Healthcare and Medical Devices, Vol. 51, 2022, 278–285
https://doi.org/10.54941/ahfe1002128

Comparison of Lab- and Remote-Based Human Factors Validation – A Pilot Study

Karoline Johnsen, Bernhard Wandtner, and Michael Thorwarth

Fresenius Medical Care Deutschland GmbH, Bad Homburg, Hesse, 61342, Germany

ABSTRACT

Conducting human factors validations remotely becomes increasingly important, not only due to the COVID-19 pandemic. However, there is a lack of research addressing the reliability of remotely obtained data in the field of medical products. This research focuses on producing and analyzing first data to compare lab-based and remote-based setups. The goal is to evaluate if and under which circumstances human factors validations could be conducted remotely and which methodological aspects must be considered. In a simulated usability test, two lab-based and two remote-based conditions were investigated for two products of different dimensionality. Observational data of five human factors professionals per condition was used for data analysis. The descriptive comparison focused on the similarity and quality of the data as well as the effect on the observers' cognitive workload. Findings do not seem to strongly favor either one of the approaches, but the remote-based setup performed better for the two-dimensional than for the three-dimensional product. Overall, initial results from the pilot study highlight the potential of remote evaluations. However, more research is needed to validate the results with a larger sample size and determine the influencing factors that might favor remote vs. lab-based approaches.

Keywords: Remote usability evaluation, Medical device usability evaluation, Human factors validation, Usability test

INTRODUCTION

For medical devices, the focus on proper usability is more than just a luxury. Normative standards and regulations require that usability testing is carried out as it relates to the safety of the product (International Electrotechnical Commission [IEC], 2015a; Food and Drug Administration [FDA], 2016). Classically, the testing for medical devices is performed in a usability laboratory that simulates the natural environment of the product (Wiklund, 2015). For this purpose, the testing is video recoded and can be reviewed if necessary (Ross, 2021). This approach is recognized by the FDA to generate data in a Human Factors (HF) Validation, on which basis a certification can be decided (Geis and Johner, 2020). However, due to the COVID-19 pandemic and consequently the need for social distancing, it is clear that there must also be alternatives for this to continue to allow studies to operate (Lourenco and Tasimi, 2020). Users selected according to the normative requirements

are either patients who are exposed to a particular risk during a pandemic, or personnel in the clinical environment who are especially needed in pandemic times and whose time resources must therefore be handled with care. Due to their work environment, they are also exceptionally exposed to the risk of an COVID-19 infection (Forkey and Clark, 2021).

A solution to the problem could be found in the application of remote testing. Greater flexibility in terms of time and spatial separation are obvious advantages of this method (Fidas et al., 2007). Also, it opens up the possibility to address more people and thus to test a more representative sample, to conduct faster data collection, to conduct cross-cultural studies and it can be economically advantageous (Woods et al., 2015).

However, it is questionable to what extent remote evaluation complies with regulatory requirements. While the FDA recognizes that due to the COVID-19 pandemic an in-person testing regarding HF might neither be feasible or appropriate, they do also not suggest a remote setup as an alternative in general and "are unable to provide a general statement at this time about whether remote HF testing for drug products could potentially be an acceptable approach" (Chan et al., 2021, p. 4). This is because "the agency is currently not aware of any data that supports the use of remote HF validation testing or of any consensus scientific guidelines or standards that can inform as acceptable remote testing approach" (Chan et al., p. 4).

While there is some empirical research in the field of website and consumer-products (e.g. Duh et al., 2006, Andreasen et al., 2007; Sauer et al., 2019), McLaughlin et al., (2020) stated that in their literature research "No studies were found comparing laboratory and remote testing of medical devices" (McLaughling et al., 2020, p. 3). These authors could therefore mark the first publication in the named area. However, they focused on a theoretical approach and did not acquire experimental data.

