INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS 83 (2014) 27–36

journal homepage: www.ijmijournal.com

Merits of usability testing for PACS selection

Wiard Jorritsma[a,*], *Fokie Cnossen*[b], *Peter M.A. van Ooijen*[a]

[a] *Department of Radiology, University Medical Center Groningen, Hanzeplein 1, 9713 GZ Groningen, The Netherlands*
[b] *Department of Artificial Intelligence, University of Groningen, Nijenborgh 9, 9747 AG Groningen, The Netherlands*

ARTICLE INFO

Article history:
Received 24 May 2013
Received in revised form
10 October 2013
Accepted 11 October 2013

Keywords:
Radiology
Picture Archiving and Communication System
User–computer interface
Software validation
Usability

ABSTRACT

Objectives: To compare the usability of different Picture Archiving and Communication System (PACS) workstations, determine whether a usability test has added value with respect to the traditional way of comparing PACSs based on functional requirements, and to evaluate the appropriateness of a task-based methodology for a PACS usability test.

Methods: A task-based usability test of four PACS workstations was performed. Radiologists' subjective responses to the PACSs and their performance on the tasks were measured. To mimic the traditional PACS selection process, two functional requirements were defined which the PACSs met in varying degrees. The focus of the usability test was on the aspects of the PACS related to these requirements. The usability results were compared to the PACSs' ability to meet the functional requirements.

Results: One PACS outperformed the other PACSs both in terms of subjective preference and task performance, indicating its superior usability. There were differences in usability between PACSs with identical functionality. Also, a PACS with theoretically advantageous functionality for a given task did not necessarily have better usability for this task than a PACS without this functionality. There was a discrepancy between participants' subjective preferences and their task performance, which indicates that it is vital to include performance measures in the usability assessment so that it accurately reflects the efficiency of interaction.

Conclusions: The differences in usability between PACSs with identical functionality indicate that functional requirements alone are insufficient to determine a PACS's overall quality. A usability test should therefore be used in addition to a functional requirement list in a PACS selection process to ensure that a hospital buys the PACS with the highest quality. A task-based usability evaluation methodology, which yields both subjective preference data and objective performance data of radiologists interacting with the PACS, is very suitable for such a usability test.

1. Introduction

Many hospitals currently have a Picture Archiving and Communication System (PACS) that has been in operation for many years. Several of these hospitals have decided to put out bids for a new PACS. Traditionally, the decision regarding which PACS to buy is based on the vendor's ability to meet a list of requirements defined by the hospital. These requirements range from desirable features to knock-out criteria and are based on demands of the IT staff, radiologists, PACS administrators, management, and other stakeholders within the hospital.

* Corresponding author. Tel.: +31 50 3612288.
E-mail address: w.jorritsma01@umcg.nl (W. Jorritsma).
1386-5056/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ijmedinf.2013.10.003

A limitation of this method is that it assesses a PACS's quality solely on the basis of functionality, and does not consider *how* the functionality is implemented. This implementation determines the PACS's *usability*: the effectiveness, efficiency, and satisfaction with which radiologists can interact with it [1,2]. Because the PACS plays such a crucial role in the radiological workflow, the quality of interaction between radiologists and the PACS is very important, and usability should therefore be a major decision criterion in the PACS selection process.

Because of the importance of the interaction between health care professionals and technology, usability evaluations are becoming increasingly common in the medical domain [3]. Most of these evaluations (e.g. [4–6]) are performed on a single system and aim to determine how its usability might be improved. Comparative usability studies require a different approach, because they need to yield a quantitative representation of usability that allows for a comparison between multiple systems. These kinds of studies are much rarer. To our best knowledge, there are only two previously published multi-vendor PACS usability evaluations [7,8]. Both these studies assessed usability by measuring radiologists' subjective responses to the PACS. This is a valid and widely used method to measure usability, but it is not optimal for a PACS usability test, because subjective measures do not always reflect the efficiency of interaction with a system. Several studies have shown that subjective measures correlate with performance measures [9–11], but there are also studies that show that these measures do not correspond [12,13]. Even when the measures do correlate, the correlation is never extremely strong, so measuring performance data provides additional information about a system's usability that is not contained in the subjective measures [10]. Due to the increasing workload radiologists face [14,15], the efficiency with which they can work with the PACS is essential and should therefore be accurately represented in a PACS usability assessment.

In this study, we performed a task-based usability test of four different PACS workstations, in which four radiologists completed a set of predefined prototypical tasks on each PACS. In addition to subjective usability data, this methodology allowed us to obtain objective performance data (e.g. the time needed to perform certain tasks), which reflect the efficiency with which radiologists can work with the PACS. We aimed to compare the usability of the PACSs, determine whether a usability test has added value with respect to the traditional way of comparing PACSs based on functional requirements, and to evaluate the appropriateness of a task-based methodology for a PACS usability test.

2. Methods

2.1. Participants

Three board certified radiologists and one fifth-year radiology resident participated in the test (age range 33–46, mean age: 36.5).

