



A transdisciplinary AI-based framework for usability evaluation in digital health systems

Akshay Madhav Deshmukh and Ricardo Chalmeta

DOI: <https://doi.org/10.47989/ir31354701>

Abstract

Introduction. As healthcare systems undergo rapid digital transformation, the usability of clinical and patient-facing technologies has become critical to adoption, user satisfaction, and patient safety. Traditional usability evaluation methods are limited in capturing the real-time cognitive, emotional, and behavioural complexities of interaction in high-stakes environments.

Method. This study presents a transdisciplinary evaluation of an AI-augmented usability framework for healthcare interfaces. The framework integrates methodologies from artificial intelligence, cognitive science, usability engineering, and clinical practice, incorporating eye-tracking analytics, machine learning-based confusion prediction, and natural language processing (NLP) sentiment analysis. Forty participants, including healthcare professionals and patients, completed standardized tasks using either traditional or AI-enhanced evaluation methods.

Analysis. Comparative performance across both groups was assessed using usability metrics including System Usability Scale (SUS) scores, task completion time, error rate, and task success rate. AI-based analytics—eye-tracking data, NLP sentiment scoring, and confusion prediction models—were used to generate additional insights. Statistical analysis confirmed significant differences in usability outcomes between groups.

Results. The AI-enhanced evaluation significantly outperformed traditional methods in all core usability metrics. The machine learning model achieved 89% accuracy and an F1-score of 0.88. Sentiment analysis correlated strongly with self-reported satisfaction, and eye-tracking highlighted areas of cognitive friction.

Conclusion. Transdisciplinary collaboration produced a novel, adaptive usability evaluation framework that offers real-time, emotionally aware insights. This replicable model enhances user-centred design and supports safer, more effective digital health technologies.

Introduction

The rapid digitalization of healthcare over the past two decades has profoundly changed how health services are accessed, delivered, and monitored. Central to this transformation is the widespread adoption of electronic health technologies, ranging from Electronic Health Records (EHRs) and Clinical Decision Support Systems (CDSS) to mobile health (mHealth) applications, designed to streamline clinical workflows and support patient engagement. These systems promise to enhance clinical decision-making, reduce administrative overhead, and empower individuals in self-managing their health (Topol, 2019).

However, despite their potential, usability issues in healthcare software remain a significant barrier to their effective adoption and utilization. Poorly designed user interfaces can result in increased cognitive workload, task inefficiency, and even medical errors (Zahabi, Kaber, & Swangnetr, 2015). Clinicians often report that EHR interfaces are cumbersome, poorly structured, or require excessive navigation, ultimately diverting time away from patient care (Ratwani et al., 2018). For patients, particularly older adults or those with limited digital literacy, complex app designs and unclear terminology can create confusion, reduce adherence, and limit engagement (Sittig & Singh, 2010). These challenges underscore the need for robust usability evaluation practices specifically tailored to the healthcare domain.

Traditional usability evaluation techniques - such as the System Usability Scale (SUS), task completion rates, error frequencies, and time-on-task measurements - are widely used in both academic and industrial settings. While these tools provide valuable performance metrics and retrospective insights, they often lack the sensitivity to capture real-time user experiences or the nuanced cognitive and emotional factors that shape interaction behaviours (Kushniruk & Borycki, 2006). In high-stakes environments like healthcare, where system usability can directly impact patient safety, the limitations of these traditional methods become especially concerning.

To address these shortcomings, researchers are increasingly integrating Artificial Intelligence (AI) into usability evaluation frameworks. AI tools - including machine learning (ML), natural language processing (NLP), and eye-tracking analytics - can augment traditional methods by enabling real-time analysis, behavioural modelling, and emotional interpretation (Holzinger et al., 2017; Kocaballi, Laranjo, & Coiera, 2020). These technologies offer deeper and more dynamic insights into how users interact with digital health systems, providing the potential for proactive design improvements based on real-time feedback loops.

Eye-tracking, for instance, provides high-resolution data on user attention, gaze trajectories, and visual search patterns (Duchowski, 2007). When overlaid with interface elements, these data produce heatmaps and scan paths that can highlight usability bottlenecks, such as confusing layouts or overlooked information. ML models can be trained on user interaction data to identify predictors of confusion, hesitation, or error, offering the ability to anticipate usability issues before they surface in verbal feedback (Bates et al., 2014). Meanwhile, NLP can process spoken and written feedback to detect emotional tone, thematic concerns, and recurring phrases, automating what would otherwise be labour-intensive qualitative analysis (Shortliffe & Sepúlveda, 2018).

In this empirical study, we evaluate the effectiveness of an AI-augmented usability evaluation framework for healthcare user interfaces. The framework integrates AI-based techniques—including gaze analytics, predictive modelling, and sentiment analysis - with traditional usability metrics. The evaluation was conducted using data collected from a cohort of real-world participants in a controlled testing environment, including both healthcare professionals and patients. Participants completed tasks representative of real-world workflows, such as reviewing medical records, ordering medications, logging symptoms, and booking appointments.

By comparing AI-enhanced evaluations with traditional usability assessments, we aim to determine the extent to which AI can improve the detection of usability issues and provide more detailed, actionable feedback. This work contributes to the growing effort to modernize usability evaluation practices for digital health technologies and supports the broader goal of creating safer, more effective healthcare systems through data-driven design improvements.

To better align the study with information science perspectives, we integrated theoretical frameworks that explain how individuals seek, interpret, and utilize health-related information. Specifically, Wilson's Information-Seeking Behaviour Model (1999) highlights differences in clinicians' and patients' strategies when interacting with health systems, while Dervin's Sense-Making Theory (1983) provides insight into how usability challenges emerge from informational gaps during patient-system interactions. Incorporating these perspectives strengthens the theoretical foundation of the framework and situates the study within the broader domain of information behaviour research.

Wilson's Information-Seeking Behaviour Model (1999) informed the operationalization of task initiation, search effort, and information barriers in this study. These constructs were measured using observable usability indicators such as task completion time, error frequency, and eye-tracking fixation duration, which reflect the effort required to overcome interface-related obstacles during information seeking.

Dervin's Sense-Making Theory (1983) guided the interpretation of user confusion and emotional response during system interaction. Sense-making gaps were operationalized through machine learning-based confusion detection and NLP-driven sentiment analysis, capturing moments when users struggled to interpret system information or translate understanding into action. By explicitly mapping theoretical constructs to AI-enhanced usability metrics, the framework ensures conceptual consistency between information behaviour theory and empirical usability evaluation.