CONSIDERATIONS FOR EVALUATING THE SUITABILITY OF THE SETUPS

To evaluate whether remote HF testing can be a valid approach, it is important to clarify the rationale for such a decision. A closer reading of the requirements from the FDA Guideline reveals that particular emphasis is placed on observation in a Human Factors Validation. This is highlighted by excerpts such as "Some data is best collected through observation; for example, successful completion of or outcome from critical tasks should be measured by observation rather than relying solely on participant opinions" (FDA; 2016, p. 24) and "The human factors validation testing should include observations of participants performance of all the critical use scenarios (which include all the critical tasks)" (FDA; 2016, p. 25).

It is assumed that the remote-based observation must at least be as reliable as the currently used approach (on-site observation combined with video and audio recording), to be considered as suitable. This classical or traditional (TRAD) approach is however not fully independent of a "remote component" due to the possibility to review video and audio recordings post-session. By reducing the possibility to check unclear facts post-hoc on the basis of video

and audio recordings made, the observation would solely rely on the live observation in the usability lab (LAB). The remote-based equivalent to these lab-based condition would be the observation of the recordings without alterations (REMOTE) and the opportunity to re-watch the recordings to clarify uncertainties (REPLAY). These four conditions were used in the pilot study to enable a precise comparison of the different components of lab-based and remote-based study setups.

In order to compare the setups in a quantitative manner, the observations need to be described in a standardized format. A structured protocol with all actions to be performed (based on a task analysis) ensures that the focus is about observation rather than environmental factors. Classifications established in the field of software evaluations and consumer industry seem to not fit well with requirements on medical devices, therefore it appears to be obvious to rather consider classifications derived from usability engineering standards and guidelines. This leads to the approach of distinguishing the observation of the expected user action as a success, use difficulty, use error and artifact (IEC, 2015a; IEC, 2015b; FDA, 2016; IEC, 2020). Further, the categories not applicable and missing should be added to account for actions which were not performed or observations not protocolled.

In the field of medical technology, a distinction is also made a-priori between the tasks themselves according to the severity of possible use errors, in the form of critical and non-critical tasks (FDA, 2016). For the certification of a medical product, it is ultimately important that all use errors in critical tasks are assessed during the Human Factors Validation (FDA, 2016). This differentiation should therefore also be considered and can even be applied on the action level of the task analysis.

To carry out the remote observation, it has to be decided which method comes to use. According to available literature, the option of video recording the interaction with a three-dimensional (3D) medical product is the only one imagined to be suitable for Human Factor Validations (Mejía-Gutiérrez and Carvajal-Arango, 2017, as cited in McLaughlin et al., 2020). McLaughlin et al. (2020) further point out that distinguishing between a two-dimensional (2D) and a 3D product is a relevant factor when considering medical devices.

Lastly, cognitive workload might differ between lab- and remote-based observations. Multiple streams of information can lead to an overload of human cognitive capacity, and the consequence of this cognitive overload is performance errors (Gevins & Smith, 2003). Keeping these performance errors low seems to be a logical goal when evaluating the usability of medical devices.

PILOT STUDY: COMPARISON OF LAB- AND REMOTE-BASED OBSERVATIONS

Goal of the pilot study was to collect empirical data while accounting for the considerations mentioned above. The lab- and the remote-based conditions should be compared regarding the similarities of the observational data. Further, it was of interest how accurate the observations resembled the real situation, describing the quality of the observation in general. To account

for the possibly influencing factor of cognitive workload, data targeting this variable was also collected and analysed.

Methods

In total, 10 observers participated in the study and directly produced the data to be analyzed. The age ranged from 27 to 56 years ($M = 39.9$, $SD = 10.46$). According to the theoretical considerations, a 2D and separately a 3D product was chosen to be in focus for the usability test. The decision was made to evaluate a medical device and a medical software from the field of hemodialysis. Prior experience by the observer with the products was self-estimated on a five-point scale ranging from 1 = *not at all* to 5 = *very much*.

