JHIA - online.org
Journal of Health Informatics in Africa

11th Health Informatics in Africa Conference (HELINA 2018)
Peer-reviewed and selected under the responsibility of the Scientific Programme Committee

A Comparative Usability Study of Two Touchscreen Clinical Workstations for Use in Low Resource Settings

Timothy Mtonga*, Menna Abaye, Samuel Charles Rosko, Gerald Paul Douglas

Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, United States

Cost of implementation is one of the biggest barriers to scale up and sustainability of electronic medical record systems in low resource settings. Several approaches can be used to overcome this barrier. In this comparative usability study, we assess whether a lower cost 10-inch touchscreen clinical workstation (TCW) is a suitable alternative to a 14-inch TCW that has been widely deployed in Malawi. A total of 27 participants performed a patient registration task using three different TCWs, a 14-inch device and two 10-inch devices. One of the 10-inch devices used the same size of interface controls as the 14-inch TCW with reduced space between buttons, and the 10-inch device used a modified interface with controls shrunk down to accommodate the smaller screen size. We measured task completion times and error rates as metrics for assessing performance with each workstation. We also captured user perceptions using a usability survey and an exit survey. We compared the mean task completion time, number of errors, and survey scores between the three devices. In addition, a codebook was created to conduct a thematic analysis of the free-text responses to the exit survey. Our results suggest that the 10-inch TCW is a suitable alternative to the 14-inch TCW based on task completion time. Nonetheless, modifications to the user interface are necessary to reduce error rates on the 10-inch TCW before implementation.

1 Introduction

Electronic medical record systems (EMR) have shown potential to improve patient care and outcomes in low-resource settings [1, 2]. However, cost of initial implementation and maintenance remains a significant barrier to the adoption and scaling up of EMRs [3]. Due to the high cost of implementation, the rate of adoption for EMRs has been slow in low-resource settings.

The Center for Health Informatics for the Underserved (CHIU) in the Department of Biomedical Informatics (DBMI) at the University of Pittsburgh and Baobab Health Trust (BHT) have been working towards reducing the cost of health information technology deployments. Since 2001, BHT has deployed low-cost touchscreen workstations in over 100 health facilities in Malawi for various applications including patient registration, managing antiretroviral therapy, and managing chronic non-communicable disease [4, 5]. These implementations have used a 650 USD, 14-inch touchscreen workstation whose cost continues to present a barrier to scaling up into additional health facilities.

One possible approach to reducing this cost is through the use of an alternative hardware platform, such as a Raspberry Pi mini-computer combined with a 10-inch touchscreen display. At a cost of less than 200 USD, this technology solution offers significant cost savings in comparison with the 14-inch TCW currently in use. However, the cost of the workstations is not the only factor that should be considered. Before the existing workstations can be replaced, other factors that affect user satisfaction should also be considered. The most important of these factors is the need for technology to enhance and not hinder the work process. Technology can often be a hinderance if it introduces new errors and is unreliable.

To determine if a Raspberry Pi workstation can replace the 14-inch TCW, we conducted a comparative usability study to assess if users can achieve comparable task performance and efficiency with either

* Corresponding author address: tmm113@pitt.edu

J Health Inform Afr. 2018;5(2):78-84. DOI: 10.12856/JHIA-2018-v5-i2-209

workstation. We hypothesized that there will not be a statistically significant difference in task completion time or number of errors when using either workstation.

2 Methods

2.1 Setting and Materials

To measure task completion times and errors, we used three TCWs, a 14-inch TCW and two 10-inch Raspberry Pi TCWs. On each workstation, we ran the Baobab Health Trust Patient Registration (BHT-PR) system which is available at no cost as open-source software on GitHub [6]. Two workstations, the 14-inch TCW and one 10-inch Raspberry Pi TCW (10-inch-Normal) used the current user interface configuration for the patient registration task. The size of the buttons and other controls was the same for these two devices but the spacing between them was reduced for the 10-inch workstation due to reduced screen real estate. The third workstation, another 10-inch Raspberry Pi device (10-inch-Resized), used a customized user interface where the button and control sizes were decreased in size to maintain the overall feel and look of the user interface. A screenshot of each of the three workstations is provided in **Supplement 1**. All workstations were set up in an isolated room at DBMI offices for the entire duration of this study.

