Healthcare and Medical Devices, Vol. 79, 2023, 204–215
https://doi.org/10.54941/ahfe1003489

Measuring Subjective Usability of Medical Devices - Questionnaire Development and Evaluation

Marisa Koopmann[1], Bernhard Wandtner[1], Michael Thorwarth[1], and Karsten Nebe[2]

[1]Fresenius Medical Care Deutschland GmbH, Bad Homburg v.d. Höhe, Germany
[2]Rhine-Waal University of Applied Science, Kamp-Lintfort, Germany

ABSTRACT

Safety and effectiveness are major usability concerns for the development process of medical devices. Other relevant factors like user satisfaction or overall user experience (UX) are sometimes neglected as they are not required from a regulatory perspective, nor can they be evaluated well through the classical approach of usability testing. Usability/UX questionnaires can measure these subjective variables, however, only few researchers have addressed the development of standardized questionnaires for medical products. This two-parted research aims to further close this gap. First, numerous attributes of usability/UX were researched and then critically evaluated by usability experts ($N = 9$) with practical experience in healthcare. The constructs relevant and applicable were then divided into clusters and items were newly created or carefully chosen from existing questionnaires and then condensed to a 70-item raw version of the questionnaire. In the second part of this study, nurses ($N = 106$) from South Africa, UK, and USA evaluated a dialysis device, providing responses to the questionnaire statements alongside the System Usability Scale (SUS) for validation purposes. Psychometric analysis showed that the average internal reliability across the eleven subscales was $\alpha = 0.70$ and ranged from 0.48 to 0.84. Seven items were chosen to be eliminated because of their weak item discrimination and difficulty which would lead to an increase of internal reliability. The initial scores for 9 out of the 11 subscales moderately correlated with the SUS ($r = 0.53$ to 0.60) with a significance of $\alpha < 0.05$. Overall, the results indicate that the newly developed questionnaire could be feasible to close the identified gap. Nevertheless, the modified questionnaire ought to be validated with a larger sample size and across a broader range of medical products.

Keywords: Usability, Human factors, User experience, Questionnaire development, Healthcare, Medical devices, Usability engineering

INTRODUCTION

While medical devices are vital for healthcare environments and patient wellbeing, they are not without risk: thousands of patients in the world are exposed to misdiagnosis, improper treatment, hospitalization, or death due to the incorrect use of medical devices (Roma & de Vilhena Garcia, 2020). To encounter these use-related hazards, it is essential to ensure good usability as incorrect device use is often related to design deficiencies (Obradovich

& Woods, 1996). While the International Standard for Ergonomics of Human-System Interaction, which defines 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" (International Organization for Standardization [ISO], 2018, p. 2) can be followed to improve the usability of any product, the development and validation of medical devices must comply with the Usability Standard for Medical Devices, the IEC 62366–1 (International Electrotechnical Commission [IEC], 2015). This specific regulation guides medical device manufacturers in the proper application of the usability engineering process. Here, the main concern is to achieve appropriate usability as it relates to safety and effectiveness by identifying and minimizing use errors (IEC, 2015; Food and Drug Administration [FDA], 2016). The most frequently used technique to achieve this goal is the classical approach of usability testing (Hedge, 2013), which is also the normative requirement that usually must be fulfilled to obtain certification (Geis & Johner, 2015; Conley, 2015). During usability testing, real users are observed in a simulated environment while interacting with the product to show evidence that the device can be used safely and effectively (Wood, 2018) and fulfills its requirements (IEC, 2015; FDA, 2016).

