Human Behavior and Emerging Technologies

WILEY

RESEARCH ARTICLE OPEN ACCESS

Expert-Based Usability Evaluation of an mHealth Application for Older Adults: A Mixed-Method Approach

Agnese Augello[1] | Giuseppe Caggianese[1] | Sotiria Antaranian[2] | George Zissis[2] | Giuseppe De Pietro[1,3] | Luigi Gallo[1,3]

[1]Institute for High Performance Computing and Networking (ICAR), National Research Council (CNR), Rende, Italy | [2]Innovation Lab (iLab), Athens Technology Center SA (ATC), Athens, Greece | [3]Department of Information Science and Technology, Pegaso University, Naples, Italy

Correspondence: Giuseppe Caggianese (giuseppe.caggianese@icar.cnr.it)

Received: 16 September 2024 | **Revised:** 24 February 2026 | **Accepted:** 27 March 2026

Academic Editor: Puspa Setia Pratiwi

Keywords: mHealth | older adults | sentiment analysis | smart health | usability evaluation

ABSTRACT

Demographic aging presents new challenges and highlights the growing need for smart healthcare applications tailored to older adults. This paper presents the assessment of a mobile health application designed to facilitate the communication between healthcare providers and older adults, in the context of the Smart Bear Horizon 2020 project. The study describes a task-oriented usability assessment, applying a mixed-method approach through the combination of quantitative metrics of performance, standardized usability assessment questionnaires (ASQ and PSSUQ), and interaction logs, along with qualitative user comments. For this purpose, 50 ICT specialists in digital health were engaged who conducted a structured, expert-based diagnostic evaluation aimed at identifying usability barriers relevant to older adults prior to end-user testing. Usability was evaluated in terms of effectiveness, efficiency, and satisfaction, following the ISO 9241-11: 2018 standard. Results show an overall task completion rate of 83.5% and a PSSUQ score of 3.16 on a 7-point scale (lower scores indicating higher perceived usability). The findings reveal some critical usability issues, particularly in navigation efficiency and discoverability of specific system features. The study underlines the importance of an expert-based, mixed-methods usability evaluation as an initial step in the design of mHealth applications for older adults. This process is crucial in identifying and resolving key usability challenges before engaging vulnerable user populations. The proposed method provides actionable guidance for developing and evaluating more inclusive digital health technologies, supporting expert-first usability testing as a best practice in the eHealth domain.

1 | Introduction

The proliferation of mobile health (mHealth) applications with features such as remote monitoring, personalized interventions, and ongoing support has transformed health care delivery, particularly for elderly users. Technology for monitoring health and fostering physical and cognitive well-being, while also lessening the burden on caregivers and reducing health care costs, is a growing necessity for those aged 65 and older. However, to promote health and well-being in older populations, such tools must be usable and accessible [1–5]. This suggests a potential imbalance between supply and demand in assistive technologies. Despite their promising features, their value depends on actual uptake and, most critically, the ease of use by the target population.

There is a diverse landscape of available solutions, highlighting benefits such as cost savings and reduced caregiver stress, as well as concerns and challenges, including privacy, cybersecurity, data ownership, infrastructure, and resource-related obstacles [6]. In particular, the integration and utilization of data from smart home sensors and mobile devices that detect physiological parameters can enhance daily task management and interventions, particularly for memory-impaired older adults [7]. These

technologies are reshaping traditional healthcare services, improving quality of life and well-being, and promoting healthy aging through remote monitoring [7]. However, it is crucial to explore their effectiveness and acceptance among the aging population, and to analyze significant variations in the usage and demand for these apps, focusing on user-friendliness and safety [8]. Several studies have already demonstrated the promising role of these technologies in supporting older adults in health-seeking behavior [2, 9] and monitoring their activities in cases of cognitive impairment [10], with significant implications for future health services and policies.

Although the integration of Internet of Things (IoT) and smart home sensors offers significant potential for remote monitoring, the growing interest in this domain is evidenced by the shift toward discussing the suitability of traditional usability evaluation methods for older adults. Researchers emphasize the need for measures that better reflect aging characteristics and recommend adapting usability evaluation methods to better suit the elderly [11–14]. However, to the best of our knowledge, existing studies often focus on specific aspects such as satisfaction or learnability, and it remains unclear how best to combine expert evaluation with end-user feedback to optimize application design. Specifically, there is a lack of structured, multimethodological approaches that precede large-scale deployment to identify critical usability barriers before they impact the final user. The literature shows substantial heterogeneity in how task-based evaluations are designed and reported for older adults, and studies do not always make explicit how task outcomes map to ISO 9241-11:2018 dimensions, using both objective interaction measures and standardized questionnaires. This methodological variability is particularly relevant given the rapid diversification of digital health solutions for older adults, in which increasingly complex app ecosystems must be evaluated not only for potential benefits but also for usability.

To bridge the gap between the initial design and final user needs, the study presented here provides a task-based usability evaluation of a mobile app for elderly health monitoring, aiming to assess and improve the first release of an mHealth app and to investigate its efficiency and effectiveness. Although focused on a specific mHealth application, the methodology and insights discussed are broadly applicable and can inform best practices for future mHealth solutions. The objective of this study was twofold: (i) to implement a rigorous expert-based diagnostic evaluation aimed at identifying usability barriers relevant to older adults prior to app deployment, and (ii) to establish a repeatable best practice methodology for mHealth development. The evaluation involved 50 information and communication technologies (ICT) experts experienced in developing technological solutions to support older adults. A mixed-method evaluation was conducted, including standard tools, open-ended questions, and interaction log analysis. The results, obtained from data collected between October 2022 and December 2022, showed that the app was generally well received, with participants appreciating its clear, pleasant look and feel. However, several areas for improvement were identified, including navigation efficiency and the accessibility of certain functionalities. Overall, the evaluation provides valuable insights into the app's usability and underscores the importance of conducting usability evaluations throughout the design and development process.

In what follows, the mHealth application and the rationale for the evaluation are introduced along with the metrics and the tools applied to the usability evaluation in Section 2. The results and feedback obtained by technical experts are discussed in Section 3; then, the redesign of the app driven by the evaluation outcomes is outlined in Section 4. Finally, after discussing the conclusive remarks, limitations, and future direction in Sections 5, the two paragraphs in the appendix include all the details related to the statistical analysis performed and of the qualitative feedback collected.

2 | Material and Methods

The evaluation process described in this paper represents the first step in a comprehensive assessment of the Smart Bear [15] 2.1 mobile app. This phase, undertaken prior to its deployment to participants in the project pilots, was conducted to ensure that the app met the usability requirements of its target audience. For this reason, we conducted a task-based evaluation by involving ICT experts. The experts performed specific tasks to identify potential areas of concern or shortcomings, highlighting aspects of the app's functionality or interface design that may negatively impact its usability. The adopted approach is a mixed quantitative–qualitative evaluation methodology based on the administration of standard tools, open-ended questions, and the analysis of data extracted from log files, where the involved ICT experts act as proxies for the elderly users to gain a thorough understanding of how users may interact with the app, identifying strengths and areas for improvement in the app design. At the same time, this evaluation process gave us the opportunity to define and experiment with a repeatable best practice methodology for mHealth development, and the results, discussed in the next sections, have been used to drive and enhance the redesign of the app.

2.1 | Smart Bear Project

Smart Bear [15], a flagship project funded under the Horizon 2020 program, has defined a robust IoT platform integrating wearable and medical devices, smart home sensors, and advanced methodologies for big data analytics. The platform enables continuous monitoring of older adults' health conditions and, in synergy with the Smart4Health [16] and Holobalance [17] platforms, provides evidence to support personalized interventions [18–20]. Through large-scale studies conducted across five European countries and involving thousands of older adults, the project demonstrated how such technologies can prevent, slow the progression of, or effectively manage these impairments, thereby improving quality of life while also yielding substantial reductions in healthcare costs.

2.2 | Smart Bear Mobile App

The mHealth application investigated in this study was designed to bridge the gap between complex clinical monitoring and the daily life of older adults. The design framework was driven by the need to balance robust medical data collection with the cognitive and physical requirements of older adults. The primary

objective was to ensure high usability, addressing age-related challenges such as reduced visual acuity, possible difficulty interacting with small screens, and potential cognitive overload. Recognizing these factors, it was crucial to prioritize an interface that minimizes mental effort and prevents interaction errors. Due to the specific needs of the app's target users, who often manage cardiovascular, cognitive, and sensory impairments, the app must serve as a simplified entry point into a complex ecosystem. This complexity requires a systematic usability evaluation to ensure that technical interoperability does not lead to user exclusion.

An essential function of the app is collecting and harmonizing health data from diverse medical devices (e.g., oximeters, blood pressure monitors, and smartwatches) into a unified, FHIR-compliant format. From a design perspective, this centralizes information that would otherwise be fragmented across multiple interfaces, directly addressing the "app fatigue" and confusion often experienced by elderly users when managing multiple health tools. The collected data remains available to the user at all times, empowering them through self-monitoring—a key factor in promoting health literacy and autonomy.

