The Role of Artificial Intelligence in Usability Testing and its Potential Impact on Equity and Medical Solution Design: A Narrative Literature Review

Selena Lombardi, Enid Montague

Submitted to: Journal of Medical Internet Research
on: October 07, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

The Role of Artificial Intelligence in Usability Testing and its Potential Impact on Equity and Medical Solution Design: A Narrative Literature Review

Selena Lombardi[1] BASc; Enid Montague[1] BS, MS, PhD

[1]Department of Mechanical and Industrial Engineering University of Toronto Toronto CA

Corresponding Author:
Selena Lombardi BASc
Department of Mechanical and Industrial Engineering
University of Toronto
800 Bay St.
Toronto
CA

Abstract

Background: Usability testing is a critical component in the development of medical devices and services to minimize potential use errors that result in injury and death. Despite its importance, usability testing is often resource- and time-intensive, and traditional approaches have demonstrated inconsistencies in producing reliable results across evaluators. These limitations have raised concerns regarding whether usability evaluations adequately address the needs of diverse user groups. With the growing integration of artificial intelligence (AI) to augment usability testing practices, there is increasing uncertainty surrounding the extent to which these advancements will improve the quality of usability testing results and ensure equitable outcomes, particularly for marginalized communities.

Objective: The purpose of this study is to understand how AI will be implemented into usability testing processes and its potential impact on the inclusivity and equity of usability testing practices.

Methods: A narrative literature review was conducted to summarize the current state of AI in usability evaluation practices. An adapted version of the Society of Automotive Engineers (SAE) five levels of automation was used to assess the extent to which AI automates usability testing processes. The literature was also evaluated on a five-level scale for their contribution to equitable testing practices, ranging from no explicit consideration of equity (Level 0) to comprehensive equity integration (Level 5).

Results: 46 studies were reviewed, of which 35 focused on the development of AI tools designed to evaluate digital products, including mobile applications. These AI tools primarily supported data analysis for various usability testing methods, with particular emphasis on think-aloud protocols. The three most frequently developed tools were those aimed at identifying usability issues, performing sentiment analysis, and evaluating interface usability, leveraging innovative AI techniques. Most tools were designed to assist UX experts by automating specific stages of usability testing (Levels 1 and 2 automation). Six studies utilized AI to simulate human participants or substitute evaluators under specific conditions (Levels 3 and 4 automation). Despite these advancements, 23 studies did not address the inclusion of diverse user groups (Level 0 equity consideration), while 22 studies acknowledged equity considerations without incorporating them into their methodologies (Levels 1-3 equity consideration). Only three studies specifically explored the application of AI-supported usability tools with marginalized communities.

Conclusions: The literature review reveals a significant trend toward integrating novel AI models into usability testing, particularly over the past few years. However, despite these advancements, questions about the equity and inclusivity of these tools remain insufficiently explored. Future research efforts should develop standardized and equitable usability testing guidelines for traditional and AI-informed methods. Additionally, comparing these findings with evaluations of current AI-informed usability tools could provide insights into the market’s direction concerning equitable medical device and service development.

(JMIR Preprints 07/10/2024:67285)
DOI: https://doi.org/10.2196/preprints.67285

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

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

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

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

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

Original Manuscript

Paper Type: Original Paper

The Role of Artificial Intelligence in Usability Testing and its Potential Impact on Equity and Medical Solutions Design: A Narrative Literature Review

Authors:

Selena Lombardi, SL, Master of Applied Science Candidate, Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, Ontario, Canada

Enid Montague, EM, Associate Professor, Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, Ontario, Canada

Keywords: Usability Testing; Artificial Intelligence; Equity; Human Factors Engineering; Mhealth; Uhealth

Abstract

Background: Usability testing is a critical component in the development of medical devices and services to minimize potential use errors that result in injury and death. Despite its importance, usability testing is often resource- and time-intensive, and traditional approaches have demonstrated inconsistencies in producing reliable results across evaluators. These limitations have raised concerns regarding whether usability evaluations adequately address the needs of diverse user groups. With the growing integration of artificial intelligence (AI) to augment usability testing practices, there is increasing uncertainty surrounding the extent to which these advancements will improve the quality of usability testing results and ensure equitable outcomes, particularly for marginalized communities. *Objectives:* The purpose of this study is to understand how AI will be implemented into usability testing processes and its potential impact on the inclusivity and equity of usability testing practices. *Methods:* A narrative literature review was conducted to summarize the current state of AI in usability evaluation practices. An adapted version of the Society of Automotive Engineers (SAE) five levels of automation was used to assess the extent to which AI automates usability testing processes. The literature was also evaluated on a five-level scale for their contribution to equitable testing practices, ranging from no explicit consideration of equity (Level 0) to comprehensive equity integration (Level 5). *Results:* 46 studies were reviewed, of which 35 focused on the development of AI tools designed to evaluate digital products, including mobile applications. These AI tools primarily supported data analysis for various usability testing methods, with particular emphasis on think-aloud protocols. The three most frequently developed tools were those aimed at identifying usability issues, performing sentiment analysis, and evaluating interface usability, leveraging innovative AI techniques. Most tools were designed to assist UX experts by automating specific stages of usability testing (Levels 1 and 2 automation). Six studies utilized AI to simulate human participants or substitute evaluators under specific conditions (Levels 3 and 4 automation). Despite these advancements, 23 studies did not address the inclusion of diverse user groups (Level 0 equity consideration), while 22 studies acknowledged equity considerations without incorporating them into their methodologies (Levels 1-3 equity consideration). Only three studies specifically explored the application of AI-supported usability tools with marginalized communities. *Conclusions:* The literature review reveals a significant trend toward integrating novel AI models into usability testing, particularly over the past few years. However, despite these advancements, questions about the equity and inclusivity of these tools remain insufficiently explored. Future research efforts should develop standardized and equitable usability testing guidelines for traditional and AI-informed methods. Additionally, comparing these findings with evaluations of current AI-informed usability tools could provide insights into the market's direction concerning equitable medical device and service development.

Introduction

Early detection and prevention of patient harm in healthcare is an international policy priority, as it is a leading cause of morbidity and mortality internationally, with one in 20 patients subjected to *preventable* harm [1]. Human Factors engineering has become a practice used in the medical design process to successfully reduce preventable medical errors [2]. In one study, a surgical safety checklist based on human factors principles reduced the death rate from 1.5% to 0.8% and the complications rate decreased from 11% of patients to 7% of patients. As such, improved usability of medical devices and services is required to minimize potential use errors that result in injury and death [3].

