Evaluating the Usability and Quality of a Clinical Mobile Application for Assisting Physicians in Head CT Scan Ordering: A Think Aloud and Mobile Apps Rating Scale (MARS) Study

Shiva Meidani, Aydine Omidvar, Hossein Akbari, Fatemeh Asghari, Reza Khajouei, Zahra Nazemi, Ehsan Nabovati, Felix Holl

Submitted to: JMIR Human Factors
on: January 01, 2024

Table of Contents

Evaluating the Usability and Quality of a Clinical Mobile Application for Assisting Physicians in Head CT Scan Ordering: A Think Aloud and Mobile Apps Rating Scale (MARS) Study

Shiva Meidani[1] PhD; Aydine Omidvar[1] MD; Hossein Akbari[1] PhD; Fatemeh Asghari[1] MSc; Reza Khajouei[2]; Zahra Nazemi[1] MSc; Ehsan Nabovati[1] PhD; Felix Holl[3] PhD

[1]Kashan University of Medical Sciences Kashan IR
[2]Kerman University of Medical Sciences Kerman IR
[3]Neu-Ulm University of Applied Sciences, Neu-Ulm DE

Corresponding Author:
Ehsan Nabovati PhD
Kashan University of Medical Sciences
5th of Qotb –e Ravandi Blvd. Kashan, IRAN
Kashan
IR

Abstract

Background: Among the numerous factors contributing to healthcare providers' (HCPs) engagement with mobile apps (apps), including user characteristics (dexterity, anatomy, and attitude) and mobile features (screen and button size), usability and quality of apps were introduced as the most influential factors.

Objective: Therefore, this study aims to investigate the usability and quality of the Head CT Scan Appropriateness Criteria mobile application (HAC app) for physicians' CT scan ordering.

Methods: Our study design was primarily based on methodological triangulation by utilizing mixed methods research involving quantitative analysis of the Mobile Apps Rating Scale (MARS) for quality assessment, quantitative and qualitative Think Aloud (TA) usability testing, and debriefing across three phases. Sixteen medical interns participated in quality assessment and testing usability characteristics, including efficiency, effectiveness, learnability, error, and satisfaction with the HAC app.

Results: The effectiveness of the HAC app was deemed satisfactory, with a rating of 96.9%. MARS assessment scale also indicated the overall favorable (82 out of 100) quality score of the HAC app. Scoring four MARS subscales, "Information" (73.37 out of 100) and "Engagement" (73.48 out of 100) had the lowest scores. Analysis of the items in each MARS subscale revealed that in the "Engagement" subscale, the lowest score of the HAC app was "customization" (63.6 out of 100). In the "Functionality" subscale, the HAC app's lowest value was "Performance" (67.4 out of 100). Qualitative TA usability testing of the HAC app found eight notable usability issues also highlighted in the MARS quality assessment.

Conclusions: Evaluating the quality and usability of mobile apps using a mixed methods approach provides valuable information about the functionality and disadvantages of the mobile apps and complements each other. It is highly recommended to embrace a more holistic and mixed methods strategy when evaluating mobile apps, as relying solely on one method proves imperfect to reflect trustworthy and reliable information regarding the usability and quality of apps.

(JMIR Preprints 01/01/2024:55790)
DOI: https://doi.org/10.2196/preprints.55790

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
Please make my preprint PDF available to anyone at any time (recommended).
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
✓ **No, I do not wish to publish my submitted manuscript as a preprint.**

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Evaluating the Usability and Quality of a Clinical Mobile Application for Assisting Physicians in Head CT Scan Ordering: A Think Aloud and Mobile Apps Rating Scale (MARS) Study

Zahra Meidani[a,b], Aydin Omidvar[c], Hossein Akbari [d], Fatemeh Asghari[a], Reza Khajoueie[e], Zahra Nazemi[a], Ehsan Nabovati[a, b, 1], Felix Holl[f]

[a] *Health Information Management Research Center, Kashan University of Medical Sciences (KAUMS), Kashan, Iran.*
[b] *Department of health information technology and management, Kashan University of Medical Sciences (KAUMS), Kashan, Iran.*
[c] *Department of Neurosurgery, Kashan University of Medical Sciences, Kashan, Iran.*
[d] *Department of Epidemiology & Biostatistics, Kashan University of Medical Sciences, Kashan, Iran.* [e]*Department of Health Information Sciences, Faculty of Management and Medical Information Sciences, Kerman University of Medical Sciences, Kerman, Iran.*
[f] *DigiHealth Institute, Neu-Ulm University of Applied Sciences, Neu-Ulm, Germany*

Abstract

Background: Among the numerous factors contributing to healthcare providers' (HCPs) engagement with mobile apps (apps), including user characteristics (e.g. dexterity, anatomy, and attitude) and mobile features (e.g. screen and button size), usability and quality of apps were introduced as the most influential factors. Therefore, this study aims to investigate the usability and quality of a Head CT Scan Appropriateness Criteria mobile application (HAC app) for physicians' CT scan ordering.

Methods: Our study design was primarily based on methodological triangulation by utilizing mixed methods research involving quantitative and qualitative Think Aloud (TA) usability testing, quantitative analysis of the Mobile Apps Rating Scale (MARS) for quality assessment, and debriefing across three phases. Sixteen medical interns participated in quality assessment and testing usability characteristics, including efficiency, effectiveness, learnability, error, and satisfaction with the HAC app.

Results: The efficiency and effectiveness of the HAC app were deemed satisfactory, with rating of 97.8% and 96.9%, respectively. MARS assessment scale indicated the overall favorable quality score of the HAC app (82 out of 100). Scoring four MARS subscales, "Information" (73.37 out of 100) and "Engagement" (73.48 out of 100) had the lowest scores; while aesthetics had the highest score (87.86 out of 100). Analysis of the items in each MARS subscale revealed that in the "Engagement" subscale, the lowest score of the HAC app was "customization" (63.6 out of 100). In the "Functionality" subscale, the HAC app's lowest value was "Performance" (67.4 out of 100). Qualitative TA usability testing of the HAC app found notable usability issues grouped into eight main categories: lack of finger-friendly touch targets, poor search capabilities, input problems, inefficient data presentation and information control, unclear control and confirmation, lack of predictive capabilities, poor Assistance and support, and unclear navigation logic.

Conclusion: Evaluating the quality and usability of mobile apps using a mixed methods approach provides valuable information about the functionality and disadvantages of the mobile apps and complements each other. It is highly recommended to embrace a more holistic and mixed methods strategy when evaluating mobile apps, as relying solely on one method proves imperfect to reflect trustworthy and reliable information regarding the usability and quality of apps.

Keywords: Mobile Applications, User-Centered Design, User-Computer Interface, Physicians, Tomography, X-Ray Computed

Introduction

[1] Corresponding Author, Ehsan Nabovati. Health Information Management Research Center, Kashan University of Medical Sciences, Kashan, Iran. Email: Nabovati@kaums.ac.ir

Mobile devices and mobile health (mHealth) applications (apps) have equipped the healthcare system with a strategy to improve health through enhanced self-management among patients and access to educational materials for healthcare professionals [1]. Considering their advantages regarding the fastest and most convenient ways to access health care services, they have been introduced as effective e-health technology to address health priorities [2]. Recently, a global initiative has been launched to apply mobile technologies to provide healthcare services and manage various diseases [3]. A 2015 World Health Organization (WHO) survey revealed that 15,000 mobile apps are available for healthcare usage [4]. However, the continuity in the use of apps is highly challenging, and existing evidence presented poor user engagement and relatively high drop-out rates on apps among patients and healthcare providers (HCP) [5]. Earlier research revealed that nearly half of mHealth app users avoid continuous use of them [6]. The drop-out rates in app-based interventions for chronic diseases were reported at 43% (95% CI 29-57) in a meta-analysis by Meyerowitz-Katz et al [7].

Usability has been introduced as a surrogate marker for app quality and user engagement with them to address this challenge [8-10]. Given the significance, assessing the usability and quality of mobile apps occupies a crucial part of app development and users' overall assessment of app quality [6, 8]. However, emerging research has debated that mobile apps suffer from usability and quality issues and are limited by their ability to address users' needs [9, 11, 12]. Physicians use many mobile apps to access a wide range of knowledge and information in educational materials, drug reference guides, x-ray results, laboratory test information, and clinical guidelines [13]. Medical apps were positively perceived, with physicians reporting increased dependency on the apps. The use of apps in the medical setting has steadily grown in recent years [14]. While a considerable number of physicians now use mobile devices and apps for clinical practices globally (ref), there are also reports of drop-out rate and short-term engagement among physicians with these mobile apps [1]. Arguably, no clear understanding exists of the physicians' motivations and interests in adopting and long-term use of mobile apps [8]. A variety of factors, from organizational and social factors [15] to users' characteristics (e.g., user dexterity and anatomy, positive attitude) [16-19] or mobile features (e.g., screen and buttons size, poor resolution, usability) [6, 20-22] would influence on the successful adoption of mobile apps among physicians.

Hence, usability and quality issues have been reported as central to user engagement with mobile apps [6, 21]. The research team previously developed a mobile application aimed at assisting physicians in prescribing head CT scans based on appropriateness guidelines, Head CT scan Appropriateness Criteria mobile app (HAC app) [1]. However, during that study, neurology and

neurosurgery residents expressed concerns about usability issues despite their interest in utilizing the app. Therefore, before proceeding with full implementation, it is essential to identify and address usability problems of the app using mixed methods that involve participation from final users.
In the current study, we seek to investigate the usability of HAC app using mixed methods research involving quantitative analysis of the Mobile Apps Rating Scale (MARS) for quality assessment, quantitative and qualitative think-aload (TA) usability testing, and debriefing across three phases..