2.2.1 | Description of App Sections and Rationale for Evaluation

The architecture of the app is divided into modules that map directly to the specific vulnerabilities associated with advancing age. The app's main screen serves as a central hub for navigating all functionality. Figure 1 shows the different button icons on the main screen, which allow the user to access the distinct app sections. To facilitate mental models, the interface employs a color coded semantic system: Blue icons represent passive medical monitoring (e.g., Heart and Balance), whereas red icons denote active behavioral interventions (e.g., Diet and Medication). This distinction is intended to help users categorize tasks quickly, reducing the search time and cognitive load. The medical information varies according to the user profile, providing the user with the latest measurements received from the relevant monitoring device, along with a comprehensive history of those measurements. For instance, within the Heart section, users can navigate to screens for heart rate, pressure, weight, medication, activity, oximeter, and temperature, where both the most recent measurement and historical user measurements retrieved from the connected devices are shown (see Figure 2). Meanwhile, other app sections, such as diet, medication, and rehabilitation for physical activity promotion, are displayed by default for all users, constituting the core components of independent living and active aging and thus requiring the highest level of cross-demographic usability. These core functionalities were selected as the primary focus for our usability study because they represent high-frequency, high-cognition tasks. Medication management is a critical daily need for seniors, where errors can have immediate health consequences, making its interface a priority for error-prevention testing. Similarly, dietary logging, intended to capture eating habits and daily food intake, involves data entry and navigating among different food subsections, posing a significant challenge for users and leading to cognitive fatigue. Furthermore, the rehabilitation module, which delivers tailored exercise programs via instructional videos, presents unique usability challenges requiring users

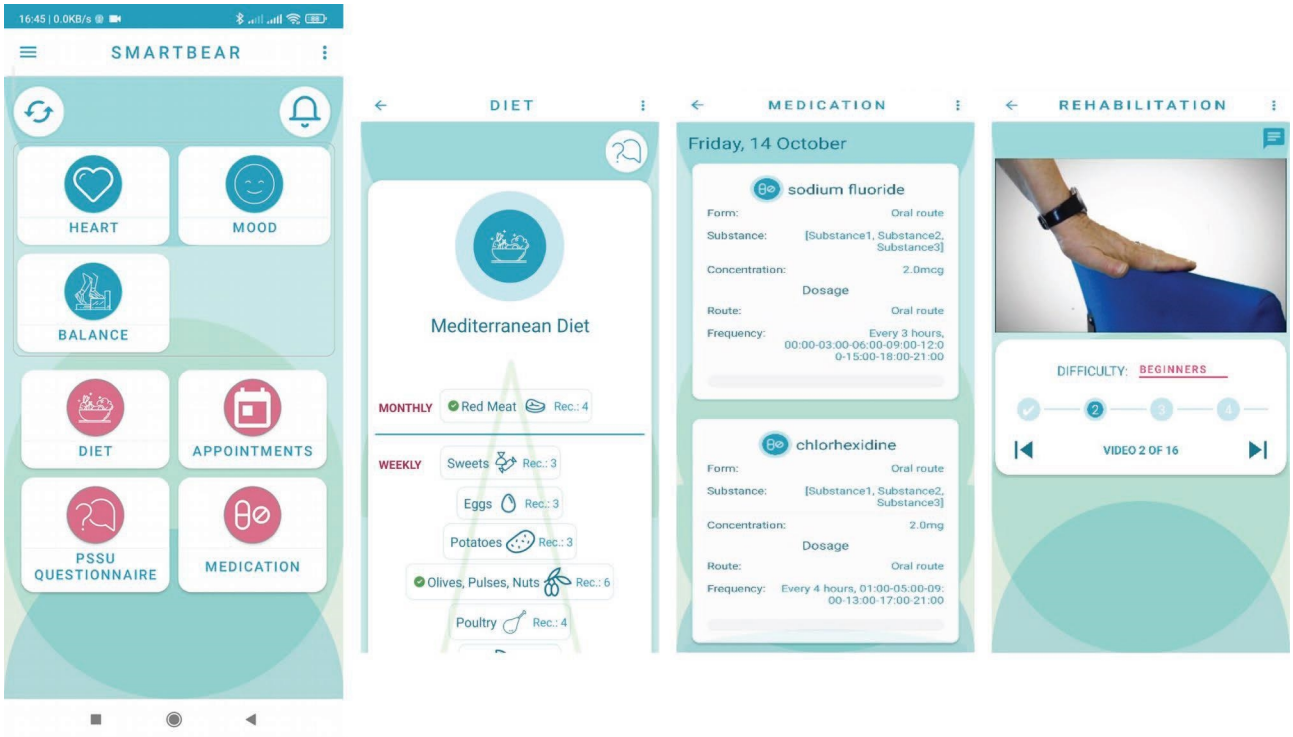

FIGURE 1 | On the left is the main screen of the app used for the study. The buttons allow the user to access medical information (blue icons) and other sections related to healthy behaviors (red icons). On the right are the three sections of the app involved in the evaluation study: Diet, Medication, and Rehabilitation, which can be reached using the balance button.

FIGURE 2 | Example of usage of the Smart Bear mobile app: On the left, a participant measures temperature using a thermometer, whereas on the right, an operator demonstrates how to access measurement history.

to maintain focus on a digital screen while performing physical movements, thereby demanding high levels of cognitive–motor coordination.

2.3 | Adopted Usability Dimensions and Used Tools

In the study, we evaluate the usability by relying on the ISO 9241-11:2018 definition [21]. Therefore, we considered the three dimensions of usability: effectiveness, efficiency, and satisfaction based on the participants' approach to the app sections and offered functionalities. These dimensions were evaluated for all tasks proposed to the participants and presented in Section 2.6.

- *Effectiveness* was assessed by the completeness and correctness of task completion, regardless of the effort put into completing the task. We operationalized this metric as the product of the tasks completion rate and the tasks correctness rate (*effectiveness* = *completion* × *correctness*), indicating whether users successfully navigate the app or encounter critical failures. In cases where outcomes lack a binary correctness criterion, effectiveness is simplified to the task completion rate, a standard practice in usability research [22].
- *Efficiency* depends on the effort to complete a task in relation to the resources expended (number of clicks to solve the task and time taken to solve the task). These metrics were used as indicators of performance, that is, a measure of the outcome of the user's interaction.
- *Satisfaction* was considered in terms of perceived satisfaction and acceptance of technology. Two of the most commonly used questionnaires in the literature were chosen: the 3-item After-Scenario Questionnaire (ASQ) [23] and the 16-item Poststudy System Usability Questionnaire (PSSUQ) [23]. The two questionnaires, ASQ and PSSUQ, were administered at the end of each task and after all tasks were completed, respectively. Specifically, the ASQ is used to assess users' experience regarding the facility in performing the task, the time required to complete it, and the level of assistance received during the process. The PSSUQ is used to collect subjective evaluations regarding system usefulness (SYSUSE), information quality (INFOQUAL), and interface quality (INTERQUAL). For both instruments, responses were recorded on a 7-point scale, where 1 corresponds to *strongly agree* and 7 to *strongly disagree*. The scale also includes a "not applicable" (N/A) point. Finally, both questionnaires have been integrated with open-ended questions to collect suggestions for app improvements.

2.4 | Participants

The study involved a sample of 50 ICT experts (age: mean 42.09, SD12.04), recruited from the Research and Industrial partners of the Smart Bear project. These participants were experienced in digital health technologies and in the design and development of solutions tailored to the older adult population. They were selected for their expertise in assessing the technical aspects of the app and in understanding specific challenges faced by older adults, such as limitations in vision, motor skills, and technology literacy, which informed their evaluation of the app. The participants were instructed to perform the task scenarios under conditions intended to reflect common usage contexts for older adults, in order to identify potential usability barriers, interaction breakdowns, and navigation inefficiencies. In this context, the experts' technical background enabled a systematic and informed assessment of interaction design choices, interface consistency, and functional complexity, which are known to affect older users. This expert-first approach aligns with established human–computer interaction practices, in which early-stage analytical and task-based evaluations precede testing with the target end-user population.

The participants were recruited indirectly to ensure their anonymity. The experts belonging to the partner responsible for the development of the app were excluded from the study. We are aimed at involving at least 50 participants since scientific literature highlighted potential limitations of the Nielsen "10 ± 2" sample size rule, recognizing that larger samples tend to identify more usability problems [24–27]. In particular, we considered the work of Schmettow [28], which, for a formative evaluation using a qualitative method, recommends at least 30 participants. Even though this number of participants allows identifying most interface problems, we decided to increase it to 50 to account for a possible dropout rate due to errors in performing the evaluation remotely.

During the study, participants were asked to perform a technical assessment of the app. They did this by completing the proposed task scenarios (see Section 2.6), considering typical cognitive, perceptual, and motor constraints associated with older adult users. Afterward, they filled out questionnaires to provide their impressions of the app.

In addition to being ICT experts with experience in the health domain, participants in the study were required to be Android users willing to install the app being tested on their devices. This is because, as indicated in the following section, the app was only developed for Android devices.

2.5 | Apparatus and Data Collection

The study opted not to involve IoT devices due to logistical constraints and the focus on evaluating app functionalities and interface usability rather than device operation or integration. Providing IoT devices to all participants would have been impractical and costly. Evaluating their integration with the app would not have significantly contributed to assessing the core objectives of the study. Therefore, a mock-up version of the app was used exclusively to evaluate user interaction and usability aspects without introducing the complexities of IoT device integration.

The mock-up version of the app was designed to work on an Android smartphone since the tested app was developed for the Android mobile operating system. No special equipment was needed for the study.

The version of the app used for the study was customized to simplify the remote execution. The app was able to track user actions during task execution, collecting information such as the number of clicks and the execution time anonymously and allowing questionnaire filling and answer collection directly in the app.

2.6 | Task Scenarios

Four task scenarios were defined in the study on top of the app sections: Diet, Medication, and Rehabilitation, since those are displayed by default for all users, regardless of their conditions, constituting the core components of the app. The instructions provided to participants were intentionally kept vague to avoid directing users toward specific sections of the app. In the following, each task is presented by first reporting in italics the text provided to participants, followed by a detailed description of the actions participants are tasked to perform and the criteria for task completion.

Task 1 *Using the app, fill in your eating habits and, in the end, answer the proposed questionnaires. In case of difficulties, use the Smart Bear mobile app manual*. In this task, participants indicate their eating habits by accessing the Diet section and completing the diet questionnaire. The task is considered complete once all questionnaire items have been answered.

Task 2 *Using the app, fill in your daily food intake and, in the end, answer the proposed questionnaires. In case of difficulties, use the Smart Bear mobile app manual*. In this task, participants report their daily food intake. Upon accessing the Diet section, they specify the quantities consumed for each listed category. This is accomplished by tapping on the corresponding icons and entering the consumed quantities. The task is considered complete once all categories in the daily section have been completed.

Task 3 *Using the app, perform (watching the video) all the rehabilitation exercises proposed by the session, imagining you are in a situation of "no pain, no recent falls, no shortness of breath, and level of unsteadiness less than 5." In the end, fill out the proposed questionnaires. Use the Smart Bear mobile app manual in case of difficulties*. In this task, participants simulate a rehabilitation session by first accessing the Balance section and selecting the rehabilitation icon. This action initiates the rehabilitation workflow, through which participants answer a series of questions regarding their perceived health status (as specified in the task text) to authorize access to the rehabilitation exercises. Once the session begins, participants acknowledge the safety instructions and proceed through the instructional videos associated with the exercises. The rehabilitation session consists of 17 exercises, each accompanied by a video that can be watched or skipped using the designated control. Between consecutive videos, the app presents questions concerning the participant's current health status, allowing them to continue or interrupt the session. The task is considered complete once all 17 videos have been either watched or skipped.

Task 4 *Using the app, modify the timeframe of all the medications shown in the app by selecting 8:00 and 17:00 as new timeframes. In the end, fill out the proposed questionnaires. Please rely on the Smart Bear mobile app manual in case of need*. In this task, participants access the Medication section and select each listed medication to modify the administration timeframe as specified. The task is considered complete once all medications have been updated to the new timeframes.