Usability testing throughout all stages of development is critical to verifying that a given design is safe, efficient, and easy to use for diverse user groups, uses and use environments [4]. The U.S. Food and Drug Administration (FDA) emphasizes the importance of human factors engineering throughout the product development lifecycle, to minimize errors and adverse events, as well as gain a deeper understanding of patients and their conditions [5]. Depending on the classification of a medical device, the FDA mandates a pre-market review, including a human factors assessment based on their recognized standards, before the product's launch [6][7].

However, user testing is often a bottleneck in the development cycle, as it requires a lot of time and resources to run sessions with enough participants and analyze large amounts of data [8]. Moreover, research has raised concerns about the adequacy of *traditional* usability testing methodologies and their capacity to accurately represent the diversity of user populations. Molich et al. found variations of usability results of the same product tested between different organizations - questioning the assumption that current usability testing practices are uniform [9].

This finding raises concern as to whether current usability testing practices are equitable and inform inclusive and accessible medical device and service development. If a product/service is poorly designed, interventions can become more accessible to, adopted more frequently by, adhered to more closely by, or more effective to socioeconomically advantaged groups with greater access to resources– increasing health disparity between patient groups [10]. This healthcare disparity is often linked to social conditions such as level of education, occupational status, and income; residential segregation; environmental barriers; stigmatization; and discrimination. Healthcare disparities are associated with poor health outcomes and premature death of vulnerable groups, as well as increased health care costs [11]. Therefore, inadequate usability testing practices of medical devices and services can exacerbate the health gaps in marginalized communities.

Efforts have been made to improve equitable usability testing practices through expert discussions and equity assessment checklists [12][13]. Artificial intelligence (AI) also presents an opportunity to revolutionize usability testing by improving efficiency and providing less biased outcomes. Many companies involved in user testing are now integrating AI into their evaluation workflows to streamline the process [14][15]. Grey literature has explored how AI can enhance the efficiency and quality of usability evaluations throughout each phase, including automating the generation of test scripts, analyzing data from testing sessions, and even using AI-driven agents to simulate human behaviour [14][16][17]. Despite these promising advancements, there remains a lack of formal research on how AI-supported usability tools account for diverse user groups, particularly marginalized communities, and whether AI improves current testing practices or inadvertently perpetuates existing biases and inequalities.

The study described in this paper uses a narrative literature review of 46 papers to assess the current and foreseeable purposes and implications of AI in usability testing practices. This paper aims to evaluate the role of AI in usability testing found in the literature and how it will impact equity, particularly for marginalized communities - especially since traditional usability practices may not provide consistent results.

Methods

The following section details the narrative literature review search methods, including the key research questions, search query, strategy, inclusion and exclusion criteria, and evaluation methods.

Research Questions

This study will evaluate how AI integration in usability testing will impact equitable practice using the following questions.

- **RQ1**. How is and will AI be implemented in usability testing processes?
- **RQ2**. How does AI impact the efficiency of traditional usability testing practices?
- **RQ3**. How will AI impact the inclusivity and equity of usability testing, especially for marginalized communities?

Search Query

The following search query generated the most relevant results to answer the three research questions.

(“Artificial intelligence” OR “AI” OR “artificial general intelligence” OR “artificial superintelligence” OR “natural language processing” OR “machine learning” OR “image recognition” OR “neural network” OR “reinforcement learning” OR “speech translation” OR “recommend* engines” OR “recommend* systems” OR “automation”) AND

(“usability testing” OR “user testing” OR “human factors testing” OR “usability evaluation” OR “user evaluation” OR “human factors evaluation” OR “usability assessment” OR “user assessment” OR “human factors assessment” OR “usability validation” OR “user validation” OR “human factors validation” OR “user test” OR “usability test” OR “human factors test” OR “moderated usability test*” OR “unmoderated usability test*” OR “remote usability test*” OR “in-person usability test” OR “explorative usability test*” OR “comparative usability test*” OR “guerilla” OR “ usability interviews” OR “ergonomics test*” OR “ergonomics evaluation” OR “ergonomics assessment” OR “human factors engineering” OR “formative usability test*” OR “summative usability test*” OR “usability survey”)

Please note that medical-related keywords were not included in this search, as AI-supported usability tools can be generalizable to different types of products and services.

Moreover, terms related to equity and marginalized communities were removed from the search query, as there were no relevant papers which addressed all key terms (i.e. AI, usability testing, equity, and marginalized communities). Equity and marginalized communities are now variables coded for during the analysis. See Appendix A for the descriptions of usability testing, artificial intelligence, and equity and marginalized communities, in the context of this paper. See Appendix B for all keywords and synonyms tested throughout the development of this search query.

Search Strategy

Six databases were selected to search for relevant papers. Interdisciplinary databases, such as Scopus, Web of Science, and Google Scholar, were chosen to collect as many relevant papers about the topic. Engineering-related databases, such as IEEE and ACM Digital Library, were selected for topics related to AI and general usability testing. Lastly, the PubMed database aimed to collect papers specific to the usability testing of medical or healthcare products. The search query was applied to the title/abstract and keywords, if available, in the databases.

Inclusion and Exclusion Criteria

Table 1. The inclusion and exclusion criteria for relevant papers.

Inclusion Criteria	**Exclusion Criteria**
Must mention AI and usability testing.	*Manual* usability testing is used to evaluate AI-based design.
Must mention AI facilitating any phase of usability testing.	The usability testing method is best suited for explorative and early design phases.

Must be written between 2000 to 2023.	Written before the year 2000.
Must be translatable to English.	

Search Results

The following PRISMA diagram (Figure 1) outlines the screening process. SL conducted the initial screening of 3448 papers and SL and a second reviewer (refer to Acknowledgements) managed the full-text review of 189 papers. SL made the final decision for any conflicts. 46 papers were included in the final literature review. See the Reference list for all included papers ([P1-P46].

