*Improving Usability, Safety and Patient Outcomes with Health Information Technology*
*F. Lau et al. (Eds.)*

*doi:10.3233/978-1-61499-951-5-430*

# Usability Analysis of Contending Electronic Health Record Systems

Mari TYLLINEN[a,b,1], Johanna KAIPIO[a] and Tinja LÄÄVERI[b,c]
[a] *Department of Computer Science, Aalto University, Finland*
[b] *Oy Apotti Ab, Finland*
[c] *Inflammation Center, University of Helsinki and Helsinki University Hospital, Finland*

**Abstract.** In this paper, we report measured usability of two leading EHR systems during procurement. A total of 18 users participated in paired-usability testing of three scenarios: ordering and managing medications by an outpatient physician, medicine administration by an inpatient nurse and scheduling of appointments by nursing staff. Data for audio, screen capture, satisfaction rating, task success and errors made was collected during testing. We found a clear difference between the systems for percentage of successfully completed tasks, two different satisfaction measures and perceived learnability when looking at the results over all scenarios. We conclude that usability should be evaluated during procurement and the difference in usability between systems could be revealed even with fewer measures than were used in our study.

**Keywords.** usability, evaluation, measurement, procurement, electronic health record system

## 1. Introduction

Usability of electronic health record (EHR) systems is a continuous topic of discussion. Usability is defined as effectiveness [1], efficiency [1, 2], satisfaction [1, 2], learnability [2], and errors [2]. Usability testing is an established method for analyzing the usability of an IT system [2] and different measures can be used to quantify usability [3, 4]. Usability evaluation methods have been used increasingly in health informatics field since 2005 [5]. However, EHR systems' usability continues to be inadequate and physicians are dissatisfied with them [6]. In a recent study, two leading EHR systems' usability was compared in ordering tasks and differences were found both between vendors and within different implementations of the same system [7].

Comparisons of user performance between EHR system implementations are not widely available. The purpose of this study was to measure and compare usability of two contending EHR systems. Usability testing was performed as part of a large-scale procurement of a client and patient information system for specialized and primary health care as well as social care in autumn 2014 in Finland [8]. In this paper, we report the measured usability of the two systems based on the data gathered during usability testing in procurement. Based on the data, we discuss the main usability challenges users face when using two of the leading EHR systems.

[1] Corresponding Author, Mari Tyllinen, E-mail: mari.tyllinen@aalto.fi

## 2. Materials and Methods

### *2.1. Study Design, Setting and Participants*

A within-participants usability testing study was conducted between two contending EHRs (Epic and Cerner) during a governmental IT system procurement in Finland. This study was part of the second phase of the procurement, and comprised three user scenarios and user groups. The scenarios and the tasks in them were selected to represent typical tasks of physicians and nurses, which are relevant and of high importance from the viewpoint of EHR system use and patient safety. Clinical experts and usability specialists from the procurement office prepared the three scenarios: ordering and managing medications by an outpatient physician (19 tasks), medicine administration by an inpatient nurse (12 tasks) and scheduling of appointments by nursing staff (14 tasks). The contending vendors configured their existing systems to best meet the needs of the scenarios. Identical fictitious patient data was given to the vendors to be used in the test.

A paired-user usability testing method was used [9], because the users were not familiar with the EHRs beyond seeing demonstrations. Participants received standardized instructions to complete each task as a team. They watched a training video prepared by the vendor of maximum 7 minutes. The pairs were instructed to switch the person using keyboard and mouse when the first half of tasks were completed. This order was reversed for the second EHR. The moderator, who was a usability specialist, read each task aloud and gave it in written format, but did not assist with task performance.

Testing was done in two small conference rooms at the procurement office. Both vendors provided the software and hardware setup for testing. The procuring organization provided wifi connection, identical monitors, mouse and keyboard. Audio, screen capture, mouse click, keystrokes and satisfaction rating data was collected during testing. The test moderator documented user performance using a separate device [10]. The same usability specialist moderated the tests for the same scenario on both vendors' EHR systems. Before testing, the usability specialist prepared herself for the testing by using the EHR system utilizing detailed instructions for correct performance of tasks provided by the vendor, and with the help of a clinical expert.