2.2. Apparatus

The workstation consisted of a Barco 6MP MDCC-6130 30.4 in. monitor and a Barco 2MP MDRC-2120 20.1 in. monitor controlled by a PC workstation with a Barco graphics card. This is the same setup radiologists at our hospital normally work with. Participants interacted with the system through a standard mouse and keyboard. RUI (Recording User Input) version 2.02 [16] was used to log the mouse movements, mouse clicks and key presses. The large screen was recorded by a Logitech Webcam Pro 9000 and the small screen by a Logitech Webcam C110. Each webcam was connected to a different PC. Both PCs recorded with Logitech Webcam Software. Both the webcams and the key/mouse logger recorded the system time of the workstation PC so that their outputs could be synchronized. We used webcams instead of screen capture software because screen capturing of such high resolution monitors could affect the performance of the workstation.

2.3. Materials

The usability test consisted of six scenarios and a questionnaire. Each scenario consisted of a set of tasks participants had to perform. The scenario set was carefully designed, in consultation with several radiologists, to be as representative of a radiologist's daily work as possible. The scenarios and questionnaire aimed to evaluate common interactions with a PACS and were not specifically designed for any of the four PACSs used in this study. All scenarios were based on studies from the top ten most frequently occurring studies in our hospital (shown in Appendix A). The scenarios are shown in Table 1.

The patient images used in the scenarios were collected from our hospital's database. The DICOM header was anonymized and the images were subsequently stored in a local database. During the anonymization process, the patient's name and patient identification number were replaced with the scenario number. Each scenario contained multiple studies of the same patient. In addition to the images necessary for completing the scenarios, each scenario except Scenario 6 also contained images of that same patient that were irrelevant to the scenario. These studies were included to make the selection of relevant studies realistic and not trivial.

The questionnaire consisted of 19 statements regarding the usability of the PACS. Participants were asked to indicate the extent to which they agreed with each statement on a five-point Likert scale with the following levels: strongly disagree, disagree, neutral, agree and strongly agree. There were also two text fields in which participants could indicate additional positive and negative aspects of the PACS. The entire questionnaire is shown in Appendix B.

2.4. Design and procedure

Four PACSs were compared in the test. For legal reasons, we anonymized the names of the PACSs and refer to them as PACS A, B, C, and D. To mimic the process of PACS selection based on a list of functional requirements, we defined two requirements for the PACSs, one which was met by only two PACSs (automatic retrieval of images from relevant previous studies) and one which was met by all four (measurement tools)

Table 1 – Scenarios used in the usability test.

Scenario	Modality	Body part	Clinical information	Task
1	CR/DX	Knee	Osteoarthritis?	• Retrieve the new and old images – Does this patient have osteoarthritis? – Zoom in on the patella in the old and new lateral images
2	CR/DX	Thorax	Follow-up pleural fluid, infiltrates	• Retrieve the new and old images – Is there an increase in pleural fluid? ○ Determine the cardio-thoracic ratio in the old and new images – Place the relevant older study next to the other studies. Maintain the temporal order of the studies
3	CT	Thorax	Hypoxemia with syncope, unknown cause. Indications of a pulmonary embolism?	• Retrieve the new images ○ Determine the CT density – Adjust the window level based on the CT density - Does this patient have a pulmonary embolism?
4	CT	Brain	Planning of intracystic treatment with interferon alfa	• Retrieve the new and old images – Align the old and new images with the 3D viewer ○ Measure the lateral ventricle size in the old and new images
5	MRI	Brain	Lesion around aqueduct L > R, DD inflammation or malignancy. Patient received high-dose steroid therapy. Follow-up lesion; nature?	• Retrieve the new and old images, do not retrieve the localizers – Determine the nature of the lesion
6	MRI	Brain	Tuberous sclerosis, frontal lobe epilepsy. Last MRI is from 2010. Follow-up tuberous sclerosis: increase in tumor size?	• Retrieve the new and old images, do not retrieve the localizer and position marker – Has the tumor increased in size? ○ Measure the tumor – Make a screen capture of the tumor

The clinical information was copied from the diagnosis request letter of the most recent study in the scenario. Participants' performance was measured on the image retrieval tasks (indicated by a solid circle) and the measurement tasks (indicated by an empty circle)

(Table 2). The main focus of our study was on the usability of the aspects of the PACSs relevant to these two functionalities (retrieving images and performing measurements). This was done in order to be able to evaluate the benefits of a usability test relative to a functional requirement list. If there would be no differences in usability of the measurement tools, and only a difference between the two PACSs that could automatically retrieve previous studies and those that could not, the usability test would not add any discriminating power to the requirement list.

Participants received a sheet of paper containing the instructions and clinical information for each scenario. The tasks shown in the right column of Table 1 were explicitly stated on the instruction sheet. Participants were instructed to perform the tasks as well and fast as possible. In all scenarios, participants had to retrieve the most recent study and, with the exception of Scenario 3, a relevant previous study of the scenario's patient. The date of the previous study was included in the instructions to ensure that it was perfectly clear to the participants which study they were supposed to retrieve. The studies were presented in the PACS's worklist.

Because this study is not concerned with diagnosis quality, participants did not have to make an official diagnostic report. Instead, for the diagnosis participants simply had to answer a yes or no question (in Scenarios 1–3 and 6), or a question to shortly describe the patient's lesion (in Scenario 5). Scenario 4 did not contain a diagnostic question. Participants wrote the answers to the diagnostic questions on an answer sheet. When

Table 2 – Functional requirement list.