In contrast to the medical device industry, usability practices for consumer products are not as restrictive and allow for a broader approach. While effectiveness, efficiency, and satisfaction (ISO, 2018) are usually considered as the main concerns, products are also evaluated concerning learnability, memorability, error and error prevention, simplicity, reliability, and many more (Nielsen, 2012; Bitkina, Kim & Park, 2020). In addition, the related concept of User Experience (UX) has become more relevant over the years. Most researchers are aligned in believing that UX is a complex, context-dependent, and subjective construct (Zarour & Alharbi, 2018; Law, Roto, Hassenzahl, Vermeeren & Kort, 2009), however, there is no common agreement upon a concrete definition. Some view it as a holistic concept, including all types of emotional, cognitive, and physical responses when using a product (ISO, 2018), others focus on the multidimensionality and split the general notion of UX into several distinct criteria (Schrepp, Hinderks & Thomaschewski, 2014), which allows to measure each quality criterion independently. A common approach has been to differentiate between pragmatic quality criteria (also referred to as instrumental, goal/task-oriented) such as controllability or learnability and hedonic quality criteria (non-instrumental, non-goal/task-oriented), like stimulation, emotions, or aesthetics (Schrepp et al., 2014). It is argued that the classical approach of usability testing ignores the hedonic quality dimensions of a system as they have no clear relation to the task the user wants to accomplish (Hassenzahl, Platz, Burmester & Lehner, 2000). Therefore, usability is often measured with both objective and subjective methods, where objectively obtained data mainly includes measures of the participants' performance (e.g., task success, error rate, etc.) and subjective measures are related to the participants' opinions or attitudes toward the perceived usability (Lewis, 2009).

An efficient and cost-effective metric to measure subjective usability are psychometrically validated questionnaires which began to be developed in the late 1980s and early 1990s (Sweeney & Dillon, 1987). One of the most commonly known and used questionnaire is the System Usability Scale (SUS), a 10-statement survey initially developed by Brooke (1996). The SUS can be applied to a variety of products and services and presents respectable validity and reliability measures (Kortum & Bangor, 2013). Other validated usability/UX questionnaires include the CSUQ and PSSUQ (Lewis, 1995), SUMI (Kirakowski & Corbett 1993), UEQ (Laugwitz, Schrepp, & Held, 2008), and QUIS (Chin, Diehl, & Norman, 1988), to name a few. While some of them focus specifically on pragmatic qualities, others include both hedonic and pragmatic aspects. They further differentiate in length, subscales vs. global score and response format.

Even though the use of subjective usability metrics has shown significant value for consumer products, they have not been applied much in the development and validation of medical products. To comply with the strict regulations, the focus remains on ensuring safety and effectiveness through usability testing and not user satisfaction and perceived usability. While there has been some research towards the development of a questionnaire specifically for the assessment of a medical device's usability, and there are existing first exploratory questionnaire versions (Parreira et al., 2020; Müller & Backhaus, 2019), there is still a lack of a highly reliable and validated solution.

Therefore, the goal of this two-parted research study is to develop a first version for a new usability/UX questionnaire specifically for medical devices that potentially can serve as a reliable and validated solution to the aforementioned problem.

To ensure an unambiguous vocabulary, the term usability will refer to concepts of usability and UX in the following sections.

QUESTIONNAIRE DEVELOPMENT

Methods

The questionnaire was developed in four steps. First, a pool of constructs associated with usability was identified. This was done through the revision of several existing usability/UX questionnaires to ascertain what specific aspects of usability are measured. This process included well-validated questionnaires as well as initial draft usability questionnaires that have not been validated yet. In addition, usability constructs were identified from various definitions, models, and standards available in the literature. For the literature search the following keywords were used in Google Scholar and there referenced data bases as well as the online library of the Rhine-Waal University of Applied Science: usability; user experience; questionnaire; medical device usability.

In the second step, experts from the field who have worked as registered nurses in the past were recruited to participate in individual expert interviews. The goal was to obtain qualitative and quantitative data regarding the constructs' relevance and applicability for medical devices. The identified constructs were presented to the experts including a definition and a few sample items. First, the experts provided specific feedback on the relevance and applicability of the constructs. To obtain quantitative data, the constructs were then rated for their relevance on a scale from 1-5, where 1 indicated no relevance and 5 indicated high relevance. Regarding the applicability, participants chose between "applicable" or "not applicable", for each of the constructs. At the end of each interview, the experts were asked for any further constructs that are relevant for medical products' usability that were not yet considered. Exclusion criteria for a construct was an average relevance rating below 2.5 or a total of four or more "not applicable" ratings for medical products.