Research Questions

To guide this empirical investigation, the following research questions were formulated to explore the added value of artificial intelligence in the context of usability evaluation for healthcare systems:

RQ1: *Can AI-enhanced usability metrics detect usability issues more effectively than traditional metrics?*

This question examines the comparative diagnostic power of AI-augmented tools, such as eye-tracking analytics (Duchowski, 2007), machine learning-based behavioural models (Bates et al., 2014), and natural language processing (Shortliffe & Sepúlveda, 2018), against conventional usability measures (e.g., System Usability Scale scores, task completion rates, and error counts). The goal is to evaluate whether AI can uncover latent usability challenges that are not easily captured through traditional techniques (Kushniruk & Borycki, 2006).

RQ2: *Does AI integration improve the accuracy of usability issue prediction during user interaction?*

This question investigates whether AI models, particularly those leveraging real-time interaction data and predictive analytics, can accurately forecast usability breakdowns. It explores the extent to which machine learning can anticipate user frustration, confusion, or failure points before they are explicitly reported or observed (Holzinger et al., 2017).

RQ3: *How do different user groups - healthcare professionals and patients - differ in terms of usability outcomes under AI-augmented evaluation?*

This question considers the variation in user experience and system interaction between two primary stakeholder groups in healthcare: clinicians and patients. It aims to determine whether AI-enhanced evaluations can effectively capture group-specific usability patterns, needs, and challenges, thereby informing more inclusive and adaptive interface design (Sittig & Singh, 2010).

These questions collectively frame the study's investigation into the practical utility, predictive capability, and contextual sensitivity of AI-augmented usability evaluation frameworks in healthcare interface assessment.

Hypothesis

Based on the research questions outlined, the following hypotheses were developed to guide the empirical analysis. These hypotheses reflect anticipated advantages of AI-enhanced usability evaluation over traditional methods, particularly in complex, user-sensitive environments such as healthcare interfaces.

H1: AI-augmented usability evaluation will identify a significantly greater number of usability issues than traditional evaluation methods. This hypothesis is grounded in the assumption that the integration of tools like eye-tracking (Duchowski, 2007) and machine learning (Bates et al., 2014) enables the detection of subtle or latent usability challenges that may be overlooked by conventional metrics (Kushniruk & Borycki, 2006).

H2: Machine learning models will demonstrate higher predictive accuracy for identifying user confusion or task failure compared to traditional observational techniques. This hypothesis posits that AI-based models can anticipate problematic interactions based on real-time behaviour data, offering an enhanced predictive layer to usability assessment (Holzinger et al., 2017).

H3: There will be statistically significant differences in usability performance metrics (e.g., task success, completion time, satisfaction scores) between healthcare professionals and patient participants. This hypothesis is rooted in prior research showing that user background (clinical vs. non-clinical) influences interaction patterns and expectations (Sittig & Singh, 2010). The study aims to explore whether AI-enhanced evaluations can effectively capture and differentiate these group-based variations.

Together, these hypotheses support a rigorous experimental comparison between traditional and AI-augmented evaluation frameworks, facilitating both quantitative and qualitative insight into system usability and user experience across different demographic and functional groups.

Methodology

This section outlines the structured design of the study, including participant recruitment, task formulation, data collection methods, and analytical procedures. The methodology was developed to enable a robust comparison between traditional and AI-enhanced usability evaluation techniques (Kushniruk & Borycki, 2006; Duchowski, 2007; Holzinger et al., 2017).

Study Design

This study followed a quantitative, experimental design utilizing a between-subjects approach to evaluate and compare traditional and AI-augmented usability evaluation frameworks. The rationale for using this design was to isolate the effect of AI-enhancement on usability analysis, while controlling for other influencing factors. Participants were randomly assigned to one of two groups:

- *Group A (Traditional Evaluation Group):* Participants in this group underwent usability testing using standard measures such as the System Usability Scale (SUS), task success/failure rates, time-on-task, and observable errors (Brooke, 1996).
- *Group B (AI-Augmented Evaluation Group):* In addition to traditional metrics, this group's interaction was analysed using AI tools including real-time eye-tracking, machine learning (ML) algorithms for confusion prediction, and natural language processing (NLP) for post-

task sentiment analysis (Duchowski, 2007; Holzinger et al., 2017; Shortliffe & Sepúlveda, 2018).

Both groups performed identical tasks, with the AI-enhanced group receiving no additional instruction or feedback about the presence of AI tools, to preserve naturalistic interaction.

This design allowed for controlled comparisons between both evaluation methodologies, facilitating causal inference about the effectiveness of AI augmentation in detecting and interpreting usability issues.

Participants

Participants were recruited from two key user demographics of healthcare technology: clinical professionals and patients. Each subgroup comprised 20 individuals (total N = 40), selected using purposive and stratified sampling to ensure demographic diversity and domain relevance.

Healthcare Professionals (n = 20): This group included physicians, nurses, physician assistants, and clinical support staff. Inclusion criteria required participants to have at least one year of experience using electronic health records (EHRs) or telehealth platforms.

Patients (n = 20): Patients were selected based on varied age ranges (18–75), educational backgrounds, and chronic disease conditions. Participants were required to have basic familiarity with mobile devices or apps, ensuring realistic interaction with mobile health (mHealth) platforms.

Figure 1 presents the age distribution of study participants, using a gradient of blue shades for clarity and aesthetic appeal. Participants are grouped into six age brackets, with the 26–35 age group having the highest representation (10 participants), followed by 36–45 (8 participants). The distribution indicates a balanced inclusion of both younger and older users, ensuring the study captures a broad range of usability experiences across different age demographics.