Functionality	PACS A	PACS B	PACS C	PACS D
Automatic retrieval of previous studies	No	Yes	No	Yes
Measurement tools	Yes	Yes	Yes	Yes

The degree to which the different PACSs met the functional requirements we defined. The focus of the usability test was on the aspects of the PACS related to these requirements.

participants had to make a measurement, they also wrote the value they measured on the answer sheet.

Each scenario was divided into several tasks. Some tasks were strictly defined and aimed to probe a certain aspect of the PACS. The other tasks, regarding the diagnosis, were loosely defined and gave participants the opportunity to interact with the PACS in a natural and unconstrained way. Performance was measured on the tasks involving image retrieval and measurements.

We used a within-subject design with PACS as within-subject factor. The experiment was divided into four sessions, one for each PACS. The sessions took place over the course of two months. It was not possible to counterbalance the order of the PACSs due to their limited availability for this study. Participants' first session was with PACS A, the second with PACS B, the third with PACS D and the fourth with PACS C.

Prior to each session, participants attended a presentation in which a representative of the PACS vendor demonstrated the workings of the PACS interface and explained how to perform the actions required to complete all tasks in the scenarios. Participants did not interact with the PACS during this presentation.

2.5. *Data analysis*

We measured the execution times and number of button presses (i.e. mouse button clicks and key presses) of the tasks involving image retrieval and measurements, the answers to the diagnostic and measurement questions, and the answers to the questions on the questionnaire.

We examined the webcam videos of participants to determine the start and end times of the tasks. The start time was defined as the moment the participant initiated the execution of the task. The end time was defined as the moment the last action of the task had been performed and, in the image retrieval tasks, the correct images were displayed on the large screen. To determine the number of button presses for each task, we wrote a script that counted the number of mouse clicks and key presses in the log file between a certain start and end time.

To allow for quantitative analysis of the questionnaire data, the levels of the Likert scale were converted to numerical values as follows: strongly disagree = 0, disagree = 1, neutral = 2, agree = 3 and strongly agree = 4.

One-way repeated measures ANOVAs with PACS as within-subject factor were used to test for differences in image retrieval performance, measurement performance, and subjective responses between the PACSs. Paired *t*-tests were used to assess between which PACSs the differences occurred.

3. Results

Fig. 1 shows the total task execution time and number of button presses (i.e. mouse clicks and key presses) of the image retrieval tasks per participant for all PACSs. Visual inspection of the data indicates that all participants performed the image retrieval tasks fastest on PACS B. A repeated measures ANOVA of task execution time with PACS as within-subject factor showed that there was a significant difference between the PACSs ($F(3,9) = 12.997$, $p = .001$, $\eta^2 = .812$). Paired *t*-tests showed that this difference was between PACS A and B ($t(3) = 3.671$, $p = .035$), PACS B and C ($t(3) = -4.569$, $p = .020$), and PACS C and D ($t(3) = 7.830$, $p = .004$). There were no other significant differences between the PACSs.

A repeated measures ANOVA of number of button presses with PACS as within-subject factor also showed a significant difference between the PACSs ($F(3,9) = 8.647$, $p = .005$, $\eta^2 = .742$). Paired *t*-tests showed that this difference was between PACS A and B ($t(3) = 5.263$, $p = .013$), PACS B and C ($t(3) = -3.400$, $p = .042$), and PACS C and D ($t(3) = 0.015$, $p = .015$). There were no other significant differences between the PACSs.

Fig. 2 shows the total task execution time and number of button presses of the measurement tasks per participant for all PACSs. A repeated measures ANOVA of task execution time with PACS as within-subject factor showed that there was a significant difference between the PACSs ($F(3,9) = 4.003$,

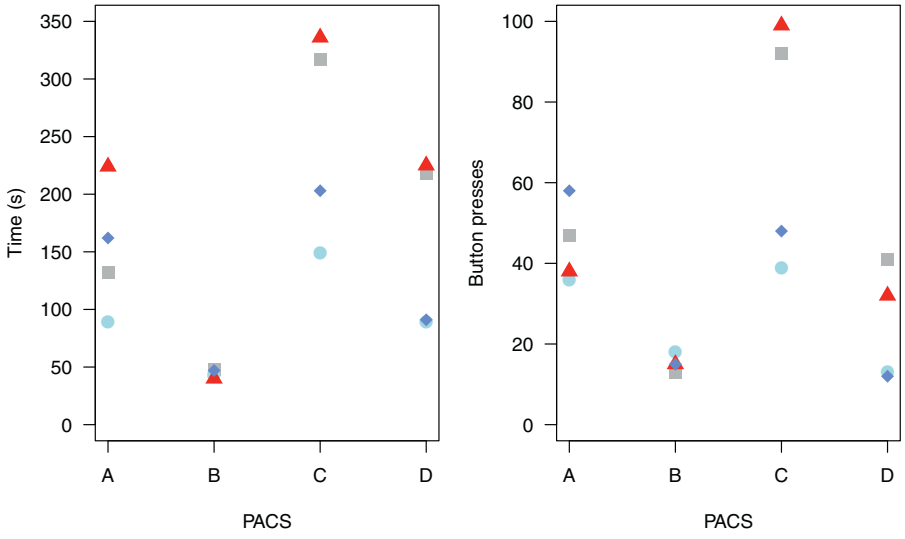


Fig. 1 – Total task execution time (left) and number of button presses (right) of the image retrieval tasks per participant for all PACSs. Each plotting character represents one participant.