2.7 | Procedure

The study was conducted remotely with the participation of 50 individuals. Each technical partner designated process supervisors who were responsible for identifying and involving three or four of their expert colleagues in the study.

Subsequently, each supervisor was asked to send each participant detailed instructions for the execution of the test, which included (i) the description of the tasks to be performed, (ii) the APK file of the app to be installed on their device, (iii) the instructions for the installation, (iv) the user manual of the SmartBear mobile app, (v) the user id to be used in the app, chosen from those associated with their institution, and (vi) the order in which to perform the tasks to counterbalance the executions using a balanced Latin square.

The participants were first instructed to install and configure the app on their Android devices through the provided documentation. Additionally, the instructions clarified that (1) the app used for testing would simulate interactions with monitoring devices; (2) all interactions would be logged; and (3) at the conclusion of each task, the app would automatically administer after-task questionnaires (ASQ), along with a final questionnaire (PSSUQ) to rate the overall experience across all tasks.

Before beginning the task executions, participants were advised to take a tour of the manual to become familiar with the app interface and to keep the manual easily accessible during task execution for reference if needed. Following this, the participants proceeded with the execution of the task scenarios as indicated in Section 2.6.

After completing each task, the subjects filled in the ASQ questionnaire and the included open-ended question for the elicitation of functionalities and the deepening of issues. Moreover, after the completion of all the tasks, each participant filled in the PSSUQ questionnaire and an additional item to keep track of the age of each participant. The total time for the execution of the user study can be summarized as follows:

- App installation—about 7 min.
- Review the user manual to familiarize yourself with the app interface—about 10 min.
- Task execution and task-level questionnaires repeated four times—about 28 minutes subdivided as follows:
  - task execution—about 2 min.
  - completion of task-level questionnaires—about 5 min.
- Poststudy questionnaires—about 10 min.

The entire procedure and the tools used have been administered in English since the participants involved in the task-based evaluation are all professionals accustomed to using English in their work practices.

2.8 | Collected Data and Evaluated Measures

Data have been collected anonymously to ensure no participant's activities and scores will be individually attributable. As mentioned earlier, the role of participants consisted of a technical assessment of the app by accomplishing the task scenarios and providing their impression by filling in the ASQ and PSSUQ questionnaires. In addition to such an explicit evaluation, each activity (and its respective timeframe) undertaken during the execution of each task was tracked in order to obtain a quantitative evaluation (in terms of effectiveness and efficiency).

For such a quantitative analysis aimed at understanding how users performed the tasks, the following data were collected:

- user id used to configure the app, user_id;
- age of the user, user_age;
- starting and ending timestamp for each task:
  - starting point—> opening the Smart Bear App; task_start_timestamp;
  - ending point —> all planned actions have been performed (end condition is different for each task), task_end_timestamp;
- clicked button and event timestamp for each task, to track all user actions and possible errors from the beginning to the end of the task and verify that the task is performed with the least number of clicks without navigating unnecessary app sections, button_clicked_event_timestamp.

To operationalize the interaction data, we defined the following specific constant values to categorize the click counts into discrete performance levels. In this context, the "right section" is the functional area of the application that hosts the specific features or information necessary to satisfy the requirements of a given task.

- the expected number of clicks to complete the task (14 for Task 1, 15 for Task 2, 32 for Task 3, and 4 for Task 4), number_of_clicks_expected_total;
- the number of clicks needed to reach the right app section, number_of_clicks_expected_to_right_section;
- the expected number of clicks to complete the task after reaching the right section, number_of_clicks_expected_forTask;

Additional metrics included in the evaluation were as follows:

- the number of clicks made by each user to reach the right section, number_of_clicks_to_right_section;
- the number of clicks made by each user after reaching the right section, clicks_after_right_section;
- the ratio between the number of clicks and *number_of_clicks_expected_total*, number_of_clicks_total_ratio;
- the ratio between *number_of_clicks_to_right_section* and *number_of_clicks_expected_to_right_section*, number_of_clicks_to_right_section_ratio;
- the ratio between *clicks_after_right_section* and *number_of_clicks_expected_forTask*, number_of_clicks_for_task_ratio;
- the time spent to complete a task, task_duration;
- the ratio between the time spent to complete a task and the shortest recorded time among the participants, task_duration_ratio;

These metrics were selected to provide insights into user interaction within the app during task execution and to capture different dimensions of usability. To quantitatively assess the

app's usability, we selected a set of metrics designed to distinguish between navigational effort and task-processing workload, both of which are critical factors for older adults. Specifically, the number of clicks to reach the "right section" indicates how effortlessly users navigate to the intended area of the app. This information is particularly relevant for older adults, as complex navigation paths can exacerbate cognitive load. Conversely, the number of clicks performed after the right section is reached reflects the operational complexity of specific tasks, the "depth of user interaction." By depth of user interaction, we refer to the intrasection workload required to complete the task, represented by the number of clicks and the engagement with different interface elements. We chose these metrics because they allow us to isolate both navigation efficiency and interaction depth, providing a granular view of where usability barriers occur, whether in navigating between app sections or in the interaction needed to execute the task. This distinction also allows for identifying if a failure is due to poor signposting (vision/cognition) or overly complex input requirements (motor skills/precision). The number of clicks made by participants was also compared with the expected total for each task and for each task phase. In this way, the ratio of total clicks made to expected total clicks evaluates overall efficiency; meanwhile, the ratio of clicks to reach the right section and the ratio of clicks after reaching the right section assess navigation and interaction efficiency more specifically. For the analysis of task duration, because no fixed timeframe was defined for each task, we used the shortest observed completion time as a conservative reference. Since all participants were first-time users, no one benefited from prior familiarity with the app. We therefore computed task_duration_ratio as each completion time divided by the minimum observed time, enabling cross-comparisons and revealing insights. The considered metrics (click count and task duration) and their normalization at the task level allowed different tasks to be compared correctly. This was needed to account for the inherent variations in task length and complexity. By addressing these structural differences, the resulting indicators provide a more precise representation of user effort, regardless of each task's intrinsic characteristics.

Regarding explicit feedback, to evaluate the user experience according to the chosen standardized scales, we collected the scores for the ASQ and the PSSUQ questionnaires. Specifically, after each task we collected the scores on the three items of the ASQ questionnaire, related to the easiness of execution (*easySatisfactionAnswer*), the amount of time to complete the task (*timeSatisfactionAnswer*), the level of support received throughout the process (*infoSatisfactionAnswer*). The total ASQ score is then obtained by averaging the three responses. Together with the ASQ score, in relation to each performed task, we also collected the participant's answer to this open-ended question: *Describe your overall experience with the app highlighting strengths or weaknesses and any suggestions for improvement*. For what concerns the feedback on the whole scenario, we collected the scores on the 16 items of the PSSUQ questionnaire, evaluating then the scores on its subscales, named SYSUSE (System Usefulness as the average scores of Questions 1-6), INFOQUAL(Information Quality as the average scores of Questions 7-12), INTERQUAL (Interface Quality as the average scores of Questions 13-15), and an open-ended question. The overall score is evaluated as the average of Questions 1–16. The PSSUQ questionnaire also included a final open-ended question, namely *Please write your suggestions to improve the app*, used to collect suggestions from the participants. For the computation of the average score, NA responses were handled according to the standard method outlined by the authors of the used questionnaires [23]. The NA responses were treated as nonresponses and excluded from the final average score because if an item is not appropriate for a specific task and users choose not to answer it, the questionnaire is still useful. Table 1 summarizes collected and evaluated data.

With regard to the operationalization of the usability dimensions, the computation procedures adopted to evaluate effectiveness and efficiency are detailed in the following. For what concerns the effectiveness, in our study, the tasks did not involve producing "correct" or "incorrect" responses, but rather performing concrete actions that demonstrated the ability to use the app as intended. Therefore, task completion alone was considered a sufficient indicator and effectiveness was then calculated as the ratio between the total number of tasks completed by all users and the total number of tasks assigned ($4 \times N_{users}$) (eq. 1 in Table 2). Since each task is representative of a specific app functionality, the effectiveness has also been evaluated for each of the considered subsections of the app by evaluating each task completion rate (the percentage of users that were able to successfully complete the given task) (eq. 2 in Table 2).

Efficiency was assessed in terms of navigation effort, operationalized through the ratio between expected and observed user interactions, eq. 3 in Table 2. We did not consider efficiency in terms of time spent on the task due to the absence of a reliable baseline for expected task duration. Nevertheless, task duration

TABLE 1 | Quantitative and qualitative collected data and corresponding considered measures.

Collected Quantitative Data	**Quantitative Measures**
user_id	number_of_clicks_total_ratio
user_age	number_of_clicks_to_right_section_ratio
task_start_timestamp	number_of_clicks_for_task_ratio
task_end_timestamp	task_duration
number_of_clicks	task_duration_ratio
Collected Qualitative Data	**Qualitative Measures**
ASQ scores	ASQ overall score
PSSUQ scores (16 items)	PSSUQ SysUse score
ASQ open questions	PSSUQ InfoQual score
PSSUQ open questions	PSSUQ InterQual score, PSSUQ overall score

was recorded as discussed in Section 2.8 to support a cross-task comparison, as further specified in Section 3.1.3.4.

3 | Results and Discussion

3.1 | Quantitative Evaluation

Among the 50 participants, only 29 completed all the tasks (3 users completed just one task, 6 users completed two tasks, 12 users completed 3 tasks, and 29 users completed 4 tasks). Figure 3 displays the number of users who completed the tasks. We evaluated the app in the task-based scenario in terms of quantitative measures of effectiveness and efficiency.

3.1.1 | Effectiveness

Overall effectiveness, across 167 completed task observations (47 users completed Task 1, 36 Task 2, 42 Task 3, and 42 Task 4), was 83.5%, with Task 2 exhibiting the lowest completion rate (72%), as shown in Figure 4. These results suggest usability issues primarily related to diet-management interactions.

3.1.2 | Efficiency

Overall, navigation efficiency across all tasks was 45.2%, as shown in Figure 5, indicating a substantial deviation from the optimal interaction paths. Efficiency varied across tasks, with Task 1 exhibiting the highest navigation efficiency and Task 4 the lowest. These results highlight difficulties in locating the appropriate sections and navigating between the features, particularly for tasks involving multi-step interactions.

3.1.3 | Differences Among the Tasks

This subsection describes the differences in interaction workload and duration across tasks. We evaluate (i) overall navigation effort, (ii) effort sin reaching the target section, (iii) effort within target section, and (iv) task duration (best case normalization). Overall, Task 1 consistently required fewer interactions, whereas Tasks 2 and 4 exhibited higher workload and variability.