1- Methods

2-1- Study setting

This study was conducted as part of a broader effort to develop a mobile app called the Head CT Scan Appropriateness Criteria (HAC app) based on clinical guidelines. The development occurred at an academic hospital with Kashan University of Medical Sciences (KAUMS) in Iran, which has 510 beds. This newly developed HAC app allows end-users to search for appropriate CT scans based on diseases, signs, symptoms, and modalities, such as CT, CTA, and MRI. Appropriate CT scans refers to imaging studies that are deemed clinically justified and indicated based on established medical criteria, including patient symptoms, signs, and relevant clinical history, in accordance with evidence-based guidelines and best practices in diagnostic radiology.

The study involved 16 medical interns from an academic hospital at KAUMS. For this study, the focus was on assessing the end-user usability through a TA approach [23, 24], evaluating the quality of the HAC app using the Mobile Application Rating Scale (MARS), [25] and conducting informal debriefing sessions to gather insights and opinions for the medical interns regarding the HAC app.

2-2-The profile of HAC app: HAC app content and functionality

The HAC app was developed using applied four-tier architecture, including presentation, data service, business logic, and data access layers. The app was designed using JavaScript in such a way that allows it to be installed and compatible with the latest version of Android in 2021 (version 12) as well as earlier versions. . The HAC app encompasses essential criteria arranged by Care Core guidelines for head CT scans. Care Core provides a list of disease titles, for example, head trauma, which is supplemented by the list of clinical criteria in terms of signs and symptoms of the given disease. Clicking the plus sign (+) provides the detailed clinical criteria. Under each main heading or in front of each condition, the appropriate imaging procedure in MRI, CT, and computed tomography

angiography (CTA) is provided. A shortlist menu is designed to organize and quickly find frequently used diseases or clinical criteria. It enables users to add common diagnoses to the shortlist menu. Screenshots of the functionalities of the HAC app are presented in Figures 1 and2.

2-3- Approaches to conduct the study

Three approaches have been employed to conduct the research and achieve the study objectives.

2-3-1- The TA usability testing

In this phase, we tested the HAC app's effectiveness, efficiency, error, and learnability. The study objectives have been determined to ensure the accurate fulfillment of the tasks, the correct selection of the icons and buttons, and end users' use of the mobile app without errors in an efficient way. The TA approach set out to determine the following measures to achieve the objectives:

- The effectiveness of the participants' navigating app was measured by the accurateness and completeness of the HAC app on CT scan ordering based on diseases, signs, symptoms, and modalities, e.g., CT, CTA, and MRI.
- The efficiency of participants was specified by the number of touch targets on the app screen and the task completion time.
- The simplicity and learnability of the HAC app were measured by the number of tasks that were easily completed and the severity of errors made by the users.
- Errors were identified as the number of user mistakes when using the HAC app.

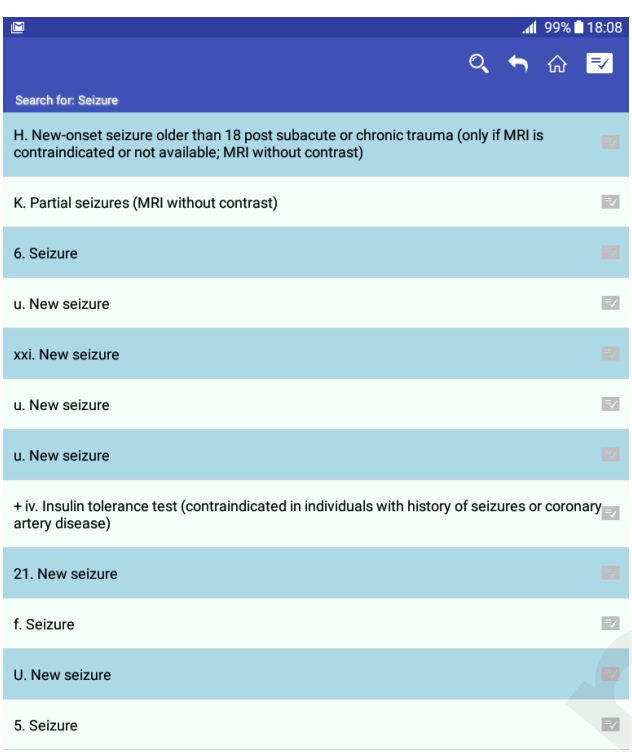
2-3-2- MARS quality assessment

To evaluate the HAC app quality in terms of engagement, functionality, aesthetics, information, and subjective quality, we applied the MARS tool [25], and the following dimensions were addressed:

- The overall quality score of the HAC app and its subscales, including engagement, functionality, aesthetics, information, and subjective quality.
- A statistically significant difference between MARS subscales quality score
- A statistically significant difference between two sets of pairs of MARS subscales (e.g., engagement and functionality or functionality and aesthetics)
- The correlation between MARS subscales for the HAC app
- The significant relationship between medical interns' characteristics (i.e., age, gender, Interest in using mobile apps for learning and clinical practice) with MARS subscales.

2-3-3- Debriefing

An informal debrief was conducted to review and digest interns' general ideas about using mobile apps and physicians' expectations for a suitable mobile app. It was also applied to collect underexplored facts for further revision of the HAC app.



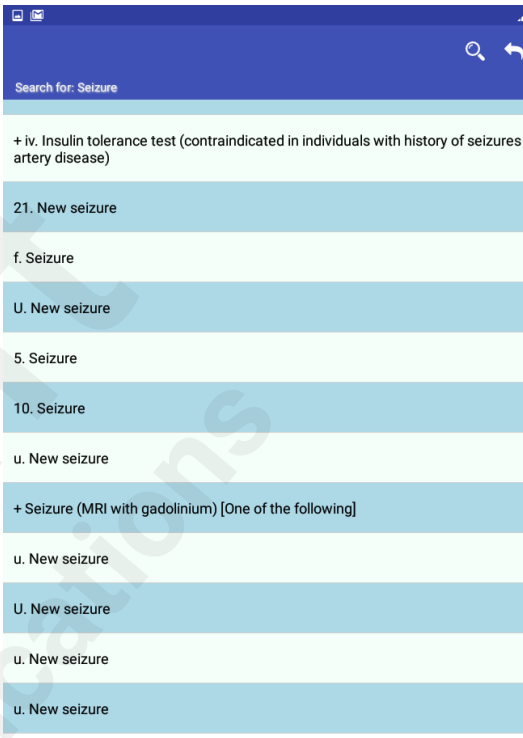


Figure 1: HAC app search results for seizure

2-4- Study design and data analysis

Our study design was primarily based on methodological triangulation through the utilization of mixed methods research, and investigator triangulation to enhance the understanding and interpreting the results [26].

The mixed methods study involved quantitative (MARS quality assessment, TA quantitative usability testing) and qualitative methods (TA qualitative usability testing, debriefing) across three phases. By employing the technique of investigator triangulation, a variety of researchers, such as medical practitioners, experts in health information technology, and professionals in health information management, involved in the gathering, and analyzing of the data. The details of each phase will be discussed below.

Phase 1: TA usability testing approach

2-4-1- 1- Design

The study employed a TA study design to explore the user's cognition, including feelings, thoughts, and whatever else comes to mind while interacting with a system to perform a task. This standard data collection method for assessing users' cognitive behavior during system interaction helps identify errors and necessary changes [23, 24].