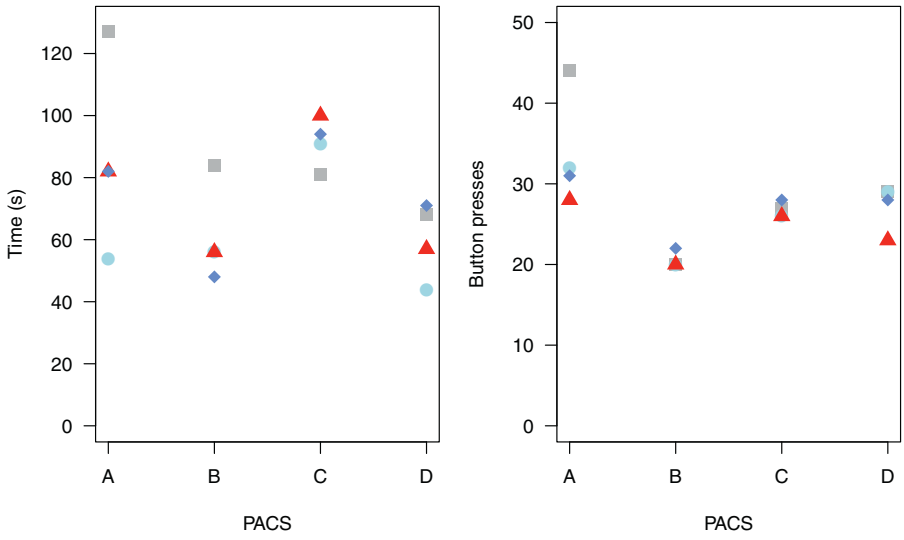


Fig. 2 – Total task execution time (left) and number of button presses (right) of the measurement tasks per participant for all PACSs. Each plotting character represents one participant.

$p = .046$, $\eta^2 = .572$). Paired t-tests showed that this difference was between PACS C and D ($t(3) = 3.890$, $p = .030$). There were no other significant differences between the PACSs, although visual inspection of the data shows that all participants performed faster on PACS D compared to PACS A, and all participants except one performed faster on PACS B compared to PACS A and C.

All participants needed the fewest button presses to complete the measurement tasks on PACS B. A repeated measures ANOVA of number of button presses with PACS and task as within-subject factors showed that there was a significant difference between the PACSs ($F(3,9) = 9.373$, $p = .004$, $\eta^2 = .758$). Paired t-tests showed that this difference was between PACS A and B ($t(3) = 3.598$, $p = .037$), PACS B and C ($t(3) = -25.000$, $p < .001$), and PACS B and D ($t(3) = -4.700$, $p = .018$). There were no other significant differences between the PACSs.

Fig. 3 shows the mean subjective response per participant for all PACSs. All participants except one rated PACS B higher than the other PACSs. A repeated measures ANOVA of mean subjective response with PACS as within-subject factor showed no significant difference between the PACSs ($F(3,9) = 2.033$, $p = .180$, $\eta^2 = .404$).

Fig. 4 shows the results of the questionnaire items regarding image retrieval (Appendix B, Item 18) and measurements (Appendix B, Item 16). There was a large between-participant variability in the image retrieval responses. Each participant rated PACS C differently, and two participants gave PACS D a strongly negative rating, while the other two rated it positively. All but one participant rated PACS B positively. A repeated measures ANOVA of subjective response regarding image retrieval with PACS as within-subject factor showed that there was no significant difference between the PACSs ($F(3,9) = 1.548$, $p = .268$, $\eta^2 = .340$).

There was a large discrepancy between participants' preference and their performance on the image retrieval tasks. Two participants (light blue circle and red triangle) rated PACS C higher than PACS A, while their performance was better with PACS A. One participant (red triangle) rated PACS C and D higher than PACS B, while his performance was far better with PACS B. One participant (gray square) rated PACS C higher than D, while his performance was better with PACS D. One participant (blue diamond) gave PACS A, C and D the same rating, while he performed much better on PACS D. One participant (light blue circle) gave PACS B, C and D the same rating, while he performed better on PACS B and D than on C.

