dHealth 2020 – Biomedical Informatics for Health and Care
G. Schreier et al. (Eds.)

doi:10.3233/SHTI200081

Comparison of Formative Evaluation Methods in the Usability Process on the Example of a Medical App

Michael SCHOLTES[a,1], Katharina NOLL[a], Keywan SOHRABI[a] and Volker GROSS[a]
[a] *TH Mittelhessen University of Applied Sciences – Faculty of Health Sciences,*
Wiesenstraße 14, 35390 Gießen, Germany

Abstract. Background: The process of developing a medical device has become increasingly complex, not least because of the Medical Device Regulation. Some of the associated changes affect in particular the software development companies. One of the challenges to provide safe software products is the usability engineering process. Adequate methods must be identified to find severe usability issues effectively and efficiently. Objectives: The aim of this study was a comparison of normative recommended formative evaluation methods. Methods: Two expert-based methods and two user-based methods were compared regarding effectiveness in finding issues, the level of detail and the temporal effort. Results: The heuristics and the isometrics, both showed a significant advantage regarding the effectiveness and similarity in level of detail and temporal effort. Conclusion: Based on this paper, the heuristics and the isometrics, both can be recommended, but further studies are necessary to consolidate this statement. Also, further methods should be considered, and more test persons must be involved to give an overall comparison of formative evaluation methods in the usability process of medical apps.

Keywords. ergonomics, medical device regulation, mobile apps

1. Introduction

1.1. Development of regulatory affairs in computer applications

The way of establishing software applications in medicine has been long and difficult. In 1989, Hafner et al. already stated that using computer applications in medicine may be less harmful than not using them [1]. Today, we take computers and software in medicine for granted and their usage became an integral part of modern medicine. Nevertheless, there are differences in quality of such software. Consequently, there have always been incidents that triggered discussions about safety, regulation, and registration of software application in medicine. For example, Lancet published in 1996 a small software error that led to reduced detection rates of Down's syndromes during pregnancies [2]. The most experience in regulating software in medicine hail from the USA, where the Food and Drug Administration regulated medical device related software in 1976 for the first time [3, 4]. Over time, the basic idea of regulations has changed.

[1] Corresponding Author: Michael Scholtes, TH Mittelhessen, Wiesenstraße 14, 35390 Gießen, Germany, E-Mail: michael.scholtes@ges.thm.de.

1.2. Admission of mobile medical apps in Europe

In 2007, the Council of the European Union has published the amending directive 2007/47/EC [6], that underlines the explicit applicability of the European regulations for medical devices to software. Thereupon, the discussion about the regulation of stand-alone software and mobile medical apps (MMA) raised. This is not surprising, because according to industry estimates [7], more than 1.7 billion smartphone and tablet users will have downloaded MMAs at the end of 2018. These users include health care professionals, consumers, and patients.

In 2013, the FDA published the guidance document "Mobile Medical Applications: Guidance for Food and Drug Administration Staff", a new point of reference for the regulation of MMAs also for European manufacturers. The FDA had to update this guidance within two years to be consistent with the guidance document "Medical Devices Data Systems, Medical Image Storage Devices, and Medical Image Communications Devices" [8].

But the development of the regulations continues: In 2017, the Medical Device Regulation (MDR) was published [9]. One of the relevant updates is the classification of stand-alone software, such as MMAs. So far, a lot of stand-alone software could be classified as Class I product. The new classification rule 11 of the MDR Annex VIII states that software "…intended to provide information which is used to take decisions with diagnosis or therapeutic purposes…" must be at least Class IIa. If there might be a serious deterioration of a person's state of health, it must even be Class IIb [9]. Based on the definition of a medical device (Article 2, MDR), it can be postulated, that most medical stand-alone software will be classified as Class IIa or higher. This leads to an enormous additional effort for the software manufactures and a prolonged development time.

1.3. Formative evaluation methods in the usability process of medical devices

Especially in medicine, usability and ergonomics play an important role, because in decisive moments, life might depend on it. For example, an ambiguous user interface of an infusion pump could lead to an overdosage of drugs. [10]

Therefore, a risk-based usability engineering process is mandatory for the development of medical devices. The IEC 62366-1 describes such process and distinguishes minimum requirements (called "regulatory usability") which must be fulfilled and user experience (called "market usability") which might be fulfilled. The usability engineering process includes iterative usability tests during development with the goal to detect and eliminate usability problems – the formative evaluation. [11]

1.4. Aim of this paper

The aim of this paper is the comparison of various methods for the formative evaluation based on a development of an MMA. This work shall investigate to what extent the requirements of the standard IEC 62366-1 are applicable in practice to the development process of an MMA and provide suggestions which methods might be preferred.