3.1.3.1 | Navigation Effort. Table 3 and Figure 6 summarize the total-clicks ratio across tasks. Task 1 showed the lowest

TABLE 2 | Equations considered to evaluate effectiveness and efficiency.

Ref.	Dimension	Equation
Eq. 1	Effectiveness	$\text{Effectiveness} = \text{N Completed tasks by all users} / 4 \times \text{N users}$
Eq. 2		$\text{Task}_i \text{ effectiveness} = \text{N Task}_i \text{ completed by all users} / \text{N users}$
Eq. 3	Efficiency	$\text{Navigation efficiency} = \frac{\text{expected clicks total}}{\text{observed clicks}}$

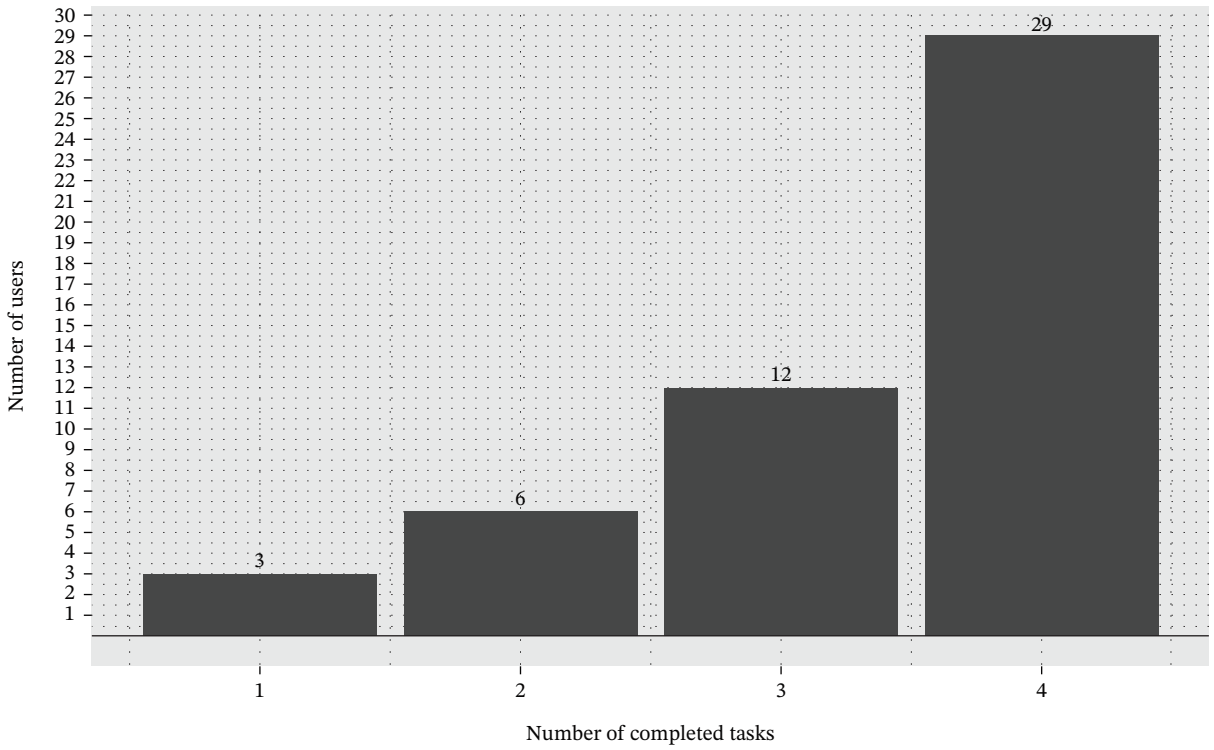

FIGURE 3 | Tasks completion.

effort (mean = 1.469), whereas Tasks 2 and 4 required substantially more interactions (means = 2.594 and 2.609, respectively), with higher variability.

A Kruskal–Wallis test confirmed a significant effect of task on navigation effort ($\chi^2(3) = 37.583$, $p = 3.46 \times 10^{-8}$). Post hoc Dunn tests (Appendix A.1) indicated that Task 1 differed significantly from Tasks 2–4, whereas no significant differences emerged among Tasks 2–4.

3.1.3.2 | Navigation Effort to the Right Section. Navigation effort to reach the target app section is summarized in Table 4 and Figure 7. Task 1 showed the highest effort to locate the correct section (mean = 3.349), whereas Task 4 required the fewest interactions (mean = 1.297).

A Kruskal–Wallis test indicated a significant effect of task on this metric ($\chi^2(3) = 18.356$, $p = 3.714 \times 10^{-4}$). Post hoc Dunn tests with Bonferroni correction (Appendix A.2) showed significant differences for Task 1 versus Task 4 and Task 3 versus Task 4, whereas the remaining comparisons were not significant after correction.

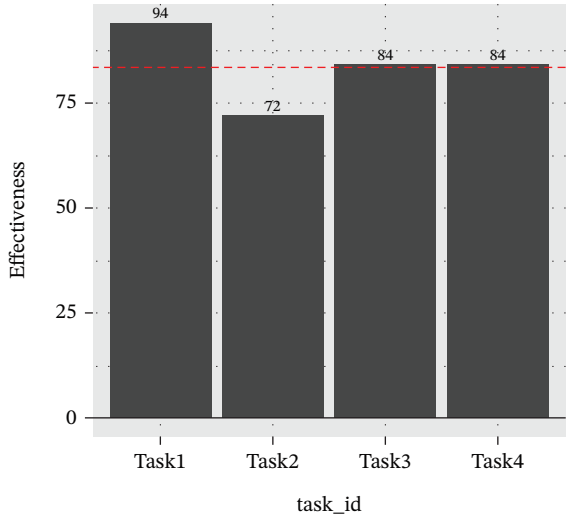

FIGURE 4 | Effectiveness results per task.

3.1.3.3 | Navigation Effort Within the Right Section. Effort after reaching the correct section is reported in Table 5 and Figure 8. Task 1 showed the lowest within-section effort (mean = 1.201), whereas Tasks 2 and 4 required substantially more interactions (means = 2.613 and 2.386).

A Kruskal–Wallis test confirmed a significant task effect ($\chi^2(3) = 69.496$, $p = 5.473 \times 10^{-15}$). Dunn post hoc tests with Bonferroni correction (Appendix A.3) indicated that Task 1 differed significantly from Tasks 2–4, whereas Tasks 2–4 did not differ significantly from each other after correction.

3.1.3.4 | Task Duration. Task duration was analyzed as a descriptive measure for cross-task comparisons using best case normalization (minimum observed time), defined as the ratio of each observed completion time to the shortest observed completion time (*task_duration_ratio*), as discussed in Section 2.8. Descriptive statistics for each task are reported in Table 6, and the corresponding distributions are shown in Figure 9.

TABLE 3 | Navigation effort per task: mean, standard deviation (SD), and 95% confidence intervals (CI).

Task	Mean (*m*)	SD	95% CI (lower)	95% CI (upper)
Task 1	1.469	0.463	1.332	1.606
Task 2	2.594	0.988	2.260	2.929
Task 3	2.002	0.604	1.809	2.196
Task 4	2.609	1.755	2.040	3.178

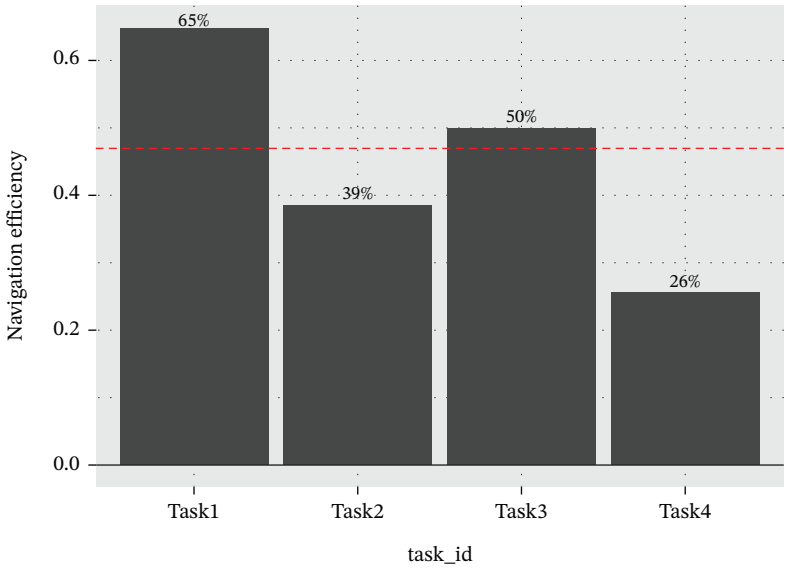

FIGURE 5 | Navigation efficiency results per task.

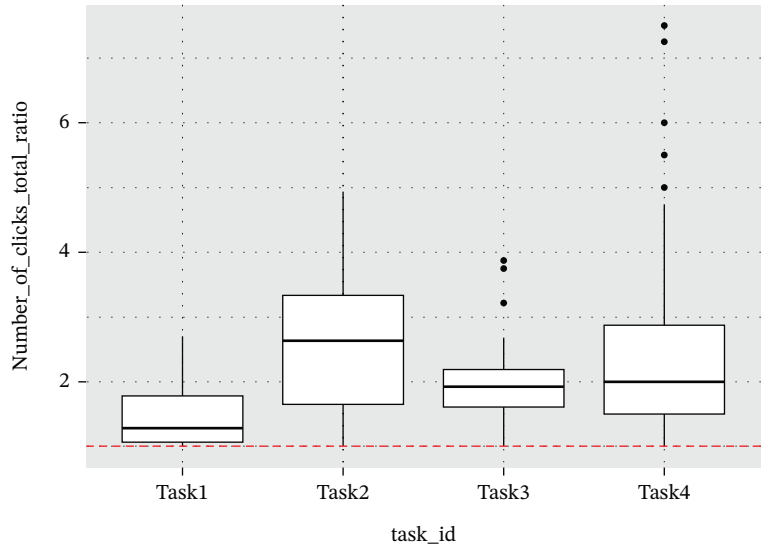

FIGURE 6 | Data distribution with respect to *number_of_clicks_total_ratio*.

TABLE 4 | Navigation effort to the right section per task: mean, standard deviation (SD), and 95% confidence intervals (CI).