Third, a workshop with usability experts was held to review the results from the expert interviews with the overall goal to cluster the remaining constructs into subcategories for the questionnaire. Special attention was paid to the interrelationships of the constructs and how they might influence each other (e.g., if construct A has a good perceived usability rating, how does that affect the perceived usability rating for construct B).

In the last step, items were selected from existing questionnaires or were newly created. An initial pool of items was presented to the usability experts and each item was reviewed for its suitability and proper wording.

Results

Based on the literature review of both existing questionnaires and definitions, a total of 59 unique constructs was identified. Constructs with different wording but identical meaning (e.g., ease to learn, learnability) were grouped together and only accounted for one in the total count. 31 final constructs were selected for the expert interviews mainly based on their frequency in occurrence. A few constructs were chosen due to their perceived importance to usability for medical products.

A total of nine experts from the field were recruited, including four international training consultants, two application specialists, and three human factors engineers. All experts had at least five years of nursing experience in dialysis, with four of them over 25 years. The interviews were conducted remotely via Microsoft Teams and lasted 65 minutes on average. Overall, the results showed that the constructs were all rated highly regarding their relevance. A total of 26 constructs had an average rating of 3.5 or higher, with 13 of them being above 4.5. Only two constructs (Stimulation and Connectedness) were rated below a 3.0. Those two constructs plus the constructs "Intention to use" (relevance score = 3.0) and "Intuitivity" (3.78) received a total of two "not applicable" ratings from the experts. None of the constructs met the exclusion criteria, therefore all 31 were carried over to the workshop as well as five newly proposed constructs from the experts. Despite the lack of traceback to the literature, the five constructs were added because of the participants' high expertise in the medical field.

The workshop was conducted in person with three usability experts. First, the results from the expert interviews for each construct were reviewed and discussed extensively. It was universally decided that the constructs "Connectedness" and "Intuitivity" should be excluded due to their relatively low relevance and applicability ratings in comparison to the other constructs. Only one of the five newly proposed constructs (Handling) was kept, leading to a total of 30 constructs. Next, the constructs were clustered into subcategories. This was done by placing the constructs one by one on a whiteboard based on their definition and sample items and how similar and related they were content wise. The placement was discussed with the usability experts until there was a universal decision. As a result, eleven subcategories were formed: Subjective Usability (with 3 constructs), Emotional Aspects (2), Attractiveness (2), Controllability (3), Learnability (3), User Interface Quality (3), Error Handling (3), Information Quality (2), Goal and Goal Achievement (3), Efficiency (2) and Ergonomics (2).

For the item selection, a total of 334 items from existing questionnaires were accumulated and each item was mapped to one of the 30 remaining constructs based on its content. In the process, the original subscale of each item was disregarded due to inconsistencies in the existing questionnaires. Similar items with different wording and items that did not fit any of the subcategories were excluded. Additional items were newly created in two cases: first, when existing questionnaire items did not represent specific facets of a construct that were emphasized in the expert interviews or when a construct did not have any of the existing items mapped towards it. The preprocessed item pool for each of the subcategories (n= 120) was then presented to the usability experts and each item was reviewed for appropriateness regarding the content and wording. Overall, 52 items were taken from existing questionnaires with some of them being modified and 18 items were newly created. The direction of wording was altered for 29 items to prohibit potential response bias of participants. A total of 70 items remained for the first raw version of the questionnaire that was used for the questionnaire evaluation. The chosen response format for the questionnaire was a 5-point Likert scale ranging from “strongly disagree” to “strongly agree”, including a middle category being “neither agree nor disagree”. To determine the test scores, the average rating per subcategory was calculated. An overall test score was not considered.

QUESTIONNAIRE EVALUATION

Methods

For the evaluation study nurses from South Africa, the United Kingdom (UK), and the United States of America (USA) were recruited via email, flyers hung in staff rooms, and a posting on a web portal for nurses, respectively. The selection criteria included the ability to speak fluent English and a background in nursing, specifically for dialysis. Approval was granted for the study by the Fresenius Medical Care Compliance committee. Participation was on a voluntary basis and consent was gained from participants, assuring them confidentiality and anonymity regarding their responses.