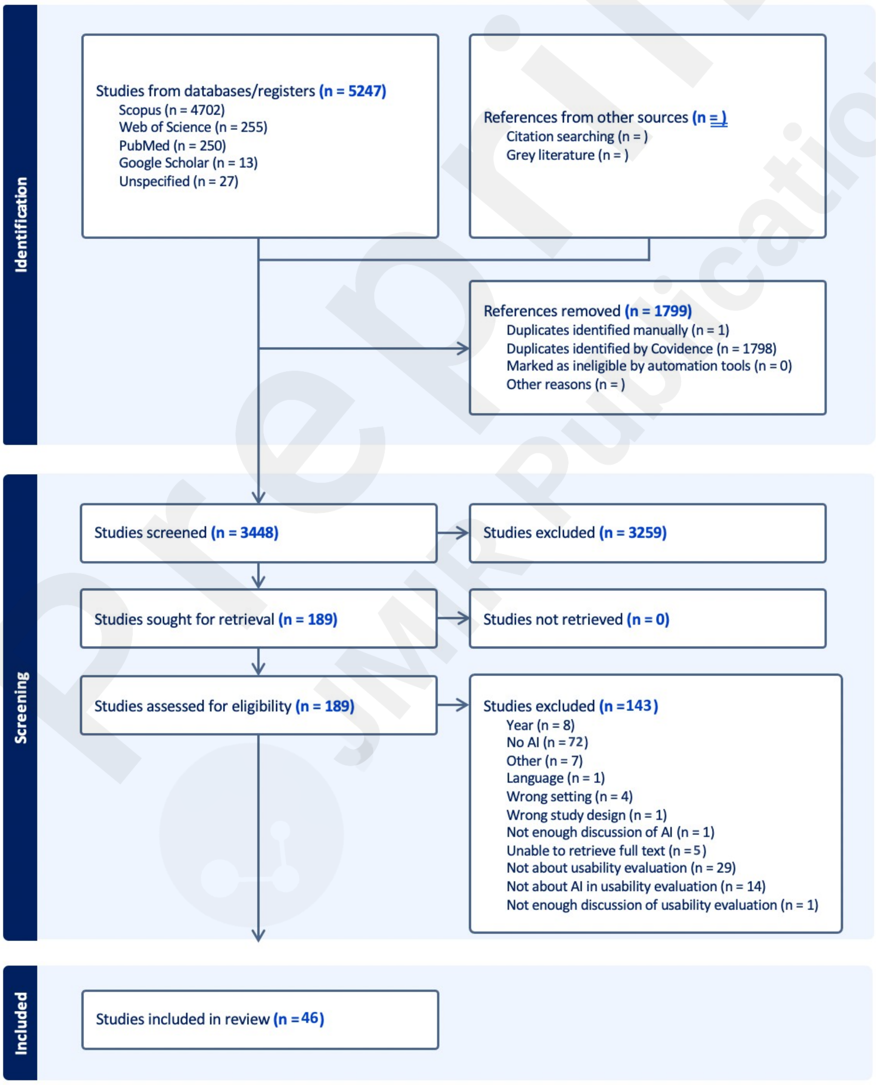
Identification
Studies from databases/registers (n = 5247)
Scopus (n = 4702)
Web of Science (n = 255)
PubMed (n = 250)
Google Scholar (n = 13)
Unspecified (n = 27)
References from other sources (n =)
Citation searching (n =)
Grey literature (n =)
References removed (n = 1799)
Duplicates identified manually (n = 1)
Duplicates identified by Covidence (n = 1798)
Marked as ineligible by automation tools (n = 0)
Other reasons (n =)
Screening
Studies screened (n = 3448)
Studies excluded (n = 3259)
Studies sought for retrieval (n = 189)
Studies not retrieved (n = 0)
Studies assessed for eligibility (n = 189)
Studies excluded (n =143)
Year (n = 8)
No AI (n = 72)
Other (n = 7)
Language (n = 1)
Wrong setting (n = 4)
Wrong study design (n = 1)
Not enough discussion of AI (n = 1)
Unable to retrieve full text (n =5)
Not about usability evaluation (n = 29)
Not about AI in usability evaluation (n = 14)
Not enough discussion of usability evaluation (n = 1)
Included
Studies included in review (n =46)

Figure 1. The PRISMA diagram of the screening process.

Evaluation Methods

The selected papers were evaluated on a five-point scale based on the Society of Automotive Engineers (SAE) levels of automation, as shown in Table 2. This scale measures the impact of AI on traditional usability testing practices (RQ2) [18].

Table 2: The five SAE levels of automation applied to usability testing.

Level of Automation	**Description**
Level 0 - No Automation	Evaluator conducts all phases of testing manually.
Level 1 - Assistance	Evaluator is required for all phases of the test. Some tasks are supported by AI, but the evaluator is in charge of making decisions.
Level 2 - Partial Automation	Rule-based decision-making to evaluate specific aspects of the design. AI is dependent on the evaluator for all input and interventions.
Level 3 - Conditional Automation	Inferred decision-making to evaluate the general usability of the design, including recommendations and actions. AI can handle most operations with some exceptions.
Level 4 - High Automation	AI automatically takes action to achieve all service-level objectives of usability testing. AI can handle all operations with few exceptions.
Level 5 - Full Automation	All operations can be handled by AI with or without human input without exceptions.
Not Applicable	Literature Review

Each study was evaluated on a five-point scale, as shown in Table 3, regarding how much their AI- informed usability features incorporated equity, inclusion, and diversity (RQ3).

Table 3: The five levels of equity consideration.

Level of Equity Consideration	**Description**
Level 0 - No Discussion	No mention of diversity in the study or

	discussion related to equity.
Level 1 - Mentioned Briefly	1-2 sentences about the impact of the study on equity or diversity. Any mentioned problems are not elaborated.
Level 2 - Problems Identified and Described	Specific equity problems are thoroughly discussed, but no solutions for future research are provided or accounted for.
Level 3 - Problems Elaborated and Solutions Offered	Specific equity problems are thoroughly discussed and some solutions for future research are provided or accounted for.
Level 4 - Insightful Discussion	Specific problems are thoroughly discussed with elaborated solutions on how future work can be inclusive and diverse. If the paper is an empirical study (with human participants), a diverse group of participants is considered for equitable and representative results.
Level 5 - Inclusivity, Diversity, and Equity Incorporated into Paper Topic	Study considers and tests elements of inclusivity, diversity, and equity as it pertains to AI in usability testing.

Results

The following section synthesizes the search results related to each research question. Figure 2 represents the number of articles published per year. 24 of the 46 papers included in this review were published in or after 2020, with the majority of papers from 2023 [P2,P4,P7,P8,P15,P16-P21,P26-P37,P39]. There has been a growing trend of research relevant to this field since 2018, apart from the spike in research in 2014. These results align with a 2023 systematic review study which found that machine learning (ML) algorithms for eye-tracking devices were overwhelming and solely reported after 2018 [P37].