2. Methods

2.1. “FoodLogger” – a mobile medical app still under development

The application "FoodLogger" is still under development in the faculty of health sciences at the TH Mittelhessen University of Applied Sciences. The MMA prototype is used for the formative evaluation of this work. The application is currently designed to offer the user the possibility to document consumed food and its nutrients such as carbohydrates, fats, fiber, water, glycemic index and calories via the user interface. An MMA prototype, such as the “FoodLogger”, will probably have many usability issues. Thereby, it works well for the comparison of different evaluation methods.

2.2. Usability test methods

The *cognitive walkthrough* (expert-based) is a usability evaluation method in which one or more evaluators work through a series of tasks and ask a set of questions from the perspective of the user [12].

The *heuristic evaluation* (expert-based) is a usability inspection method for computer software that helps to identify usability problems in the user interface (UI) design. It specifically involves evaluators examining the interface and judging its compliance with recognized usability principles (the "heuristics") [12].

The *long isometrics* questionnaire (user-based) is a standardized questionnaire consisting of 75 items and a 5-point rating scale. The user can add a weight to each item from 1 (very low severity) to 5 (very high severity). For the analysis, the weight and the frequency of denomination are combined for categorization (see Table 1).

The *plus-minus method* (user-based) is intended to enable the subjective opinion of a test person about a product to be documented. The test person can decide for himself which aspects of the product he wants to evaluate and rates these either positively with a "plus" or negatively with a "minus". [13]

The *system usability scale* (user-based) is a simple, ten-item attitude Likert scale giving a global view of subjective assessments of usability. It works well for the comparison of different systems or a before and after comparison. But it does not work well to identify usability problems.

The expert-based test Heuristic Evaluation was performed with eight software-experts, the Cognitive Walkthrough with four and the Evaluation according to Guidelines with one software-expert. The user-based tests were performed each with eight potential users of the MMA.

Table 1. Categorization of the isometrics findings [14]

Category	Explanation
A(W)	mean weight < 3
B(W)	mean weight ≥ 3
A(F)	Frequency ≥ 25% of the users
B(F)	Frequency < 25% of the users

2.3. Comparison of the test methods

The **expert-based** methods (cognitive walkthrough and heuristic evaluation) were compared regarding effectiveness. Therefore, it was investigated how many usability issues, subdivided into degrees of severity (marginal, moderate and severe), could be found with each method.

The **user-based** methods (long isometrics and plus-minus method) were also compared regarding effectiveness. Therefore, it was investigated how many usability issues, subdivided into degrees of severity (marginal = A(W) & B(F), moderate = A(W) & A(F) or B(W) & B(F) and severe B(W) and A(F)), could be found with each method.

Additionally, the level of detail and the temporal effort for **all methods** were assessed by the authors. The temporal effort was assessed in relation to each other (low, medium, high). The level of detail was defined as “low”, if the method only reflects the opinion of the test person, as “medium”, if the method additionally describes the problem, and as “high”, if the method additionally describes location and possible effect of the issue.

In order to get a better picture of how the test participants themselves judge the methods used, they were asked two questions at the end of the test, following the procedure of Figl [15]. Subjects were interviewed after the isometrics questionnaire, the SUS questionnaire and the plus/minus method had been completed. The two questions were which method the users found to be most comfortable to use and which method the users felt was the best way to express their opinion.

3. Results

The comparison of the effectiveness of the expert-based methods is displayed in Figure 1. The cognitive walkthroughs could only identify 5 of the 28 moderate issues, none of the 3 severe and none of the 10 marginal issues.

The comparison of the effectiveness of the user-based methods is displayed in Figure 2. The plus-minus-method could only identify 2 of the 6 marginal issues, 7 of the 41 moderate issues, and not the one severe issue.

The author’s assessment regarding level of detail and temporal effort are displayed in Table 2. These results are based on the consideration of the relation between the used methods in this paper. They do not reflect the relation to other methods.

Table 2. Subjective author’s assessment

Method	Level of Detail	Temporal Effort
Cognitive Walkthrough	high	moderate
Heurisitcs	high	high
Isometrics	high	high
System Usability Scale	low	low
Plus-Minus-Method	moderate	moderate