The entire study was conducted online over a five-week period. Participants were directed to a website via a link or Quick Response code (QR-Code) and competed the survey on their own time. After acknowledging their agreement to participate on the online consent form, participants were asked for demographic data including age, occupation, and years of experience in dialysis. Further, participants were asked to indicate the specific dialysis device they were going to evaluate and how many years of experience they have had with that device. Then, the participants completed the newly developed 70-item questionnaire concerning their selected device. At the end, participants were asked to complete the SUS questionnaire for the same device.

After the data collection, psychometric item analysis was conducted to identify items that may not be appropriate for or discriminate enough between respondents. Therefore, item discrimination and item difficulty indices were calculated and analyzed. The item discrimination index is the correlation coefficient between the item scores and test scores (scale-based), indicating how well an item differentiates correctly among participants in the characteristic that the scale is designed to measure. Corrected-item-total-correlations (CITC) were calculated, meaning that the item score was removed from the test score before correlation. A threshold of CITC < .3 was used as an indicator to further inspect that item (Kline, 1993). For the item difficulty index, which is the quotient of the item's average score and the maximum score possible multiplied by 100, a threshold of < 20 & > 80 was used to mark psychometric questionable items (Priest at al., 1995). In noncognitive tests item difficulty refers to the likelihood or endorsement of answering an item in keyed direction rather than how easy or difficult it is to answer the item correctly.

To evaluate the internal consistency, values for Cronbach alpha were calculated for each of the subscales, including "what if item was deleted" calculations. The rule of thumb for Cronbach's alpha is that a coefficient with an absolute value higher than.70 indicated a high degree of internal reliability (Kline, 1993; Nunnally, 1978).

In addition, Pearson correlation coefficients between the SUS score and the average test score from each subscale were calculated to determine the criterion validity of the new questionnaire. For the SUS standard score conversion procedure was used, adapting participants' answers to a score between 1 and 100.

Results

A total of 138 nurses participated in the online study. 32 participants did not complete key parts of the survey and were excluded from further analysis. Out of the 106 remaining nurses, 78 came from South Africa, 9 from the UK and 19 from the USA. 101 participants indicated their job title including dialysis nurse (92%), nephrologists (1%), and other (7%) such as head nurse or therapy specialist nurse. 16% had more than 25 years of experience in a clinical environment, specifically dialysis, 5% had experience for 21–25 years, 6% for 16–20 years, 22% for 11–15 years, 34% for 5–10 years and 18% for less than 5 years. A total of 76 participants filled out the SUS questionnaire at the end of the survey.

Once data was collected, scores for all inverted items were reversed to remove the positive/negative keying of the language in the questionnaire. All positive items were scored 0, 1, 2, 3, 4 and all negative items were scored 4, 3, 2, 1, 0.

The overall distribution of the 7420 responses for the whole questionnaire was the following: 174 (2.35%) total responses for the score of "0", 526 (7.09%) for "1", 578 (7.79%) for "2", 3674 (49.51%) for "3", and 2468 (33.26%) for "4". This distribution shows an overall tendency for positive ratings regarding the items.

Item analysis was conducted individually for each of the eleven subscales to estimate the psychometric quality of each item in measuring the target construct. Item difficulty and item discrimination indices for all items are displayed in Table 1. Overall, item difficulty values were relatively high with an average item difficulty of 76.06 across all items, ranging from 46.23 to 89.15, with one outlier of 24.06. This indicates that none of the items were below the threshold of 20. A total of 24 items had a difficulty index above 80 and therefore fell above the established upper limit. For the item discrimination, CITC were calculated with an average value of .46. Eleven items fell below the goal threshold of .3, of which two had a negative correlation coefficient. The other items were distributed as follows: eleven items in the range of $.3 \leq$ CITC $< .4$; twelve items in $.4 \leq$ CITC $< .5$; 19 items in $.5 \leq$ CITC $< .6$; twelve items in $.6 \leq$ CITC $< .7$; and five items CITC $\geq .7$. From the item analysis, a total of 34 items were marked for further investigation, as one item was outside the threshold for both, item difficulty and discrimination.