Physicians and nurses from each user group were recruited from procuring organizations' staff. Testing was conducted during their normal working hours without additional compensation. Three pairs of users tested each scenario, totaling 18 participants. All had several years of experience with clinical work and use of EHRs.

### *2.2. Measurements, Outcomes and Analysis*

The participants filled a consent form and a demographic survey. A maximum of 12 minutes was given for each task. Task time, task success and errors during the user performance were recorded. Both users used feedback devices (positive and negative responses) during the tasks and once more after each task. All data was captured with a hardware solution [10]. After completion of all tasks or reaching maximum testing time (90 minutes), the users answered a standardized usability questionnaire (SUS [11]).

Measures reported in this paper are: (1) percentage of successfully completed tasks (effectiveness), (2) errors made during task completion (deviations from optimal path), (3) perceived learnability (learnability factor from SUS questionnaire [12]), (4) user satisfaction during testing, and (5) after testing was completed (SUS). Successful completion of a task was based on the goal of each task and the optimal path provided by the vendor; errors were allowed during task completion. The percentage of successfully completed tasks were calculated out of those completed within testing time. Errors were defined as deviations from the optimal path. The errors were divided into small and large (0.5 and 1 points, respectively) and reported for successfully completed tasks. User satisfaction during testing was defined as the percentage of tasks with more positive than negative markers from both users combined. Perceived learnability and SUS score were standard measures from the SUS questionnaire [11,12]. To quantify usability, all data was analyzed for each task, test session and scenario separately. Task success, errors and satisfaction data during testing (incl. SUS) was transformed into and calculated with MS Excel.

## 3. Results

We refer to the EHR systems with letters X and Y to prevent disclosure of the specific vendors. An overview of the results is presented in Table 1.

### *3.1. Scenario 1: Ordering and Managing Medications*

The first scenario comprised 19 tasks. Within given time, the pairs completed on average 13 tasks with system X and 17 tasks with system Y. Effectiveness for system X (59.0%) was lower than for system Y (90.4%).

**Table 1.** Results of measures for all three tested scenarios.

| **Scenario** | **Usability measures** | **System X** | **System Y** |
|---|---|---|---|
| | | **Mean (SD)** | **Mean (SD)** |
| Ordering and managing medications | Tasks time to complete (count) (N=19) | 13 (2.8) | 17 (2.4) |
| | Successfully completed tasks (%) | 59.0 | 90.4 |
| | Errors (deviations from optimal path) (points/task) | 2.6 (0.6) | 1.6 (0.5) |
| | Perceived learnability (score) | 22.9 (19.7) | 58.3 (20.0) |
| | Satisfaction during testing (%) | 28.2 | 78.8 |
| | Overall satisfaction (SUS score) | 32.5 (12.2) | 70.8 (9.5) |
| Medication administration | Tasks time to complete (count) (N=12) | 12 (0) | 12 (0.5) |
| | Successfully completed tasks (%) | 50.0 | 62.9 |
| | Errors (deviations from optimal path) (points/task) | 3.0 (0.7) | 2.9 (0.2) |
| | Perceived learnability (score) | 27.1 (13.3) | 35.4 (15.2) |
| | Satisfaction during testing (%) | 47.2 | 71.4 |
| | Overall satisfaction (SUS score) | 25.4 (16.1) | 70.0 (11.6) |
| Scheduling appointments | Tasks time to complete (count) (N=14) | 14 (0) | 14 (0) |
| | Successfully completed tasks (%) | 52.4 | 76.2 |
| | Errors (deviations from optimal path) (points/task) | 1.3 (0.3) | 0.7 (0.3) |
| | Perceived learnability (score) | 27.1 (13.3) | 50.0 (23.9) |
| | Satisfaction during testing (%) * | 26.2 | 40.5 |
| | Overall satisfaction (SUS score) | 41.3 (12.1) | 64.6 (10.2) |
| Overall (all scenarios) | Successfully completed tasks (%) | 53.8 | 78.3 |
| | Errors (deviations from optimal path) (points/task) | 2.3 | 1.6 |
| | Perceived learnability (score) | 25.7 | 47.9 |
| | Satisfaction during testing (%) | 33.3 | 64.3 |
| | Overall satisfaction (SUS score) | 33.1 | 68.5 |

* Data completely unavailable for one pair, and partly available for one pair due to technical problems with data gathering. The results are calculated for both systems based on the same pairs.