PACS B and C did not receive any negative ratings for their measurement tools, while the other PACSs did. PACS B was the only PACS to receive strongly positive ratings. A repeated measures ANOVA of subjective response regarding measurements with PACS as within-subject factor showed

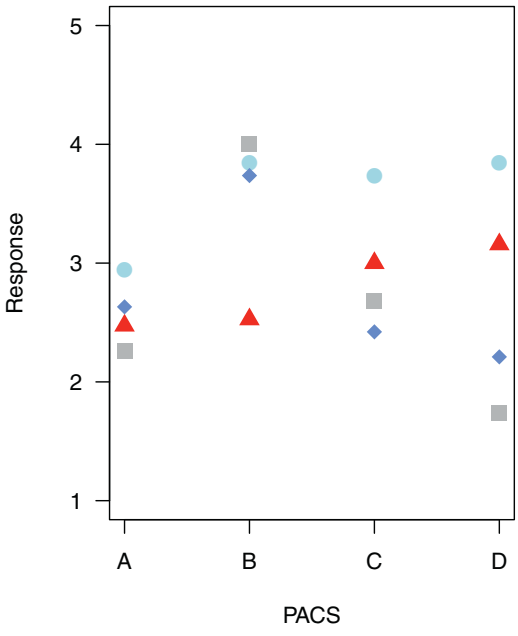


Fig. 3 – Mean subjective response per participant for all PACSs. Each plotting character represents one participant.

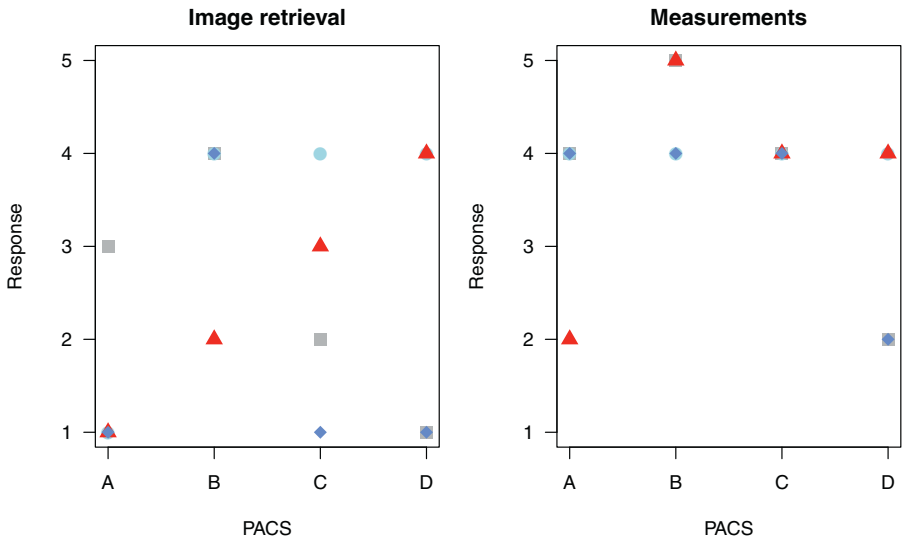


Fig. 4 – Subjective responses regarding image retrieval and measurements. Each plotting character represents one participant.

no significant difference between the PACSs ($F(3,9) = 2.000$, $p = .185$, $\eta^2 = .400$).

There was also a discrepancy between participants' preference and their performance on the measurement tasks. One participant (gray square) rated PACS D negatively and the other PACSs positively, while his performance was by far worst on PACS A. One participant (blue diamond) gave PACS A, B and C the same rating, while he performed best on PACS B. One participant (red triangle) rated PACS C and D the same, while he performed better on PACS D.

4. Discussion

In this study, we used a task-based usability test to compare the usability of four PACS workstations. The mean subjective responses showed that all participants except one preferred PACS B to the other PACSs. The performance data showed that participants performed the image retrieval tasks and the measurement tasks more efficiently on PACS B than on the other PACSs. These results indicate that PACS B had better usability than the other PACSs.

The subjective preference for PACS B can be attributed to several factors. Firstly, the information on the screen (e.g. images, patient information, and function icons) was presented in a very organized way, producing a calm and uncluttered working environment. Participants liked the PACS's menu structure, which was formed by a toolbar consisting of multiple tabs, each with a different set of functionalities (similar to the *Ribbon* in Microsoft Office 2007 and later). They also indicated that the functionalities were easy to use and that the interface was very responsive: providing quick feedback to their actions. The PACS also provided effective display protocols and allowed participants to easily compare multiple studies.

Prior to the test, we defined two functional requirements: automatic retrieval of images from previous studies (which PACS B and D provided and the other PACSs did not) and measurement tools (which all PACSs provided). The performance results showed that participants performed the image retrieval tasks fastest on PACS B. The difference between this PACS and PACS A and C would be expected based on the fact that it had useful image retrieval functionality that PACS A and C did not have. The difference between PACS B and D, however, could not have been predicted by functionality requirements and can only be explained by a difference in usability between these PACSs. This difference was due to the fact that in PACS D, participants had to select the patient, wait for a pop-up window with the diagnosis request letter to appear, and then click on the 'view' button. In PACS B, participants could retrieve all relevant images by simply double clicking on the patient, which is more efficient. Participants even performed as fast in PACS A as in D, even though they had to manually retrieve the previous study in this PACS.

The button presses data showed a similar pattern as the task execution time data, although there was no clear difference between PACS B and D. The subjective responses regarding image retrieval did not show any significant differences, although PACS B was rated positively most often.

There were also differences in performance between the PACSs on the measurement tasks. Participants performed faster on PACS B and D than on PACS A and C. The slower performance on PACS A was caused by the fact that participants had to activate viewports (by clicking in them) in order to measure in them, and the fact that the measurement tools were deactivated after each use so that when participants had to make consecutive measurements, they had to reselect the measurement tool each time. Also, the tool to measure the CT density was placed at an illogical location in the menu (it was not in the 'measurement' submenu), which made it more difficult to find.