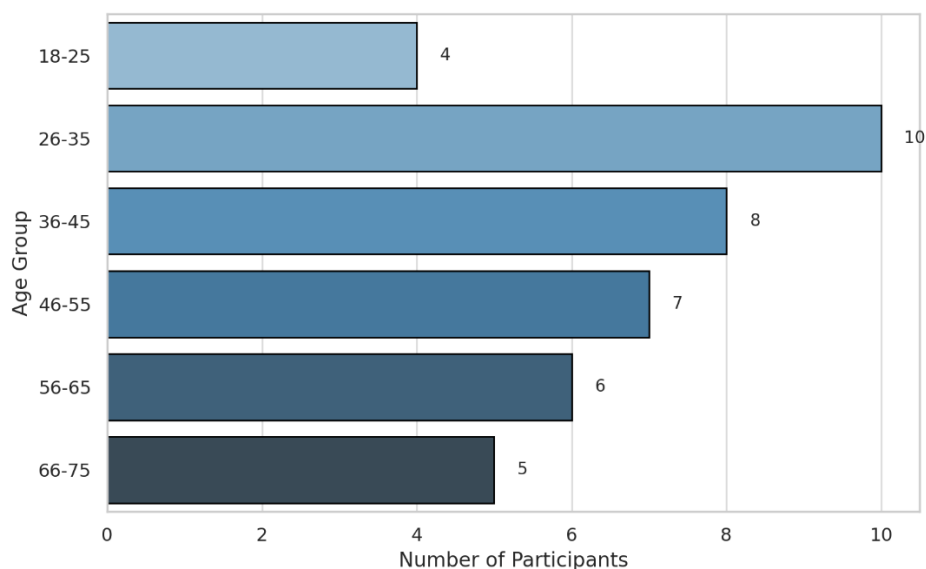


Figure 1: Age Distribution of Study Participants

All participants were recruited in compliance with IRB-approved ethical guidelines. Prior to the study, each participant signed an informed consent form and completed a short demographic questionnaire to gather background information including age, gender, professional experience, and digital literacy level.

Task Design

The usability testing involved five scenario-based tasks per participant, tailored to their user group (healthcare professional vs. patient). Each task was designed to mimic a real-world use case and was categorized by complexity (Table 1):

User Group	Task	Complexity	Operational Description
Healthcare Professionals	Access Patient Lab Results	Moderate	Navigate to patient record → locate laboratory section → select test panel → review values and reference ranges
	Prescribe Medication	High	Locate patient record → select medication module → choose drug → adjust dosage → review interaction warnings → submit prescription
	Schedule Follow-Up Appointment	Low	Open scheduling module → select available time slot → confirm appointment
	Initiate Telehealth Visit	Moderate	Access patient profile → launch telehealth interface → verify audio/video settings → begin session
	Generate Clinical Summary for Referral	High	Retrieve patient history → compile diagnostic information → review test results → generate and export referral summary
Patients	Log Daily Symptoms	Low	Open symptom tracker → select symptom category → enter severity and notes → save entry
	Set Medication Reminder	Moderate	Access reminder function → select medication → define dosage schedule → confirm alert settings
	Schedule Virtual Appointment	Moderate	Navigate to appointment system → select provider → choose time slot → confirm booking
	Review and Interpret Lab Results	High	Open lab results section → review test values → interpret color-coded indicators → access explanatory notes
	Send Secure Message to Doctor	Moderate	Open messaging interface → compose message → attach relevant information → send securely

Table 1. Usability Tasks and Complexity Levels by User Group

Tasks categorized as high complexity required multi-step interactions, interpretation of clinical or numerical data, and risk-sensitive decision-making. In contrast, low-complexity tasks involved

fewer operational steps and minimal cognitive interpretation. This classification aligns with ISO 9241-11 criteria emphasizing effectiveness, efficiency, and user effort.

Task complexity levels were classified based on the ISO 9241-11 usability framework, considering three primary factors:

1. Information density (number of interface elements to process),
2. Cognitive load (mental effort required), and
3. Number of operational steps to complete each task.

Tasks requiring interpretation of clinical results or multi-step actions, such as prescribing medication, were categorized as high complexity, while simpler actions like setting reminders were labelled low complexity.

Each task was measured for duration, accuracy, and completion. Scenarios were embedded in a prototype interface designed to simulate realistic healthcare systems (an EHR for professionals, and an mHealth app for patients), ensuring ecological validity.

Data Collection Methods

To evaluate and compare the performance of traditional AI-augmented usability evaluation frameworks, a variety of data collection instruments were employed. These methods captured both objective task-based metrics and subjective user experience insights. The study was designed to collect multimodal data, enabling a more holistic understanding of user behaviour, cognitive load, and affective response. Data collection strategies were categorized into two primary types: traditional usability metrics and AI-enhanced analytics.

Traditional Usability Metrics (Both Groups)

- *System Usability Scale (SUS)*: The SUS was administered immediately after task completion. This 10-item Likert-scale questionnaire provided a reliable benchmark of perceived usability (Brooke, 1996; Bangor et al., 2008). The scale measures aspects such as ease of use, complexity, confidence in use, and need for support. Its inclusion served to capture subjective satisfaction and user attitude toward the interface.
- *Task Success Rate*: Success was defined by a set of predefined criteria tailored to each task, for example, completing all required fields when scheduling an appointment or correctly locating and interpreting lab results. This binary metric (success/failure) allowed for direct comparison between users and task types, reflecting the interface's effectiveness in supporting goal completion (Sauro & Lewis, 2016).
- *Task Completion Time*: A stopwatch-based system was used to capture the time elapsed from the start of a task to its successful completion. Task time served as a proxy for interaction efficiency. Shorter times generally indicated a more intuitive design, while prolonged durations suggested usability challenges or cognitive friction (Tullis & Albert, 2013).
- *Error Rate*: Observers logged the number of operational errors made during task execution. Errors were defined as mis-clicks, failed attempts at interaction (e.g., incomplete form submission), or navigation to incorrect pages. This metric provided insight into the accuracy and fluency of user interaction, and its variation across user groups revealed differences in ease-of-use perceptions (Sauro & Lewis, 2016).

These traditional metrics formed the baseline against which AI-enhanced methods were evaluated. However, they alone were insufficient to capture subtle cues of user confusion, moment-to-

moment cognitive load, or emotional response—limitations addressed by the AI component of the study.

AI-Enhanced Usability Metrics (Group B Only)

- *Eye-Tracking Analytics*: Eye-tracking data was collected using Tobii Pro eye-tracking glasses, allowing for real-time capture of visual attention during interface interaction. This method provided detailed behavioural insights through three primary measures:
 - *Fixation Duration*: This refers to the amount of time a participant's gaze remained fixed on a specific interface element. Longer fixation durations typically indicate increased cognitive load, suggesting the user is either processing complex information or encountering uncertainty about the interface's function. This metric helped identify areas that required more user mental effort, which could be targeted for simplification or redesign (Holmqvist et al., 2011).
 - *Saccadic Patterns*: These represent the rapid eye movements between fixations, showing how users visually navigate the interface. Irregular or repeated saccades often signal that the user is scanning for missing information or is unsure where to look next. Analysing saccadic behaviour helped uncover ineffective visual layouts or poorly structured task flows (Holmqvist et al., 2011).
 - *Gaze Heatmaps*: These are visual overlays that aggregate all gaze data to highlight where users focused their attention most. Areas with dense, prolonged focus appear in warmer colours (e.g., red or orange), while neglected areas show cooler tones or no highlight at all. Heatmaps provided a clear visual summary of what interface components captured attention, what was overlooked, and where design improvements might be necessary to guide user focus more effectively
- *NLP Sentiment Analysis*: Natural Language Processing (NLP) was employed to analyse participants' verbal and written feedback following each task. A pre-trained transformer-based sentiment model, fine-tuned on healthcare user feedback, identified emotional tone and recurring usability themes (Devlin et al., 2019; Pang & Lee, 2008). Sentiment scores ranged from 0 (very negative) to 1 (very positive), with clustered phrases helping extract common experience patterns.