2.2 Participant Recruitment

A convenience sampling method was used to recruit study participants from DBMI. Participants were required to be adults who had no prior experience using the patient registration application. Familiarity with touchscreen devices was not required for participants to be eligible for the study. Participants in the study were not compensated in any way. The study was approved as non-human subjects research by the Institutional Review Board (IRB) at the University of Pittsburgh.

To determine the number of participants that were required for our study, we performed a power analysis calculation. Due to the limited time we had to perform the study, we opted for a power level of 0.80 and a significance level of 0.10 in our power calculation. We estimated that a 30% increase in task completion time would not significantly increase the total time it takes to register patients over the course of a day. Therefore, we used an effect size of 0.3 in our power calculation, which yielded a required sample size of 30 participants.

2.3 Data Collection

Each participant was required to perform a patient registration task using a standardized task list. The task involved entering demographic information for a simulated patient, including patient name, date of birth, place of origin, and current address. All demographic details for the simulated patient were provided at each workstation. Additional information on the specifics of the task list can be found in **Supplement 2**.

To calculate task completion time for each participant, we logged button presses as the participants performed their task. This was done using an automated process that recorded each button that was pressed and the exact time that it was pressed. In addition to task completion time, we also recorded errors that the participants made while performing the task. This was done through review of screen capture videos that were recorded as participants worked through the patient registration task. In this study, errors are defined as the cases where a user fails to press the right target on the touchscreen. Other literature will refer to this as slips [7]. We did not record cases where the user did not know what action to perform as errors as this was not of interest for our study.

Upon completion of the task on each workstation, the participants completed a 10-question survey that was aimed at gathering user opinions about the workstation they had just finished using (**Supplement 3**). We adapted our survey from the System Usability Scale by modifying it to make sure that the questions were focused on the performance of the touchscreens rather than the system as a whole [8]. Questions were rated using a Likert-scale ranging from 1-5 where 1 was “Strongly Disagree” and 5 was “Strongly Agree”. In addition, participants were given an exit survey after completing the task on all three workstations that asked them to discuss the advantages and disadvantages of the workstations, in addition to general comments about their experience as a whole.

Since study participants performed the task on the three workstations in one sitting, we were concerned about learning effects when using the workstations. To address this concern, we utilized a Latin square design to assign the order in which participants performed tasks on each workstation [9, 10].

2.4 Data Analysis

To test our hypotheses, we first checked whether our data assumed a normal distribution using the Shapiro-Wilk test for normality. If our data was not normally distributed, we performed outlier analysis using a boxplot to identify data points that fell outside the 75th percentile and removed them. If the removal of these outliers resulted in a normal distribution, we ran a repeated measures Analysis of Variance (ANOVA) test to compare the difference among the means of the workstations. Otherwise, we ran the Kruskal-Wallis Ranked Sign test, which is the non-parametric equivalent of the repeated measures ANOVA if the removal of the outliers did not result in a normal distribution.

In addition to these tests, we analyzed our 10-question system usability survey by calculating a Cronbach's alpha for each workstation in order to evaluate their reliability. Our exit surveys were analyzed by generating a codebook from the comments in each section of the survey. The codebook compiled a list of main themes in the participants comments with a description and an example. We then used the codebook to go back and code the comments in the exit survey to identify those that came up most often to get an idea of user perceptions of the workstations.

3 Results

3.1 Participant Recruitment

To ensure the viability of our study protocol, we recruited two trial participants to perform the patient registration task on all three touchscreen workstations. These trial participants identified several flaws with our protocol, which we addressed before beginning our data collection. Following this trial run, we recruited 25 participants from DBMI to perform our patient registration task and collected data for each of them. Due to time constraints, we were unable to recruit 30 participants as determined by our power calculation. This reduced the power of our study from 80% to 74% as calculated in our post-hoc power analysis.

3.2 Analysis of Task Completion Times

Completion times were recorded for all 25 participants as they completed tasks across all three touchscreen devices. Table 1 shows the average times and standard deviations for each device.

Table 1. Summary statistics about task completion time for the patient registration task in seconds.

Metric \ Device	14-inch TCW	10-inch-resized	10-inch-normal
Mean	106.97	110.40	109.24
Std. Deviation	28.61	32.40	27.79

We performed a Shapiro-Wilk test for normality to check if task completion times were normally distributed for each device. Table 2 shows the results of this test.

Table 2. Results of test normality for task completion times with Shapiro-Wilk test.