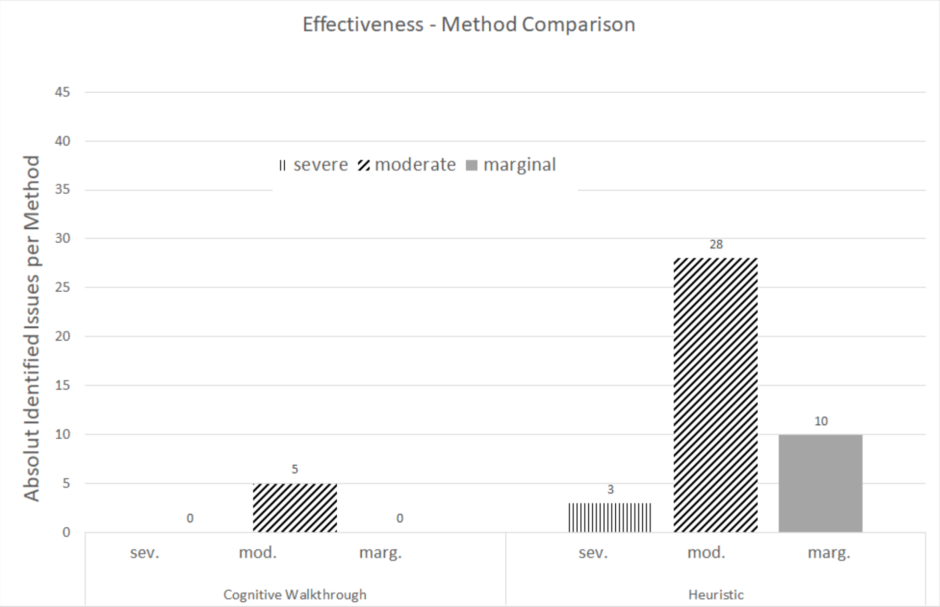

Figure 1. The comparison of the effectiveness of the expert-based methods (Cognitive Walkthroughs and Heuristics) subdivided into degree of severity.

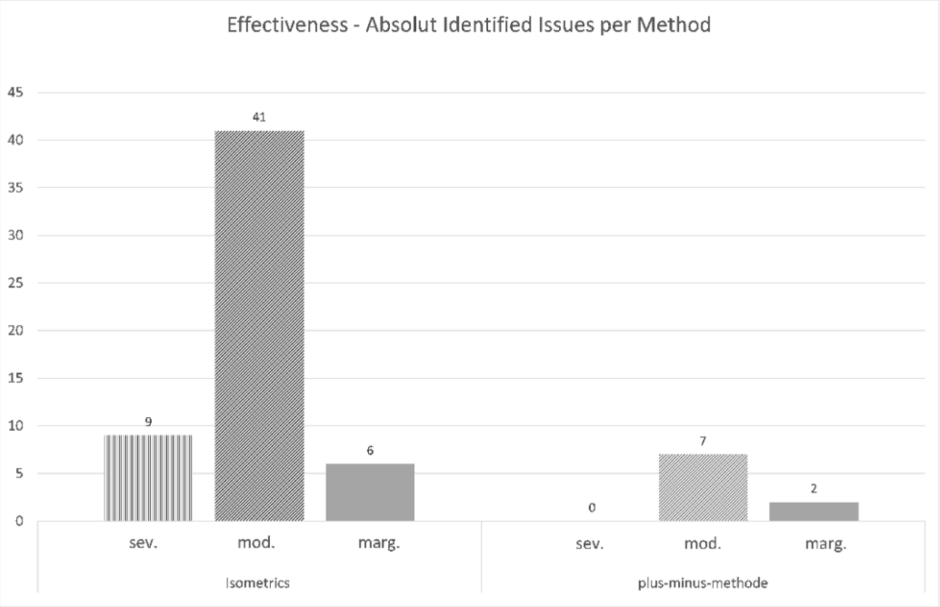

Figure 2. The comparison of the effectiveness of the user-based methods (isometrics and plus-minus-method) subdivided into categorization referring to Table 1.

The judgement of the test participants themselves is displayed in Figure 3. In total, 5 of 8 participants judged the SUS as most comfortable to use, no one chose isometrics (Figure 3 – A). The plus-minus-method and the isometrics, both were judged by 4 of 8 participants to be the best method to express their opinion (Figure 3 – B).