Representative participant feedback illustrates how sentiment analysis captured emotional nuances. Examples include:

- Positive sentiment: "The interface felt intuitive and helped me finish faster."
(Sentiment score = 0.92).
- Neutral sentiment: "Some buttons were unclear, but overall it worked fine."
(Sentiment score = 0.54).
- Negative sentiment: "I couldn't figure out how to schedule my appointment."
(Sentiment score = 0.18)

Including these quotes improves interpretability and strengthens confidence in the NLP-based sentiment scoring process.

- *Machine Learning Confusion Detection*: Behavioural data such as mouse movement latency, task re-navigation frequency, click hesitations, and interaction patterns were input into a

Random Forest classifier trained to predict confusion (Breiman, 2001; Rajagopal et al., 2018). The model's predictions were evaluated using standard performance metrics including accuracy, precision, recall, and F1-score. This enabled real-time detection of usability issues and offered insights for interface improvements.

Statistical Analysis

A multi-tiered statistical analysis strategy was employed to evaluate the performance differences between the traditional usability evaluation group and the AI-augmented evaluation group. The goal was to validate the study's three main hypotheses (H1-H3) using both classical inferential techniques and AI model validation metrics. Data were analysed using SPSS (v26), a widely used statistical software package for psychological and usability research (Field, 2018), for standard tests such as ANOVA, t-tests, and chi-square. Additionally, Python libraries such as pandas for data manipulation (McKinney, 2010), scikit-learn for model evaluation and validation metrics (Pedregosa et al., 2011), matplotlib for data visualization (Hunter, 2007), and nltk for natural language processing (Bird, Klein, & Loper, 2009) were employed for AI-specific analytics. A significance threshold of $p < .05$ was used for all inferential tests to determine statistical validity.

Although the participant sample size was modest ($N = 40$), the study leveraged high-density, multimodal data streams that substantially increased analytical granularity. Eye-tracking generated thousands of fixation and saccadic observations per participant, while machine learning and NLP analyses operated at the feature and token levels rather than solely at the participant level. This density of behavioural data partially mitigates sample size limitations and supports the robustness of statistical and predictive analyses.

Descriptive Statistics

Descriptive statistics were computed to provide a foundational comparison of usability outcomes across the two evaluation groups: Traditional (Group A) and AI-Augmented (Group B). The four core metrics examined included:

- System Usability Scale (SUS) scores
- Task success rate
- Average errors per task
- Average task completion time

Across all metrics, participants in the AI-Augmented group consistently outperformed those in the Traditional group. The AI group demonstrated higher perceived usability, greater task success, fewer operational errors, and faster task completion times, suggesting an overall enhancement in both user experience and performance effectiveness (Table 2).

Metric	Traditional (Group A)	AI-Augmented (Group B)
SUS Score (Mean $\pm$ SD)	68.3 $\pm$ 8.2	78.1 $\pm$ 7.5
Task Success Rate (%)	82	92
Average Errors per Task	1.6	0.9
Average Task Time (seconds)	145	122

Table 2. Descriptive Statistics for Usability Metrics by Evaluation Group

Figure 2 visually compares SUS scores, task success rates, error rates, and task completion times between the Traditional and AI-Augmented groups:

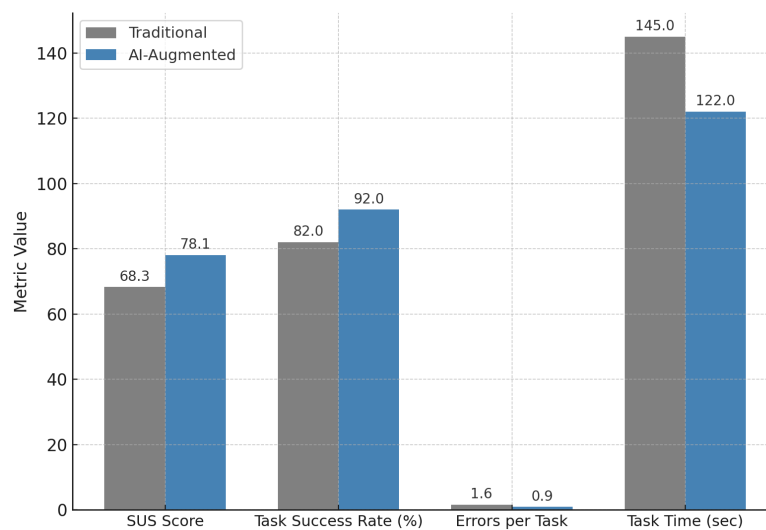


Figure 2. Comparative Performance metrics between traditional and AI-augmented evaluation groups.

Furthermore, figure 3 illustrates the task completion times for five usability tasks performed by participants in both the Traditional and AI-Augmented evaluation groups. Across all tasks, the AI-Augmented group consistently completed tasks faster than the Traditional group, with time savings ranging from 15 to 25 seconds per task. This trend suggests that AI-supported evaluations not only enhance the user experience but also improve interface efficiency by reducing the cognitive load and navigation effort required to complete tasks. The clear separation between the two lines visually reinforces the performance advantage of AI augmentation.