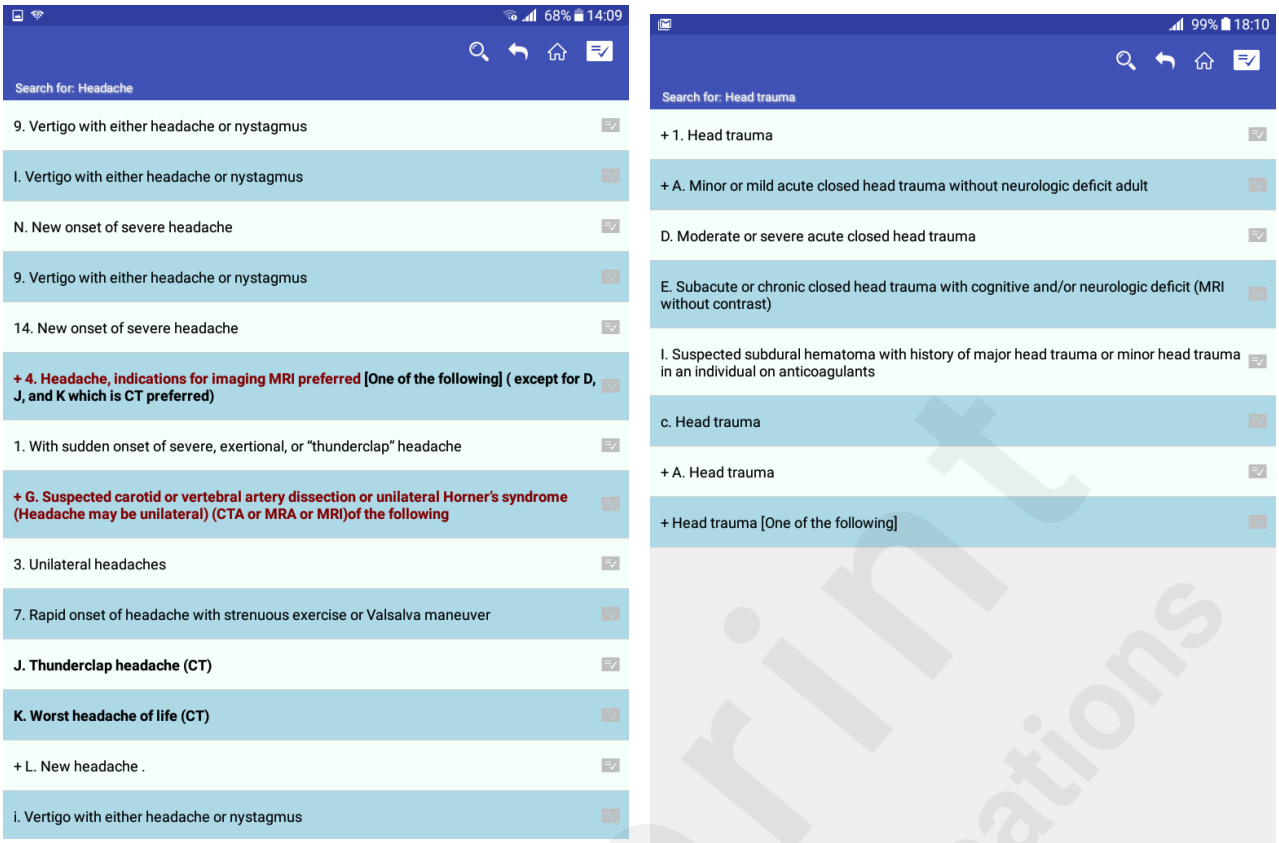


Figure 2: HAC app search results for head trauma and headache

There are two fundamental usability testing methods: qualitative and quantitative [27]. Qualitative methods primarily aim to explore users' interaction experiences with a product and describe possible issues they encounter [28]. In contrast, the quantitative methods employ various metrics, such as task times, completion rates, and errors, to measure and categorize the errors and problems users encounter during usability testing [29]. Both qualitative and quantitative methods were applied in the current study to reach the research objectives. Usability evaluation is also conducted in the different stages of a product development life cycle. Formative evaluation is done in the early product development life cycle to shape the design direction. Summative assessment is done toward the end of the product development (final product) to evaluate its performance against a set of metrics (e.g., time on task, success rate) [28]. The participants implemented summative usability testing in the current study to evaluate the performance of the HAC app.

2.4.1.2- **Participant recruitment**

Previous evidence presented that between 5 and 15 participants are sufficient to perform TA to enhance the expected level of problem discovery [30]. We recruited 16 medical interns who participated in three phases of the study. We applied social media to attract medical interns to join the study. We posted our research profile on the medical students' academic and social media

channels, including the study's title, research team, and overall study objectives. We invited those who finished their clinical internship in emergency medicine to participate in the current study. Our multidisciplinary research team, including clinicians, significantly streamlined the recruitment process. The research team, consisting of members from diverse disciplines, encouraged their previous students to participate in the research. No rewards or compensations were paid to the participants.

2.4.1.3- Protocol

The TA usability testing was conducted in multiple sessions of the same activity. Developing a study protocol to ensure the consistency of each activity in each session helps the facilitator give all the necessary information to the participants, which seems imperative [28]. The study protocol in the current study consists of session introduction, information capture methods (observation, videotaping), task scenarios, user interactions with the product and any identified product problems and difficulties, and measurement criteria to be discussed below.

2-4-1-4- Session Introduction

Interested volunteers were contacted to schedule face-to-face visits. TA sessions were held in the physicians' actual workplace. It is widely believed that evaluations conducted in the field resemble the setup that matches the user's real work context, providing “ecological validity” to the study and accurately reflecting the users' context [28].

Once the researcher arrived in the field, they gave the participants an overview of the session and the overall goals. They let them know about the presence of any facilitator or observers in the session and the rules for conducting the usability testing. We informed them that the Ethics Review Board at KAUMS [Code #IR.KAUMS.MEDNT.REC.1399.075] had approved the current study at the KAUMS and emphasized the voluntary nature of participation, assuring them of the confidentiality of information. The non-evaluative environment of the TA session is also explained by a trained moderator (researcher). Then, participants who attended the face-to-face meeting consented to participate in the study and TA session run.

2-4-1-5 Data Collection

The usability data collection protocol is generally implemented via two approaches: concurrent think-aloud (CTA) and retrospective think-aloud (RTA) protocols [31].

Since CTA is more objective and less dependent on users’ memory and prior experience of completed tasks compared with RTA, CTA was adopted as a standardized method to conduct usability testing of the HAC app [28, 32].

Considering that most users are uncomfortable installing something on their devices (mobiles, computers) [28] and the importance of using the same tool to capture usability-testing data, interaction with the HAC app was done via an Android mobile phone dedicated only to the research purposes. A portfolio of methods, including video screen-recording software, audio/screen recording, and notetaking, were applied to collect data. Four scenarios containing four to six tasks were given to the participant to interact with the HAC app. All the activities to accomplish the scenario, including the number of touch targets on the app screen, the task completion time, and the elapsed time, were captured using the free AZ Screen Recorder for Android [33].

Three evaluators facilitated the testing sessions and analyzed the results. The researchers adopted the verbal protocol to collect the data. Although the verbal protocol is the most traditional protocol with limited probing methods compared with active users' participation methods, such as communication-based and coaching protocols, it resembles an authentic context experience by not offering any external assistance to the users [34].

Thus, one researcher supervised the evaluation session, but neither user received instruction during the task performance stage. Attention was given to shortening the testing process and keeping a participant on the phone for less than 10 minutes [28]. Each TA session lasted 20-30 min.

2-4-1-6- Tasks scenario

The scenarios (including their goals and actions) were designed to examine different parts and functions of the HAC app and covered the most common tasks that a clinician may use in a typical working application. Usability problems are detected by researchers from analyses of user behavior and expressions during interactions with the system.

2-4-1-7- Measurement

A coding framework was developed according to five usability characteristics and based on the International Organization for Standardization (ISO) and Nielsen's definitions to recognize the specific user-computer interaction problems in detail to define the measurement criteria [35-37]. Nielsen put forward five usability attributes: learnability, efficiency, memorability, errors, and satisfaction [37].

Combining ISO and Nielsen usability attributes yields the following six criteria: efficiency, effectiveness, learnability, memorability, errors, and satisfaction. Since the participants only used the HAC app in this study, and there was no need to remember the options for the next session, we omitted memorability in our evaluation. The remaining five attributes are composed of our coding framework [23].

We used the TA method to measure effectiveness, learnability, errors, and efficiency characteristics,

and the MARS questionnaire was used to measure satisfaction.

Errors/ usability problems: They were detected based on the analysis of the “critical issues” encountered by the participants during the interactions detected from the video reviews. Critical issues’ were defined as those issues that prevented task completion, ‘‘severe issues’’ were defined as those issues that caused significant slowdown or frustration, and ‘‘cosmetic issues’’ were the ones that were left and caused minimal issues [38].

Learnability: It was evaluated by measuring the number of quickly completed tasks.

2-4-1- 8- Data Analysis

Phase I: Think aloud usability testing

2-4-1- 8- 1- TA quantitative part

Data analysis and measurements of usability metrics were addressed based on a coding framework mentioned in the study design and protocol section. The usability characteristics and problems and their severity rating are described as follows:

Efficiency: It was measured by two metrics: 1) the number of touches targeted and 2) the task completion time. The mean time taken for the users to perform each task is based on the following equation:

Efficiency = [(total of full completion of a task (1) or non-completion (0) / (time spent on a task)] / [(total number of tasks*number of users)]*100

Effectiveness: It was measured by the number of completed tasks (task completion rate), indicating the task’s success rate. The extent to which the user can fully and accurately achieve his task goals. Effectiveness was measured using the following equation.

Effectiveness = [(number of successfully completed tasks) / (total number of tasks performed)]*100

The range of effectiveness was taken as ‘awful’ (0–50%), ‘bad’ (50–75%), ‘normal’ (75–90%), and ‘good’ (90–100%) [24].

Learnability: It was evaluated by measuring the number of quickly completed tasks.

Errors were identified as the number of user mistakes when performing the tasks.

Satisfaction: It was measured based on the user's total score on the MARS questionnaire.

2-4-1-8-2- TA qualitative part

The video reactions of the participants were transcribed verbatim. Usability, characterized by users’ comments, silences, repeated actions, and error messages, was collected through the recordings.

Three members of the research team analyzed the obtained content. Transcripts and usability problems were also reviewed to identify the most common concerns. In any case of discrepancy in content analysis, a third-party reviewer was consulted.

These differences were categorized based on the tasks in the scenarios (measurements, zoom and magnifying, and contrast and window level).

Data collected during the think-aloud tasks (phase 1) was analyzed using fundamental inductive content analysis consisting of data reduction, data grouping, and the formation of concepts to answer research questions [39].

The inductive process is a bottom-up process that looks at all the issues as a whole by aggregating like issues together until all of the issues have been sorted into groups. Once all the groups (subcategories) had been sorted, they were labeled to create more significant categories [40, 41]. Thus, at the end of this process, we identified significant usability category issues and the specific problems associated with each one.

Phase II: Evaluation of the quality of the HAC app using the MARS questionnaire

2-4-2- 1- Design

The participants (16 medical interns) were asked to complete the MARS questionnaire immediately after the TA session. MARS is the most popular scale and highly reliable tool designed to assist researchers, professionals, and clinicians in classifying and assessing the quality of mHealth apps [25].

