- Detailed analysis in a team supported by a device expert (e.g. designers, product specialist): this phase began with a brief training of our usability team. The training was conducted confirming and updating the task list drafted in the preliminary phase and then describing and simulating the core user experience scenarios, enabling the product experts to identify any interface pitfalls, bugs or other details in the device that were not yet addressed by the development team. The training session provided the usability experts with the proper knowledge to:
 - Evaluate the primary operating functions, defined as functions that are directly related to the device safety or that are frequently used.
 - Execute the task analysis, which is a technique aimed to understand the process of learning of ordinary users by observing them in real-life situations; it describes in detail how they perform their tasks and achieve their intended goals.

The evaluation of the primary operating function was completed by assigning to each function a score representative of the interface problems encountered during the function analysis. The score ranges from 0 to 4, where 0 regards no problem, while 4 is the value assigned to the highest risk related to the device. The following scale was used. This scale was used for answers both in the heuristic questionnaire and in the primary operating function evaluation, to minimise training of participants, ensure consistency and allow the comparison of the scores.

- 0 = no problem
- 1 = Before using it I have to spend some time figuring out how to do it
- 2 = Complicated use and makes me nervous
- 3 = Impossible use and/or incomprehensible instructions
- 4 = Possible risk for the patient (patient misidentification; clinical pathway interrupted)

We analysed the results of the evaluation through a radar plot to have a glimpse of the usability risk profile of the device. This plot, presented in Figure 1, allows for immediate comprehension of the approval of the design interface and whereas the device is intuitive and requires minimal effort to complete the main task. As the area underlying the dots increases, the graph shows that the task is not well understood and accepted by the evaluators.

At the end of the formative evaluation, we designed two novel questionnaires intended to ease the data collection during the summative phase:

- **Heuristic evaluation:** this is a useful, efficient, and low-cost method that we proposed to evaluate patient safety features of medical devices through the identification of usability problems. Furthermore, it provides an estimation of the severity of these problems. The questionnaire is intended to be composed by carefully formulated questions and closed answers. The questions are designed by the usability experts so that the user can assess Zhang's heuristics, without the need for training regarding the underlying theory (Zhang et al., 2003). Each relevant heuristic is represented by at least one question formulated to lead users to identify any heuristic violation. The questions should have a scope broad enough to allow the user to answer the question on the base of its own experience, without the bias given by the moderator experience. Therefore, we designed the questionnaire tackling the specific heuristic defined by Zhang with proper terms

for the device type, but without including any reference to specific situations (e.g., Are the icons and interactions consistent with devices you habitually use?). Then, the severity of the heuristic violation is assessed through the provision of a score; scores are presented with meanings associated with the single user experience and perception of risk. In this way, the user could quickly answer without any additional training, and the moderators could relate the scores given by the users with the violation severity.

- **Interview:** The questionnaire was designed so that each question was consistent with one primary operating function of the device and to be representative of the user interactions. It is intended to provide an evaluation of the primary operating functions as perceived by the intended users.

For each technique, we analysed and represented the best response, worst, mean and median for completeness.

When all these activities are concluded, and the results of the formative evaluation are correctly reported, the usability assessment can proceed to the summative evaluation.

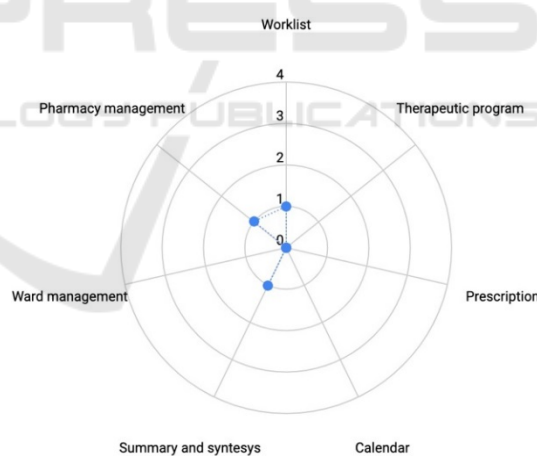


Figure 1: Representation of an interview response radar plot, the significance of the scores is reported in the text.