The calculations of Cronbach alpha for each of the subscales ranged from .48 to .84, with seven subscales exceeding the goal of .70 (see Table 1). The average internal consistency across all subscales was $\alpha = .70$. Cronbach's alphas after removing each of the 34 items with questionable psychometric values were calculated. For subscales that contained multiple items, the internal consistency was calculated for all possible cases of item exclusion.

After reviewing all statistical parameters, seven items were identified for elimination mainly because of their poor item discrimination and because of the substantial increase of internal consistency when removed. The other items were kept despite their questionable psychometric values because they cover important aspects of usability and were deemed as important. Two items were excluded from the "Subjective Usability" scale, 2 items from "Attractiveness" and three items from "Error Handling", leading to an increase of α from .76 to .85, .48 to .63, and .51 to .72, respectively. The average internal consistency increased to .73.

The correlation coefficients among the average values of the eleven subscales with the total SUS score are shown in Table 2. The table shows that nine out of the eleven subscales moderately correlated with the SUS, with values ranging from $r = .53$ to .60, with a significance of $\alpha < .05$. Only one of the subscales (Attractiveness) had no significant correlation ($p = .48$). After eliminating the seven items, the correlation coefficient for the subjective usability subscale decreased from 0.55 to .51 and increased for the "Attractiveness" and "Error Handling" subscale from .23 to .33 and .39 to .48, respectively. The "Attractiveness" subscale correlated significantly with the SUS, after the item exclusion.

Table 1. Item characteristics for the 70 items of the newly questionnaire, divided in the 11 subscales. *Removed item after item analysis. (Cronbach Alpha after item exclusion in parentheses).

Items	P$_i$	CITC
Subjective Usability (Cronbach's Alpha = 0,76 (0,85))	89.15	.62
1. I feel comfortable using this product.	79.48	.17
2. *I do not feel confident using the product.*	87.97	.73
3. Overall, I am satisfied with the product.	85.38	.64
4. I think the product is easy to use.	80.66	.64
5. The product allows me to complete my tasks easily.	78.77	.51
6. It is complicated to use the product.	79.01	.27
7. *I would consider this product as useful for my tasks.*		.47
8. *The product does not fulfil its purpose.		
Emotional Aspects (Cronbach's Alpha = 0,84)	77.12	.62
9. Working with this product is a frustrating experience.	80.90	.71
10. *Working with this product makes me angry.	76.18	.66
11. Working with this product is motivating.		.71
12. Working with this product is discouraging.		
Attractiveness (Cronbach's Alpha = 0,48 (0,63))	68.87	.31
13. The product does not present itself in an attractive way.	73.82	.39
14. The product is aesthetically pleasing.	75.24	.35
15. In my opinion the product is not innovative.	24.06	-.14
16. *I find the product conventional.*	69.34	.20
17. *I would not like to use this product every day.*		.39
18. I would not swap this product for any other.		
Controllability (Cronbach's Alpha = 0,80)	76.42	.50
19. When I use this product, I feel in control.	77.83	.59
20. It is easy to make the product do exactly what I want.	70.75	.40
21. *The product is not always operating how I intended.	78.30	.47
22. The product is impractical.	80.90	.51
23. I find the various functions in the product are well integrated.	77.36	.59
24. *The product does not integrate well into my workflow.	82.08	.65
25. The product allows flexible usage according to my needs.	76.65	.65
26. *The product is easy to adjust to better perform my tasks.		.21
27. This product is awkward when I want to do something which is not standard.		
Learnability (Cronbach's Alpha = 0,80)	81.37	.74
28. It is easy to learn how to use this product.	79.95	.50
29. I imagine that most people would learn to use this product very quickly.	77.59	.52
30. Sometimes it is not clear what to do next when performing tasks with the	80.66	.40
product.	78.30	.48
31. Performing an action with this product leads to a predictable result.	81.37	.41
32. It is easy to forget how to do things with this product.		.67
33. *I keep having to look for assistance when I use this product.		
34. Remembering terms used by this product is easy.		
User Interface Quality (Cronbach's Alpha = 0,84)	75.47	.58
35. The user interface indicates clearly what steps I have already done and what I	80.42	.56
have yet to do.	74.53	.55
36. I like the user interface of this product.	79.01	.59
37. *User interface elements (buttons, levers, switches, etc.) are not easy to use.	79.95	.62
38. The arrangement of steps in a task seems logical to me.	77.36	.45
39. It is easy to find the information I need.	81.37	.72
40. I think there is too much inconsistency within this product.		.62
41. The information is arranged and displayed in a consistent way.		
42. *The terminology within the product is consistent.		