2-4- 2-2- Data collection

A validated and reliable Persian language version of MARS was used to collect the HAC app quality data [42].

MARS consists of 23 items of five objective quality subscales:

- Engagement encompasses five items and mainly focuses on entertainment and interest features of mobile apps
- Functionality includes four items and addresses the ease of use and functional capabilities of mobile apps
- Aesthetics consist of three items and discuss mobile app layout and visual appeal
- The information part encloses seven items and mainly considers quality, quantity, credibility, and visual enhancement of included information.

- The subjective quality subscale of MARS focuses on the overall rating of the app, its benefits, and its value.

2-4-2-3- Data Analysis

Each subscale item is rated a five-point score from 1 (inadequate) to 5 (excellent). Usually, the mean score and standard deviation (SD) were used to rate the quality of apps. Since the number of items in each subscale was different, we also used this formula [(mean of subscale/ number of items in subscale) * (100)] to compute the score out of 100 and compare the subscales. To calculate the total HAC app score, the [(total mean of HAC app/ total MARS items * (100)].

Friedman test was applied to compare the users' scores in five MARS subscales. The Wilcoxon test investigated the mean difference between two sets of pairs of MARS subscales. Spearman's rank correlation coefficient was used to analyze the positive correlation between MARS subscales. Kruskal-Wallis and one-way ANOVA tests were used to assess differences between medical interns' characteristics and MARS' subscales. All statistical analyses were performed using Statistical Package for Social Sciences 16.0 (SPSS Inc., Chicago, IL, USA) at a significant level of 0.05.

We applied inductive content analysis consisting of data reduction, data grouping, and the formation of concepts to analyze TA qualitative data and transform physicians' ideas into categories in the debriefing phase.

Phase III: Debrief Participants

A debrief session is an informal conversation to collect users' experiences and [43] any features of the app that they particularly like or dislike, how easy or difficult it is to use, and what they think about the content and design of the app was discussed in the debriefing session. The medical interns' general opinion regarding the effective mobile apps to assist HCPs in education or clinical practice was also investigated in this phase. During the analysis of the recorded videos and voices, it came to our attention that the debrief sessions were carried out with the participation of clinicians research team were the most active and engaging ones. To analyze and present the debriefed data, a narrative analysis method was utilized [44].

3- Results

The findings of each phase will be presented under the same heading in the methods section, including TA quantitative, TA qualitative, MARS quality assessment, and then a debriefing session.

Phase I*: TA usability testing*

Table 1 illustrates the scenarios, goals, and actions needed to complete the tasks.

Table 1: Descriptions of the scenarios used in the usability testing

Scenarios	Goals	Actions
1. A head trauma patient was admitted to the emergency department. Please check if the CT scan indicated a patient with "minor or mild acute closed head trauma without neurologic deficit adult."	According to the guidelines, search for appropriate imaging procedures for a given diagnosis.	1- Selecting the search icon 2- Typing the disease title in the search box 3- Finding the head trauma from the query list 4- Clicking on the plus button 5-Check if the imaging procedure is recommended for the patient.
2. A patient was admitted to the emergency department with "new onset of seizures older than 18 following acute trauma". Please select the appropriate imaging procedure for the case.	To use appropriate imaging procedures for seizures.	1-Opening the search query 2- Typing the seizures into the search box 3- Navigating between items in the search list 4-Selecting the appropriate imaging procedure based on the patient's symptom
3. Headache and vertigo are common symptoms at the emergency department. Please add headache to the shortlist for forthcoming queries.	To apply a shortlist menu to collect appropriate imaging procedures for common diseases and symptoms.	1- Adding headache to shortlist menu 2- Backing to the first page 3- Opening the shortlist 4- Deselecting the items you are not interested in anymore
4. A patient with proven subarachnoid hemorrhage (negative angiogram) was admitted to the hospital for follow-up. Please check for appropriate imaging procedures.	To use the CT/ CTA button to access subarachnoid hemorrhage.	1- Opening list of diseases under the title of CT 2- Finding subarachnoid hemorrhage 3- Moving one step backward 4- Selecting the CTA button 5- Navigating between items in the search list 6- Click on the plus sign to search for detailed information on subarachnoid hemorrhage and its subgroups.

3-1-1- TA quantitative

Efficiency

Based on the equation, the HAC app's relative overall efficiency was 97.8 %. The average time spent for each scenario was 97.5 seconds, and the number of additional clicks was 0.93. The highest average of performing scenarios belonged to scenario three (109.25 seconds), and the lowest average was related to scenario four (83.875 seconds). Among the users, the highest total average time for four scenarios was related to user three (user#3) (161.8 seconds), and the lowest time was for user ≠ 11 (58.0 seconds).

Effectiveness

The HAC app's effectiveness in assisting users in performing the scenarios based on the equation was good (96.9%). Of 16 users, 14 (87.5%) completed all four scenarios, two (12.5%) completed three, and two users had difficulty performing scenario two, which was focused on searching for "new onset of seizures older than 18 following acute trauma". The characteristics of this scenario that

caused usability issues have been discussed under the heading TA qualitative**,** "inefficient data presentation and information control," and "poor searching capabilities" (Figures 1 and 2).

Learnability

Out of 16 users, 11 users (68.8 percent) managed to complete four scenarios (100 percent), and four users (25 percent) managed to complete three scenarios (75 percent) without encountering critical issues. Two users (12.5 percent) faced critical issues to complete two scenarios.

Errors

Out of 16 users, 10 users (62.5%) did not make any errors while doing the scenarios, and six users (32.5%) were able to do the scenarios with more than one error.

Table 2: Matrix of efficiency and effectiveness of the HAC app

Users	Efficiency					Effectiveness
	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Total Average	Total Scenarios Completed
≠ 1	98	189	150	210	**161.8**	**3**
≠ 2	75	101	134	82	**98.0**	**4**
≠ 3	192	135	129	108	**141.0**	**4**
≠ 4	92	86	115	57	**87.5**	**4**
≠ 5	74	125	63	84	**86.5**	**4**
≠ 6	115	73	129	53	**92.5**	**4**
≠ 7	115	83	82	52	**83.0**	**4**
≠ 8	69	93	106	127	**98.8**	**4**
≠ 9	60	109	89	80	**84.5**	**4**
≠ 10	70	117	170	82	**109.8**	**3**
≠ 11	34	57	79	62	**58.0**	**4**
≠ 12	40	97	125	45	**76.8**	**4**
≠ 13	109	132	87	64	**98.0**	**4**
≠ 14	72	83	95	67	**79.3**	**4**
≠ 15	101	185	95	107	**122.0**	**4**
≠ 16	110	59	100	62	**82.8**	**4**
Average for each scenario	**89.125**	**107.75**	**109.25**	**83.875**	**97.5**	

3-1-2- TA qualitative

The results of the inductive content analysis regarding usability issues were grouped into eight main

categories and discussed below.

3-1-2-1-Lack of finger-friendly touch targets

Most participants had difficulty tapping the target buttons, like the shortlist button, or icons like the plus sign (+) on the screen, and it was an intensive task to perform successfully. The participants stated that the given features are inappropriate for finger-touch targets. It might be due to the wrong size of the buttons or the need for more padding between buttons and icons around the edge of the screen. Consequently, it led to selecting the bad screen portion, and they frequently faced mistapping the shortlist menu. Most participants often used this statement, "I can't get the button.”Failure to press the targeted button and retouching the icons multiple times occurred frequently, resulting in a long time on the task and decreased efficiency. Moreover, it caused a failure of task completion by two users and reduced the effectiveness of the HAC app.

3-1-2-2- Poor search capabilities

Navigating the diseases and signs and symptoms was case-sensitive to the upper case. It made the searching diagnosis and signs & symptoms keywords awkward." The participant struggled to find diseases and signs & symptoms that had not been typed in upper case. Some participants forgot the “case-sensitive” feature every time they started the new scenario. Thus, participants backed out and jumped over the navigation process or tried to find the given case from a long list of search results. Both situations made it time-consuming and inefficient and caused participants frustration.

3-1-2-3- Input problems

The main complaint by the participants was that the font size was inappropriately amplified with the limited mobile size. The participants mentioned that typing on the mobile phone screen was an intensive task. We found some difficulty in typing on the small screen; all the participant's attention was focused on what they had typed. The “case-sensitive” feature in searching data amplified the problem. The lack of finger-friendly touch targets also made the typing more cognitive load and distracted from their main concerns, interacting with the patients.

3-1-2-4- Inefficient data presentation and information control

Another usability issue that caused frustration among users was inefficient data presentation and information control. To apply the HAC app, users enter specific diseases, signs, or symptoms enclosed in the Care Core guideline in the "Index" box. However, the list of clinical criteria under the disease heading is grouped using the plus sign (+) to provide a proper data presentation. A long list of conditions in the form of a dropdown menu enclosing the common signs and symptoms makes it

confusing for the participant. Since the mobile screen is too small, providing a long list of search results makes it time-consuming and inefficient. The lack of proper information layering and data categorization made it difficult for the participant to scroll the list. The participants commented, "It requires much attention and is very inconvenient since we need to interact with patients, other colleagues, and clinical settings environment." The critical issue was related to bringing cognitive load to the participants.