2.2 Summative

The summative evaluation aims to assess the adequateness of the user interface by considering the outcome in terms of the risk of potential user errors and by providing evidence that all minimization of known causes of use error is in place.

The core technique that our group employed was the user test. We recruited 15 participants, a practical

minimum number of participants for human factors validation testing (Health, 2019). The tests were carried out in a simulated-use environment to ensure adequate observation. Additionally, to ensure patient privacy, we created adequate simulated patients profiles according to the mode principle (A. Ravizza et al., 2020). By doing so, patient privacy was ensured while we also allowed the test participants to interact with realistic data. The mode principle allows describing simulated patients using the data that are most frequent in the patient population, which are considered more representative than mean values because the latter can be inconsistent with real data (A. Ravizza et al., 2020). The test scenario was designed by referring to the task analysis conducted in the formative evaluation.

The task list, that is the main script for the user tests, was implemented to test all of the primary functions. Thus, within the same scenario, the user might be asked to do multiple tasks per feature (e.g., inserting a new patient in the EHR by searching her from a contact list or by inserting the personal information in a search bar). By allowing the presence of tasks sharing similar sub-steps, the test participant had the option to understand the navigation pathway better and conclusively give an informed opinion on the interface characteristics based on multiple interactions rather than a single one.

At the beginning of the user test, due to the complex interface of the medical device, we invited the device expert to conduct an introductory speech and a brief training session to give proper information about the intended use of the device and the purpose of the user test. The speech aims to ease the participants into the experience, by providing them with a basic introduction on what they will later see. More importantly, the speech helps them focus on the crucial aspect of their contribution, which is to report what they perceive, what they reason about, and which action they will take accordingly. This decomposition allowed the interviewer, during the test, to assess the level of individual user interaction with the specific task. Moreover, by using the PCA approach (International Electrotechnical Commission, 2016), the interviewer was able to identify the main categories of use errors which stemmed from perception, cognition and action errors. Besides, potential use problems can be targeted by asking the user about the consequences of a failed task. As prescribed by the standard, we trained the moderators to not intervene during the user test, but to limit the intervention only if the user could not complete the task autonomously.

The participants tested the functionalities of the devices following the task list with the supervision of the moderators and, for each user task use, the moderators evaluated the user actions with the following policy:

- ok: the task was completed without error
- ue (user error): the user was not able to complete the task and requires help from the moderator, or the user made an error that had no impact on the patient (e.g., the input of the password with the caps lock on), or the user knowingly neglected to complete a task
- ce (critical error): the user made an error that has an impact on clinical risk. e.g., ignored a notification regarding critical clinical risk (for example, drug interaction); skipped the patient identification.
- te (technical error): task not completed due system failure.

The purpose of analysing the task performed by users was to evaluate the presence of ce (critical error) and to identify which task may have caused uncertainty or confusion. The result of the task completion is an informative source of improvements for technical manuals and training sessions, allowing the designers to understand which require additional clarity in the instructions and more examples during training. Additionally, it can provide feedback on the unsolved technical issues occurring during normal use.

Additionally, during the user test, the participants may comment on the device performance (in terms of usability), and the moderators may propose open-ended questions to the users, which may lead to additional problems and uncertainty information and further product improvement. We encourage collecting the notes from the user impression; once vetted, they can be a valuable source for further product improvement.

At the end of the simulated use, we let users evaluate the devices with three different metrics: interview, heuristic questionnaire, and System Usability Scale (SUS). The first two are the techniques designed during the formative evaluation, while the SUS questionnaire provides a "quick and dirty," reliable tool for measuring usability. It consists of a 10-item questionnaire with five response options for respondents; from Strongly agree to strongly disagree (Jordan et al., 1996).

The questionnaires were briefly described by the moderators to the participants and then filled autonomously by the participants.