Continued

Table 1. Continued.

Items	P_i	CITC
Error Handling (Cronbach's Alpha = 0,51 (0,72))	83.25	.04
43. **I do not feel safe when I use the product.*	51.89	.04
44. *The product is not associated with large error possibility in its use.*	83.49	.33
45. *I trust that the product protects me from unsafe usage.	73.35	.53
46. Whenever I make a mistake using the product, I can recover easily.	70.99	.34
47. The product provides instructions that clearly tell me how to fix problems.	82.31	.36
48. *The product indicates immediately when something went wrong.	46.23	-.18
49. **The product can continue functioning despite of invalid inputs.*	61.79	.50
50. This product sometimes stops working unexpectedly.		.45
51. *The product is not reliable in stressful situations.		
Information Quality (Cronbach's Alpha = 0,72)	79.01	.61
52. The content of information (e.g., on-screen messages, labels, user manuals,	80.90	.56
symbols etc.) provided with the product is easy to understand.	74.29	.41
53. The product keeps you informed about what it is doing.	75.24	.51
54. *The amount of information displayed is inappropriate.		.33
55. The system sounds are appropriate.		
56. *Auditory alarms are distinctive and recognizable.		
Goal & Goal Achievement (Cronbach's Alpha = 0,78)	84.20	.48
57. I cannot sufficiently complete my tasks using this product.	82.55	.65
58. The product facilitates the achievement of my tasks.	79.25	.57
59. I find the product unnecessarily complex.	76.89	.56
60. This product's capabilities do not meet my requirements.		.56
61. This product has all the functions I expect it to have.		
Efficiency (Cronbach's Alpha = 0,57)	89.86	.30
62. I am able to efficiently complete my tasks using this product.	69.34	.23
63. The number of steps needed to accomplish my tasks is appropriate.	73.11	.34
64. This product seems to disrupt the way I normally like to arrange my work.	71.46	.36
65. This product responds too slowly to inputs.		.53
66. The product gives an immediate response when an action is initiated.		
Ergonomics (Cronbach's Alpha = 0,55)	70.52	.45
67. *When using the product, I can effortlessly access anything I need to complete	82.31	.48
my tasks.	66.98	.22
68. *It is easy to operate the product when using protective equipment, such as		.25
gloves, face mask, glasses, etc.		
69. It does not require physical effort to use the product.		
70. *When using the product, I experience physical discomfort.		

Note. Eliminated items are in italics. Self-created items are marked with asterisks. Pi = item difficulty index; CITC = corrected-item-total-correlations indicating the item discrimination. Cronbach Alpha after item exclusion in parentheses.

Table 2. Correlation coefficients among questionnaire subscales with SUS-score before item exclusion.