3-1-2-5- Unclear control and confirmation

Another failure dealt with providing feedback and confirmation. The participants expected the HAC app to inform them about what was happening, using appropriate feedback. For instance, when they were asked to add a given disease to the shortlist, they waited for a dialogue to let them know the conditions were added. The absence of the appropriate resulted in the users being moved to the shortlist and checked on if the command was run. The exact process occurred when they were asked to remove the given disease from the shortlist. The users awaited a confirmation dialog regarding spoken questions, such as a “yes” or a “no,” to remove the disease from the shortlist before executing the removing command. Without a physical response, users do not know the current system status and are not confident about the consequences of their prior actions. They feel confusion and frustration.

3-1-2-6- Lack of predictive capabilities

Some participants expect more predictive capabilities and automation to optimize manual tasks and increase efficiency across various functions. For example, the participant stated, “We prefer the HAC app to automatically move the most visited diseases or signs and symptoms to the shortlist menu. They believed the sole manually supported feature for making a shortlist menu could be more efficient and time-consuming.

3-1-2-7- Poor Assistance and support

The participants thought some features on the HAC app, like shortlists highlighted in red or items with the plus sign, were difficult to recall or interpret and caused cognitive load. The participants need assistance or information to learn more about these features, like tooltips, which display informative text, such as a description of its function when users hover over, focus on, or tap an icon. They were looking for a help tab and found it unclear because it was at the bottom of the “About us” tab." It caused the HAC app to be less self-descriptive and more dependent on external help, which needed to be clarified and made clearer.

3-1-2-8- Unclear navigation logic

Some fundamental navigation control issues (e.g., “back” function) were also reported during usability testing. For example, the participants tended to click the back button to return to the previous page, but it actually led them back to the home page. This drawback can lead to work duplication and frustration in task completion.

Phase II: The quality of the HAC app using MARS

3-2-1- Analysis of overall quality scores of the HAC app

Table 1 indicates that the overall quality score of the HAC app was favorable (82 out of 100). The four MARS subscales assessed present "information" (73.37 out of 100), and "engagement" (73.48 out of 100) had the lowest score; while aesthetics had the highest score (87.86 out of 100) (Figure 3).

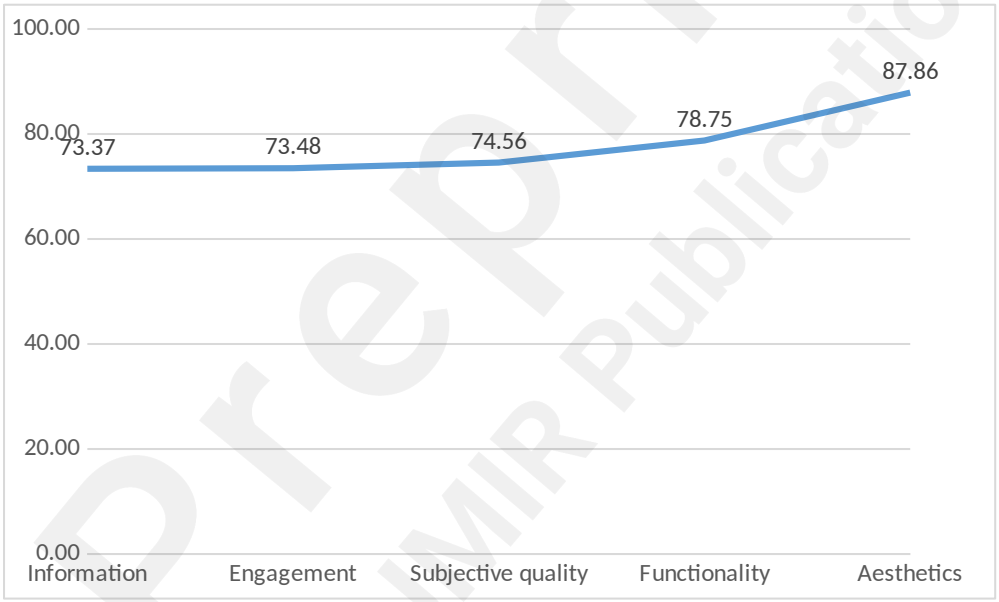


Figure 3: Overall quality scores of the HAC app

3-2-2- Analysis of significant differences and correlation between MARS’ subscales

Using the Friedman test, the users’ scores in five MARS subscales were compared, and the result revealed a significant difference (P <0.001).

Wilcoxon test was applied to investigate the mean difference between two sets of pairs of MARS subscales. The results indicated a significant relationship between the aesthetics subscale and engagement (P=0.001), information (P=0.003), subjective quality (P=0.004), and functionality (P=0.02). A significant relationship was also found between functionality and information subscales (P=0.013).

Table 3: The means difference between two sets of pairs of MARS subscales

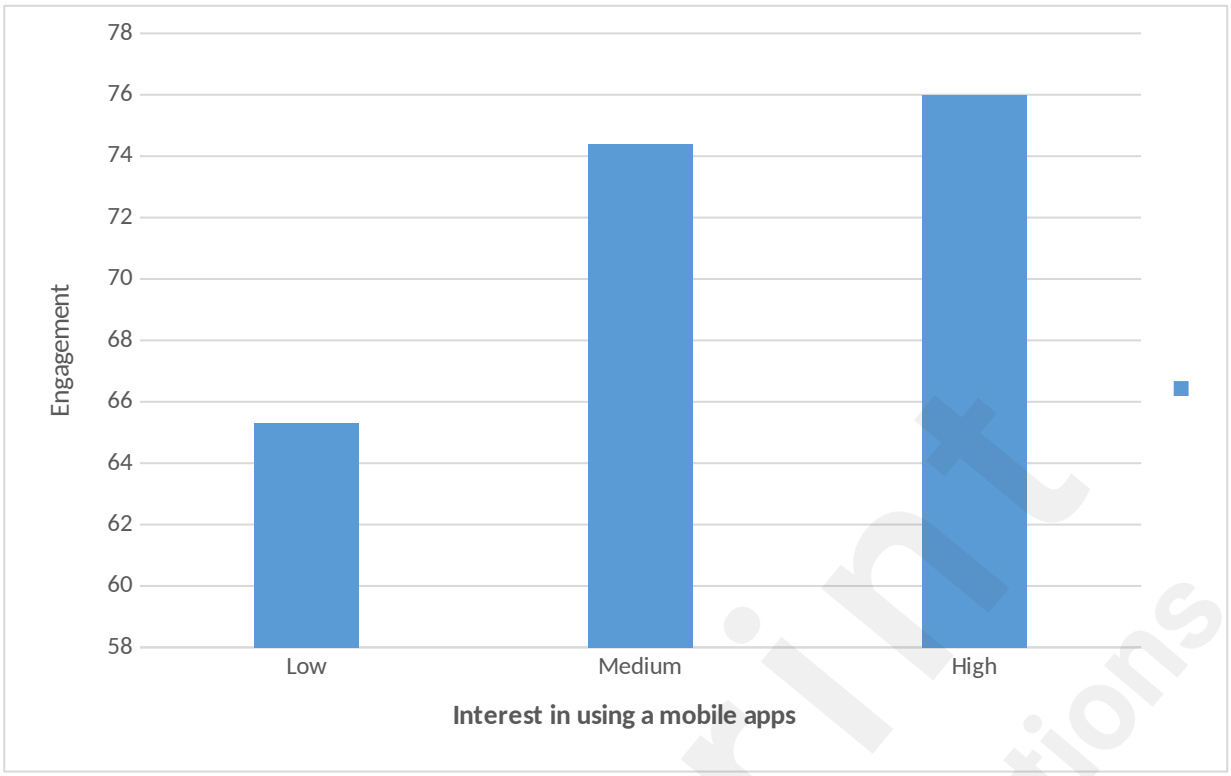
MARS Subscales	Engagement	Information	Subjective quality	Functionality	Aesthetics
Information	0.909				
Subjective Quality	0.53	0.900			
Functionality	0.057	0.013	0.32		
Aesthetics	0.001	0.003	0.004	0.02	

Spearman's rank correlation coefficient presented a positive correlation between information with functionality subscales, r (0.588), p =.017. A positive correlation also was seen between information and satisfaction, r (0.648), p =.005. Table 4 indicates in the subscale "information," the lowest score of the HAC app is "evidence base" (66.2 out of 100), and the highest score is "visual information" (82 out of 100). In the subscale "engagement," the lowest score of the HAC app is "customization" (63.6 out of 100), and the highest score is Interest (90 out of 100). In the subscale "functionality," the lowest score of the HAC app is "performance" (67.4 out of 100), and the highest score is "ease of use" (91.2 out of 100). In the subscale "aesthetics," the lowest score of the HAC app is "visual appeal" (83.6 out of 100), and the highest score is "graphics" (91.2 out of 100).

3-2-3- Medical interns' characteristics and MARS subscales

Of the 16 users participating in the study, none had used the HAC app before, and only one person (6.2%) had used similar applications. Among them, eight users (50%) believed using mobile applications for learning and clinical practice is helpful and were interested in using them. Figure 4 presented a significant difference between medical interns' Interest in using mobile apps for learning and clinical practice (low, medium, high) with engagement subscales using the Kruskal-Wallis (P=0.033).

Figure 4:



Significant difference between engagement and Interest in using the mobile app

Figure 5 also indicated a significant difference between the medical interns' Interest in using mobile apps with subjective quality subscales using a one-way ANOVA test (P=0.04).

Phase III: Debrief

We explored how useful they perceive the app to be, any features they particularly like or dislike, how easy or difficult it is to use, and what they think about the content and design of the app, which was discussed in the debriefing session. Although all users appreciated the high simplicity and learnability of the HAC app, they debated that navigation between pages and search capabilities need serious consideration.