Usability test Process

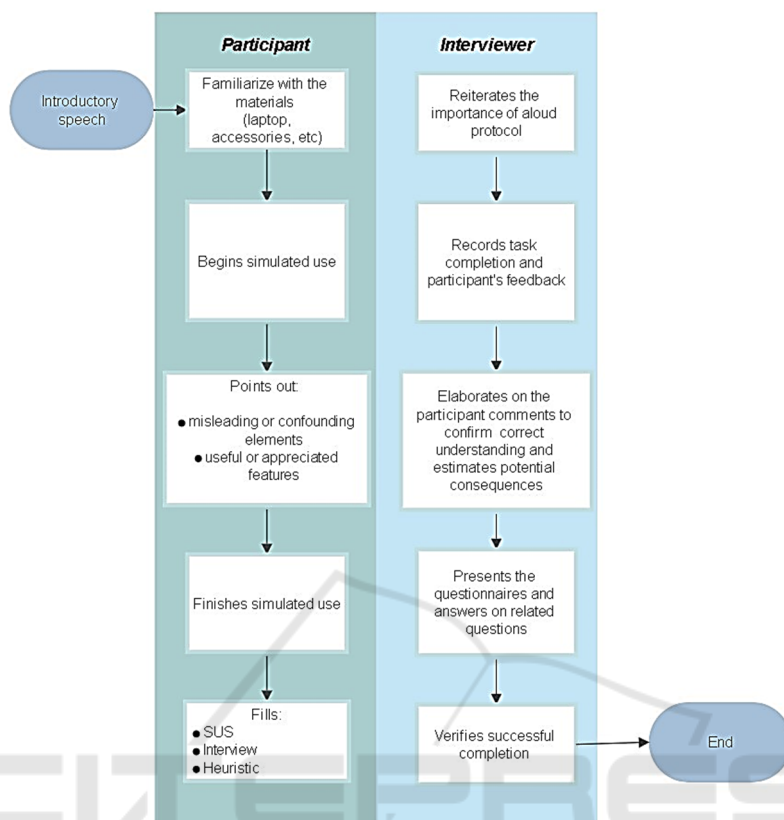


Figure 2: This caption has more than one line so it has to be set to justify.

After the testing phase, all the results are collected and analysed. The analysis and the result report complete the analysis of the medical device interface and to assess its usability. A summary of the test process is presented in Figure 2.

3 RESULTS

The methodology described above was applied to the usability evaluation study of ten products manufactured by Dedalus SPA. The study was aimed at the usability evaluation of 10 different SaMDs designed to help the management of the care path of hospital patients, in different wards. Considering that the work domains varied between groups, we analysed different work domain ontologies that are associated with different levels of intrinsic complexities and corresponding risks. They were grouped in families according to the intended use:

- EPMA, identified by the intended use “ Electronic prescribing and medicine

administration” . The intended users are oncologists, nurses, nuclear medicine physicians, radiologists, radiology technicians and general physicians.

- AID, identified by the intended use “ Operating and emergency room assistance” . The intended users are trauma surgeons, orthopaedics, general surgeons, anaesthesiologists, nurses and perioperative nurses.
- ERH, identified by the intended use “ Electronic health record and screening” . The intended users are general practitioners.

The recruitment of the user group included people with different level of familiarity with the software. Some of them used similar devices; some had previously used the specific device, while others only used paper records in their administrative and medical operations. The difference level of experience with different age group allowed for a complete result that can reflect the real-use applications.

Table 1: Summary of most relevant use-errors.