Scale	1	2	3	4	5	6	7	8	9	10	11
SUS	.55	.60	.23*	.60	.53	.56	.39	.59	.59	.53	.52

Note. *p > .05. 1 = Subjective Usability; 2 = Emotional Aspects; 3 = Attractiveness; 4 = Controllability; 5 = Learnability; 6 = User Interface Quality; 7 = Error Handling; 8 = Information Quality; 9 = Goal & Goal Achievement; 10 = Efficiency; 11 = Ergonomics

DISCUSSION AND CONCLUSION

According to the literature review, a highly reliable and validated usability questionnaire specifically designed for medical devices does not yet exist. While some researchers have started to investigate this issue, none of the proposed solutions have been widely adopted. The aim of this study was to develop and evaluate a new questionnaire for the assessment of a medical device's usability. Input for relevant constructs was gathered from various published journal articles, including specifically the review of existing validated usability questionnaires and usability definitions. Considering the uniqueness and specificity of medical devices, experts from the medical field rated the usability constructs for their relevance and applicability. After grouping the constructs into subscales, items were selected from validated metrics or created anew if existing items did not represent relevant facets of a construct. That process was accompanied by usability experts integrating their knowledge and experience into discussion and item selection. The involvement of relevant stakeholders ensured the content validity of the questionnaire. In the second part, the questionnaire was tested in an online study with 106 participants evaluating an individually chosen dialysis device. The data analysis results showed overall acceptable reliability measures for an initial questionnaire version, which was reflected in Cronbach alpha values. Seven items were eliminated because of their poor item characteristics, which led to an increase in reliability. The performance of the new questionnaire was compared with the SUS. The correlation coefficients among the questionnaire subscales with SUS score are positive and statistically significant after item exclusion which points to the assumption that the individual scales of new developed questionnaire measure usability aspects.

During the development of the questionnaire, some decisions were based on the subjective assessment and judgement from experts from the medical and usability fields rather than being pulled from the literature. Their practical experience and high level of expertise is a crucial and valid factor that should be taken into account. However, had a larger or different group of experts been consulted in the making of the questionnaire, the result might have differed in some ways. As the market for medical devices is very large and highly heterogenous, it is possible that some of the subscales might not be applicable for all medical products. Therefore, the questionnaire was designed to calculate test scores for each subscale individually and does not provide one single overall score. This allows the disregarding of subscales without affecting the outcome of the overall evaluation. In this study, criterion validity analysis was limited to correlation analysis with the SUS score. To further investigate aspects of criterion validity the questionnaire could be applied in conjunction with a usability test. Task-success rate, frequency of error, and time to complete task measures could be correlated with the new questionnaire. The evaluation study of this work was performed on dialysis devices only. It is possible that the results might have been different for other medical devices. In addition, the questionnaire was applied post-hoc, as the participants have been working with the evaluated device for multiple years. An ad-hoc application of the questionnaire could reveal further insights. 3 of the 11 subscales had internal consistencies below the goal of $\alpha < 0.7$ and need further investigation. An increase of number of items in the subscale could be a possible solution, as the number of items is directly related to internal consistency measures.

Taking the limitations of this work into consideration, directions for future research are suggested to further enhance the new metric. To further investigate validity aspects, studies with larger and heterogenous sample sizes and different medical devices in various stages of the development process are necessary. Future validation studies should include additional statistical approaches, particularly factor analysis. Nevertheless, the results of the present study are already promising and provide valuable insights towards a feasible solution to further close the gap of missing usability questionnaires for medical devices.

REFERENCES

Bitkina, O. V., Kim, H. K., & Park, J. (2020). Usability and user experience of medical devices: An overview of the current state, analysis methodologies, and future challenges. International Journal of Industrial Ergonomics, 76, 102932.

Chin, J. P, Diehl, V. A., & Norman, K. (1988). Development of an instrument measuring user satisfaction of the human–computer interface. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '88). Association for Computing Machinery, New York, NY, 213–218.

Conley, D. (2015). Two Paths for Medical Device Approval: FDA vs CE. Health Management. 15(2).

Food and Drug Administration. (2016). Applying Human Factors and Usability Engineering to Medical Devices.

Geis, T. & Johner, C. (2015). Usability Engineering als Erfolgsfaktor: Effizient IEC 62366- und FDA-konform dokumentieren. Beuth Verlag, Berlin.

Hassenzahl, M., Platz, A., Burmester, M., & Lehner, K. (2000). Hedonic and ergonomic quality aspects determine a software's appeal. Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI '09). Association for Computing Machinery, New York, NY, 201–208.