The observers in the lab-based setup were on average slightly younger ($M = 37.8$, $SD = 9.68$) than the observers in the remote setup ($M = 42.0$, $SD = 11.90$) and had a higher proportion of females (3 females) than the remote setup sample (1 female). The lab-based observers estimated their prior experience to be between one and five for the medical device ($M = 2.4$, $SD = 1.95$), which is lower than the remote observer group ($M = 3.6$, $SD = 1.95$). However, they rated their experience with the medical software higher with utilization of the scale between levels one and five ($M = 2.0$, $SD = 1.64$) than the remote-based group who reported scores of one or four ($M = 1.6$, $SD = 1.34$). The observers therefore had mixed product expertise and there was tendency of lower expertise with the medical software overall.

All observers were human factors professionals, which was crucial because unexperienced observers are considered to be unable to precisely categorize usability problems according to Andreasen et al. (2007).

The observations were based on test participants' interaction with the products. The sample of the usability test participants consisted of two females and three males. Their ages ranged from 24 to 27 years ($M = 25.4$, $SD = 1.14$), and they had no prior experience with the products used in the usability tests other than a short introduction to the main principles. The absence of a proper training was purposely, to be able to provoke use errors in course of the evaluation and therefore generate enough data for comparison. Each test participant was assigned to one lab- and one remote-based observer and therefore interacted with the two products just once.

The lab-based observer was present in the test room during the evaluation (LAB condition). Afterwards, the session's recording could be reviewed as a second variant of the lab-based observation (TRAD condition). The remote-based observer had the recording as a resource for observation only (REMOTE condition) and the chance to review it afterwards as a second condition (REPLAY condition). The observations were based on a simulated human factors validation for two different medical products (device and software). The main basis for data analysis was a pre-defined observation protocol in which the individual actions to be performed were listed and then categorized by the two observer groups. The categories to choose from were success, use difficulty, use error, artifact, missing or not applicable.

The recordings were realized with four cameras for the medical device and two video sources for the medical software. While for the 3D product the

test participant's face, the machine display, the machine body and the participants perspective via a head-mounted camera was recorded, for the medical software inputs from the laptop camera pointing at the test participant's face and capturing of the screen were used. It is to be noted that the head mounted camera was first realized via eye tracking glasses, before being replaced with a head-mounted webcam due to technical failure of the glasses. This however did not influence the recorded videos much as they differed only marginally. The positions of the cameras in the test room as well as the positions and sizes of the video inputs to the final picture displayed to the observers were determined based on preliminary tests of the setup.

To have an objective measure for the quality of the observation, a sample solution was created after each usability tests took place. The scores were based on prior determined success criteria and an independent observation of the usability evaluation in-person and remotely by a person not involved in the data acquisition.

To record the observers' cognitive workload, it was decided to use the NASA-TLX questionnaire. For data analysis, "RAW-TLX" scores were calculated (Hart, 2006).

Results

Due to the small sample size, all analyses were performed descriptively. Generally speaking, the opportunity of re-watching the recordings in the lab- and remote-based condition seldom accounted for changes in the protocol. While in the lab-based condition 6% of the actions were re-categorized for each product, no categorization was changed in the lab-based condition for either one of the products. Most actions were in both cases and for both products categorized as a success, study artifacts were observed in only 1% of all actions performed. The usability tests differed in their length and number of actions between the products, with the medical software being less than half as long and containing less than half as many actions compared to the 3D product. From subjective ratings, the head mounted camera and the capturing of the screen were perceived most relevant for the remote observations of the respective products, compared to the other screens available.

To evaluate the similarity of the lab- and remote-based observations, the any-two agreement and Cohen's κ were calculated. Descriptive analysis show differences in observations of the lab-based vs. remote-based setup that become smaller when potentially critical actions are in focus. For the medical software less than 10% of the observations differ compared to around 15% for the medical device considering only critical use errors. Focusing on actions which were observed differently between the lab-based and the remote-based setup, no pattern could be identified which they had in common.