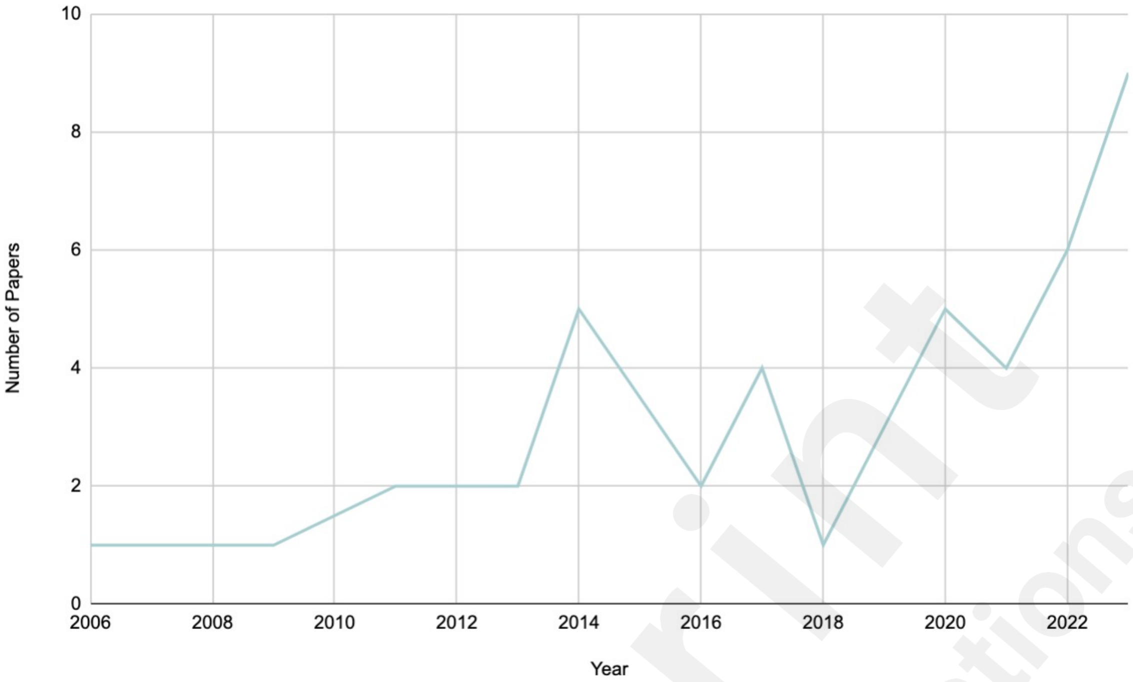

Figure 2.

The number of included papers by year (2006 – 2022).

Figure 3 displays the proportion of included papers according to publication type (journal articles, conference papers, or conference proceedings). Just over half the included papers (24) were empirical studies [P1-P9,P11,P14-P17,P19-P21,P25-P34,P36,P38-P43,P45,P46], and two papers were systematic reviews [P23,P37].

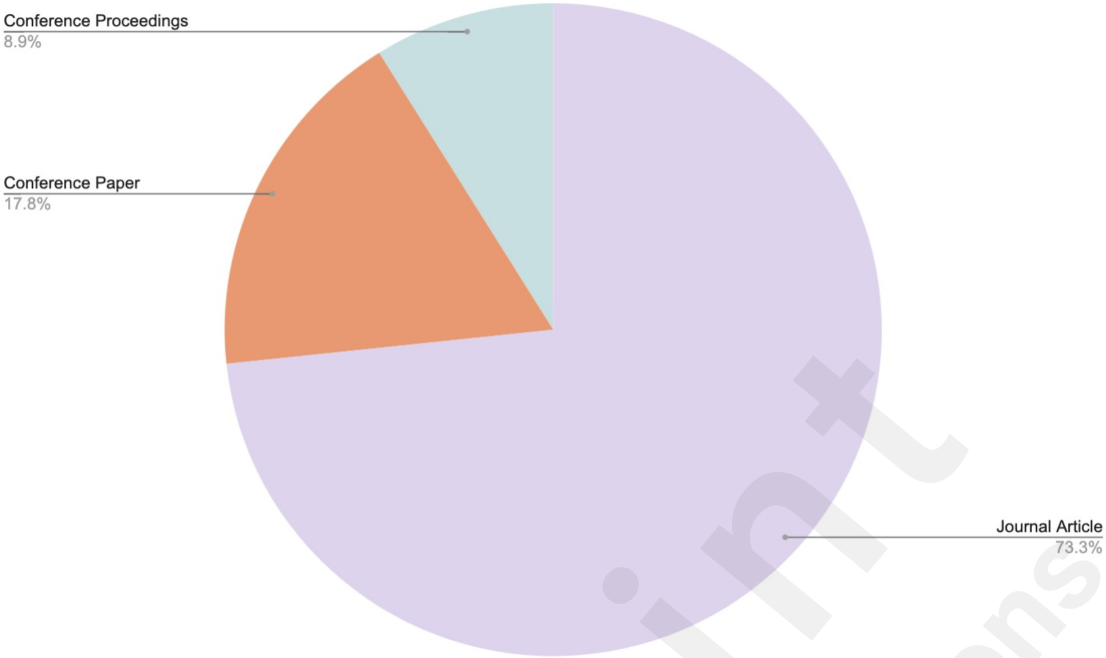

Figure 3. The proportion of included papers by publication type.

AI Implementation in Usability Testing Practice

35 of the 46 papers implemented AI usability testing tools intended to test digital products. These products included mobile applications [P15,P20,P21,P34,P36], tablets [P41], websites [P1,P3-P5,P10],

[P15,P17,P19,P20,P26,P27,P29,P46], e-learning systems [P22], video games [P2,P24], and virtual reality systems [P6,P33,P43]. Tools designed to assess the usability of physical products were specifically tailored for these products and could not be used for evaluating other physical devices. Physical products tested were coffee machines [P4,P16,P17,P20], remote controls [P4,P16], water faucets [P25], and thermostats [P9].

The AI tools were integrated into various types of usability evaluations including moderated [P1,P4,P6-P8,P10-P20,P27,P29-P33,P40-P46], and unmoderated testing [P7,P11-P14,P20,P21,P25-P32,P38-P39,P4-P44,P46], formative [P3-P5,P7,P9,P11-P15,P17,P18,P21,P22,P24-P36,P38,P40-P46] and summative testing [P4,P7,P11-P13,P15,P17,P18,P21,P22,P24-P33,P40-P46], in-person

[P4,P7,P11-P13],P15,P17,P18,P21,P22,P24-P33,P40-P46] and remote testing [P17,P38,P43]. Think-aloud usability testing was a primary focus of seven papers, especially those that incorporated audio and transcript information as input into their models [P4,P7,P8,P16,P19,P20,P41].

40 papers explored various AI techniques to support usability testing data analysis [P1,P3-P23,P25-P36,P38-P43]. This includes the 2023 systematic study in which half of the ML algorithms were used for data analysis or processing raw data [P37]. Table 4 synthesizes the six most studied purposes of AI in usability testing and the papers that corresponded to these techniques. Refer to Appendix C for the entire list of AI roles in usability testing found in the literature.

Table 4. The six most studied purposes of AI in usability testing and corresponding papers.