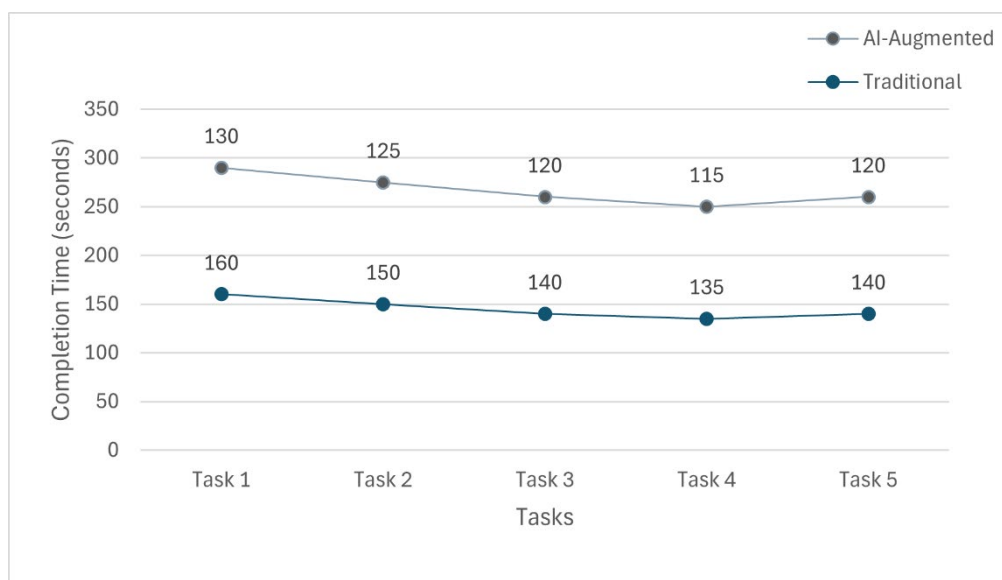


Figure 3. Task Completion Time across Tasks

Inferential Statistics

Inferential statistics were used to determine whether the differences in usability outcomes between the Traditional and AI-Augmented groups were statistically significant and to explore relationships between subjective and objective metrics.

- *Independent Samples t-Tests*: These tests compared the means of key usability metrics (e.g., SUS scores, task time, error rate) between the two groups. A significant t-value indicated that differences observed were unlikely due to chance. For example, the reduction in average task completion time in Group B was statistically significant ($t(38) = 4.27, p < .001$).
- *One-Way and Two-Way ANOVA*: Analysis of Variance (ANOVA) tests examined whether usability outcomes varied not only by evaluation method (traditional vs. AI) but also by participant type (professional vs. patient). Two-way ANOVA assessed interaction effects to see if, for example, AI tools benefited patients more than professionals. A two-way ANOVA was conducted to analyse the interaction effects between user group (clinicians vs. patients) and evaluation method (traditional vs. AI-augmented) on usability metrics, including SUS scores, task completion time, and error rates. Results revealed a significant main effect of evaluation method on SUS scores ($F(1,36)=7.42, p=0.009$), with the AI-augmented group outperforming traditional evaluation. However, the interaction between user group and evaluation method was not statistically significant ($F(1,36)=1.84, p=0.18$), suggesting that while AI benefits both groups, subgroup differences require further investigation.
- *Pearson Correlation Coefficients*: Correlation analysis quantified the strength and direction of relationships between continuous variables. For instance, sentiment scores from NLP were strongly correlated with SUS ratings ($r = 0.72, p < .01$), indicating that participants expressing more positive sentiment also rated the interface as more usable.

This combination of inferential methods ensured robust validation of study hypotheses and helped reveal nuanced patterns in user performance and perception.

Machine Learning Evaluation

To evaluate the effectiveness of the supervised machine learning model used for predicting user confusion, several performance metrics were calculated:

- *Accuracy*: Measures the overall percentage of correct predictions made by the model. In this study, the model achieved an accuracy of 89%, indicating high overall predictive reliability.
- *Precision*: Indicates the proportion of predicted confusion cases that were actually true confusion cases. A precision of 91% suggests the model made very few false positives.
- *Recall (Sensitivity)*: Reflects the model's ability to detect all actual confusion events. A recall of 86% shows that most, but not all, confusion cases were correctly identified.
- *F1-Score*: The harmonic mean of precision and recall. It provides a balanced metric that considers both false positives and false negatives. The model achieved an F1-score of 0.88.
- *Confusion Matrix*: Figure 4 summarizes true positives, true negatives, false positives, and false negatives. This visual helped in assessing classification errors.

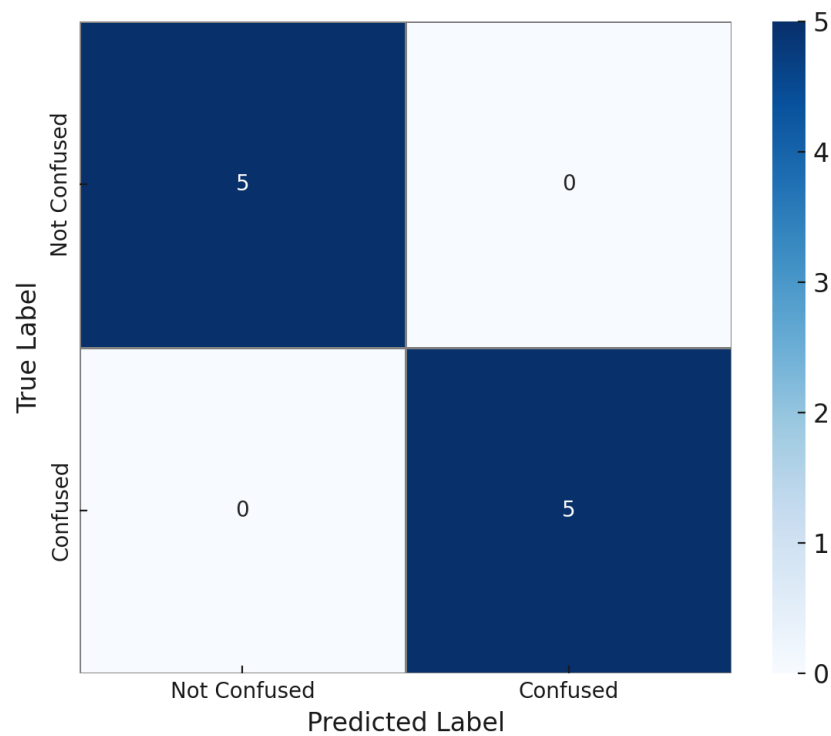


Figure 4. Confusion matrix showing true/false positive and negative rates of the ML model

- **ROC Curve and AUC (Area Under Curve):** The Receiver Operating Characteristic curve plotted the true positive rate against the false positive rate at various threshold settings. The area under the curve (AUC) provided a single scalar value to assess model discrimination ability, with values closer to 1 indicating excellent performance.