Task	Mean (*m*)	SD	95% CI (Lower)	95% CI (Upper)
Task 1	3.349	3.779	2.186	4.512
Task 2	2.029	1.823	1.402	2.655
Task 3	2.118	1.066	1.746	2.490
Task 4	1.297	0.777	1.038	1.556

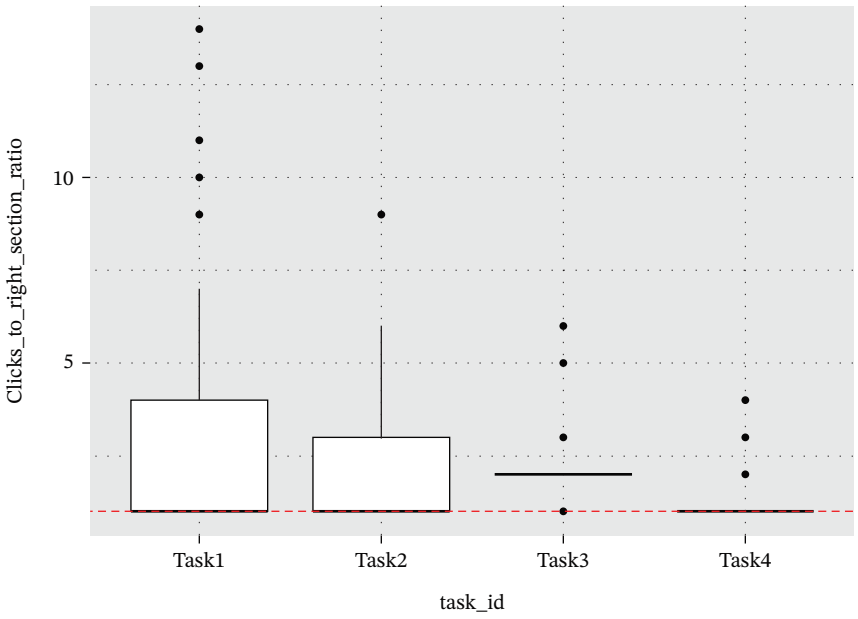

FIGURE 7 | Data distribution with respect to *number_of_clicks_to_right_section_ratio*.

Overall, Tasks 1-3 exhibited comparable duration profiles, whereas Task 4 showed a higher mean duration and greater variability. However, a Kruskal–Wallis test did not reveal statistically significant differences among tasks ($\chi^2(3) = 4.4424$, $p = 0.2175$), indicating that task duration differences should be interpreted cautiously in this expert-based evaluation setting (Appendix A.4).

Table 7 provides a comparative overview of performance, workload, and duration metrics across tasks.

3.2 | Qualitative Evaluation

This section synthesizes the qualitative data from the posttask and postscenario questionnaires and the feedback from participants in the open-ended questions. This analysis complements the quantitative findings and helps to identify the usability concerns from the participants, which focus primarily on navigation, the clarity of interaction, and visual accessibility, which informs the redesign of the application (Section 4).

3.2.1 | After Task Evaluation

The ASQ was used by the participants to give feedback on their experience. The ASQ overall scores showed there was an overall moderate-to-high satisfaction across all of the tasks, and Task 2 was the one that received the best satisfaction scores in all the categories of satisfaction from ease of use, quality of information, and time saving (Figure 10 and Table 8).

The open-ended feedback from the participants showed that there was low satisfaction and concern regarding clarity of navigation and how they are to interact with the system. These concerns from the participants were used to set the redesign priorities in Section 4.

3.2.2 | Postscenario Evaluation

The PSSUQ was used to analyze the perception of overall usability. The system received generally positive evaluations. The evaluation of usefulness (SYSUSE = 3.21), INFOQUAL (INFOQUAL = 3.36), and INTERQUAL (INTERQUAL = 2.92) were positive. The overall system usability score is 3.16 (Figure 11 and Table 9).

However, despite the overall positive perception, feedback indicated that there is still navigation and visual accessibility on smaller devices to improve. These aspects are further examined through qualitative feedback and informed the redesign process.

3.2.3 | Qualitative Feedback and Sentiment Analysis

Open-ended responses collected through ASQ and PSSUQ were analyzed to identify recurring usability themes. A BERT-based sentiment analysis was applied to summarize the overall polarity of the feedback (Table 10).

TABLE 5 | Navigation effort within the right section per task: mean, standard deviation (SD), and 95% confidence intervals (CI).

Task	Mean (*m*)	SD	95% CI (lower)	95% CI (upper)
Task 1	1.201	0.227	1.133	1.268
Task 2	2.613	1.076	2.249	2.977
Task 3	1.944	0.567	1.765	2.123
Task 4	2.386	1.436	1.914	2.858

TABLE 6 | Task duration per task: mean, standard deviation (SD), and 95% confidence intervals (CI).

Task	Mean (*m*)	SD	95% CI (lower)	95% CI (upper)
Task 1	6.651	6.992	4.550	8.752
Task 2	5.215	4.366	3.641	6.789
Task 3	6.143	6.186	4.191	8.096
Task 4	12.520	15.483	7.568	17.472

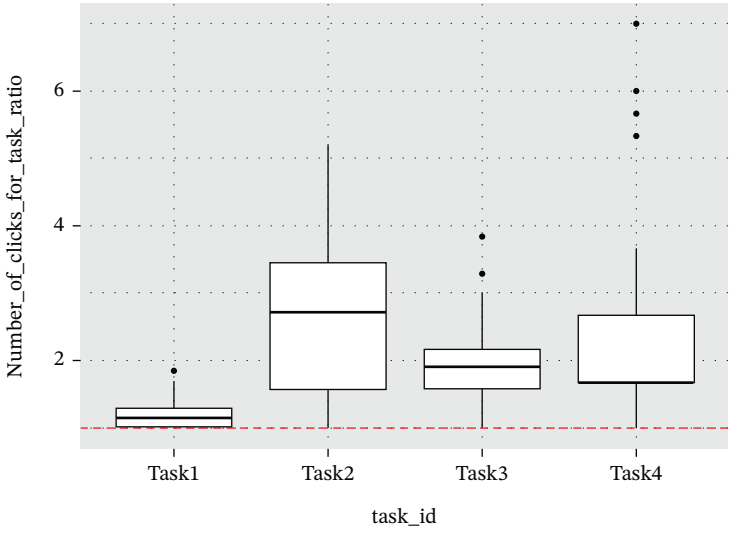

FIGURE 8 | Data distribution with respect to *number_of_clicks_for_task_ratio*.

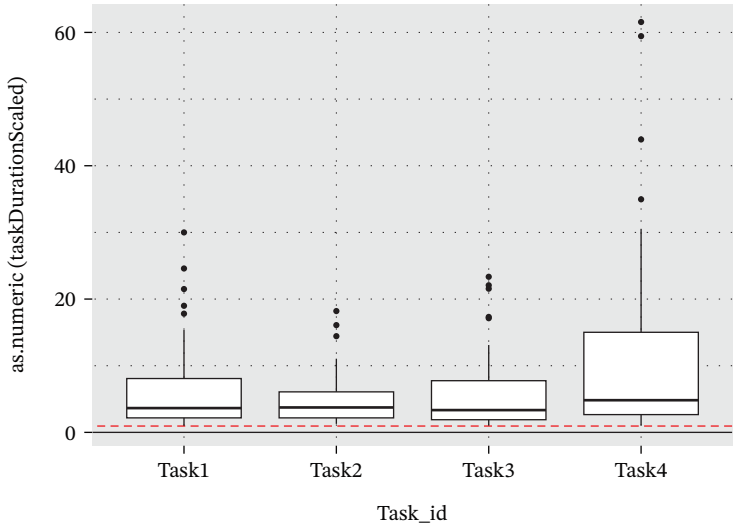

FIGURE 9 | Data distribution with respect to *task_duration_ratio*.

TABLE 7 | Comparison among the four tasks. Higher values in effectiveness and efficiency indicate better performance, whereas higher values in effort and duration indicate, respectively, greater workload and longer completion time.

Metric	Task 1	Task 2	Task 3	Task 4
Positive performance indicators				
Effectiveness (%)	94	72	84	84
Navigation efficiency (%)	65	39	50	26
Workload and duration indicators				
Navigation effort	1.469 ± 0.463	2.594 ± 0.988	2.002 ± 0.604	2.609 ± 1.755
Navigation effort to the right section	3.349 ± 3.779	2.029 ± 1.823	2.118 ± 1.066	1.297 ± 0.777
Navigation effort within the right section	1.201 ± 0.227	2.613 ± 1.076	1.944 ± 0.567	2.386 ± 1.436
Task duration	6.651 ± 6.992	5.215 ± 4.366	6.143 ± 6.186	12.520 ± 15.483

Across tasks, negative comments were primarily associated with navigation difficulties, nonintuitive icons, and layout issues on smaller smartphone screens, whereas positive feedback highlighted the usefulness of core functionalities and the clarity of content once accessed. Task-specific feedback revealed that navigation issues were particularly prominent in Tasks 1 and 3, aligning with the quantitative findings on navigation effort and efficiency (Section 3.1.3).

Representative examples of positive and negative comments are reported in Appendix B. The consolidated qualitative findings directly informed the redesign decisions summarized in Section 4.

4 | App Redesign After the Usability Evaluation

Following the evaluation results and the collected comments, the app underwent a comprehensive redesign, incorporating significant improvements based on user feedback. Specifically, quantitative and qualitative findings were treated as complementary components of the evaluation rather than as independent analyses. Although quantitative metrics pinpointed *where* usability issues were discerned (e.g., tasks with increased navigation complexity and/or prolonged completion times), the constructive feedback collected through the ASQ and PSSUQ questionnaires supported in explaining *why* the issues occurred. Table 11 summarizes how quantitative usability metrics and qualitative user feedback were jointly used to modify the design.