Role of AI	**Details**	**Papers**
Detects occurrence of a usability issue	Tool identifies occurrences of potential usability issues after a session.	[P4,P7,P8,P16,P19, P20,P25,P33,P41, P45]
Sentiment analysis	Tool gauges users' emotional reactions while using a product. It detects positive, neutral, and negative reactions, which can be indicative potential usability issues.	[P27,P28,P31,P35,P40, P42]
Scores interface usability	Tool provides a quick estimate of the general usability of a digital interface.	[P21,P38,P46,P13, P46]
Detects types of usability issues	Tool detects particular usability issues given a set of guidelines.	[P3,P5,P6,P9,P36]
Uses intelligent agent participants	Intelligent agents simulate the behaviours of human participants in usability testing.	[P2,P18,P24,P44]
Determines percent contribution of	Tool identifies what features of their product can be further improved	[P12,P22,P34]

usability features	based on user perceptions in questionnaires.	

Refer to Appendix D for the list of AI algorithms used for the purposes listed in Table 4.

Detection of Usability Issue Occurrences

Ten papers developed and evaluated an AI-supported usability tool that detected the occurrence of usability issues [P4,P7,P8,P16,P19,P20,P25,P33,P41,P45]. Six studies extracted verbalization data (speech loudness, speed, keywords, pitch, structure patterns and sentiment) from sessions to train their models to detect whether a user experienced a usability issue at a certain point in time [P4,P7,P8,P19,P20,P25]. Other studies used visual data from recordings (head and hand movements) [P25,P33], electroencephalogram (EEG) signal patterns [P33], written reviews to predict when usability issues occurred [P41], and user interaction logs [P45]. The list of AI models implemented is provided below.

- Natural language processing (NLP) [P4,P41], typically used to convert natural language from transcripts to a comprehensible computer input,
- Google speech recognition [P8],
- Neural networks [P25], convolutional neural networks [P4,P16], recurrent neural networks [P4,P16],
- Keyword matching algorithm [P16],
- Classifier algorithms: supervised ML (SVM) [P4,P16,P25,P33], random forest [P4,P16,P33], logistic regression [P25,P33], k-nearest neighbour [P33], and multilayer perceptron [P33], used to learn which features contribute to the identification of a usability issue,
- Closed sequential pattern mining [P45].

Three articles compared the accuracy of their algorithm when trained on different combinations of input features. They discovered that their models achieved greater accuracy when the most input features were considered [P4,P16,P33]. Examples of input features included transcripts [P4], verbalization and speech features [P4,P33], head and hand movements [P33], EEG signals [P33], questionnaires [P33], and speed while completing tasks [P33].

Four articles compared the accuracy of various classifier algorithms detecting usability issue occurrences [P4,P16,P25,P33]. Two papers found SVM to perform the best considering all three metrics: precision (76%), recall (70%), and F-score (73%), even though other models (i.e. random forest, convolutional neural networks, and recurrent neural networks) performed better in specific categories [P4,P16]. However, one paper found random forest to have the greatest overall accuracy (84.23%), over SVM (80.82%) [P33]. Two papers compared the accuracy of their algorithms to the results of manual testing [P16,P25]. One paper found that, on average, neural networks underestimated usability issues by 9.1 percentage points, while SVM and logistic regression classifiers overestimated by 13.6 and 31.0 percentage points, respectively. Overall, the accuracy of AI algorithms implemented to detect usability issues differed depending on the input features considered.

One paper visualized their AI model into a voice and/or text UX assistant interface. During data analysis, UX evaluators would pose questions to the AI assistant about the session, and the AI assistant would provide their analysis of the session recording. The study found that evaluators mostly asked questions about participant actions, mental models, and suggestions. UX evaluators were more trusting of factual and objective information about the session from the AI assistant, which was also easier to miss [P7]. Another tool was integrated into an interface that displayed a timeline of the session recording and time-stamped points in which the AI detected usability issues. Additionally, the tool highlights the video segments that have the same set of problem features as the currently paused timestamp, as shown in Figure 4 [P16].

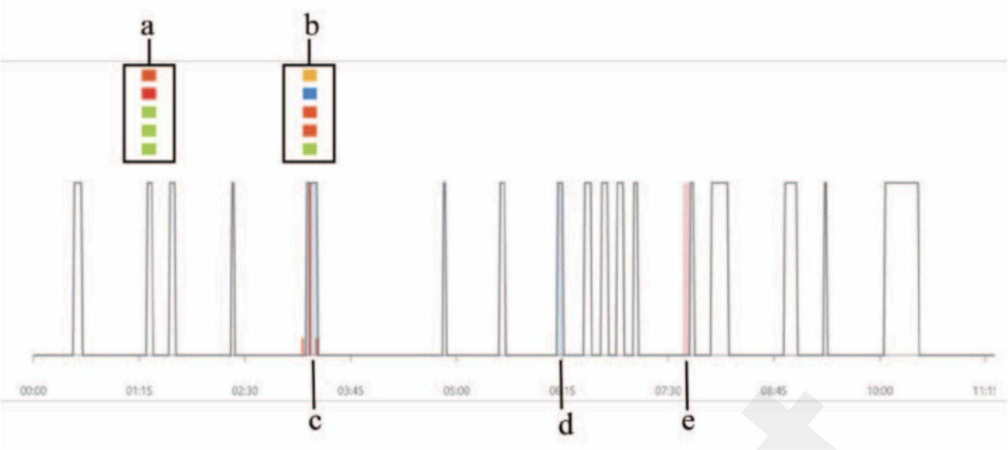

Figure 4. VisTA Problem Timeline [P16]

Sentiment Analysis

Six papers implemented AI to conduct a sentiment analysis of user experiences in usability testing sessions. These tools used EEG signals to determine the valence and arousal of participants during sessions [P27,P40], video recording data [P28,P31], and written reviews [P35,P42]. The list of specified AI models used is provided below.

- Recurrent neural network [P27]
- Convolutional neural network [P28,P31]
- NLP [P35]
- Max entropy and k-nearest neighbour [P42]

Two articles reported their algorithms achieved an average overall accuracy of approximately 85%, detecting positive reactions more accurately than negative sentiments [P31,P42]. Additionally, one paper reported an overall accuracy rate of 90%, with no specific dimension of accuracy (i.e., precision, recall, specificity, true positive rate, F1-Score, and decision) below 80% [P27]. In contrast to other paper results, this AI technique was better at detecting negative valence (92.13%) than positive valence (90.19%). Lastly, a paper employing a two-part algorithm for sentiment analysis reported a final recognition accuracy of the seven basic emotions and neutral state of 84.1%, with fatigue resulting in the lowest accuracy of 74.5% [P28].

Impact of AI on Usability Testing Practice

Table 5 provides an overview of the degree of automation achieved through the integration of AI into usability testing practices across the included research papers. In the context of usability testing, automation refers to leveraging AI technologies to execute tasks, thereby decreasing UX evaluator intervention.

Table 5. An overview of the degree of automation achieved through the integration of AI into usability testing practices.