As can be seen in Figure 1, the quality of observations was slightly higher when the observer was on-site, and better overall for the medical device compared to medical software regarding percentual agreement with the sample solution. The hit- and correct rejections rate of critical use errors, which is most crucial, is visibly higher in the TRAD condition compared to the remote-based conditions for the medical device. For the medical software,

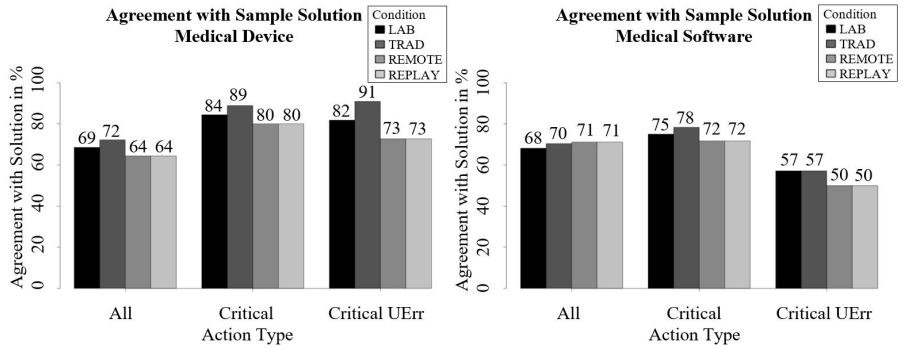

Figure 1: Agreement between the Observer's categorization of the actions with the sample solution in percent. Distinguished between the medical device (left) and the medical software (right). All = all actions, Critical = critical actions, Critical UErr = critical use errors.

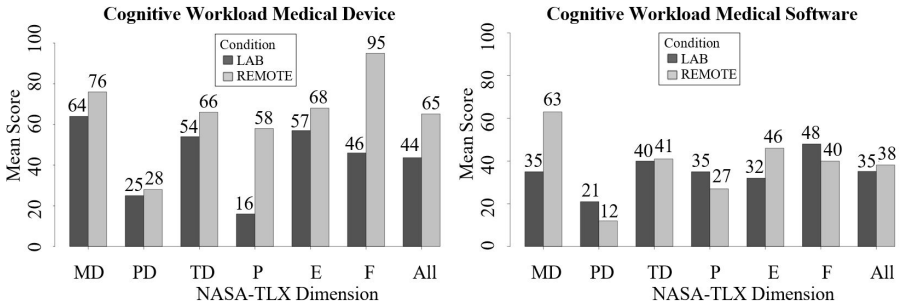

Figure 2: Mean NASA-TLX scores for the medical device (left) and the medical software (right), comparing the lab-based and the remote-based observation. MD = mental demand, PD = physical demand, TD = temporal demand, P = performance, E = effort, F = frustration, All = overall score.

the difference is smaller, favoring the lab-based conditions over the remote-based conditions. However, when considering all actions for the 2D product, the remote-based setup produced higher quality observations than the lab-based observation. Regarding actions which were not classified according to the sample solution, no common factor could be extracted.

Cumulated data of cognitive workload for the observation of the two product categories can be seen in Figure 2. Interestingly, a particularly high cognitive workload occurred when the medical device was observed remotely comparing the total NASA-TLX scores between the setups.

DISCUSSION AND CONCLUSION

The descriptive study results implicate that the TRAD condition in the lab-based setup is superior to the alternative approaches when having the total detection rate of critical use errors in mind. However, the observations are rather similar regarding the agreement rates between the observations, especially when comparing if a critical use error was observed or not between the conditions. For 2D products, it is more likely that a remote-observation

could be a feasible approach, as the data outperformed on some occurrences the lab-based observation. For remote observation of a usability test with a 3D medical device, the realization of the observation method should be improved, as indicated by the relatively high cognitive workload.

Systematic patterns accounting for different observations between the conditions could not be identified. Future research should clarify which type of tasks might be especially difficult to assess remotely compared to on-site and vice versa.

It has to be considered that the results of the study at hand are limited by the small sample size, preventing inferential statistical analyses. Furthermore, the quality of the camera recordings could be improved, as some observer stated that it hindered their observation. The angles for the camera recordings in the test room could further be an aspect to improve the remote-observation setup.