Hedge, V. (2013). Role of human factors / usability engineering in medical device design. Proceedings Annual Reliability and Maintainability Symposium (RAMS), 1–5.

International Electrotechnical Commission (2015). Medical devices - Part 1: Application of usability engineering to medical devices (IEC Standard No. 62366-1).

International Organization for Standardization. (2018). Ergonomics of human-system interaction - Part 11: Usability: Definitions and concepts (ISO Standard No. 9241–11).

Kirakowski, J., & Corbett, M. (1993). SUMI: The Software Measurement Inventory. British Journal of Educational Technology, 24(3), 210–212.

Kline, P. (1993). The Handbook of Psychological Testing. Routledge, London.

Kortum, P., & Bangor, A. (2013). Usability Ratings for Everyday Products Measured With the System Usability Scale (SUS). International Journal of Human-Computer Interaction, 29(2), 67–76.

Laugwitz, B., Held, T., Schrepp, M. (2008). Construction and Evaluation of a User Experience Questionnaire. In: Holzinger, A. (Eds.), HCI and Usability for Education and Work. USAB 2008. Lecture Notes in Computer Science: Vol: 5298. Springer, Berlin, Heidelberg.

Law, L. C., Roto, V., Hassenzahl, M., Vermeeren, A., & Kort, J. (2009). Understanding, scoping and defining user experience: A survey approach. Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI '00). Association for Computing Machinery, New York, NY, 201–208.

Lewis, J. R. (2009). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. International Journal of Human-Computer Interaction, 7, 57–78.

Müller L., Backhaus C. (2019). Entwicklung eines Fragebogens zur ergonomischen Bewertung von Medizinprodukten innerhalb des Beschaffungsprozesses in Gesundheitseinrichtungen. Dokumentation zum 65. Kongress der Gesellschaft für Arbeitswissenschaft in Dresden. Dortmund: GfA-Press. B.2.8. ISBN: 978-3-936804-25-6.

Nielsen, J. (2012, January 12). Usability 101: Introduction to Usability. Nielsen Norman Group. https://www.nngroup.com/articles/usability-101-introduction-to-usability/.

Nunnally, J. C. (1978). Psychometric Theory. New York, NY: McGraw-Hill.

Obradovich, J. H., & Woods, D. D. (1996). Users as designers: how people cope with poor HCI design in computer-based medical devices. Human Factors, 38(4), 574–592.

Parreira, P., Sousa, L. B., Marques, I., Santos-Costa, P., Cortez, S., Carneiro, F., … & Oliveira, A. (2020). Usability Assessment of an Innovative Device in Infusion Therapy: A Mix-Method Approach Study. International Journal of Environmental Research and Public Health, 17(22).

Priest J., McColl B. A., Thomas L. & Bond S. (1995) Developing and refining a new measurement tool. Nurse Researcher, 2(4), 69–81.

Roma, M. S. G., & de Vilhena Garcia, E. (2020). Medical device usability: literature review, current status, and challenges. Research on Biomedical Engineering, 36, 163–170.

Schrepp, M., Hinderks, A., & Thomaschewski, J. (2014). Applying the User Experience Questionnaire (UEQ) in Different Evaluation Scenarios. In: Marcus, A. (Eds.), Design, User Experience, and Usability. Theories, Methods, and Tools for Designing the User Experience. DUXU 2014. Lecture Notes in Computer Science: Vol. 8517. Springer, Cham.

Sweeney, M., & Dillon A. (1987). Methodologies employed in the psychological evaluation of HCI. In: Proceedings of the Second IFIP Conference on Human–Computer Interaction (INTERACT '87). Stuttgart, Germany, 376–373.

Wood, J. (2018, November 26). Usability testing 101. IDR Medical. https://info.idrmedical.com/blog/medical-device-market-research-usability-testing

Zarour, M., & Alharbi, M. (2018). User Experience Framework that Combines Aspects, Dimensions, and Measurement Methods. Cogent Engineering, 4(1).