The redesign phase was directly informed by the combined interpretation of quantitative indicators and qualitative user feedback. This integration enabled the design team to translate empirical usability findings into targeted and actionable design changes. User comments directly informed revisions of the app's content, visualization, and interaction, generally and, more specifically, for the three app sections involved in the study. Noteworthy enhancements encompassed refined UI elements, clearer indicators for scrollability, resolution of visualization bugs that occurred on some smartphones (see Figure 12 on the left), improvement of visualization preventing the keyboard from obstructing screen content, and introduction of a more dynamic display of screens. Contextually,

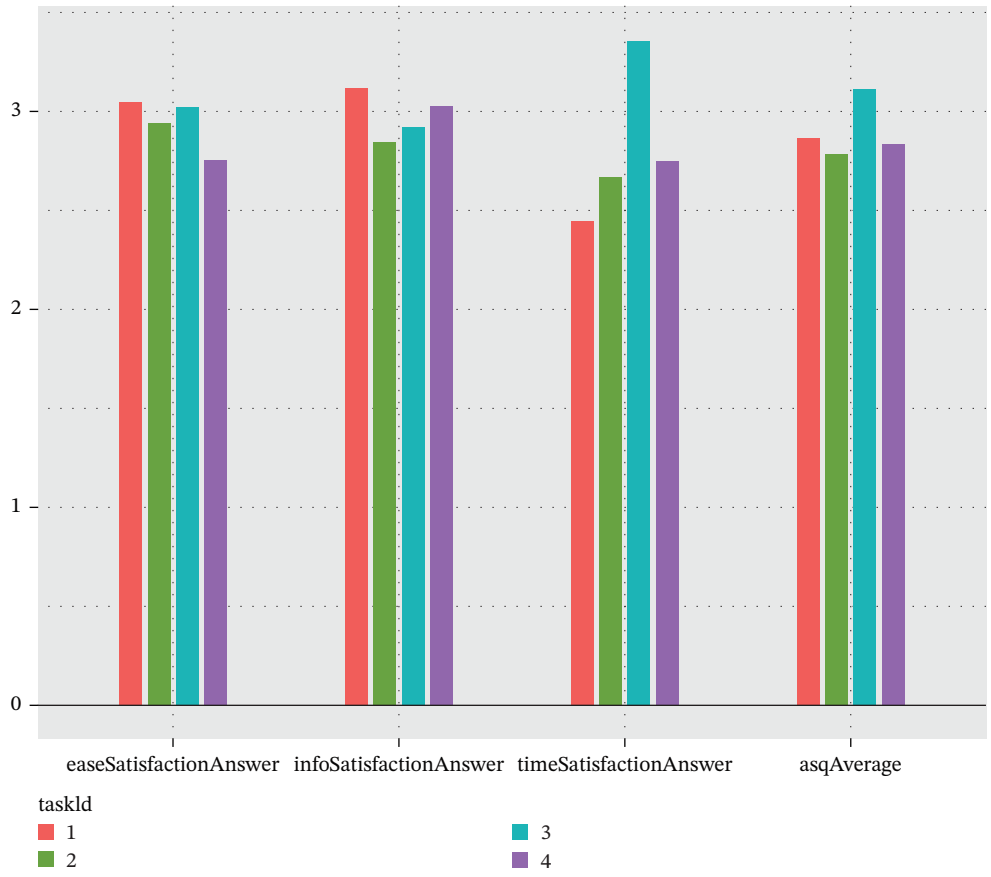

FIGURE 10 | ASQ scores.

TABLE 8 | ASQ scores.

Score_type	Task 1 score	Task 2 score	Task 3 score	Task 4 score
ASQ_easySatisfaction_score	3.04	2.94	3.02	2.76
ASQ_infoSatisfaction_score	3.12	2.85	2.92	3.03
ASQ_timeSatisfaction_score	2.45	2.67	3.36	2.75
ASQ_overall_score	2.87	2.82	3.10	2.84

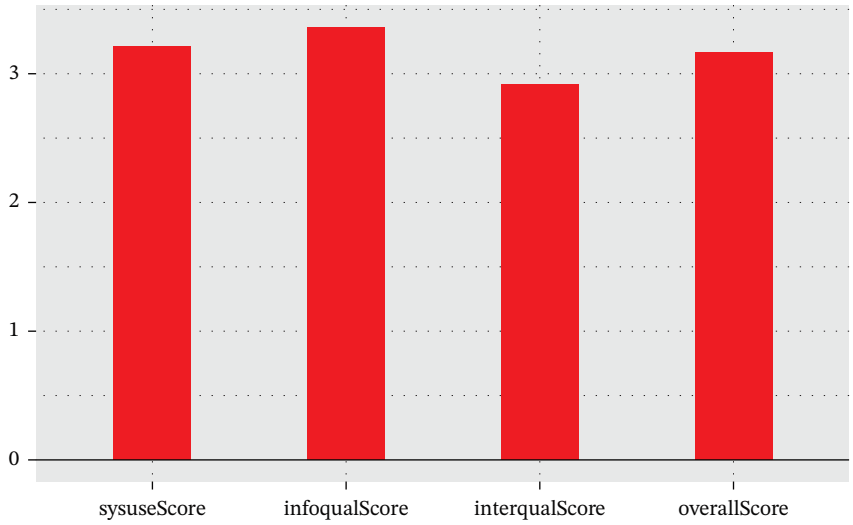

FIGURE 11 | PSSUQ scores.

the user manual received an upgrade with clearer instructions to facilitate more straightforward navigation, contributing to a more user-friendly interaction with the app.

Significant enhancements were implemented in the section of the app dedicated to diet information, driven by insights gleaned from user questionnaires. The process of recording daily food intake underwent a thorough streamlining, eliminating redundant screens, and incorporating numerous user recommendations to elevate the overall user experience. In the original design, users would tap on the icon of the desired food category to enter the number of portions, triggering a pop-up dialog. Within this dialog, users could adjust portion amounts by tapping on corresponding buttons and then submitting their input. However, the redesigned version presents all food categories with buttons to adjust portions directly, eliminating the need for a pop-up dialog. Users can seamlessly tap buttons and submit their input without leaving the diet app (see Figure 12 on the middle). Additionally, in the diet questionnaire, where each question initially occupied the full screen, users faced difficulty in realizing the need to scroll to the next question. The redesigned version addresses this by featuring a thicker and always visible scroll bar on the right side of the screen, clearly indicating scrollability. After providing an answer, a button with a downward arrow appears at the bottom right corner, enabling users to effortlessly navigate to the next question. Although users can still manually scroll to the next question without using the button, this intuitive design improvement ensures a more user-friendly interaction with the app (see Figure 12 on the right).

Not all of the redesign suggestions were implemented because the app design and its features must comply with clinical guidelines, safety requirements, and operational constraints. Design examples include the Heart Rate, Weight, and Activity features, which are implemented in multiple areas of the app for users with different access profiles. No vegan options are present in the diet module because of adherence to Mediterranean diet criteria and the Mediterranean Diet Score questionnaire. Although multiple users suggested the ability to pause and resume exercises within the rehabilitation section, this omission aligns with guidelines stating exercises must be completed in one uninterrupted session. These guidelines also require continuous monitoring to mitigate injury risk associated with the exercises and include self-assessment questions to monitor physical status.

Overall, the redesign process integrated suggestions aimed at improving usability while preserving safety, adherence to clinical guidelines, and consistency with the app's functional and clinical requirements.

TABLE 9 | PSSUQ scores.

Score_type	Score_value
PSSUQ_sysuse_score	3.21
PSSUQ_infoqual_score	3.36
PSSUQ_interqual_score	2.92
PSSUQ_overall_score	3.16

5 | Conclusive Remarks, Limitations and Future Directions

In this work, we discuss the evaluation results of a mobile app designed to assist older adults with their healthcare-seeking behavior. The paper outlined a task-based user study executed by 50 ICT experts from the digital health domain to obtain feedback on app sections and their general functionalities and to inform usability improvements. Specifically, the participants completed four representative tasks covering diet, medication, and rehabilitation features. A mixed-method evaluation protocol was employed. Usability was assessed, combining performance metrics, standardized questionnaires (ASQ and PSSUQ), qualitative comments, and interaction logs.

The results show that the app was appreciated overall, especially for its clear and pleasant look and feel. Contextually, the evaluation highlighted areas for improvement, particularly in navigation efficiency and accessibility of functionalities, leading to a comprehensive redesign of the app. Overall, the study demonstrates how a combined analysis of quantitative usability metrics and qualitative user feedback can effectively identify critical interaction issues and inform targeted design improvements in mHealth applications for older adults.

The findings align with prior work and support the use of an experts-first strategy to improve inclusivity while maintaining methodological rigor in eHealth usability studies. Several

TABLE 10 | Sentiment analysis on participants' comments.

	Positive		Negative		Neutral	
	Number	Value	Number	Value	Number	Value
ASQ Task 1	19	0.404687	25	0.531484	3	0.063830
ASQ Task 2	19	0.514018	15	0.404901	3	0.081081
ASQ Task 3	12	0.279919	23	0.553414	7	0.166667
ASQ Task 4	24	0.571163	11	0.262171	7	0.166667
PSSUQ scenario	15	0.318427	26	0.553913	6	0.127660
External comments	18	0.221393	64	0.778607	0	0

TABLE 11 | Evaluation findings and redesign actions.

App section/task	Quantitative evidence	Qualitative feedback (Appendix B)	Redesign actions implemented
Diet questionnaire (Task 1)	High navigation effort to reach task section; moderate completion rate	Effort in locating the task entry point; uncertainty about icons, scrollability and portion quantities	Inclusion of more explicit entry points and labeled icons; addition of a persistent, thicker scroll bar; addition of an explicit navigation button to navigate through the questions
Daily food intake logging (Task 2)	High effectiveness but low efficiency in navigation	Need to reduce interaction steps to improve efficiency; confusion caused by pop-up dialogs	Removal of pop-up dialogs; modification button added to the main screen to increase the efficiency of interactions
Rehabilitation exercises (Task 3)	High navigation effort; high interaction variability	Difficulty in locating the rehabilitation section; too long videos; lack of pause/resume and indexing features	Enhanced video controls and improved navigation; improved structure and feedback for exercise flow and interaction
Medication management (Task 4)	Longest task duration; high variability	Unclear timeframes; difficulty in distinguishing start and end of the tasks; terminology issues	Improved timeframe visualization; clearer color coding; refined terminology and interaction cues
Global UI and navigation	Consistently low navigation efficiency across tasks	Non-intuitive icons; hidden buttons; poor visibility on some devices	Refined UI elements; increased button visibility; resolution of visualization bugs; responsive layout improvements