Software	Use error	Severity	Principle violated	Prevalence	Recommended solution
Software A	User failed to add a product to the warehouse	not critical	Giving control and freedom to the user about reversible actions	25%	Add another selection option and provide a double-check with a summative message (pop-up windows) when the user confirms the action
Software B	The user was unable to correctly find and fill the mandatory tabs to require and exam	not critical	Encourage recognition rather than memory	40%	Target this scenario in the training session and modify the wording associated with the task
Software C	The user (nurse) misread the information "reported allergies" and read "unknown allergies". Then she thought that the allergies of the patient were already asked the patient and that there is no known allergy, while the allergy section included several allergies.	critical	Provide a simple and natural language: any data that the user has to insert must Be presented in a completely similar way to the paper format.	20%	The information about the presence of this section and the meaning of the keywords used in summary included in the dashboard must be for future installations at customer sites

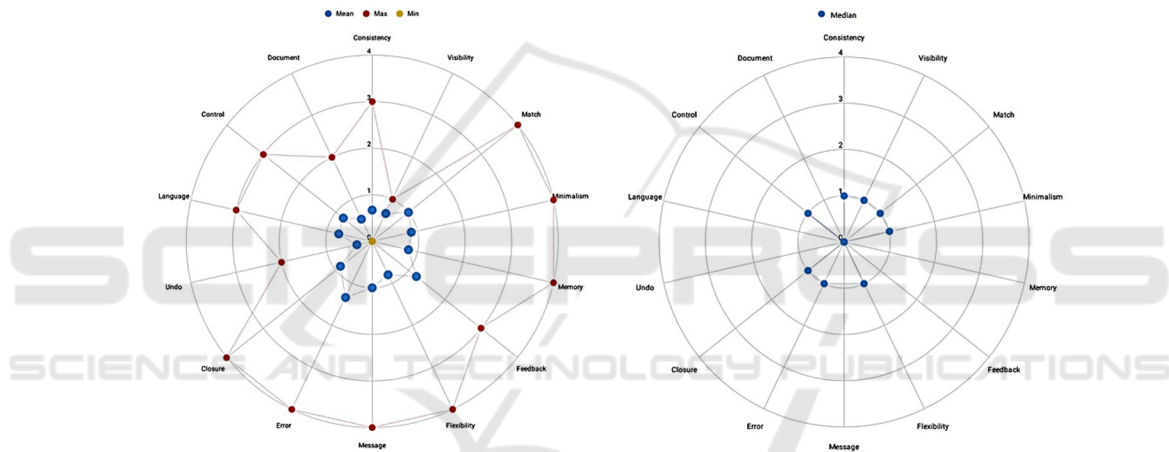


Figure 3: Example of the heuristic response radar plot. Score 0, No problem, 1 The use is a bit complicated, the user needs time to get used to the device use, 2 The use is complicated, the user could get nervous during the device use, 3 The use is impossible due to design issue or non-understandable instructions, 4 Possible patient risk.

3.1 Task Completion

During the analysis of the task completion, it is possible to quantitatively estimate the percentage of the correctly executed tasks, technical errors and evaluate the severity of the use error. The evaluation of the severity of a use-error is typically not uniquely defined and strongly relies on the interviewer's judgement. It may lead to a modification of the risk-analysis or just to a suggestion for further product improvement. In this study, the moderators were trained in advance, to minimise bias, to assign the “critical error” class to actions that could expose the patient to serious risk. Excluding the technical failures, all other errors are then classified as user errors. We summarised the most significant examples of user errors and their classification in Table 1.

3.2 Heuristics

As cited in the methods section, the results were evaluated according to their median, average, best and worst values. When the mean score is lower than one, it represents a consensus opinion related to the specific heuristic that is partially favourable for the product. The score equal to one is considered as a threshold to identify efficacy problems related to the specific heuristic. Any score equal to or higher than two should require further analysis since it may be a source of patient risk. In the reported wards example, related to the software for emergency wards management, by observing the worst-case evaluation, we detect multiple at-risk categories. To better understand the meaning of the heuristic tag, we reported the

Table 2: Example of primary operating function and testable requirement.