Level of Automation	**Description**	**Papers**
Level 0 - No Automation	Evaluator conducts all phases of testing manually.	
Level 1 - Assistance	Evaluator is required for all phases of the test. Some tasks are supported by AI, but the evaluator is in charge of making decisions.	[P4,P7,P8,P11, P13, P14,P15,P17,P19,P20, P21,P25,P26,P27,P28,P29, P30,P31,P32,P34,P35,P39, P45,P46]
Level 2 - Partial Automation	Rule-based decision-making to evaluate specific aspects of the design. AI is dependent on the evaluator for all input and interventions.	[P1,P3,P5,P6,P9,P10, P12, P16,P22,P33,P38]
Level 3 - Conditional Automation	Inferred decision-making to evaluate the general usability of the design, including recommendations and actions. AI can handle most operations with some exceptions.	[P2,P18,P24,P36,P43]
Level 4 - High Automation	AI automatically takes action to achieve all service-level objectives of usability	[P44]

	testing. AI can handle all operations with few exceptions.	
Level 5 - Full Automation	All operations can be handled by AI with or without human input without exceptions.	
Not Applicable	Literature Review	[P23,P37]

Based on the results from Table 5, AI primarily assisted evaluators, by offering a second opinion for usability errors or performing tedious tasks. These results align with a 2023 systematic study that found around 8% of all ML-based eye-tracking articles reported using automatic processes [P37]. Further, high automation was often used for specific use cases and in a narrow area of applicability [P37]. Studies identified as Level 2 - Partial Automation used AI to identify particular usability issues of a product given a set of guidelines, such as Nielsen Norman Heuristics [P3][P5][P6][P9]. Level 3 - Conditional automation included studies that implemented intelligent agents to mimic the behaviours of human participants in usability studies or replaced the evaluator under certain conditions [P2,P18,P24,P36]. The Level 4 - High automation paper implemented intelligent agents in video games [P44].

Impact of AI on Equity

Each article was evaluated on a zero-to-five-point scale regarding how much they incorporated inclusion and diversity into their study design and/or considered the implementation of equitable usability testing practices in their discussion. Table 6 summarizes the results.

Table 6. An overview of the degree of equity consideration in study design of papers.

Level of Equity Consideration	**Description**	**Papers**
Level 0 - No	No mention of diversity in the	[P2,P3,P5,P6,P7,P8,P9,P10

Discussion	study or discussion related to equity.	, P11,P12,P13,P21,P27,P29, P30,P31,P34,P35,P38,P40, P42, P44,P46]
Level 1 - Mentioned Briefly	1-2 sentences about the impact of the study on equity or diversity. Any mentioned problems are not elaborated.	[P1,P14,P18,P22,P24,P26, P28,P32,P33,P36,P37,P39, P41,P43]
Level 2 - Problems Identified and Described	Specific equity problems are thoroughly discussed, but no solutions for future research are provided or accounted for.	[P4,P15,P23,P45,P16,P17]
Level 3 - Problems Elaborated and Solutions Offered	Specific equity problems are thoroughly discussed and some solutions for future research are provided or accounted for.	[P16,P17]
Level 4 - Insightful Discussion	Specific problems are thoroughly discussed with elaborated solutions on how future work can be inclusive and diverse. If the paper is an empirical study (with human participants), a diverse group of participants is considered for equitable and representative results.	
Level 5 - Inclusivity,	Study considers and tests elements of inclusivity, diversity, and equity	[P19,P20,P25]

Diversity, and Equity Incorporated into Paper Topic	as it pertains to AI in usability testing.	

As shown in Table 6, half of the papers (23) did not consider the impact of AI on equity, diversity, and inclusion (EDI) of usability testing methods. Several papers briefly mentioned that AI could capture and analyze large datasets in a shorter amount of time, generating more accurate results [P18,P20]. Some studies noted that AI can increase the objectivity of usability testing results through the measurement of quantitative data [P22,P23,P33,P34,P38], while other papers indicated that the output from AI tools was still subjective [P17,P23,P45]. The level of subjectivity is dependent on the purpose of the AI-supported usability tool and what data it collects. For example, a sentiment analysis AI model trained on written reviews will likely result in more subjective results than a model trained on participant answers on a user experience Likert scale questionnaire. Similar conclusions were drawn in a systematic review conducted in 2017. The article argued that there were ongoing ambiguities in quantitatively measuring usability, as well as a gap between its more objective dimensions (e.g. effectiveness and efficiency) and subjective aspects (e.g. user satisfaction) [P23].

One paper noted that there is a lack of overlap between the issues found by the AI model and traditional user testing, noting the same concern about the adequacy of current testing methodologies and their capacity to accurately represent the diversity of user populations [9,P41].

Several papers highlighted that AI serves as a valuable second opinion for UX evaluators during data analysis, thereby mitigating the evaluator effect [P16,P20]. This effect occurs when distinct groups or individuals conducting a usability evaluation identify vastly different issues. However, AI has the potential to perpetuate the evaluator effect, depending on how it was trained and the degree of reliance on it. If the model was trained on a limited dataset from a small group of UX evaluators or learns predominantly from the inputs of a specific evaluator, it introduces biases. Additionally, if the evaluator overly relies on the AI tool, it influences their perceptions of the observed session, exacerbating the evaluator effect [P17].

Three articles thoroughly integrated aspects of EDI in their studies. One paper evaluated the verbalization patterns of Chinese non-native English speakers in different think-aloud testing protocols. Comparing native and non-native English speakers against verbalization categories (i.e. what they talked about: procedure, reading, design and other), the study found that despite subtle differences in verbalization patterns, there was no statistically significant effect on either the number of problem encounters or the number of actual problems identified between language groups. They concluded that intermediate-level non-native English speakers can be as effective as native speakers in identifying UX problems. As such, verbalization data from non-native speakers can be used to train AI that aims to detect usability problems [P19]. Led by the same author, a similar study compared the verbalization patterns of young adults versus older participants. They found older adults used more negative sentiments and noted their observations, spoke aloud more frequently while reading, and low speech rate was not a precise indicator of usability issues like young adults [P20]. The third study evaluated an AI-supported usability tool that detected the occurrence of usability issues for older adults with dementia while using three types of water faucets. The AI models first segmented clips of interest from the recordings and identified the specific action (e.g. turning on the faucet) where the usability issue occurred [P25].

Ten studies evaluated their tool with university students and five studies included UX evaluators of a specific educational and national background [P1,P5-P12,P17,P19, P22,P29,P32,P34,P40]. There was a notable underrepresentation of elderly individuals, people with disabilities, and those from diverse ethnic and economic backgrounds. These results align with the findings from a 2023 systematic review study, in which nearly all studies included young adults and students, while only a few studies mentioned including people with disabilities in their samples

[P37]. Articles that evaluated AI-supported usability tools with UX evaluators and professionals had a smaller sample size than studies with other participant demographics. Refer to Appendix E for a summary of the participant demographics noted in experimental studies.