In PACS C, viewports also needed to be activated before they could be measured in. Furthermore, multiplanar reconstructions (MPRs) were generated in separate tabs so that when

participants had to measure in MPRs of different series, they had to switch between the tabs. Selected tools in the toolbar were not identical between tabs, so participants had to reselect the measurement tool when switching to another tab.

In PACS B and D, viewports were automatically activated when they were measured in. As in PACS A, the measurement tools were deactivated after each use in PACS D, but this did not cause participants to perform slower on this PACS than on PACS B.

Because PACS B required fewer actions to perform measurements than the other PACSs, participants needed fewer button presses in the measurement tasks on this PACS.

There were no statistically significant differences between the PACSs in subjective ratings of the measurement tools. However, only PACS B and C were unanimously given a positive rating.

In line with [12,13], we found a discrepancy between the subjective responses and the performance data. Participants' subjective ratings regarding image retrieval and measurements did not accurately reflect their performance on these aspects of the PACS. As Bailey [12] pointed out, users tend to integrate their preferences into their judgments of the efficiency of a system. This tendency was also observed in our study. The most extreme example of this is that one participant expressed a strong dislike for PACS B's image retrieval capabilities because an error occurred during one image retrieval task (the error was caused by the participant accidentally opening a second instantiation of the PACS). This caused him to rate the efficiency of this PACS's image retrieval capabilities negatively, while he actually performed the image retrieval tasks much faster on this PACS than on the other PACSs.

The discrepancy between subjective responses and performance data we found demonstrates the usefulness of obtaining performance data in a usability test, especially in a domain such as radiology, in which efficiency is very important. Collecting both subjective responses and objective performance data produces a more accurate, reliable, and complete assessment of a PACS's usability than could be achieved by measuring either of these data alone.

The validity of the results of a task-based usability test greatly depends on the quality of the scenario set and the tasks within these scenarios being used. A poorly constructed scenario set will lead to results that have a low generalizability to the real radiology work environment. It is therefore vital that the scenarios are developed in consultation with a group of radiologists to ensure that they are representative of interaction with a PACS in a natural environment and that the tasks within the scenarios represent common interactions with the PACS. Each scenario should contain a balance between low-level strictly defined tasks, on which performance can be measured, and more loosely defined tasks (such as diagnosis), which give participants the opportunity to interact with the PACS in a natural and unconstrained way. In this study we only measured performance on two types of tasks. A real-life PACS usability test should obviously measure performance on a wider variety of tasks.

As long as the scenario set as a whole is a good representation of the radiologist's daily work, variations in the specific scenarios that are used will not have a large effect on the results of the usability test. For example, for participants' ability to evaluate a PACS's usability it does not matter whether a CT thorax or a CT abdomen is used in the scenario set, because the tasks within these scenarios are similar. Variations in the scenarios also have minimal effect on the performance data. When performing a length measurement for example, it does not matter whether participants measure a lung nodule on a CT or a brain tumor on an MRI.

The data obtained from the usability test should be transformed in such a way that they can be easily integrated into the PACS selection process. Common practice in tender procedures for new products is to assign a score to each vendor for each item on the requirement list, after which a weighted average score is calculated. The different usability measures could be combined into a single usability score that can serve as an additional requirement to the functional requirement list. The weight of the usability requirement relative to other requirements can be determined by the hospital. An excellent way of standardizing usability measures and combining them into a single score is described in detail in [10]. Another (somewhat less elegant) possibility is to define each different usability measure as a separate requirement.

Besides the value of usability from a user's perspective, it is also interesting from a business perspective. To illustrate: the difference in mean execution time of the image retrieval tasks between PACS B and D was 111.25 s. These data consisted of six patients, so the difference was 18.54 s per patient. Our hospital's radiology department diagnoses approximately 200,000 patients per year. This means that having PACS B instead of D would save 1030 h per year, which is equivalent to 0.5 FTE (based on a 40-h work week), and this is only for one type of task.

A limitation of this study was that only four participants were used. The results of our statistical tests should therefore be interpreted cautiously. The data of the individual participants did show clear patterns and by presenting these data we provided the reader with a way to assess the impact of the different PACSs on each individual participant. In a real-life PACS selection process, a larger participant group should be used for the usability test in order to obtain a more reliable and more statistically powerful comparison between the PACSs.

Another limitation might be that the order of the PACSs was not counterbalanced, meaning that each participant used the PACSs in the same order, which could have led to order effects in the data. Counterbalancing the PACSs was not possible, because each PACS was only available for this study for a very short period of time. It is quite likely that order effects occurred for the tasks involving diagnosis (e.g. participants finding a lesion faster because they remembered its location from the session with a previous PACS). However, these tasks were designed to allow natural interaction with the PACS and were not used to produce performance data. Possible order effects in these tasks are therefore not harmful to the validity of the results. We believe that the order effects for the tasks on which performance was measured were negligible, because learning to perform such low-level tasks on one interface does not necessarily generalize to a different interface. Positive effects

might occur when the interfaces are very similar, but negative effects might also occur if they are very dissimilar (e.g. searching for a button in the wrong location based on expectations from a previous interface). Our results did not show any trends indicating either positive or negative order effects.