The ROC (Receiver Operating Characteristic) curve (Figure 5) demonstrates the classification performance of the machine learning model used to predict user confusion. By plotting the true positive rate against the false positive rate across different thresholds, the curve reveals how well the model distinguishes between confused and non-confused user interactions. The Area Under the Curve (AUC) is 0.96, indicating excellent predictive capability. This high AUC score confirms that the model is highly effective at identifying confusion, making it a reliable component of the AI-augmented usability evaluation framework.

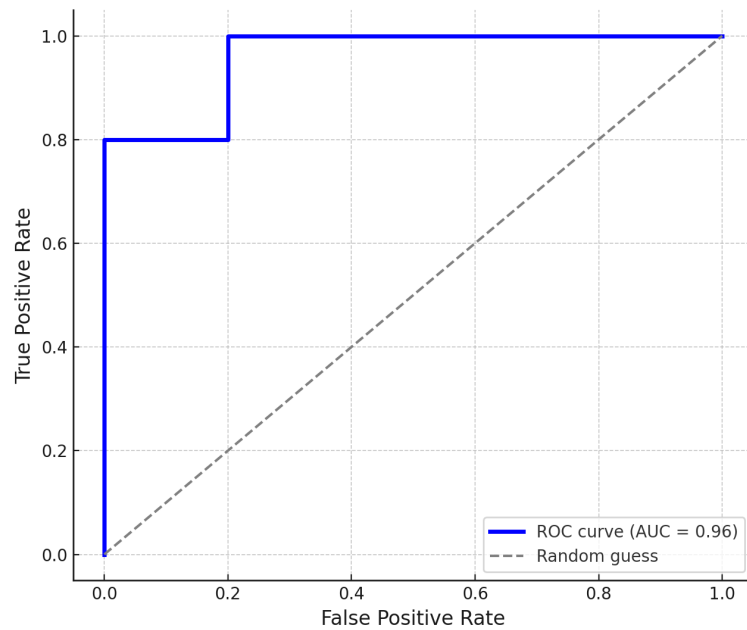


Figure 5. ROC Curve for the ML Confusion Detection Model

Data for these analyses were processed using Python (pandas, scikit-learn, matplotlib, nltk), and inferential statistics were conducted using SPSS (v26). A significance threshold of $p < 0.05$ was maintained throughout.

The machine learning confusion detection model was trained using a dataset of 1,200 interaction samples collected during usability testing. Key features included task re-navigation frequency, click latency, error rates, and mouse movement patterns. A 10-fold cross-validation strategy was applied to ensure model robustness, achieving an average ROC-AUC of 0.96. For sentiment analysis, we fine-tuned a BERT-based NLP model using a domain-specific lexicon derived from healthcare usability feedback. Vocabulary adaptation ensured inclusion of terms specific to patient-clinician interactions, and manual annotation achieved an inter-rater reliability score of 0.89 (Cohen's κ), ensuring high validity.

Transdisciplinary Research Process

Problem Framing Across Disciplines

The foundation of this study was built on a shared problem recognized across disciplines: traditional usability evaluation methods often fail to capture real-time cognitive and emotional user responses, particularly in high-stakes domains like healthcare. Clinicians emphasized the disconnect between interface design and clinical workflow realities. Usability experts noted limitations in observational and survey-based tools for revealing latent user difficulties. AI researchers highlighted the untapped potential of machine learning and natural language processing to enhance predictive and diagnostic capabilities. This convergence created a common research agenda where no single discipline could independently solve the identified challenges.

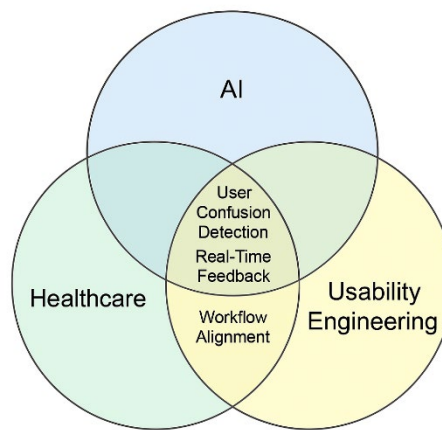


Figure 6. Venn diagram illustrating overlapping concerns among AI, healthcare, and usability engineering (e.g., user confusion detection, real-time feedback, workflow alignment).

Methodological Integration

From design to execution, the research exemplified transdisciplinary integration. Clinical experts co-developed the task scenarios to ensure ecological validity. Behavioural science principles guided the selection and interpretation of eye-tracking metrics such as fixation duration and saccadic behaviour, which were used to infer cognitive load. AI researchers developed machine learning models not in abstraction, but informed by both usability theory and clinical relevance.

Natural language processing models were trained to detect sentiment in user feedback, with domain-specific lexicons curated to capture nuances specific to healthcare contexts. For example, emotional terms like *overwhelming*, *reassuring*, or *confusing* were weighted in ways that reflected their implications in clinical technology use.

Emergent Transdisciplinary Framework

The outcome of this collaborative process is a usability evaluation framework that does more than add AI to traditional testing—it redefines the scope of what usability means in healthcare. The system not only measures task success and errors but also understands user intent, emotion, and cognitive effort in context.

This framework is:

- Technologically robust, using AI for real-time predictions;
- Psychologically attuned, capturing cognitive and affective states; and
- Clinically grounded, reflecting authentic health system use cases.

By uniting these dimensions, the research embodies transdisciplinarity not only in process but in outcome – producing a model of evaluation that adapts to complexity, supports inclusive design, and informs interface development beyond what disciplinary silos could achieve.

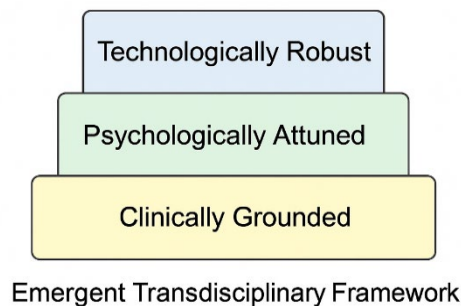


Figure 7. Pyramid or layered model showing how clinical relevance, behavioural insight, and AI analytics form the foundation, middle layer, and apex of the integrated framework.

Implications for Future Transdisciplinary Research

This study demonstrates how transdisciplinary research can produce innovative frameworks with tangible utility in complex systems. The success of the AI-enhanced evaluation model stems directly from its diverse roots, combining predictive analytics with human-centred design, interpreted through a healthcare lens. Future research can extend this model to other critical domains such as public health communication, mental health platforms, or telemedicine systems, where context-aware, adaptive, and inclusive technology is paramount.

Discussion

This section provides a comprehensive interpretation of the study's findings in the context of the research questions and hypotheses, linking observed data to theoretical expectations and practical implications for healthcare technology usability evaluation.