Metric \ Device	14-inch TCW	10-inch-resized	10-inch-normal
P-value with outliers	0.218	0.008	0.263
P-value without outliers	0.292	0.439	0.409

DOI: 10.12856/JHIA-2018-v5-i2-209

The P-values for the 14-inch TCW and 10-inch-normal devices showed that their data was normally distributed, but this was not the case for the 10-inch-resized device, which returned a P-value of 0.008. Because the distributions were not normal for all three devices, a boxplot was created to perform outlier analysis.

Using the boxplot, we identified three outliers in task completion time. We reviewed the screen capture videos for the outliers to gain an understanding of why they had extreme task completion times. All three outliers were found to have issues with completing the task related to selecting a date from the calendar. Because the issues with the calendar were unrelated to the touchscreens that were being compared in this study, the three outliers were removed and the distributions were re-tested for normality using the Shapiro Wilk test. The results of this test are shown in Table 2.

This test showed that all three distributions were normally distributed following the removal of outliers, so a one-way repeated measures ANOVA test was run to compare the mean task completion time across the three devices. The ANOVA test returned a p-value of 0.977, indicating that there was no statistically significant difference in mean task completion time among the three devices.

3.3 Analysis of Error Rates

Error rates were recorded for 20 of the 25 participants as they performed the required tasks across all three touchscreen devices. We were not able to use data from five of the participants due to issues with the screen capture software. The mean and standard deviation of number of errors for each device is shown in Table 3.

Table 3. Summary statistics about number of errors made during the patient registration task.

Metric \ Device	14-inch TCW	10-inch-resized	10-inch-normal
Mean	2.84	9.20	7.00
Std. Deviation	3.44	3.71	4.23

Before making a statistical comparison of means, we performed the Shapiro-Wilk test for normality. The results of this test are shown in Table 4.

Table 4. Results of Shapiro-Wilk test for normal distribution for number of errors.

Metric \ Device	14-inch TCW	10-inch-resized	10-inch-normal
P-value with outliers	0.00002	0.338	0.006
P-value without outliers	0.031	0.359	0.118

The results of this testing showed that the data from the 14-inch and the 10-inch-normal TCWs are not normally distributed as noted by their low p-values of 0.00002 and 0.006 respectively. On the other hand, data from the 10-inch-resized TCW was normally distributed. Because the number of errors was not normally distributed for all three devices, a boxplot was used to determine if there were any outliers that were affecting the distribution.

Our analysis of the boxplot revealed two outliers, one for the 14-inch device and one for the 10-inch-normal device. These outliers coincided with those identified in the task completion time analysis and were removed for the same reasons. Following the removal of outliers, the Shapiro-Wilk test was performed again to check if the number of errors was now normally distributed.

Despite the removal of outliers, the p-value of the 14-inch TCW remained below the desired level of 0.1. Therefore, we used the Kruskal-Wallis Rank Sum test to compare the means for the number of errors between devices. This test returned a p-value < 0.001. This indicates that the means number of errors is statistically significantly different among the three devices we tested.

3.4 Analysis of Usability Surveys

Data was collected from 24 of the 25 participants recruited for the study using a 10-question usability survey. One participant did not complete the survey for all three devices. As a result of this, their survey was not included in the data analysis. We first analyzed our survey data by calculating a Cronbach's alpha value. This calculation yielded a value of 0.452, which is considered to be below the acceptable level.

Following this reliability test, we took the survey responses and calculated a total system usability score from them, using a scale from 0-100. System usability scores above 68 are considered to be "above average" and those below 68 are considered to be "below average". The mean and standard deviations for the system usability score for all three devices are shown in Table 5.

Table 5. Summary statistics about number of errors made during the patient registration task.

Metric \ Device	14-inch TCW	10-inch-resized	10-inch-normal
Mean	69.33	62.67	65.17
Std. Deviation	6.94	7.99	9.45

Using the Shapiro-Wilk test, we found that data from all three of the devices was normally distributed at alpha = 0.1 as can be seen in Table 6.

Table 6. Results of Shapiro-Wilk test for normal distribution of usability scores

Metric \ Device	14-inch TCW	10-inch-resized	10-inch-normal
P-value	0.1899	0.7699	0.2002

Since the data was normally distributed, we proceeded to perform a repeated measures ANOVA to compare the mean score for each of the three devices. This test returned a p-value < 0.0001, which is well below the threshold value of 0.1, leading us to the conclusion that the mean system usability score was statistically significantly different among the three touchscreen devices.