Discussion

The literature explores various techniques and AI models to aid in usability testing, especially in the most time- and resource-intensive phase of data analysis. This exploration is crucial for identifying the most optimal AI models, as well as how to best present these algorithms to support HF analysts. Despite these advancements, the scope of many of these tools is not generalizable to medical devices and services, as they have not been well-evaluated with diverse user groups and scenarios. There are still research gaps that need to be addressed before or in parallel with AI development for usability testing. These claims are explored in detail below.

The Usefulness of AI-Supported Usability Tools in Usability Testing

An analysis of the usefulness of the most studied AI features from the Results section is provided below. Usefulness is defined as the degree to which the AI-supported usability tools alleviate UX evaluator workload in their tasks, as it pertains to testing medical devices and services.

Detects occurrence of a usability issue:

- *Benefits:* Trained on a wide set of input features, these tools can quickly and accurately detect the occurrence of usability issues and can find more usability issues than traditional user testing [P41].
- *Limitations:* The tools cannot identify what usability issue the user experienced and cannot explain why they experienced this issue. Some techniques inaccurately segmented the video clip with the usability issue, especially when identifying the start and end of an issue [P16,P25]. Given the current capabilities of the tool, UX evaluators are still required to assess

every issue. Moreover, the diversity of the data the tool is trained on is limited, not accounting for all user groups. As such, the tool may not significantly reduce the amount of time and cognitive load required to conduct their analysis.

Sentiment analysis:

- *Benefits:* These features are beneficial for formative user research and early prototype testing when user perceptions are of greater priority than the objective usability of a device. Studies focusing on real-time measurements of valence, arousal, and cognitive load during tasks could be especially valuable for high-fidelity usability testing. This approach allows evaluators to pinpoint specific tasks to discuss during debrief sessions, aiding in identifying whether users encountered usability issues.
- *Limitations:* Similar to tools that detect the occurrence of usability issues, AI-informed sentiment analysis is unable to explicitly explain why the user experienced certain emotions, requiring evaluators to manually assess each potential usability issue occurrence [19]. Thereby, not significantly reducing the amount of time and cognitive load required to conduct their analysis.

Scores interface usability:

- *Benefits:* These tools help identify whether an iteration of an interface design has improved the overall usability of the product. This is particularly useful for comparing the overall usability of two products or services [20].
- *Limitations:* These tools are not particularly beneficial for usability testing which requires the identification of specific usability issues and recommendations for improvement of the product [4]. Moreover, its accuracy is limited by the features it was trained on, such as interface architecture or selected questionnaires [P13,P46].

Detects types of usability issues:

- *Benefits:* UX evaluators can consult with these tools to confirm their findings, which is especially beneficial if evaluators are unable to collaborate with other colleagues.
- *Limitations:* The usefulness of automated usability problem detection is currently specific to certain guidelines or products/services, making it difficult to adopt in general usability practices where usability issues are more complex. Another challenge that may arise with the development of this feature is over-reliance [P17,21]. Especially as the feature becomes more advanced, UX evaluators may not feel the need to critically evaluate the products/services as thoroughly as they did with traditional testing. Over-reliance can result in a bias towards the AI tool and the data it was trained on.

Uses intelligent agent participants:

- *Benefits:* Intelligent participants allow designers and evaluators to identify different perspectives of using their product/service that may not be easily accessible in real contexts or apparent when testing with human participants. This was demonstrated with usability testing of video games, as simulated agents can traverse paths through a video game that a sample of human users may not have passed through [P2,P24]. It can positively impact inclusivity in user testing, as there are barriers to traditional user testing that prevent some user groups from participating. For example, usability testing sessions require participants to carve out time out of their day to participate. This disproportionately prevents low-wage participants from participating in sessions, as they tend to work substantially long hours [20]. Intelligent agents can encapsulate a basic level of the perspectives of excluded participants, still including their experiences in the design of the product/service.
- *Limitations:* Simulated agents come with the risk of oversimplifying human complexities, resulting in the introduction of biases and stereotypes. Further, evaluators must assess

whether the behaviours exhibited by intelligent agents frequently occur in real-world settings and how severe the consequences are of these behaviours.

Determines percent contribution of usability features:

- *Benefits:* These tools assist designers to prioritize which features of their product should be further improved based on user perceptions.
- *Limitations:* These tools are trained on provided questionnaires and checklists, neglecting results from scenario testing and evaluator observations. Further, these tools do not specify the usability issues associated with a particular feature, requiring the evaluator to manually conduct their root cause analysis.

Impact of AI-supported Usability Testing Tools and UX Evaluator Perceptions

Most of the AI-supported usability tools found in the literature were intended to assist UX evaluators with their practice. More studies that looked to automate aspects of the process replaced participants rather than the evaluators. These research directions align with two 2023 survey studies evaluating the perceptions of designers and UX evaluators who use AI-assisted usability tools in their work. Evaluators felt that AI tools should prioritize human oversight and control, as AI cannot replicate human capabilities [21][22]. Although these perceptions may not be true, initially developing AI-supported usability tools that align with the UX evaluator expectations will be more easily adopted in practice. Moreover, while data analysis remains the most time- and resource-intensive aspect of usability testing, another solution would be to further develop AI tools that can handle laborious tasks that do not require the same degree of problem-solving capabilities as data analysis. Examples include the development of AI models to support the generation of test plans and test scripts, as well as participant recruitment and creative work [22]. UX evaluators may be more accepting of fully automating tasks that do not require as much problem-solving capability and are more well-tested with technology such as ChatGPT. It should be noted that this finding may be a result of a limitation of this literature review, as the papers included in this study were required to develop/evaluate AI-supported tools specifically for usability testing. AI currently implemented into other experiment methods and could potentially be applied to usability testing, such as participant recruitment, were not included.

Evaluators were also concerned that AI tools may not be compatible with existing design workflows [21]. Five studies addressed this concern and sought feedback from UX evaluators regarding the design of their tools and how the tools would integrate into the workflow [P5,P7,P8,P17,P32].

However, UX evaluators from the survey studies identified four other concerns regarding the use of AI in their practices that were not addressed in the literature and were not areas of focus in this review. A list of these concerns is provided below:

1. Designers must acquire new skills to use AI tools [21].
2. Sufficient and appropriate training data is challenging to acquire [21].
3. AI algorithms have the potential to harm the user [21][22].
4. AI poses a threat to privacy and confidentiality [21][22].

These concerns should be considered when evaluating the development of future AI-supported usability tools to ensure that not only they are well-adopted by evaluators, but also provide reliable results, especially in high-risk environments such as healthcare.