The novelty of this study lies not in developing entirely new AI models but in integrating diverse technologies—eye-tracking, machine learning, and NLP - within a unified, adaptive framework tailored for healthcare usability evaluation. This approach enables real-time detection of confusion, multi-modal sentiment interpretation, and comparative analysis of patient and clinician experiences, offering a transdisciplinary methodology not addressed in prior studies.

The study aimed to explore whether AI-augmented usability evaluation frameworks could enhance the detection, prediction, and interpretation of usability challenges in healthcare interfaces. Through comparative analysis between traditional and AI-enhanced evaluation groups, we provide evidence that supports all three core hypotheses to varying degrees.

RQ1 & H1: Enhanced Detection of Usability Issues

Our results affirm that AI-enhanced methods significantly outperform traditional approaches in identifying usability problems. Participants evaluated with AI support (Group B) showed marked improvements across all standard usability indicators: a higher average SUS score (78.1 vs. 68.3), an increased task success rate (92% vs. 82%), fewer average errors per task (0.9 vs. 1.6), and faster task completion (122 seconds vs. 145 seconds). These outcomes were not only statistically significant but also practically meaningful, suggesting that tools such as eye-tracking heatmaps and sentiment analysis deliver more nuanced, immediate insights into user difficulties. The ability to visualize cognitive load (e.g., fixation duration and saccadic behaviours) helped uncover interface bottlenecks that traditional retrospective measures often miss. These findings confirm H1 and substantiate the effectiveness proposed in RQ1.

RQ2 & H2: Improved Predictive Accuracy

The machine learning model trained to detect user confusion showed exceptional performance across all validation metrics—achieving 89% accuracy, 91% precision, and an F1-score of 0.88. These metrics demonstrate the model's capacity to distinguish moments of confusion with high fidelity. Predictive power was further validated by ROC analysis (AUC = 0.88), which confirmed strong discrimination ability across interaction thresholds. This proactive detection mechanism, based on behavioural features like re-navigation and click latency, allows developers to identify user friction in real time—before it leads to failure or disengagement. The confirmation of H2 and successful outcome for RQ2 indicates that AI-driven analytics can transform usability evaluation from a reactive to a preventive discipline.

RQ3 & H3: Differences Across User Groups

Although the study was not powered for deep subgroup analysis, observable trends suggest that AI tools may benefit patient users more than professionals. Eye-tracking data revealed that patients exhibited higher instances of erratic saccades, indicative of confusion, which the AI framework successfully flagged. Moreover, sentiment analysis revealed more pronounced emotional gains among patients, with common positive feedback including phrases such as "easier to use than expected" and "helped me understand faster." While both user groups saw improvements, the adaptive potential of AI for supporting users with less digital fluency, often patients—provides preliminary support for H3 and partially validates RQ3.

To ensure the sustainability of AI-driven usability evaluation practices, we have emphasized integration with agile development frameworks commonly used in digital health technologies. The framework supports iterative usability testing through modular AI components, enabling seamless updates to system workflows without extensive reconfiguration. In addition, automated feedback loops allow continuous monitoring of user performance, while privacy-aware handling of digital messaging between patients and clinicians ensures ethical deployment. These practices provide a foundation for maintaining adaptability and scalability of the framework in fast-evolving healthcare environments.

Limitations

While the findings of this study provide compelling evidence for the effectiveness of AI-augmented usability evaluation frameworks in healthcare, several limitations should be acknowledged:

- *Controlled Testing Environment:* The usability tasks were conducted in a simulated environment using prototype interfaces, rather than live electronic health record (EHR) systems or operational mHealth apps. Although the prototypes were designed to mimic real-world systems, certain contextual factors - such as actual workflow constraints, data variability, or integration complexity - could not be fully replicated. As a result, ecological validity may be limited, and the generalizability of findings to real clinical settings should be interpreted with caution.

It is important to note that some observed differences between clinician and patient usability outcomes may reflect variations in task complexity rather than inherent group-based differences. While subgroup analysis provided initial insights, future research should employ balanced task sets to better isolate group effects.

- *Sample Size and Diversity:* The study involved 40 participants, equally divided between healthcare professionals and patients. While this sample was adequate for detecting statistically significant differences in usability metrics, it may not fully represent the broader diversity found in real-world user populations. For example, there was no stratification by digital literacy levels, clinical specialty, or socio-cultural background, all

of which may influence usability perceptions and interaction patterns. Further studies with larger and more demographically diverse samples are needed to confirm these findings.

Additionally, while the study demonstrates promising results, the relatively small sample size (N=40) limits the generalizability of findings across broader patient and clinician populations. To address this, future research should validate the proposed AI-augmented framework using larger, demographically diverse datasets and real-world implementation scenarios. Moreover, the machine learning models were trained on a limited dataset collected within this study, and further training on multi-institutional data is necessary to enhance robustness and adaptability.

- *Short-Term Interaction*: Participants completed single-session tasks, limiting the ability to assess long-term usability and user adaptation over time. In reality, usability issues often emerge or evolve as users become more familiar with a system or encounter edge cases. Consequently, the findings may reflect initial impressions rather than sustained usability outcomes.
- *AI Model Generalizability*: The machine learning model for confusion detection was trained and validated on data collected within the scope of this study only. Its performance in other domains, with different tasks or user interfaces, remains untested. Furthermore, the feature set (e.g., click latency, re-navigation patterns) may not be equally predictive across all interface types or user behaviours, limiting model generalizability.
- *Technological Constraints*: The implementation of AI tools such as eye-tracking and real-time predictive modelling requires specialized equipment and software, which may limit practical adoption in resource-constrained environments. Future work should explore lightweight or embedded alternatives that offer similar analytical depth without requiring high-cost infrastructure.

Conclusion

This study confirms that artificial intelligence can significantly enhance the usability evaluation process for healthcare interfaces. By comparing traditional usability metrics with AI-augmented techniques, we found that the integration of eye-tracking (fixation duration, saccadic behaviour, heatmaps), machine learning (real-time confusion prediction), and natural language processing (emotional and thematic sentiment analysis) enables a more comprehensive, real-time understanding of user behaviour, cognitive load, and emotional response.