3.5 Evaluation of Exit Surveys

Participant responses in the exit survey were categorized into common themes in order to better examine the user's perceptions of the advantages and disadvantages of each device. The most frequently mentioned advantages of the two 10-inch TCWs over the 14-inch TCW were that they had a better brightness level, were easier to use, and they had faster response times. Regarding disadvantages, participants commented that the 10-inch TCWs had a smaller screen size, a lower resolution, a smaller font size, a more delayed response, and a lower screen brightness in comparison to the 14-inch TCW. In the "Additional Comments" section of the exit survey, almost all of the comments were about potential improvements to the interface of the patient registration software (See Supplement 4 for the codebook with descriptions and examples).

4 Discussion

The purpose of our study was to determine whether the 10-inch device is a viable alternative to the 14-inch TCW. To this end, we hypothesized that there would be no statistically significant difference in mean task completion time and number of errors when using either device.

First, we compared the average task completion time across all three devices to determine if participants took roughly the same amount of time to complete the task on any given device. Our analysis showed that there was no statistically significant difference in mean task completion time among the three devices (P-value = 0.977). This supports our hypothesis that the 14-inch TCW could be replaced by a 10-inch TCW without significantly increasing the amount of time it would take for users to perform tasks on a day to day basis.

Next, we compared the number of errors made by participants across all three devices in order to determine if the change in devices had an effect on the number of errors that participants made when performing the task. This is particularly important in our case because a change from a larger screen to a smaller screen reduces the amount of real estate available to display information. Since the correlation between button sizes, inter-button spacing and the number of errors is well defined, we had to confirm if our alterations to the user interface and the reduction in screen size led to an increase in errors [11].

Our analysis showed there was a statistically significant difference in the number of errors that participants committed across the three devices (P-value < 0.0001). While this is contrary to our hypothesis, it confirms the understanding that button sizes and inter-button spacing have an effect on the number of errors that users make. Despite this finding, we still believe that a 10-inch TCW is a suitable replacement for the 14-inch TCW because simple modifications to the user interface of the patient registration application could significantly reduce error rates on the devices. For example, a lot of errors that were made on the 10-inch TCW involved participants trying to select one letter on the on-screen keyboard and accidentally hitting a letter next to the desired letter. If the size of the buttons on the keyboard or the spacing between the keys were increased, we believe that the error rates of the 10-inch devices would be more similar to those attained by the 14-inch TCW. While these modifications can lead to more of the screen real estate being used for the onscreen keyboard, this is the preferred scenario as opposed to users making more errors and compromising the integrity of the data that is collected.

Another part of our study focused on user perceptions of the three different workstations. We used a system usability survey to identify which workstation participants preferred. This was done to ensure that the 10-inch TCWs were comparable to the 14-inch TCW in terms of both performance and user satisfaction. Our analysis showed that our attempts to modify the system usability scale survey instrument reduced its reliability to 0.452, which is far below the universally acceptable value. This suggests that participants in our study did not understand the questions in the survey to mean the same thing. While this may be a result of different factors, it indicates that our survey may have been poorly designed.

Our comparison of the mean system usability scores derived from the survey showed that devices were not all favored equally and participants preferred the 14-inch TCW to the 10-inch TCW. While the low reliability of the survey urges caution in the interpretation and importance of this finding, we believe that this will not be a problem when the devices are used in real life as users will not be exposed to both devices as was the case in the study.

One interesting outcome of our study was that the exit survey showed contradictory responses from participants. For example, some participants thought the 10-inch TCW had a brighter screen while others thought that the 14-inch TCW had a brighter screen. This was interesting because the lighting in the room and the brightness of the devices were constant throughout the study. We therefore concluded that this could have been a function of different user preference and perceptions.

One of the main limitations of this study is that it was a task-based study, which potentially introduced confounding factors to the study. Since the tasks that participants were asked to perform had a series of activities that involved thinking and locating specific words on the screen, we were measuring both their ability to learn and their ability to perform the task.

For example, participants were asked to select a series of specific names from a list of locations in Malawi. Participants that were unfamiliar with Malawian names struggled to select these locations from the list. Because this did not occur for every participant, this could have artificially inflated task completion times for reasons unrelated to the performance of the touchscreen. Another example is that many participants struggled with the “Enter Date of Birth - 17” task from the task list. In this case, the participants were presented with a calendar with date options presented as buttons from the 1st to the 31st, as seen in Figure 1.