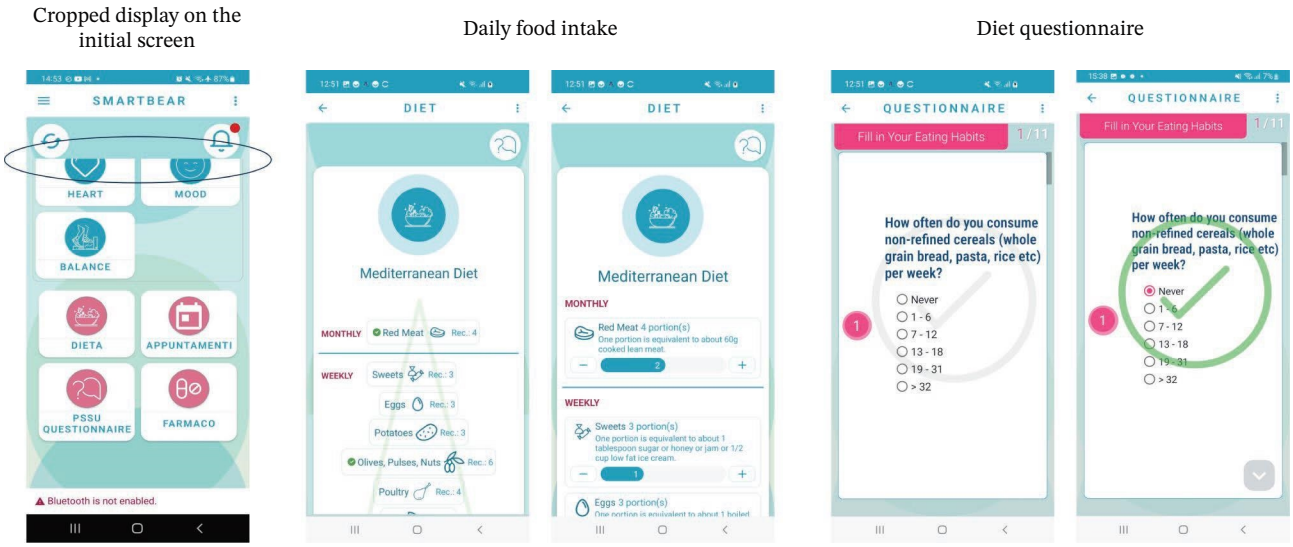

FIGURE 12 | Redesign of the app. On the left is the cropped display that occurred on some smartphones. The previous and the redesigned version of the daily food intake are in the middle. On the right is the improved navigability of the diet questionnaire, with a clear indication of the need to scroll the page.

studies suggest that expert-based methods are a helpful first step in reducing participant burden and improving user-based testing [29, 30] In practice, the approach can be applied to digital health and assistive technologies for older adults, where protecting participant well-being is of utmost importance.

5.1 | Limitations

Although the performed study has provided valuable insights to improve the app, it is important to acknowledge its limitations to fully understand the context and the potential implications

of our findings. The study involved participants from the project partners. This may have given them an advantage in understanding the app due to potential prior knowledge. As with any study based on user feedback, there is a degree of subjectivity involved. Despite our best efforts to standardize the evaluation criteria, personal biases and perceptions may have influenced the responses.

Moreover, the evaluation involved ICT experts rather than members of the target older-adult population. Although experts are well suited to identifying structural usability issues, they cannot fully replicate the interaction strategies, cognitive load, or physical constraints experienced by older adults.

The metrics used in this study were carefully selected to capture different aspects of usability. However, these metrics may not fully reflect real-world interaction patterns of the target population, as actual users might interact differently with the app due to differences in familiarity, cognitive strategies, or physical interaction capabilities. This is an inherent limitation of an expert-first evaluation and should be considered when interpreting the findings.

Accordingly, results should be interpreted as evidence of early-stage usability issues and design directions, rather than as a definitive measure of usability for the older-adult population. Acknowledging the study limitations provides context for interpreting the findings and motivates future work.

5.2 | Future Directions

Conducting an initial expert-based evaluation is particularly important before releasing the app to end users, given the vulnerable characteristics of the target population. The expert-based evaluation described in this paper represents the first step of a broader, iterative usability assessment strategy. As a result of this investigation, the app underwent a redesign process.

Subsequent steps involve assessing the redesigned app by its target population under real use conditions through a set of large-scale pilots in five different countries (Portugal, Italy, Greece, France, and Romania). The future evaluations target real user interaction data, usability, and technology acceptance, integrating metrics tailored to the specific needs of aging individuals and the healthcare context.

Future research should explore hybrid evaluation frameworks that mix expert reviews with participatory methods. This can help find a balance between efficiency, inclusivity, and real-world validity across different user groups.

Author Contributions

A.A., G.C., L.G., and G.D.P. contributed to the definition of the usability framework and the design of the study. A.A. and G.C. conducted the study by defining the task scenarios and analyzed the results. L.G. contributed to the conceptualization of the work and supervised the study. S.A. and G.Z. contributed in the design and implementation of the Smart Bear app and its redesign according to the evaluation results. A.A. and G.C. contributed equally as first authors.

Acknowledgments

We would like to express our sincere gratitude to all the participants who took part in this study. Their invaluable contributions and cooperation were essential to the success of this research.

Funding

This work is funded by the European Commission through the EU Project Smart Bear (Grant Agreement No. 857172/H2020-SC1-FA-DTS-2018-2) [15].

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

1. K. Oyibo, K. Wang, and P. P. Morita, "Using Smart Home Technologies to Promote Physical Activity Among the General and Aging Populations: Scoping Review," *Journal of Medical Internet Research* 25 (2023): e41942, https://doi.org/10.2196/41942.

2. D. Kim, "Can Healthcare Apps and Smart Speakers Improve the Health Behavior and Depression of Older Adults? A Quasi-Experimental Study," *Frontiers in Digital Health* 5 (2023): 1117280, https://doi.org/10.3389/fdgth.2023.1117280.

3. Z. Galavi, M. Montazeri, and R. Khajouei, "Which Criteria Are Important in Usability Evaluation of mHealth Applications? An Umbrella Review," *BMC Medical Informatics and Decision Making* 24, no. 1 (2024): 24, https://doi.org/10.1186/s12911-024-02738-2.

4. A. Khamaj, and A. M. Ali, "Examining the Usability and Accessibility Challenges in Mobile Health Applications for Older Adults," *Alexandria Engineering Journal* 102 (2024): 179–191, https://doi.org/10.1016/j.aej.2024.06.002.

5. Y. S. Woo, G. I. Shin, and H. Y. Park, "Empowering Older Adults: Evaluating the Impact of a Smartphone Education App on Independent Living," *Frontiers in Public Health* 12 (2024): 12, https://doi.org/10.3389/fpubh.2024.1403978.

6. S. Pahlevanynejad, S. R. Niakan Kalhori, M. R. Katigari, and R. H. Eshpala, "Personalized Mobile Health for Elderly Home Care: A Systematic Review of Benefits and Challenges," *International Journal of Telemedicine and Applications* 2023 (2023): 5390712, https://doi.org/10.1155/2023/5390712.

7. G. Facchinetti, G. Petrucci, B. Albanesi, M. G. De Marinis, and M. Piredda, "Can Smart Home Technologies Help Older Adults Manage Their Chronic Condition? A Systematic Literature Review," *International Journal of Environmental Research and Public Health* 20, no. 2 (2023): 1205, https://doi.org/10.3390/ijerph20021205.

8. J. Zhu, H. Weng, P. Ou, and L. Li, "Use and Acceptance of Smart Elderly Care Apps Among Chinese Medical Staff and Older Individuals: Web-Based Hybrid Survey Study," *JMIR Formative Research* 7, no. 1 (2023): e41919, https://doi.org/10.2196/41919.

9. Y. Zhang, E. W. Lee, and W. P. Teo, "Health-Seeking Behavior and Its Associated Technology Use: Interview Study Among Community-Dwelling Older Adults," *JMIR Aging* 6 (2023): e43709, https://doi.org/10.2196/43709.

10. H. M. Asiri, A. M. Asiri, H. F. Alruwaili, and J. Almazan, "A Scoping Review of Different Monitoring-Technology Devices in Caring for Older Adults With Cognitive Impairment," *Frontiers in Public Health* 11 (2023): 1144636, https://doi.org/10.3389/fpubh.2023.1144636.

11. Q. Wang, J. Liu, L. Zhou, et al., "Usability Evaluation of mHealth Apps for Elderly Individuals: A Scoping Review," *BMC Medical Informatics and Decision Making* 22, no. 1 (2022): 317, https://doi.org/10.1186/s12911-022-02064-5.

12. I. Sinabell, and E. Ammenwerth, "Challenges and Recommendations for eHealth Usability Evaluation With Elderly Users: Systematic Review and Case Study," *Universal Access in the Information Society* 23, no. 1 (2024): 455–474, https://doi.org/10.1007/s10209-022-00949-w.

13. S. Sharma, and B. A. Kumar, "A Systematic Review of User-Based Usability Testing Practices in Self-Care mHealth Apps," *Digital Health* 11 (2025): 11, https://doi.org/10.1177/20552076251374184.

14. M. Gomez-Hernandez, X. Ferre, C. Moral, and E. Villalba-Mora, "Design Guidelines of Mobile Apps for Older Adults: Systematic Review and Thematic Analysis," *Journal of Medical Internet Research Mhealth Uhealth* 11, no. 8 (2023): , https://doi.org/10.2196/43186e43186.

15. "Smart Bear Project," https://www.smart-bear.eu/; .[Online; accessed 23-10 2023]..

16. "Smart 4Health Project," https://smart4health.eu/; .[Online; accessed 23-10 2023].

17. "Holobance Project," https://holobalance.eu/; .[Online; accessed 23-10 2023].

18. I. Kouris, E. Vellidou, and D. Koutsouris, "SMART BEAR: A Large Scale Pilot Supporting The Independent Living Of The Seniors In A Smart Environment," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 2020 (2020): 5777–5780, https://doi.org/10.1109/EMBC44109.2020.9176248.

19. V. Peretokin, I. Basdekis, I. Kouris, et al., "Overview of the SMART-BEAR Technical Infrastructure," in *In Proceedings of the 8th International Conference on Information and Communication Technologies for Ageing Well and e-Health* (SCITEPRESS, 2022), 117–125, https://doi.org/10.5220/0011082700003188.

20. A. Cristiano, S. De Silvestri, S. Musteata, et al., *IoT Platform for Ageing Society: the SMART BEAR Project: Smart Big Data Platform to Offer Evidence-based Personalised Support for Healthy and Independent Living at Home* (IARIA, 2021)https://air.unimi.it/retrieve/dfa8b9a9-f963-748b-e053-3a05fe0a3a96/etelemed_2021_2_60_40087.pdf.

21. International Organization for Standardization, *Ergonomics of Human-System Interaction – Part 11: Usability: Definitions and Concepts (ISO 9241-11:2018)*. (ISO, 2018).

22. W. Alsabhan, *Designing a Human-Centred, Mobile Interface to Support Real-Time Flood Forecasting and Warning System* (PhD Thesis. Brunel University London, 2016).

23. J. R. Lewis, "IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use," *International Journal of Human-Computer Interaction* 7, no. 1 (1995): 57–78, https://doi.org/10.1080/10447319509526110.

24. J. Nielsen, "The Usability Engineering Life Cycle," *Computer* 25, no. 3 (1992): 12–22, https://doi.org/10.1109/2.121503.