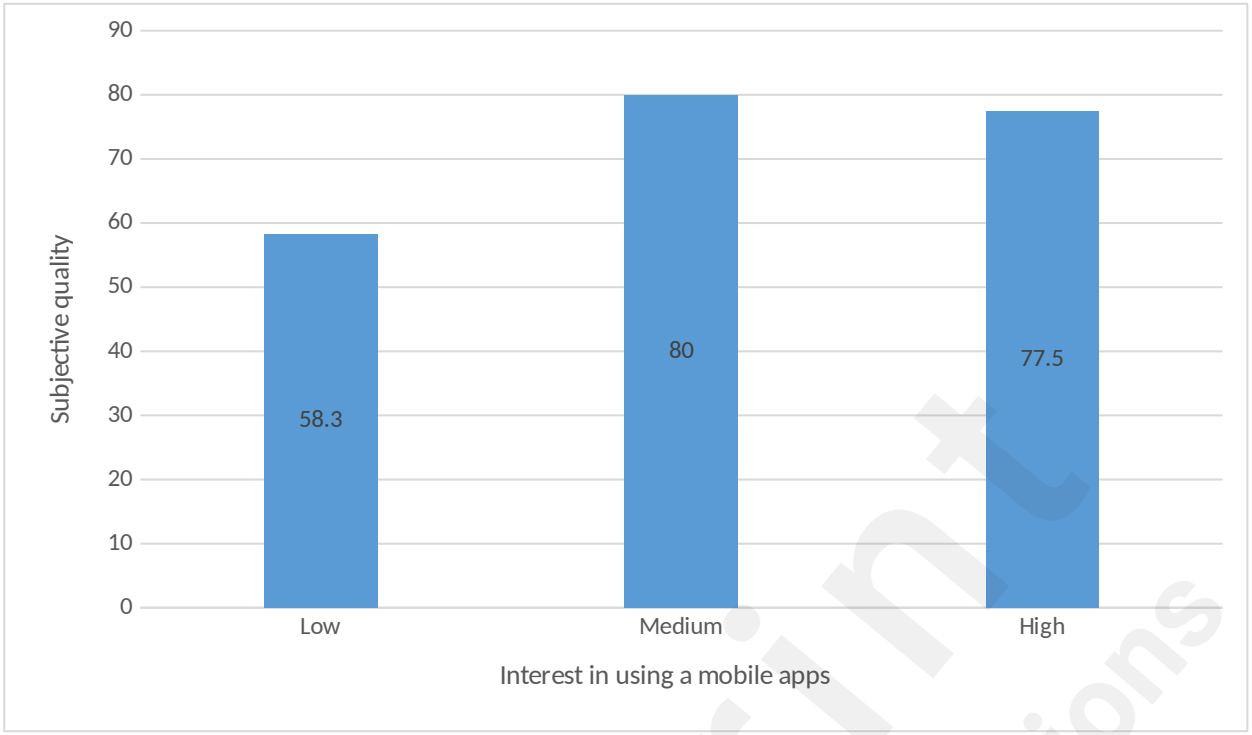


Figure 5: Significant difference between Subjective Quality and Interest in using mobile app

One of the participants wanted this tool to be equipped with voice recognition systems. *"… We use this tool while walking or moving in different parts of the hospital, and the possibility of typing or text entry increases the possibility of errors and, as a result, repeating the same action, which will reduce efficiency". [Participant 2]*

Another participant believed this tool should provide access to the app at different times and conditions. *"..I am a doctor, and my hands are bloody; I do not want to touch my mobile too much, and I prefer this app to be able to search for the proper CT scan based on voice. [Participant 15]*

Another participant expected that apps designed for students would pay more attention to the educational needs and learning styles of students.

"….I think this issue is so essential that medical education experts should also be used in the design of apps. Anyway, each of us has a style to learn. If this customization feature is not included in the design, surely some users will not be able to work with this system or at least feel comfortable and useful while working with it". [Participant 1]

Table 4: HAC app scoring based on MARS four subscales

Information	SD± Mean	Score from 100	Engagement	SD± Mean	Score from 100	Functionality	SD± Mean	Score from 100	Aesthetics	SD± Mean	Score from 100
Accuracy: The app contains what is described	3.6± 0.50	72.4	Entertainment	0.44 ± 3.25	65	Performance: accuracy/speed of the app functions and components (buttons/menus)	3.37±0.80	67.4	Layout: arrangement and size of buttons/icons/menus/content on the screen	4.43±0.72	88.6
Goals: specific, measurable, and achievable goals	3.6± 0.50	72	Interest: fun/entertaining of app	0.63 ± 4.5	90	Ease of use: Easy to learn how to use the app	4.56±0.62	91.2	Graphics: The quality/ resolution of graphics used for buttons/icons/menus/content	4.56±0.51	91.2
Quality of information: The app correct, well-written, and relevant content to the goal	3.5± 0.63	70	Customization: support all preferences for app features (e.g. sound, content)	0.54 ± 3.18	63.6	Navigation: accurate, appropriate, uninterrupted moving between screens	3.8±0.61	76	Visual appeal: Look of app	4.18±0.54	83.6
Quantity of information: the extent of coverage within the scope of the app	3.4± 0.72	68	Interactivity: Provide feedback, contain reminders, notifications	0.40 ± 3.81	76.3	Gestural design: consistency of (Taps, swipes, scrolls) across all components	3.9±0.57	78			
Visual information: visual (charts, images, videos) to describe concepts	4.1± 0.95	82	Target group	0.50 ± 3.62	72.4						
Credibility: legitimate source of app	4.06± 0.25	81.2									
Evidence base: trialed/tested of app	3.31± 1.07	66.2									

4- Discussion

Principal findings:

Our findings demonstrated that the HAC app was practical and had acceptable usability in efficiency and effectiveness. It also displayed a positive quality score based on the MARS scale. In contrast, results of the think-aloud usability test revealed that the HAC app suffers from eight notable usability issues. The results proved that despite the willingness of researchers and the simplicity of quantitative and questionnaire-based approaches to conducting usability testing [6, 45], the observational, think-aloud usability testing provided more unbiased, trustworthy, and insightful data in describing mobile app usability.

Nevertheless, through data analysis of the MARS subscales also brought to light the HAC app's usability issues, and its results support the qualitative TA results of the present study. This agreement could be explained by the fact that MARS is a scale specifically designed to assess the quality of mobile apps [46]. Typically, the available usability scales and questionnaires are not highly reliable [6], so they are general scales designed primarily for evaluating the usability of computers or websites .

Additionally, current usability and quality rating scales focus primarily on developers testing the usability of mobile apps, rather than end users who are patients or HCPs [46].

It is unlikely that usability issues will be thoroughly investigated in sole quantitative and questionnaire-based approaches [47] and need to be complemented by more objective and reliable approaches such as think-aloud methods.

In the current, HAC effectiveness assessment revealed that most users completed all four scenarios, although, two users faced problems completing scenario two, which involved finding an appropriate imaging procedure for the "new onset of seizures" case. This failure may be due to the usability issues we categorized under "poor search capabilities" and "inefficient data presentation and information control." As shown in Figure 1, “poor searching capabilities” and "poor data presentation" caused a long list of seizure conditions, confusing the participant. Since the mobile screen is too small, providing a long list of search results brings more cognitive load to select the correct item, and two participants are left to perform this scenario later. However, they never got back to the scenario again. Our results support previous research findings. In the study, Chen introduced proper navigation and searching capabilities as significant factors for users' rating of mobile health apps [48]. Schwab (2018) debated that ease of navigation is the foundation of an ideal mobile app since it smooths productivity and increases effectiveness [18]. In the study to explore the usability

of the physician-to-physician teleconsultation app in an orthopedic clinic, Choemprayong (2021) presented mobile app usability issues in terms of data entry errors, presenting large-scale data, and difficulty in selecting items from a list which arise as a result of limited mobile screen size [49].

The HAC app also indicated acceptable efficiency and meantime completion for four scenarios. However, scenario three also showed the highest mean time completion. The problem might arise due to usability issues regarding the "lack of finger-friendly touch targets." The limited screen size of mobile phones results in the inappropriate size of the buttons or lack of enough padding between the shortlist button and icons around the edge of the screen. Our results agree with previous studies that tapping the mobile phone buttons correctly and incorrect operations have also been reported frequently in previous studies [49-51]. In addition to data presentation, the low resolution of smartphone screens can lead to data input errors [49]. Existing evidence revealed highly significant differences between user effectiveness and efficiency with button sizes. In the study, Conradi reported substantial differences in error rate between button sizes (5*5 mm) compared to the other sizes (8*8 mm. 11*8 mm, and 14*14 mm. It has been debated that interaction with mobile devices due to limited screen size and resolution often requires additional considerations and a specially adapted interface [22]. Literature also claimed that key size manipulation should be considered for users' operation posture and activities (e.g., standing, sitting, walking) in mobile phone interactions [22]. However, the wide variation in optimal button size for mobile phones from 2.6 to 41.8 mm represents human-computer interaction in handheld devices. It is still in its infancy and requires more context awareness to provide assistance based on a knowledge of its environment. Another possible explanation for the highest time completion for scenario three is the usability issue categorized as "unclear control and confirmation" in the current study. The participants of the present study verbalized a lack of providing feedback on the HAC app when they were asked to add a given disease or sign and symptom to the list. The absence of the confirmation dialogue for successfully adding the given items to the shortlist resulted in the users moving to the shortlist and checking if the command was run. The exact process occurred when they were asked to remove the given disease from the shortlist. This rechecking caused work duplication and led to less efficiency. Work duplication has a significant and negative influence on physicians' performance and has been introduced as physicians' barrier to using mobile apps. In the study, Payne found that physicians would employ mobile apps to improve care workflow and productivity (38). Ely, in the study, found that physicians believed if working with IT-related tools takes more than

2 minutes, they will not be efficient and practical for the point of care (39). Therefore, the effectiveness and efficacy of mobile apps serve as critical factors for physicians' intention to use mobile apps [52, 53].