Participants had previous experience with PACS, but not with the specific PACSs used in this study. Our results might therefore have been biased by the intuitiveness of the PACSs' interface. Although ease of learning is also an important usability aspect, performance results should ideally reflect skilled performance. Experience is always a confounding factor in comparative usability tests, but it is impossible to control for. It is simply not feasible to extensively train each radiologist on each PACS. However, because we measured performance on very low-level tasks, which required minimal experience to perform efficiently, and participants received a thorough explanation of how to perform the tasks prior to the test, we believe that the confounding effects of experience were kept to a minimum.

We only measured the satisfaction and efficiency aspects of usability and ignored effectiveness. Effectiveness is often measured using task completion rates or the number of errors participants make during task execution relative to the number of opportunities for error in the task. We assumed that participants would be able to complete all tasks on all PACSs, which would make task completion rates irrelevant. We did not measure errors because identifying errors and error opportunities is time consuming and arguably subjective (there can be disagreement on what constitutes an error or error opportunity) [10]. However, including a measure of effectiveness, especially when more complex tasks are used in the test, might improve the accuracy of the usability assessment.

The quality of interaction between radiologists and the PACS workstation is affected by both the PACS software and the workstation's hardware. The PACS workstations evaluated in this study used equivalent hardware, meaning that our results only reflected differences in software, but our methodology is also appropriate for PACS workstations using different hardware. However, great care should be taken when deciding how much variability in hardware to allow in the usability test. Trivial differences in hardware (e.g. using one monitor instead of two, or a trackball instead of a mouse) should not be allowed to influence the results. However, when a PACS vendor uses a piece of hardware in a way that fundamentally changes the interaction between radiologists and the PACS, and this could not have been achieved by using the same hardware in combination with different PACS software, this piece of hardware should be allowed in the test.

We advise hospitals interested in integrating a usability test into their PACS selection process to use the following guidelines: (1) interview a group of radiologists in order to construct a set of scenarios that is representative of their daily work; (2) divide each scenario into several tasks; ensure a good balance between strictly defined tasks, aimed at evaluating performance on a certain aspect of the PACS, and more loosely defined tasks, aimed at allowing natural interaction with the PACS; (3) design a questionnaire aimed at evaluating radiologists' opinions of the PACS; (4) evaluate the vendors' responses to the functional requirement list defined by the relevant stakeholders within the hospital and invite the vendors with a sufficient score on these requirements to participate in the usability test; (5) set up a controlled testing environment with the workstations of the selected vendors; (6) combine all usability measures into a single score that can be easily integrated into the selection process.

5. Conclusion

In the traditional PACS selection process, PACSs are compared based on the functionality they possess, but not based on *how* this functionality is implemented, which determines the PACS's usability. The differences in usability we found between PACSs with identical functionality indicate that functional requirements alone are insufficient to determine a PACS's overall quality. We therefore recommend using a usability test in addition to a functional requirement list in a PACS selection process to ensure that a hospital buys the PACS with the highest quality. A task-based usability evaluation methodology, which yields both subjective preference data and objective performance data of radiologists interacting with the PACS, is very suitable for such a usability test.

Authors' contributions

W. Jorritsma contributed to the design of the study, acquired, analyzed and interpreted the data, wrote the first draft of the manuscript, and contributed to the final version of the manuscript. F. Cnossen contributed to the design of the study, data interpretation, and the final version of the manuscript. P.M.A. van Ooijen conceived the project, supervised the data acquisition and contributed to the design of the study, data interpretation, and the final version of the manuscript.

Conflicts of interest

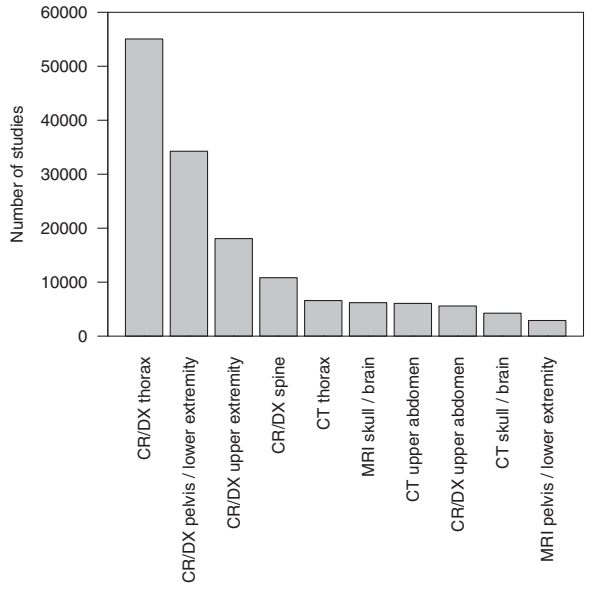
The authors declare no conflicts of interest.

Acknowledgments

We would like to thank the radiologists who helped define the prototypical tasks. We also thank Dr. L.M. Kingma and P.J.M. ten Bhömer for supporting and providing resources for this project.

Appendix A.