Primary operating function	Physician	Nurse	Interface testable technical requirements
Patient incoming	no	yes	The patient incoming function shall allow assigning a color coding according to predefined clinical criteria in an efficient manner
List	yes	yes	The list function shall allow the monitoring of the whole set of activities of the ER ward in an efficient manner
ER ambulatory patient management	yes	yes	The function of patient management shall allow all the clinical personnel, according to their privileges, to update the clinical record in the ER ward in an efficient manner

graphical representation of the results in figure 3 and the related comment:

- Icon and colors: Widely differ from the typical representations both in graphics and in colour
- Minimalism: The user interface has unnecessary and redundant information
- Memory: The user is required to remember much information about the patient and his therapeutic path while using the software
- Flexibility: The software does not accommodate user desired variation
- Message: The error message is not clear or helpful
- Closure: The user is often unsure if an operation was completed or not.
- Error: The system may be misleading

It should be noted that at least half of the user population choose values above the "low efficacy" score one.

3.3 SUS

The ten questions of the SUS questionnaire result in the System Usability score. The average value is 68. Generally, any score higher than 81 is an optimal response, while a score lower than 51 is critical and unacceptable. In this study, any of the analysed devices obtained a critical and unacceptable score, but the result showed that there is room for improvement.

3.4 Interview

As previously explained in the method section, we associate a testable requirement for each primary operating function that directly relates to safety. This is aimed to assess whether the software successfully provides a testable solution in the user-interface. This is generally done by implementing alarms, color

coding or dedicated icon. An example is provided in table 2.

Comprehensively, most of the primary operating functions were evaluated as acceptable by the participants, with most frequent scores equal to 0 or 1 (no risk area). However, any response equal to or higher than two belongs to the risk-area and needs to be addressed. We report here a few examples of the problems identified through this tool:

- One usability problem was detected by two different participants (physicians) in the same task; the physician pointed out that the clinical report at discharge was missing information regarding the drug therapies. However, the issue was already identified and resolved by the manufacturer but not available in the customization of the software designed for the test set. The contents of the clinical report at discharge are provided in a complete list by the medical device software, and the actual inclusion of one or more of the therapies (for example administered, planned, required at discharge, home therapy) is a customisation choice. Consequently, to avoid uncertainty for the users, we encouraged to highlight what is customizable and what not during the training session.
- Both our team and expert users detected a usability problem related to the drug administration task. The hazardous element was the lack of a time reference for the administration in the primary therapy window. The software already allowed access to that information, but in a different window. The results of the interview completed by the user confirmed the usability risk identified by the experts in the formative phase. The use of the same metrics and plots for the formative evaluation and interview analysis allowed us to identify the consistencies and the gaps of the formative analysis when compared to real use.

4 DISCUSSION

By implementing multiple evaluation methods for the usability evaluation, we aim to collect as many information as possible. While a questionnaire or task-completion can provide a numeric result, it cannot identify specific design problems that may need to be addressed. On the other hand, when paired with multiple qualitative methods, such as moderator observations, PCA techniques and open-ended questions, these techniques allow for valid quantification of the criticality of each issue, while qualitative methods allow the comprehension of all design flaws encountered during the usability testing.

Each one of the methods used during the study covered different aspects of the usability evaluation, and participated to the completion of the usability assessment of the different devices, providing different observations regarding the device safety and interface design. Strengths and weaknesses of each one of the results obtained are detailed in the following paragraphs.

The task performance evaluation highlighted the current weaknesses in terms of actions and part of the usability process identifying the steps and the tasks that are the most critical for the management of the medical device safety but does not provide additional insights related to the reasons and the semantic error that led to the usability pitfall. Nevertheless, with the proper integration of questionnaires and interviews, the causes related to the pitfalls can be investigated and understood.

The SUS questionnaire confirmed that is a technique that does not provide any information regarding the medical device safety, but can be used for the evaluation and quantification of the ease of use and user interface pleasurable and provides the possibility for comparison with analogous devices.

The heuristic evaluation phase, even if cannot be used directly to observe hazardous situations and potential usability pitfalls demonstrated to be an excellent tool for the identification of the weaknesses of the tested user interface and to understand how to plan and focus the improvements of the user interface.