Regarding efficiency measures, our results also indicated significant variation in scenarios' time completion between the users. For example, the user ≠3 scored the highest total average time nearly three times the user ≠ 11 (the lowest time) to perform the scenarios. Besides designing an optimal layout, significant variation in scenarios' time completion between the users may be due to the user characteristics. Xiong debated that touch accuracy in mobile phones requires proper motor skills and "hand dexterity" in the operating fingers [20]. Schwab and Ozkan also highlighted the importance of user anatomy (average index or thumb fingertip size) and user dexterity (motor skills) in users' efficiency [16]. Cho reported usability problems related to the buttons of mobile apps developed using eye- a tracking system and retrospective think-aloud usability evaluation [51].

The HAC app also showed a favorable quality score based on the MARS scale. However, HAC app quality suffers from some drawbacks in "engagement" and "information," which focuses primarily on the effectiveness of apps in terms of interactivity, customizability, sending feedback, alerts, and reminders. Our results support previous results for assessing quality apps used by HCPs. In the study on drug reference apps in Taiwan, Chen also reported poor "engagement" capabilities in terms of lack of entertainment, interactivity, and customization in studied apps in Taiwan [48]. In the study investigating influential factors in adopting a clinical photo documentation app for clinicians, Jacob discussed some drawbacks in "engagement" capabilities that need to be added for further revision of a given app [15]. Although few studies exist to use MARS to evaluate clinical apps adopted by HCPs, other relevant evidence supports our findings. In a qualitative study, Pokhrel presented that HCPS prefers mobile apps that help them in their clinical practices, including "suggestive diagnosis and treatment after entering. Reports of studies that focused on using other information technology (IT) toolkits also revealed that the IT tool would be effective among HCPs if it would support interactivity, answer physicians' questions, send feedback, and provide decision reasoning. Sandholrz (2016) also introduced "prediction capabilities of mobile apps" as the most important preferences of medical students toward specific functionalities of future mobile apps [54]. Despite the HAC app's drawbacks in engagement and information subscales, its quality in "aesthetics" has shown favorable MARS scoring. In a study of preferences and perceptions of users regarding GUI and user experience (UX), Sandesara reported that minimalist design improves UX and user control to fulfill a task in a specific

order and time. The author argued that "simplicity is the ultimate sophistication" [55]. As far as we researched in the literature, there has been no article which has even evaluated and reported the items of each sub-scales of MARS and research is lacking on evaluating adopting and usability testing of mobile app by HCPs [1]. Lack of related literature to assess the items of each sub-scales of MARS lead on poor in-depth understandings and meaningful perception of apps' quality features in previous evidence. As a result, it was impossible to compare HAC app quality rating with previous research properly. However, the results of quality assessment using MARS supports TA qualitative findings of the current study. HAC app quality scoring in the functionality subscale revealed the minimum score belonged to the performance items, which focuses on the accuracy and speed of the app functions and components like buttons/menus. Navigation also scored the minimum rating in the given subscale. In the subscale engagement the item customization which supports providing all necessary settings/for apps features, and item interactivity that allow user input, providing feedback, and containing reminders, and notifications also acquired the minimum scoring.

Our findings in the debrief session indicated that physicians with clear awareness and understanding of their clinical context and work processes tend to use other data input methods, such as voice recognition, to interact with the HAC app. The results of physicians' workflow analysis and time and motion studies presented the medical profession as a multitasking job, not only managing patient care but also spending part of their activities on indirect tasks, from doing paper works and documentation to transitioning and traveling within the clinic area, fetching or bringing something [56, 57]. Thus, in designing mobile apps, performance accuracy and time on users' tasks in different positions while walking or standing should be addressed appropriately. It has been argued that interaction with mobile devices while walking influences on people's visual acuity and suppresses this ability by nearly 20% compared to visual acuity while standing (22). Conradi debated that walking is prone to a very high number of error occurrences, which is remarkable in smaller buttons [22]. Using mobile apps with text entry methods involves physicians in various interaction issues in terms of difficulty in typing on the small screen, mistapping due to inappropriate size of the buttons or lack of spacing between buttons, poor data presentation, etc. Any poor mobile interactions are attention-grabbing and make physicians solely concentrate on interacting with the mobile app to increase their performance accuracy. It would distract them from their main concerns, interacting with the patients.

Moreover, it results in a long time on the task and decreases the efficiency and effectiveness of HCPs in clinical settings. Auditory and sonic interfaces occupy less visual attention and

make users less engaged in the sole main task. As a result, users can handle multiple tasks simultaneously [58]. Here, physicians should be equipped with an alternative input method, for example, speech recognition. Evidence revealed that speech recognition (SR) has the potential to be a more efficient and effective method to speed up the entry rates while declining the error rates. It was reported that SR supports high entry rates (speaks at a mean entry rate of 13-45 words per minute while walking around) and a low error rate of less than 2 percent [59]. Given the requirement that medical interns suggested, they emphasized the importance of "context-awareness" design in mobile apps that focuses on capturing and exploring context-based information to describe any entity (persons, places, objects, workflows) embedded in the environment to fully understand and characterize users' tasks [58].

Implications:

The evaluation framework used in this study can serve as a guide for the design and improvement of future clinical mobile apps to ensure they meet usability and quality standards for use by HCPs. Identifying usability issues through user feedback and analysis can help developers improve the usability and user satisfaction of clinical mobile apps among HCPs. Moreover, the results of the study can serve as a reference for HCPs and developers in selecting and implementing clinical mobile apps with acceptable usability and quality. It emphasizes the importance of multidisciplinary research, incorporating medical education specialists' expertise, and considering user characteristics like motor skills and hand dexterity. The mixed methods approach used in the study, including MARS and TA analysis, can be adopted to gather valuable insights into user behavior and inform the design process of future apps for HCPs and developers. The study also suggests context-awareness design as a critical factor in developing meaningful IT-based solutions like mobile apps.

Limitations:

However, our investigation is subject to some limitations. It was conducted utilizing a limited sample a specific target group (medical interns), and attending physicians and residents were not involved in the current study. No contributions from IT experts and app developers were included in the evaluation of the HAC app. The study focused on the usability and quality of the HAC app in a specific medical context in Iran, which may limit the applicability of the findings to other healthcare settings or countries.

Conclusion

A mixed methods approach in evaluating the quality and usability of mobile apps yields valuable insights into the strengths and weaknesses of mobile apps while also complementing each other. Adopting a holistic and multifaceted approach in evaluating mobile applications is highly recommended, as exclusively relying on a single methodology does not provide reliable and trustworthy information about the usability and quality of mobile apps. The results also presented that the unique characteristics of mobile devices, such as screen size, the users' anatomical characteristics, and motor skills, influence users' interaction and usability with mobile applications. Therefore, considering these characteristics and developing more tailored tools and methods for usability testing of mobile applications bring potential benefits for developers, decision-makers, and HCPs.

Acknowledgements

Research team express their sincere gratitude to anyone who participated in the current study.

Conflicts of Interest

The authors declare that they have no competing interests.

Authors Contributions

ZM, AO, EN, and R.KH made substantial contributions to the conception and design of the study. FA, ZM, AO, HA, EN, ZN, and FH participated in data collection and performed the statistical analysis. ZM, EN, FH contributed to manuscript drafting, revision, and approval. All authors read and approved the final manuscript.

References:

1. Meidani, Z., et al., *Development of clinical-guideline-based mobile application and its effect on head CT scan utilization in neurology and neurosurgery departments*. BMC Medical Informatics and Decision Making, 2022. **22**(1): p. 1-12.
2. Garner, S.L., T. Sudia, and S. Rachaprolu, *Smart phone accessibility and mHealth use in a limited resource setting*. International journal of nursing practice, 2018. **24**(1): p. e12609.
3. Matricardi, P.M., et al., *The role of mobile health technologies in allergy care: An EAACI position paper*. Allergy, 2020. **75**(2): p. 259-272.
4. Anthes, E., *Mental health: There's an app for that*. Nature, 2016. **532**(7597): p. 20-3.
5. Tarricone, R., et al., *Recommendations for developing a lifecycle, multidimensional assessment framework for mobile medical apps*. Health Economics, 2022. **31**: p. 73-97.
6. Zhou, L., et al., *The mHealth App Usability Questionnaire (MAUQ): development and validation study*. JMIR mHealth and uHealth, 2019. **7**(4): p. e11500.