The ten most frequently reviewed studies by radiologists and radiology residents in our hospital in the period December 2011–November 2012. For each modality, the studies were divided into the following body part categories: skull/brain, face/neck, spine, upper extremity, heart/aorta, thorax, upper abdomen, lower abdomen/genitals, and pelvis/lower extremity. Ultrasounds were excluded from these data, because they are reviewed directly after they are made. A review of an ultrasound on a PACS workstation is therefore not relevant.



Appendix B.

The questionnaire used in the test.

Indicate the extent to which you agree with the following statements.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
1. This PACS is user-friendly.					
2. I can work efficiently with this PACS.					
3. The images are correctly displayed on the screen automatically.					
4. It is clearly indicated which images belong to the old, and which to the new study.					
5. The screen space is used efficiently.					
6. It is clear what all the buttons are for.					
7. The PACS responds predictably to my actions.					
8. The 3D mode is user-friendly.					
9. With the 3D mode, I can execute simple 3D tasks well.					
10. I can easily find the functions I need.					
11. I can easily compare series.					
12. I can easily place images in the viewports.					
13. I can easily adjust the window level.					
14. I can easily scroll through series.					
15. I can easily zoom in and out.					
16. I can easily perform measurements.					
17. I can easily make screen captures.					
18. From the worklist, I can quickly and easily start with my diagnosis.					
19. I would like to work with this PACS/I would advise to purchase this PACS.					
20. Other positive points of this PACS.					
21. Other negative points of this PACS.					

Summary points

What was already known on the topic:

- Traditionally, hospitals buy a PACS based on functional requirements and largely ignore usability.
- Usability determines the quality of interaction between a user and a system.
- Previous PACS usability evaluations only measured radiologists' subjective responses to the PACS.
- The efficiency with which radiologists can work with the PACS is essential.

What this study has added to our knowledge:

- A usability test provides valuable additional information about the quality of a PACS with respect to a functional requirement list.
- In addition to subjective responses, a PACS usability test should measure performance so that the results of the test adequately reflect the efficiency with which radiologists can work with the PACS.

REFERENCES

[1] G. Cockton, D. Lavery, A. Woolrych, Inspection-based evaluations, in: J.A. Jacko, A. Sears (Eds.), The Human–Computer Interaction Handbook, Lawrence Erlbaum, Mahwah, NJ, 2003, pp. 1119–1138.

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

[3] M.W.M. Jaspers, A comparison of usability methods for testing interactive health technologies: methodological aspects and empirical evidence, Int. J. Med. Inf. 78 (2009) 340–353.

[4] F. Bakhshi-Raiez, N.F. de Keizer, R. Cornet, M. Dorrepaal, D. Dongelmans, M.W.M. Jaspers, A usability evaluation of a SNOMED CT based compositional interface terminology for intensive care, Int. J. Med. Inf. 81 (2012) 351–362.

[5] A. Menon, N. Korner-Bitensky, M. Chignell, S. Straus, Usability testing of two e-learning resources: methods to maximize potential for clinician use, J. Rehabil. Med. 44 (2012) 338–345.

[6] J. Chan, K.G. Shojania, A.C. Easty, E.E. Etchells, Usability evaluation of order sets in a computerised provider order entry system, BMJ Qual. Saf. 20 (2011) 932–940.

[7] N. Bazak, G. Stamm, F. Caldarone, J. Lotz, A. Leppert, M. Galanski, PACS workstations 2000: evaluation, usability and performance, in: 18th International EuroPACS Conference, Graz, 2000, pp. 133–142.

[8] T. Boehm, O. Handgraetinger, J. Link, R. Ploner, D.R. Voellmy, B. Marincek, et al., Evaluation of radiological workstations and web-browser-based image distribution clients for a PACS project in hands-on workshops, Eur. Radiol. 14 (2004) 908–914.

[9] J. Nielsen, J. Levy, Measuring usability: preference vs. performance, Commun. ACM 37 (1994) 66–76.

[10] J. Sauro, E. Kindlund, A method to standardize usability metrics into a single score, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM Press, Portland, OR, 2005, pp. 401–409.

[11] K. Hornbæk, E.L. Law, Meta-analysis of correlations among usability measures, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM Press, San Jose, CA, 2007, pp. 617–626.

[12] R.W. Bailey, Performance vs. preference, in: Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting, Human Factors and Ergonomics Society, Seattle, WA, 1993, pp. 282–286.

[13] G.V. Kissel, The effect of computer experience on subjective and objective software usability measures, in: Proceeding of SIGCHI Conference on Human Factors in Computing Systems, ACM Press, Denver, CO, 1995, pp. 284–285.

[14] M. Bhargavan, A.H. Kaye, H.P. Forman, J.H. Sunshine, Workload of radiologists in United States in 2006–2007 and trends since 1991–19921, Radiology 252 (2009).

[15] Y. Lu, S. Zhao, P.W. Chu, R.L. Arenson, An update survey of academic radiologists' clinical productivity, J. Am. Coll. Radiol. 5 (2008) 817–826.

[16] U. Kukreja, W.E. Stevenson, F.E. Ritter, RUI: recording user input from interfaces under Windows and Mac OS X, Behav. Res. Methods Instrum. Comput. 38 (2006) 656–659.