Finally the interview questionnaire evidenced its potential for the identification of the most critical functionalities of the device. With respect to the heuristic evaluation, which identifies the critical qualities of the user interface, this tool is extremely useful for the evaluation of the functionalities, how they are designed and perceived by the users.

5 CONCLUSIONS

As part of the validation activities of the usability aspects, our group acted as an external reviewer of the compliance of a series of software as medical devices with respect to the current applicable standards and Medical Device Regulation in Europe. To complete the evaluation of these devices, we used an approach derived from the applicable standards and other pertinent sources. We explicitly tailored them to ensure a complete overview of the device usability composed by structured methods. We applied this approach in this review, during the formative and summative phases of design. The proposed methodology for the activities is highly informative, repeatable, it allows for comparisons between different devices and complies with the current applicable standards. Our approach allowed a clear presentation of the results both to the developers and to the regulatory authorities. In future studies, we will analyse, improve, and standardise this methodology in order to obtain a structured workflow and a framework of techniques for the usability evaluation of SaMDs.

REFERENCES

- Health, C. for D. and R. (2019, September 2). *Applying Human Factors and Usability Engineering to Medical Devices*. U.S. Food and Drug Administration; FDA. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/applying-human-factors-and-usability-engineering-medical-devices>
- Institute of Medicine (US) Committee on Quality of Health Care in America. (2000). *To Err is Human: Building a Safer Health System* (L. T. Kohn, J. M. Corrigan, & M. S. Donaldson, Eds.). National Academies Press (US). <http://www.ncbi.nlm.nih.gov/books/NBK225182/>
- International Electrotechnical Commission. (2016). *IEC/TR 62366-2:2016* (1st ed.). IEC. <https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/91/69126.html>
- International Electrotechnical Commission. (2020). *IEC 62366-1:2015+AMD1:2020* (1.1). IEC. <https://webstore.iec.ch/publication/67220#additionalinfo>
- International Organization for Standardization. (2019). *ISO 14971:2019* (3rd ed.). ISO. <https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/07/27/72704.html>
- Jordan, P. W., Thomas, B., McClelland, I. L., & Weerdmeester, B. (1996). *Usability Evaluation In Industry*. CRC Press.
- Kwak, H., Oh, H., Cha, B., & Kim, J. M. (2021). The assessment of usability of pain medical device by physiatrists and physiotherapists. *Medicine*, 100(38),

- e27245. <https://doi.org/10.1097/MD.00000000000027245>
- Ravizza, A., Sternini, F., Giannini, A., & Molinari, F. (2020). Methods for Preclinical Validation of Software as a Medical Device: *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies*, 648–655. <https://doi.org/10.5220/0009155406480655>
- Ravizza, D., Sternini, F., Lantada, A., Sánchez, L., Sternini, F., Ravizza, D., & Bignardi, C. (2019). Techniques for Usability Risk Assessment during Medical Device Design: *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies*, 207–214. <https://doi.org/10.5220/0007483102070214>
- Schmettow, M., Schnittker, R., & Schraagen, J. M. (2017). An extended protocol for usability validation of medical devices: Research design and reference model. *Journal of Biomedical Informatics*, 69, 99–114. <https://doi.org/10.1016/j.jbi.2017.03.010>
- Sternini, F., Isu, G., Iannizzi, G., Manfrin, D., Stuppia, N., Rusinà, F., & Ravizza, A. (2021). Usability Assessment of an Intraoperative Planning Software: *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies*, 483–492. <https://doi.org/10.5220/0010252904830492>
- Zhang, J., Johnson, T. R., Patel, V. L., Paige, D. L., & Kubose, T. (2003). Using usability heuristics to evaluate patient safety of medical devices. *Journal of Biomedical Informatics*, 36(1–2), 23–30. [https://doi.org/10.1016/S1532-0464\(03\)00060-1](https://doi.org/10.1016/S1532-0464(03)00060-1)