7. Meyerowitz-Katz, G., et al., *Rates of attrition and dropout in app-based interventions for chronic disease: systematic review and meta-analysis.* Journal of Medical Internet Research, 2020. **22**(9): p. e20283.
8. Adu, M.D., et al., *The development of My Care Hub mobile-phone app to support self-management in Australians with type 1 or type 2 diabetes.* Scientific reports, 2020. **10**(1): p. 7.
9. Owen, J.E., et al., *mHealth in the wild: using novel data to examine the reach, use, and impact of PTSD coach.* JMIR mental health, 2015. **2**(1): p. e3935.
10. Singh, K., et al., *Patients' and nephrologists' evaluation of patient-facing smartphone apps for CKD.* Clinical Journal of the American Society of Nephrology: CJASN, 2019. **14**(4): p. 523.
11. Singh, K., et al., *Many mobile health apps target high-need, high-cost populations, but gaps remain.* Health Affairs, 2016. **35**(12): p. 2310-2318.
12. Losa-Iglesias, M.E., et al., *The usability of a heartbeat measuring mobile phone app: An observational study.* Journal of medical systems, 2019. **43**: p. 1-4.
13. Mickan, S., et al., *Evidence of effectiveness of health care professionals using handheld computers: a scoping review of systematic reviews.* Journal of medical Internet research, 2013. **15**(10): p. e212.
14. Al-Ghamdi, S., *Popularity and impact of using smart devices in medicine: experiences in Saudi Arabia*. BMC Public Health, 2018. **18**: p. 1-7.
15. Jacob, C., A. Sanchez-Vazquez, and C. Ivory, *Factors impacting clinicians' adoption of a clinical photo documentation app and its implications for clinical workflows and quality of care: qualitative case study.* JMIR mHealth and uHealth, 2020. **8**(9): p. e20203.
16. Ozkan, N.F. and F. Gokalp-Yavuz, *Effects of dexterity level and hand anthropometric dimensions on smartphone users' satisfaction.* Mobile Information Systems, 2015. **2015**.
17. Wu, P., et al., *Factors affecting physicians using mobile health applications: an empirical study.* BMC health services research, 2022. **22**(1): p. 1-14.
18. Schwab, T. and J. Langell, *Human factors–based mobile application design for global health.* Surgical innovation, 2018. **25**(6): p. 557-562.
19. Teferi, G.H., et al., *Smartphone medical app use and associated factors among physicians at referral hospitals in Amhara region, North Ethiopia, in 2019: cross-sectional study.* JMIR mHealth and uHealth, 2021. **9**(3): p. e19310.
20. Xiong, J., S. Muraki, and K. Fukumoto, *The effects of touch button size on touchscreen operability.* Journal of Mechanics Engineering and Automation, 2014. **4**(8): p. 667-672.
21. Islam, M.N., et al., *Investigating usability of mobile health applications in Bangladesh.* BMC medical informatics and decision making, 2020. **20**(1): p. 1-13.
22. Conradi, J., O. Busch, and T. Alexander, *Optimal touch button size for the use of mobile devices while walking*. Procedia Manufacturing, 2015. **3**: p. 387-394.
23. Esfahani, M.Z., R. Khajouei, and M.R. Baneshi, *Augmentation of the think aloud method with users' perspectives for the selection of a picture archiving and communication system.* Journal of Biomedical Informatics, 2018. **80**: p. 43-51.
24. Farrahi, R., et al., *The relationship between user interface problems of an admission, discharge and transfer module and usability features: a usability testing method.* BMC medical informatics and decision making, 2019. **19**: p. 1-8.

25. Stoyanov, S.R., et al., *Development and validation of the user version of the Mobile Application Rating Scale (uMARS).* JMIR mHealth and uHealth, 2016. **4**(2): p. e5849.
26. Thurmond, V.A., *The point of triangulation.* Journal of nursing scholarship, 2001. **33**(3): p. 253-258.
27. Pavlíček, J. and P. Pavlíčková, *Usability Testing Methods and Usability Laboratory Management*, in *Updates on Software Usability*. 2022, IntechOpen.
28. Baxter, K., C. Courage, and K. Caine, *Understanding Your Users: A Practical Guide to User Research Methods*. 2015: Elsevier Science.
29. Sauro, J., *A practical guide to measuring usability.* Measuring Usability LLC, Denver, 2010. **12**.
30. Macefield, R., *How to specify the participant group size for usability studies: a practitioner's guide.* Journal of usability studies, 2009. **5**(1): p. 34-45.
31. Cho, H. *Development and Usability Evaluation of an mHealth Application for Symptom Self-Management in Underserved Persons Living with HIV*. 2017.
32. Peute, L.W., N.F. de Keizer, and M.W. Jaspers, *The value of Retrospective and Concurrent Think Aloud in formative usability testing of a physician data query tool.* Journal of biomedical informatics, 2015. **55**: p. 1-10.
33. *google play*. Available from: https://play.google.com/store/apps/details?id=com.hecorat.screenrecorder.free&hl=en&gl=US&pli=1.
34. Olmsted-Hawala, E.L., et al. *Think-aloud protocols: a comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability*. in *Proceedings of the SIGCHI conference on human factors in computing systems*. 2010.
35. Liu, L. and M.T. Özsu, *Encyclopedia of database systems*. Vol. 6. 2009: Springer New York, NY, USA:.
36. Nielsen, J., *Usability 101: Introduction to usability (2012).* URL: http://www. nngroup. com/articles/usability-101-introduction-to-usability/[Accessed November 2016], 2012. **9**: p. 35.
37. Abran, A., et al., *Usability meanings and interpretations in ISO standards.* Software quality journal, 2003. **11**: p. 325-338.
38. Joe, J., et al., *The use of think-aloud and instant data analysis in evaluation research: Exemplar and lessons learned*. Journal of biomedical informatics, 2015. **56**: p. 284-291.
39. Kyngas, H. and K. Mikkonen, *The application of content analysis in nursing science research*. 2020, Springer.
40. Elo, S. and H. Kyngäs, *The qualitative content analysis process.* Journal of advanced nursing, 2008. **62**(1): p. 107-115.
41. Azungah, T., *Qualitative research: deductive and inductive approaches to data analysis.* Qualitative research journal, 2018. **18**(4): p. 383-400.
42. Raeesi, A., R. Khajouei, and L. Ahmadian, *Evaluating and rating HIV/AIDS mobile apps using the feature-based application rating method and mobile app rating scale.* BMC Medical Informatics and Decision Making, 2022. **22**(1): p. 1-11.
43. Henshall, C., et al., *A usability study to test the effectiveness, efficiency and simplicity of a newly developed Internet-based Exercise-focused Health App for Lung cancer survivors (iEXHALE): Protocol paper.* Health Informatics Journal, 2020. **26**(2): p. 1431-1442.
44. Smith, C.P., *Content analysis and narrative analysis*, in *Handbook of research methods in social and personality psychology.* 2000, Cambridge University Press: New

York, NY, US. p. 313-335.
45. Ye, B., et al., *Researched Apps Used in Dementia Care for People Living With Dementia and Their Informal Caregivers: Systematic Review on App Features, Security, and Usability.* Journal of Medical Internet Research, 2023. **25**: p. e46188.
46. Azad-Khaneghah, P., et al., *Mobile health app usability and quality rating scales: a systematic review.* Disability and Rehabilitation: Assistive Technology, 2021. **16**(7): p. 712-721.
47. Gibson, A., et al. *Assessing usability testing for people living with dementia*. in *Proceedings of the 4th Workshop on ICTs for improving Patients Rehabilitation Research Techniques*. 2016.
48. Chen, Y.-C., et al., *Personalized and self-management: systematic search and evaluation quality factors and user preference of drug reference apps in Taiwan.* Journal of Personalized Medicine, 2021. **11**(8): p. 790.
49. Choemprayong, S., C. Charoenlap, and K. Piromsopa, *Exploring usability issues of a smartphone-based physician-to-physician teleconsultation app in an orthopedic clinic: mixed methods study.* JMIR Human Factors, 2021. **8**(4): p. e31130.
50. Ast, K., *Diagnostic Efficacy of Handheld Devices for Emergency Radiologic Consultation: Toomey RJ, Ryan JT, McEntee MF, et al. AJR Am J Roentgenol 2010; 194: 469–74.* Journal of Emergency Medicine, 2010. **39**(2): p. 273.
51. Cho, H., et al., *A multi-level usability evaluation of mobile health applications: A case study.* Journal of biomedical informatics, 2018. **86**: p. 79-89.
52. Ely, J.W., et al., *Analysis of questions asked by family doctors regarding patient care.* Bmj, 1999. **319**(7206): p. 358-361.
53. Payne, K.F.B., H. Wharrad, and K. Watts, *Smartphone and medical related App use among medical students and junior doctors in the United Kingdom (UK): a regional survey.* BMC medical informatics and decision making, 2012. **12**: p. 1-11.
54. Sandholzer, M., et al., *Medical students' attitudes and wishes towards extending an educational general practice app to be suitable for practice: A cross-sectional survey from Leipzig, Germany.* European Journal of General Practice, 2016. **22**(2): p. 141-146.
55. Sandesara, M., et al., *Design and Experience of Mobile Applications: A Pilot Survey.* Mathematics, 2022. **10**(14): p. 2380.
56. Frey, S.M., et al., *Inter-hospital comparison of working time allocation among internal medicine residents using time-motion observations: an innovative benchmarking tool.* BMJ open, 2020. **10**(2): p. e033021.
57. Tipping, M.D., et al., *Systematic review of time studies evaluating physicians in the hospital setting*. Journal of hospital medicine, 2010. **5**(6): p. 353-359.
58. Lumsden, J., *Handbook of research on user interface design and evaluation for mobile technology*. Vol. 1. 2008: IGI global.
59. Kristensson, P.O., *Five challenges for intelligent text entry methods.* AI Magazine, 2009. **30**(4): p. 85-85.

Supplementary Files