With system X, the pairs had more deviations from optimal path of completing the successful tasks than with system Y (mean 2.6 vs. 1.6 error points). The users rated both their satisfaction during testing (28.2%) and overall satisfaction SUS score (32.5) with system X much lower than system Y (78.8% and 70.8). The perceived learnability score for system X was lower (22.9) than for system Y (58.3), although there is large deviation.

With system X, all three pairs failed in completing the same three tasks. With system Y, at least one pair completed correctly all tasks, and two or three pairs completed correctly 18/19 tasks. In both systems, two tasks had high error points (2.5 to 8 points) from more than one pair. The tasks difficult to complete were related to documenting a change in the dosing history of a medication, and marking a medication on a pause either in the past or in the future. The latter had also high error points in system Y. The pairs had a high number of errors also on tasks that related to ordering a medicine with different dosing in the morning than evening, and gradually changing dosing.

### *3.2. Scenario 2: Medication Administration*

The second scenario included 12 tasks, which almost all were completed with both systems by all pairs in time. Key measures were (system X / system Y): effectiveness (50.0% / 62.9%), error points (3.0 / 2.9), satisfaction (47.2% / 71.4%), SUS (25.4 / 70.0) and perceived learnability (27.1 / 35.4). Meaningful differences are in effectiveness and satisfaction (incl. SUS) in favor of system Y.

With system X, three tasks were either not prepared correctly or had missing system functionality, and were thus not completed successfully. All pairs failed in completing one of the tasks. With system Y, one task had missing system functionality and all pairs failed in completing one of the tasks. There were three tasks that only one pair completed correctly. The tasks the pairs had trouble with related to administering and documenting an order for an infusion, adding times for next administrations, updating the list of medications per physician orders, and documenting the administration of all oral medications at once. Two tasks in both systems had high error points (4 to 11) from more than one pair. The tasks included infusions, documenting physician orders, cancelling an already documented administration, and returning to home medications before discharge.

### *3.3. Scenario 3: Scheduling Appointments*

The third scenario included 14 tasks, which all pairs completed within the allocated time. Key measures were (system X / system Y): effectiveness (52.4% / 76.2%), error points (1.3 / 0.7), satisfaction (26.2% / 40.5%), SUS (41.3 / 64.6) and perceived learnability (27.1 ±13.3 / 50.0 ±23.9) with large deviations. Meaningful differences are in effectiveness and satisfaction (incl. SUS) in favor of system Y.

With system X, the pairs could not complete six tasks, because of either missing system functionality or problems with preparations. One task only one pair completed correctly. In system Y, one task did not have correctly prepared data and no pair completed one of the tasks. The pairs struggled in completing tasks related to checking whether patient receives notifications from appointments and to which phone number, removing a notification from an appointment, as well as scheduling a series of three appointments two days apart. With system X, one task received high error points (3.5 to 4) from two pairs. With system Y, two of the tasks had high error points (4.5 / 5.5) from one pair. The tasks were related to rescheduling an appointment, scheduling a series of three appointments and changing the default duration of the appointment.

### *3.4. Overall results*

System X received lower ratings in all measures when looking at the results over all scenarios. There were great differences with regards effectiveness (53.8% / 78.3%), satisfaction (33.3 % / 64.3%), SUS (33.1 / 68.5) and perceived learnability (25.7 / 47.9).

## 4. Discussion and Conclusion

Our paired-user usability testing study revealed a difference between the two systems. Successful task completion rates were lower for system X in all three scenarios, and lowest in medication administration (50% vs. 62.9%). Greatest differences were in ordering and managing medications. In system X, more tasks had missing system functionality or were not prepared correctly, which accounts for some of the differences in effectiveness in scenarios two and three. However, missing functionality ultimately means user goals not being met; this constitutes an important factor in usability.

The first scenario showed a clear difference in deviations from the optimal path. In the other two scenarios, the differences were not so apparent. Both systems had very high error points for certain tasks in medication administration, while in scheduling the error points were quite low. For system Y, despite low error points and high effectiveness, the user satisfaction ratings were the lowest in scheduling. System X user satisfaction ratings were low for nurse scenarios, and even lower for physicians. In contrast, system Y user satisfaction ratings were highest for physicians. The overall SUS score of 33.1 for system X is unsatisfactory while the score of 68.5 for system Y is acceptable [4].