25. J. Nielsen, *Usability Engineering* (Morgan Kaufmann, 1993), https://doi.org/10.1016/B978-0-08-052029-2.50009-7.

26. L. Faulkner, "Beyond the Five-User Assumption: Benefits of Increased Sample Sizes in Usability Testing," *Behavior Research Methods, Instruments, & Computers* 35, no. 3 (2003): 379–383, https://doi.org/10.3758/BF03195514.

27. W. Hwang, and G. Salvendy, "Number of People Required for Usability Evaluation: The 10±2 Rule," *Communications of the ACM* 53, no. 5 (2010): 130–133, https://doi.org/10.1145/1735223.1735255.

28. M. Schmettow, "Sample Size in Usability Studies," *Communications of the ACM* 55, no. 4 (2012): 64–70, https://doi.org/10.1145/2133806.2133824.

29. A. Silva, A. I. Martins, H. Caravau, et al., "Experts Evaluation of Usability for Digital Solutions Directed at Older Adults," in *DSAI '20: Proceedings of the 9th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-Exclusion* (Association For Computing Machinery, 2020), 174–181.

30. L. W. Peute, G. A. Wildenbos, T. Engelsma, et al., "Overcoming Challenges to Inclusive User-Based Testing of Health Information Technology With Vulnerable Older Adults: Recommendations From a Human Factors Engineering Expert Inquiry," *Yearbook of Medical Informatics* 31, no. 1 (2022): 074–081, https://doi.org/10.1055/s-0042-1742499.

31. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2021).

Appendix A. Detailed Statistical Analysis

The statistical analysis was performed using the software R [31].

A.1 Navigation Effort: Statistical Details

Outlier Handling

For the *number_of_clicks_total_ratio*, outliers were removed using Tukey's boxplot rule, excluding values below $Q1 - 1.5 \times \text{IQR}$ or above $Q3 + 1.5 \times \text{IQR}$, where $\text{IQR} = Q3 - Q1$. After outlier removal, the number of observations was reduced from 167 to 161.

Inferential Statistics

Task-level differences were examined using a Kruskal–Wallis test as the normality assumption was violated (Shapiro test) and homogeneity of variances was not present. The analysis revealed a statistically significant effect of task on navigation effort ($\chi^2(3) = 37.583$, $p = 3.463 \times 10^{-8}$).Post Hoc Analysis

Pairwise comparisons were conducted using Dunn's test with Bonferroni correction for multiple comparisons. The results are reported below:

- Task 1 versus Task 2: $z = 5.78$, $p = 7.68 \times 10^{-9}$, $p_{adj} = 4.61 \times 10^{-8}$ $(* \ * \ * \ *)$
- Task 1 versus Task 3: $z = 3.89$, $p = 1.00 \times 10^{-4}$, $p_{adj} = 6.01 \times 10^{-4}$ $(* \ * \ *)$
- Task 1 versus Task 4: $z = 4.25$, $p = 2.10 \times 10^{-5}$, $p_{adj} = 1.26 \times 10^{-4}$ $(* \ * \ *)$
- Task 2 versus Task 3: $z = -1.93$, $p = 5.32 \times 10^{-2}$, $p_{adj} = 3.19 \times 10^{-1}$
- Task 2 versus Task 4: $z = -1.55$, $p = 1.20 \times 10^{-1}$, $p_{adj} = 7.20 \times 10^{-1}$
- Task 3 versus Task 4: $z = 0.377$, $p = 7.06 \times 10^{-1}$, $p_{adj} = 1$

A.2 Navigation Effort to the Right Section: Statistical Details

Outlier Handling

Outliers for *number_of_clicks_to_right_section_ratio* were removed using Tukey's rule ($Q1 - 1.5 \times \text{IQR}$, $Q3 + 1.5 \times \text{IQR}$), reducing observations to 149.

Inferential Statistics

Since normality assumptions were not met, task differences were assessed using Kruskal–Wallis, showing a significant effect ($\chi^2(3) = 18.356$, $p = 3.714 \times 10^{-4}$).

Post Hoc Analysis

Post hoc Dunn tests with Bonferroni correction yielded as follows:

- Task 1 versus Task 2: $z = -1.24$, $p = 0.216$, $p_{adj} = 1$
- Task 1 versus Task 3: $z = 1.11$, $p = 0.268$, $p_{adj} = 1$
- Task 1 versus Task 4: $z = -3.14$, $p = 0.00167$, $p_{adj} = 0.01$ $(*)$
- Task 2 versus Task 3: $z = 2.23$, $p = 0.0261$, $p_{adj} = 0.156$
- Task 2 versus Task 4: $z = -1.80$, $p = 0.0726$, $p_{adj} = 0.436$
- Task 3 versus Task 4: $z = -4.04$, $p = 5.41 \times 10^{-5}$, $p_{adj} = 3.24 \times 10^{-4}$ $(* \ * \ *)$

A.3 Navigation Effort Within the Right Section: Statistical Details

Outlier Handling

Outliers for *number_of_clicks_for_task_ratio* were removed using Tukey's rule, reducing observations to 161.

Inferential Statistics

A Kruskal–Wallis test showed a significant task effect ($\chi^2(3) = 69.496$, $p = 5.473 \times 10^{-15}$).

Post Hoc Analysis

Dunn post hoc tests with Bonferroni correction yielded as follows:

- Task 1 versus Task 2: $z = 7.47$, $p = 8.10 \times 10^{-14}$, $p_{adj} = 4.86 \times 10^{-13}$ $(* \ * \ * \ *)$
- Task 1 versus Task 3: $z = 5.97$, $p = 2.36 \times 10^{-9}$, $p_{adj} = 1.41 \times 10^{-8}$ $(* \ * \ * \ *)$
- Task 1 versus Task 4: $z = 6.13$, $p = 8.82 \times 10^{-10}$, $p_{adj} = 5.29 \times 10^{-9}$ $(* \ * \ * \ *)$
- Task 2 versus Task 3: $z = -1.66$, $p = 0.0966$, $p_{adj} = 0.580$
- Task 2 versus Task 4: $z = -1.37$, $p = 0.171$, $p_{adj} = 1$
- Task 3 versus Task 4: $z = 0.272$, $p = 0.786$, $p_{adj} = 1$

A.4 Task Duration: Statistical Details

Outlier Handling

For the *task_duration_ratio*, outliers were removed using Tukey's boxplot rule ($Q1 - 1.5 \times \text{IQR}$, $Q3 + 1.5 \times \text{IQR}$), reducing the number of observations to 158.

Descriptive Statistics

Task duration was normalized using the shortest observed completion time as a best case reference. Descriptive statistics, including mean, standard deviation, and 95% confidence intervals for each task, are reported in Table 6, whereas the distributions are illustrated in Figure 9.

Inferential Statistics

Since normality assumptions were not met, differences among tasks were evaluated using a Kruskal–Wallis test. The test did not reveal a statistically significant effect of task on duration ($\chi^2(3) = 4.4424$, $p = 0.2175$), indicating that observed differences in task duration were not statistically significant.

Appendix B. Detailed Qualitative Feedback

This appendix outlines the specific qualitative information gathered from the open-ended questions from the ASQ and PSSUQ questionnaires. This adds to the findings already summarized in Section 3.2 and gives further insight into the redesign choices recounted in Section 4.

B.1 After-Task Qualitative Feedback (ASQ)

After the completion of every task, participants described their experiences using open-ended comments. This feedback is depicted in an aggregated manner and the analysis is thematically focused on key usability challenges and positive aspects for each task.

Task 1 Diet and eating habits.

The overall sentiment expressed by participants was predominantly negative. However, there were a few comments expressing appreciation toward the visual design of the app and the

simplicity of the questionnaire. The main usability issues reported by the participants were:

- difficulty in locating the task entry point within the app;
- nonintuitive icons lacking textual labels;
- unclear categorization of food items and quantities;
- limited feedback when interacting with sliders and swipe-based inputs.

Some participants suggested adding more specific labels to the icons, improving clarification of measuring units, and adding descriptive text to support data entry.

Task 2 Food intake logging.

Food intake logging feedback for Task 2 was overwhelmingly positive, with participants reaffirming the importance of the speed of the interaction and the usefulness of the functionality. Nevertheless, some reported usability problems are as follows:

- difficulty understanding the meaning of certain icons;
- the need to consult the user manual to complete the task;
- requests for quicker access from the home screen.

A few comments described personal issues (e.g., dietary preference), as opposed to issues with the app.

Task 3 Rehabilitation exercises.

Feedback for Task 3 showed a mixed sentiment. Positive comments described the exercises as relevant and of good quality, and certain elements of the interface as clear. Negative commentary focused on:

- difficulty in finding the rehabilitation section;
- excessive length of video sequences;
- lack of indexing or navigation within videos;
- issues with full-screen mode and video resumption.

Participants also reported frustration being compelled to restart videos after interruptions and suggested improving the video controls.

Task 4 Medication management.

Most participants provided a positive evaluation for Task 4 and described the task as being easy to complete and quick to complete. Nevertheless, some issues were described, mostly concerning:

- ambiguity in modifying medication timeframes;
- insufficient visual distinction between start and end times;
- unclear terminology in some interface elements.

These comments are consistent with the increased task duration and variability observed in the quantitative analysis.

B.2 Postscenario Qualitative Feedback (PSSUQ)

The open-ended responses collected through the PSSUQ questionnaire allowed participants to reflect on the overall experience with the app. The overall experience ratings were mostly negative, although many participants highlighted positive aspects. Positive feedback emphasized:

- clarity of information once content was accessed;
- pleasant visual appearance of the interface;
- perceived usefulness of core functionalities.

Negative feedback mainly concerned:

- poor usability on smartphones with smaller displays;
- nonintuitive icons and unclear clickable elements;
- difficulties in accessing some functionalities without prior guidance;
- limitations in text readability and layout responsiveness.

Participants frequently suggested to increase the size of buttons, make the design more user-friendly, and make the user interface more prescriptive.

B.3 External Comments and General Suggestions

Additional open-ended feedback was collected through an external question, asking participants to describe their overall experience and propose improvements. As expected, the majority of the feedback was critical regarding the design and layout of the app, and the focus was on negative comments. Recurring themes included as follows:

- challenges in managing rehabilitation tasks and medication schedules;
- uncertainty about which interface elements were interactive;
- lack of pause or resume options during long rehabilitation activities;
- insufficient feedback after task completion.

Redesign suggestions entailed a more structured presentation of rehabilitation exercises, as well as clearer help icons, scrolling cues, and completion indicators. The redesign choices outlined in Section 4 were informed by these insights.