Overall, the results from the pilot study highlight the potential of remote evaluations. However, more research is needed to validate the results with a larger sample size and determine the influencing factors that might favor remote vs. lab-based approaches.

REFERENCES

Andreasen, M. S., Nielsen, H. V., Schrøder, S. O. & Stage, J. (2007). What happened to remote usability testing? *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '07*, 1405–1414. https://doi.org/10.1145/1240624.1240838.

Chan, I. Z., Kontos, K. & Wiyor, H. (2021, 16. April). *FDA Panel* [Conference Slides]. Vimeo. https://vimeo.com/535691114

Duh, H. B.-L., Tan, G. C. B. & Chen, V. H.-. (2006). Usability evaluation for mobile device. *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services - MobileHCI '06*, 181–186. https://doi.org/10.1145/1152215.1152254.

Fidas, C., Katsanos, C., Papachristos, E., Tselios, N., & Avouris, N. (2007). Remote usability evaluation methods and tools: A survey. *Paper presented at Pan-Hellenic Conference on Informatics*.

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

Forkey, D. L. & Clark, S. E. (2021, 15. April). *A Case Study: Remote Usability Evaluation of a Ventilator Developed to Address the COVID-19 Pandemic* [Conference Slides]. Vimeo. https://vimeo.com/535689093

Geis, T. & Johner, C. (2020). *Usability Engineering als Erfolgsfaktor: Effizient IEC 62366- und FDA-konform dokumentieren*. Beuth Verlag.

Gevins, A. & Smith, M. E. (2003). Neurophysiological measures of cognitive workload during human-computer interaction. *Theoretical Issues in Ergonomics Science*, *4*(1–2), 113–131. https://doi.org/10.1080/14639220210159717

Hart, S. G. (2006). Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(9), 904–908. https://doi.org/10.1177/154193120605000909

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

International Electrotechnical Commission (2015b). *Medical devices - Part 2: Guidance on the application of usability engineering to medical devices* (IEC Standard No. TR 62366-2).

International Electrotechnical Commission (2020). *Amendment 1 - Medical devices - Part 1: Application of usability engineering to medical devices* (IEC Standard No. 62366-1:2015/AMD1:2020).

Lourenco, S. F. & Tasimi, A. (2020). No Participant Left Behind: Conducting Science During COVID-19. *Trends in Cognitive Sciences*, *24*(8), 583–584. https://doi.org/10.1016/j.tics.2020.05.003.

McLaughlin, A. C., DeLucia, P. R., Drews, F. A., Vaughn-Cooke, M., Kumar, A., Nesbitt, R. R. & Cluff, K. (2020). Evaluating Medical Devices Remotely: Current Methods and Potential Innovations. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *62*(7), 1041–1060. https://doi.org/10.1177/0018720820953644.

Mejía-Gutiérrez, R. & Carvajal-Arango, R. (2017). Design Verification through virtual prototyping techniques based on Systems Engineering. *Research in Engineering Design*, *28*(4), 477–494. https://doi.org/10.1007/s00163-016-0247-y

Ross, C. (2021, 29. April). *Summative Usability Testing: Why and How to Do It*. Mindflow Design. https://www.mindflowdesign.com/insights/summative-usability-testing/

Sauer, J., Sonderegger, A., Heyden, K., Biller, J., Klotz, J. & Uebelbacher, A. (2019). Extra-laboratorial usability tests: An empirical comparison of remote and classical field testing with lab testing. *Applied Ergonomics*, *74*, 85–96. https://doi.org/10.1016/j.apergo.2018.08.011

Wiklund, M. E., Kendler, J. & Strochlic, A. Y. (2015). *Usability Testing of Medical Devices* (2 New edition). Taylor & Francis Inc.

Woods, A. T., Velasco, C., Levitan, C. A., Wan, X. & Spence, C. (2015). Conducting perception research over the internet: a tutorial review. *PeerJ*, *3*, e1058. https://doi.org/10.7717/peerj.1058