Both systems are mainly used in the USA. Accordingly, the differences of the tasks in our scenarios (Finnish context) and the original workflows (US context) can be assumed similar for both systems, and are thus not likely to explain the findings. However, configurability capabilities may vary between the systems.

Two limitations deserve to be discussed. Firstly, the users had not been trained to use the systems; apart from the short introductory video and some of them had watched demonstrations earlier in the procurement. Secondly, the vendors had optimized their system for the procurement and thus the systems used did not necessarily match any existing system configuration.

User experiences and dissatisfaction have been used as a basis for EHR system usability criticism [6]. However, usability comparisons based on user performance are not widely available. A similar previous study [7] mainly focused on efficiency. We used several different usability measures to get a reliable view on the usability of the systems from different perspectives for procurement purposes. Our study revealed a clear difference in usability between the two leading systems in all three scenarios and user groups. Our results show that overall all measures can reveal a similar state of usability and similar differences between systems.

Aligned with previous literature [13], we recommend usability testing when procuring EHR systems with varying number of measures depending on desire of further use of results. Our study utilized several measures, but results indicate that also fewer measures could be considered in a pure comparison situation.

## References

[1] ISO. *Ergonomic requirements for office work with visual display terminals (VDTs)-Part 11: Guidance on usability (ISO 9241-11:1998).* International Organization for Standardization, Geneva, CH, 1998.

[2] J. Nielsen. *Usability engineering*, Academic Press Inc., San Diego, CA, 1993.

[3] K. Hornbæk. Current practice in measuring usability: Challenges to usability studies and research. *Int J Human-Computer Stud* **64** (2006), 79-102.

[4] J. Sauro, J.R. Lewis. *Quantifying the user experience.* Elsevier, Waltham, MA, 2012.

[5] M.A. Ellsworth, M. Dziadzko, J.C. O'Horo, A.M. Farrell, J. Zhang, V. Herasevich. An appraisal of published usability evaluations of electronic health records via systematic review, *JAMIA* **24** (2016), 218-226.

[6] J. Kaipio, T. Lääveri, H. Hyppönen, S. Vainiomäki, J. Reponen, A. Kushniruk, E. Borycki, J. Vänskä. Usability problems do not heal by themselves: National survey on physicians' experiences with EHRs in Finland. *Int J Med Inform* **97** (2017), 266-281.

[7] R. M. Ratwani, E. Savage, A. Will, R. Arnold, S. Khairat, K. Miller, R. J. Fairbanks, M. Hodgkins, A. Z. Hettinger. A usability and safety analysis of electronic health records: a multi-center study, *JAMIA* **25** (2018), 1197-1201.

[8] M. Tyllinen, J. Kaipio, T. Lääveri. A framework for usability evaluation in EHR procurement. In A. Ugon et al. (Eds.), *Building continents of knowledge in oceans of data: The future of co-created eHealth, Stud Health Tech Inform* **247,** IOS Press, 2018**.**

[9] G.S. Hackman, D.W. Biers. Team usability testing: Are two heads better than one? *Proc Human Factors and Ergonomics Society Annual Meeting* **36** (1992), 1205 – 1209.

[10] J. Pitkänen, M. Nieminen, M. Pitkäranta, J. Kaipio, M. Tyllinen, A. K. Haapala. UXtract – Extraction of Usability Test Results for Scoring Healthcare IT Systems in Procurement. In *Proc 14th Scandinavian Conference on Health Informatics* 2016, Gothenburg (Sweden), 2016.

[11] J. Brooke. SUS: A "quick and dirty" usability scale. In P. Jordan, B. Thomas, T. Weerdmeester, A. McClelland (Eds.), *Usability evaluation in industry*, Taylor and Francis, 1996.

[12] J. R. Lewis, J. Sauro. The factor structure of the system usability scale. In M. Kurosu (Ed.) *Human Centered Design, LNCS* **5619**, Springer, 2009.

[13] A. Kushniruk, M-C. Beuscart-Zéphir, A. Grzes, E. Borycki, L. Watbled, J. Kannry. Increasing the safety of healthcare information systems through improved procurement: Toward a framework for selection of safe healthcare systems*, Healthc Q* **13** (2010), 53-58.