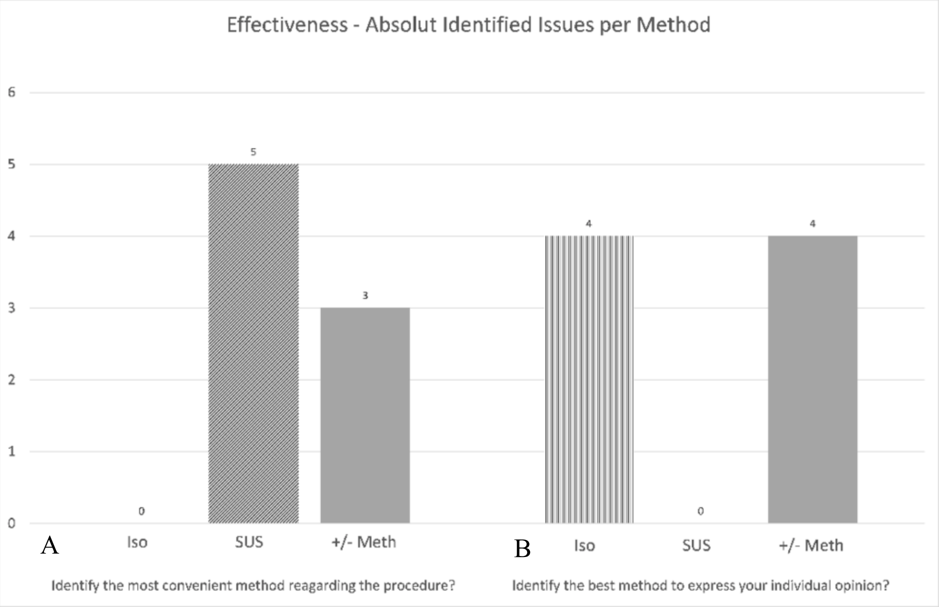

Figure 3. Participants themselves judge the methods used, they were asked two questions at the end of the test, following the procedure of Figl [15].

4. Discussion

4.1. Discussion of the results

The effectiveness comparison of the expert-based methods shows a significant advantage of the heuristics. The effectiveness comparison of the user-based methods shows a significant advantage of the isometrics. The level of detail and the temporal effort are comparable, so that there is a strong recommendation for those two methods based on this paper's results. Nevertheless, this applies only to MMAs and further limitations of the study must be considered. The prototype of the "FoodLogger" app worked well for this investigation because many minor and some major issues could be identified.

The judgement of the participants regarding the most comfortable method to use might be obvious, because the SUS is a very fast method, with only 10 questions. On the other hand, it does not give the user the possibility to express a differentiated consideration, which is reflected by the second question (see Figure 3 – B). Isometrics is a very extensive method and thereby it was not judged as a comfortable method to use.

4.2. Reflection and Limitations of the study

The statements of this study have a qualitative and descriptive character, especially due to the small number of test persons. The author's assessment might have been based on strict criteria instead of subjective evaluation. Furthermore, it is not clear if all issues have been found, because there is no reference MMA with exactly known and predefined issues. Such a reference MMA might be useful anyway, as this could also help to compare expert-based methods with user-based methods.

4.3. Perspective

This paper shows that there are differences between formative evaluation methods considering MMA. The growing regulatory effort [9] does affect in particular the software development companies, so there is a need for effective and efficient methods to support development decisions, especially in the field of usability engineering.

The findings strongly indicate the need for larger studies involving more participants and more methods as well as unambiguous criteria to allow inference statistically analysis.

References

[1] A.W. Hafner, A.B. Filipowicz, W.P. Whitely, Computers in medicine: liability issues for physicians, *Int. J. Clin. Monit. Comput.* **6** (1989), 185–194.

[2] P. Cavalli, False negative results in Down's syndrome screening, *Lancet* **347** (1996), 965–966.

[3] Food and Drug Administration (FDA), Medical Device Amendments to the Federal Food, Drug and Cosmetic Act, the May 28 (1976).

[4] R.R. Murfitt, United States government regulation of medical device software: a review, *J. Med. Eng. Technol.* **14** (1990), 111–113.

[5] J. Wyatt, Quantitative evaluation of clinical software, exemplified by decision support systems, *Int. J. Med. Inf.* **47** (1997), 165–173.

[6] The European Parliament and the Council of the European Union, Directive 2007/47/EC, *Official Journal of the European Union* (2007).

[7] U.S. Food and Drug Administration, Mobile Medical Applications, https://www.fda.gov/... MedicalDevices/DigitalHealth/MobileMedicalApplications/default.htm, last access: 28.12.2017.

[8] U.S. Food and Drug Administration, Mobile Medical Applications: Guidance for Industry and Food and Drug Administration Staff, https://www.fda.gov/downloads/MedicalDevices/... DeviceRegulationandGuidance/GuidanceDocuments/UCM263366.pdf, (2015).

[9] The European Parliament and the Council of the European Union. Regulation (EU) 2017/745. Official Journal of the European Union 2017.

[10] S. Buttron, The Importance of Human Factors & Usability Engineering in Medical Devices, https://www.namsa.com/european-market/human-factors-usability-engineering-in-medical-devices/, last access: 07.10.2019.

[11] International Electrotechnical Commission (IEC). Medical devices – Part1: application of usability engineering to medical devices (IEC 62366-1:2015 + COR1:2016); 2017-07.

[12] J. Nielsen, Usability Engineering. San Diego: Academic Press, Inc. (1993).

[13] W. Schweibenz, I. Harms, Testing Web Usability. Usability als kritischer Erfolgsfaktor des E-Business. (2000).

[14] L. Urbas, J. Ziegler, Messmethoden der Mensch-Maschine-Systemtechnik (2014).

[15] K. Figl, ISONORM 924/10 und Isometrics: Usability-Fragebögen im Vegleich. Institut für Wirtschaftsinformatik und Neue Medien, WU Wirtschaftsuniversität Wien (2009).