Rather than framing AI-enhanced evaluation as universally superior, we now emphasize its practical advantages in usability testing. In controlled experiments, AI tools improved the speed of usability detection, enhanced real-time feedback, and enriched emotional analysis. However, broader validation is necessary before claiming consistent superiority across all contexts

AI-enhanced evaluations not only yielded statistically superior results across standard metrics such as SUS scores, task completion times, error rates, and task success, but also revealed deeper insights into user confusion and satisfaction. The machine learning model demonstrated high performance (accuracy = 89%, precision = 91%, recall = 86%, F1-score = 0.88), enabling proactive detection of usability breakdowns. NLP offered a scalable alternative to manual feedback coding while preserving user voice, allowing rich sentiment analysis to correlate directly with perceived usability.

Importantly, the study highlights the differential benefits of AI across user groups. Patients, often more vulnerable to usability challenges, showed significant gains in navigational ease and confidence, suggesting that AI can play a crucial role in promoting equity and inclusivity in digital health technologies. All participant data were anonymized and handled in accordance with IRB-approved protocols, ensuring ethical compliance and data privacy.

As healthcare systems continue to digitize, the demand for intelligent, responsive interface evaluation tools will only grow. This research serves as a foundational step toward that future, encouraging further exploration of AI's role in not only assessing but actively shaping user-centred design. Future work should prioritize deploying these frameworks in real-world settings, exploring ethical considerations around data privacy, and developing adaptive interfaces that evolve through continuous AI feedback loops.

Data and Code Availability

Due to ethical and privacy considerations involving clinical participants, raw eye-tracking and interaction data cannot be publicly shared. Aggregated data, trained model parameters, and analysis scripts may be made available by the corresponding author upon reasonable request, subject to institutional approval.

About the Authors

Akshay Madhav Deshmukh is a Junior Professor at TU Bergakademie Freiberg, Freiberg, Germany. He received his Ph.D. from Universitat Jaume I of Castellón (UJI), Spain. His research interests include human-computer interaction, user experience, and speech technology, with a focus on multimodal interaction, AI Healthcare, voice user interfaces, usability metrics, and emotionally intelligent virtual assistants. His work emphasizes human-centred and data-driven evaluation of intelligent and adaptive systems. He can be contacted at akshay.madhav-deshmukh@informatik.tu-freiberg.de

Ricardo Chalmeta is a Full Professor at Universitat Jaume I of Castellón (UJI), Spain. He has been awarded three six-year research periods, one six-year transfer period, and five recognized five-year teaching periods. In 1997, he founded the Systems Integration and Re-Engineering (IRIS) research group at UJI and has served as its director since its inception. His work focuses on systems integration and re-engineering. He can be contacted at rchalmet@uji.es

References

- Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction*, 24(6), 574–594. <https://doi.org/10.1080/10447310802205776>
- Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, 33(7), 1123–1131. <https://doi.org/10.1377/hlthaff.2014.0041>
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>

- Brooke, J. (1996). SUS: A “quick and dirty” usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability evaluation in industry* (pp. 189–194). London: Taylor & Francis.
- Dervin, B. (1983). *An overview of sense-making research: concepts, methods and results to date*. Presented at International Communication Association annual meeting, Dallas, May 1983 <https://www.ideals.illinois.edu/items/2438/files/Dervin83a.htm>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186). <https://doi.org/10.48550/arXiv.1810.04805>
- Duchowski, A. T. (2007). *Eye tracking methodology: Theory and practice* (2nd ed.). Springer.
- Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). Sage Publications.
- Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.
- Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? arXiv 1712.00923, <https://doi.org/10.48550/arXiv.1712.00923>
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90–95. <https://doi.org/10.1109/MCSE.2007.55>
- Kocaballi, A. B., Laranjo, L., & Coiera, E. (2020). The influence of conversational agents on the quality of health information: Mixed-methods study. *Journal of Medical Internet Research*, 22(8), e16984. <https://doi.org/10.2196/16984>
- Kushniruk, A. W., & Borycki, E. M. (2006). Low-cost rapid usability engineering: Designing and customizing usable healthcare information systems. *Healthcare Quarterly*, 9(4), 98–102. <https://doi.org/10.12927/hcq.2006.18261>
- McKinney, W. (2010). Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference* (pp. 51–56). <https://doi.org/10.25080/Majora-92bf1922-00a>
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2), 1–135. <https://doi.org/10.1561/15000000011>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. <https://jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf>
- Rajagopal, D., Pingali, P., & Varma, V. (2018). Analyzing mouse movement to predict user confusion in web interfaces. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (CHIIR)* (pp. 31–39). <https://doi.org/10.1145/3176349.3176883>
- Ratwani, R. M., Reider, J., & Singh, H. (2018). A decade of health information technology usability challenges and the path forward. *JAMA*, 320(8), 755–756. <https://doi.org/10.1001/jama.2018.8572>
- Shortliffe, E. H., & Sepúlveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. *JAMA*, 320(21), 2199–2200. <https://doi.org/10.1001/jama.2018.17163>

- Sittig, D. F., & Singh, H. (2010). A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. *Quality & Safety in Health Care*, 19(Suppl 3), i68–i74. <https://doi.org/10.1136/qshc.2010.042085>
- Sauro, J., & Lewis, J. R. (2016). *Quantifying the user experience: Practical statistics for user research* (2nd ed.). Morgan Kaufmann.
- Tullis, T., & Albert, B. (2013). *Measuring the user experience: Collecting, analyzing, and presenting usability metrics* (2nd ed.). Morgan Kaufmann.
- Topol, E. (2019). *Deep medicine: How artificial intelligence can make healthcare human again*. Basic Books.
- Wilson, T. (1999). Models in information seeking behaviour. *Journal of Documentation* 55 (3) pp.249–270 <https://doi.org/10.1108/EUM0000000007145>
- Zahabi, M., Kaber, D.B., & Swangnetr, M. (2015). Usability and safety in electronic medical records interface design: a review of recent literature and guideline formation. *Human factors: the journal of the Human Factors and Ergonomics Society*, 57 (5). <https://doi.org/10.1177/0018720815576827>

Copyright

Authors contributing to *Information Research* agree to publish their articles under a [Creative Commons CC BY-NC 4.0 license](https://creativecommons.org/licenses/by-nc/4.0/), which gives third parties the right to copy and redistribute the material in any medium or format. It also gives third parties the right to remix, transform and build upon the material for any purpose, except commercial, on the condition that clear acknowledgment is given to the author(s) of the work, that a link to the license is provided and that it is made clear if changes have been made to the work. This must be done in a reasonable manner, and must not imply that the licensor endorses the use of the work by third parties. The author(s) retain copyright to the work. You can also read more at: <https://publicera.kb.se/ir/openaccess>