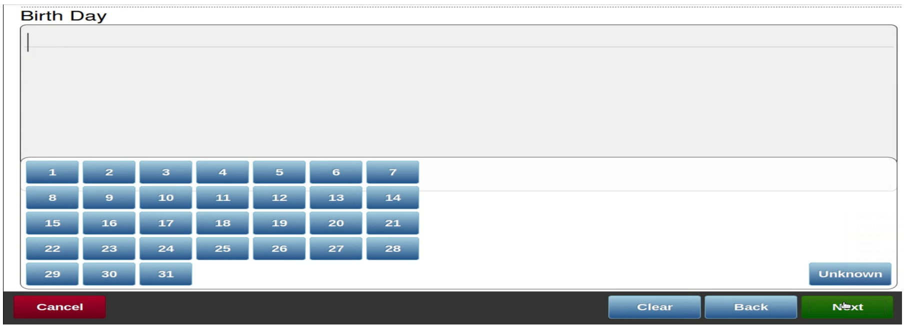

Figure 1. Calendar control used in the BHT patient registration application

Upon review of screen capture videos, we found that many participants spent a lot of time on this screen choosing "1" and "7" over and over again, not realizing that there was a "17" button below. In some cases, participants spent up to a minute doing this. These were the outliers that were removed in our analysis. This potentially introduced confounding to our measurement of task completion times and error rates. In the future, a study with a participant population that is more familiar with both the context and application could help address these issues.

5 Conclusion

In summary, we can say that the 10-inch TCW is a suitable alternative to the 14-inch TCW; however, the two devices should not use the same user interface. Our error rate findings show that modifications to the button size and spacing are necessary before replacing the 14-inch TCW with a 10-inch TCW.

References

[1] B. Jawhari, D. Ludwick, L. Keenan, D. Zakus, and R. Hayward, "Benefits and challenges of EMR implementations in low resource settings: a state-of-the-art review," *BMC Med. Inform. Decis. Mak.*, vol. 16, p. 116, Sep. 2016.

[2] J. A. Blaya, H. S. F. Fraser, and B. Holt, "E-Health Technologies Show Promise In Developing Countries," *Health Aff. (Millwood)*, vol. 29, no. 2, pp. 244–251, Feb. 2010.

[3] J. Driessen *et al.*, "Modeling return on investment for an electronic medical record system in Lilongwe, Malawi," *J. Am. Med. Inform. Assoc. JAMIA*, vol. 20, no. 4, pp. 743–748, Aug. 2013.

[4] R. C. Manjomo *et al.*, "Managing and monitoring chronic non-communicable diseases in a primary health care clinic, Lilongwe, Malawi," *Public Health Action*, vol. 6, no. 2, pp. 60–65, Jun. 2016.

[5] G. P. Douglas *et al.*, "Using Touchscreen Electronic Medical Record Systems to Support and Monitor National Scale-Up of Antiretroviral Therapy in Malawi," *PLOS Med.*, vol. 7, no. 8, p. e1000319, Aug. 2010.

[6] Baobab Health Trust, *Patient Registration*. https://github.com/BaobabHealthTrust/Registration.

[7] H. Nicolau, K. Montague, T. Guerreiro, A. Rodrigues, and V. L. Hanson, "Typing Performance of Blind Users: An Analysis of Touch Behaviors, Learning Effect, and In-Situ Usage," in *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, New York, NY, USA, 2015, pp. 273–280.

[8] J. Brooke, "SUS - A quick and dirty usability scale," p. 7.

[9] Z. L. Lewis, G. P. Douglas, V. Monaco, and R. S. Crowley, "Touchscreen task efficiency and learnability in an electronic medical record at the point-of-care," *Stud. Health Technol. Inform.*, vol. 160, no. Pt 1, pp. 101–105, 2010.

[10] R. E. Kirk, "Latin Square Design," in *The Corsini Encyclopedia of Psychology*, John Wiley & Sons, Inc., 2010.

[11] Z. X. Jin, T. Plocher, and L. Kiff, "Touch Screen User Interfaces for Older Adults: Button Size and Spacing," in *Universal Acess in Human Computer Interaction. Coping with Diversity*, 2007, pp. 933–941.