Application of AI-supported Usability Tools in Equitable Practice

Overall, the literature is in the early stages of exploration, investigating a variety of novel AI techniques to support usability testing, particularly in data analysis. However, there is a lack of robust, well-evaluated models that are ready for practical implementation, especially in medical device or service testing.

Many AI implementations studied are limited to specific product types, which restricts their applicability in broader contexts. With technical and physical components in many medical device products and services, AI-supported usability tools can currently only assist specific aspects of an evaluation, such as interface design.

Current research directions generalize their AI models to a variety of evaluation methods, with a focus on think-aloud testing. These tools might be less effective for summative usability testing, which is typically mandated by FDA guidelines, as this method requires users to evaluate the product or service in real-world settings. Consequently, users may not articulate their actions and thoughts while interacting with the product or service. It would be beneficial to develop and test the accuracy of AI models in specific evaluation contexts to gain insights into what types of models best serve certain evaluation types.

What is of more concern is that the metrics to measure the accuracy of these models are based on the traditional usability practices of a few UX evaluators. There is research to suggest that traditional usability testing is inadequate to accurately represent the diversity of user populations, as a result of the evaluator effect [5,P17,P23]. As such, we cannot be certain that AI is eliminating, perpetuating, or exacerbating the biases that exist in current testing processes if they are being tested against traditional practices. The potential for biases in AI was also a primary concern of UX evaluators found in two 2023 survey studies [21][22]. As such, ensuring that the biases of these tools are mitigated and communicated to evaluators so that they know how to mitigate them is significant when these AI-supported usability tools are deployed in practice.

Priority Research Directions

With these considerations, the priority for future research should be to develop standardized traditional usability practices that better account for equity and mitigate the evaluator effect. Collaboration between UX, AI, and sociotechnical experts is essential for expanding the understanding of usability testing beyond a purely technical view, to include the broader social systems they function within. Engaging various disciplines and stakeholders is necessary to identify equity issues in current practices and establish scientifically grounded guidelines for addressing them

[23]. These guidelines should be prioritized in foundational human factors education and regularly updated to align with the continuous evolution of AI development [21]. These guidelines will provide a more reliable baseline for researchers to compare their AI models to. Further, researchers should continue testing their models against a variety of user groups. It is recommended to test these models with individuals of various identities, including socio-economic status, educational background, sex, etc., and consider how the intersectionality of these identities can affect the accuracy and reliability of the AI algorithms.

However, as user testing companies and experts are already incorporating AI tools into their practices, it would be of interest to examine what particular AI techniques are being developed and what considerations have been made in terms of their impact on equitable product development. This may include the evaluation of existing AI-supported usability tools currently on the market, such as UserTesting AI. It is more significant for research to evaluate how these tools impact the equity of usability practices, as these tools are currently impacting the usability testing practice of medical devices and systems.

Conclusions

Equitable usability testing is critical for medical device and system testing, as it minimizes potential use errors resulting in patient injury and death. Integrating AI tools into usability testing processes only emphasizes the need to guarantee that our practices address and ideally alleviate potential equity failures. The literature has made strides to develop novel AI techniques to facilitate various usability testing methods, especially tools that detect the occurrence of usability issues and sentiment analysis. Currently, AI usability tools are intended to assist UX evaluators with their practices, aligning with UX expert preferences. However, the impact of AI on the equity of usability testing practices, especially for marginalized communities, has not sufficiently been explored in research. Overall, the scope of many of these AI usability tools is not generalizable to medical devices and services, as they have not been well-evaluated with diverse user groups and scenarios, and various types of products and services. It is recommended for future usability research to evaluate the current AI-supported usability tools on the market to ensure that they are considering the impact of their platforms on the equity of usability testing practices. More importantly, AI, UX, and ethics experts should prioritize co-developing standardized and equitable usability practices for human factors regulations and guidelines. These considerations can significantly reduce the number of poorly designed medical devices and systems, and thereby decreasing health disparities which are associated with health outcomes, premature death, and health care costs.

Acknowledgements

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada, [funding reference number RGPIN-2024-05057]. We would like to thank Ruben Tjhie for his guidance and support as the second reviewer for screening included papers.

Cette recherche a été financée par le Conseil de recherches en sciences naturelles et en génie du Canada, [numéro de référence RGPIN-2024-05057]. Nous tenons à remercier Ruben Tjhie pour ses conseils et son soutien en tant que deuxième évaluateur avec la sélection des articles inclus.

Conflicts of Interest

None declared

Abbreviations

AI: artificial intelligence

EDI: equity, diversity, and inclusion

EEG: electroencephalogram

FDA: Food and Drug Administration

ML: machine learning

NLP: natural language processing

SAE: Society of Automative Engineers

SVM: supervised modeling

UX: user experience

References

1. Panagioti M, Khan K, Keers RN, et al. Prevalence, severity, and nature of preventable patient harm across medical care settings: systematic review and meta-analysis. *BMJ*. Published online July 17, 2019:l4185. doi:10.1136/bmj.l4185
2. Carayon P. Human factors in patient safety as an innovation. *Applied Ergonomics*. 2010;41(5):657-665. doi:10.1016/j.apergo.2009.12.011
3. Carayon P, Wood KE. Patient safety - the role of human factors and systems engineering. *Stud Health Technol Inform*. 2010;153:23-46.
4. Food and Drug Administration. Applying Human Factors and Usability Engineering to Medical Devices Guidance for Industry and Food and Drug Administration Staff. Published online June 21, 2011. https://www.fda.gov/media/80481/download
5. Food and Drug Administration. Human Factors and Medical Devices. https://www.fda.gov/medical-devices/device-advice-comprehensive-regulatory-assistance/human-factors-and-medical-devices
6. Food and Drug Administration. Classify Your Medical Device. https://www.fda.gov/medical-devices/overview-device-regulation/classify-your-medical-device
7. Food and Drug Administration. Human Factors: Premarket Information - Device Design and Documentation Processes. https://www.fda.gov/medical-devices/human-factors-and-medical-devices/human-factors-premarket-information-device-design-and-documentation-processes#standards
8. Thompson KE, Rozanski EP, Haake AR. Here, there, anywhere: remote usability testing that works. In: *Proceedings of the 5th Conference on Information Technology Education*. CITC5 ’04. Association for Computing Machinery; 2004:132-137. doi:10.1145/1029533.1029567
9. Molich R, Ede MR, Kaasgaard K, Karyukin B. Comparative usability evaluation. *Behaviour & Information Technology*. 2004;23(1):65-74. doi:10.1080/0144929032000173951
10. Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. *J Am Med Inform Assoc*. 2018;25(8):1080-1088. doi:10.1093/jamia/ocy052
11. Gibbons MC, Lowry SZ, Patterson ES. Applying Human Factors Principles to Mitigate Usability Issues Related to Embedded Assumptions in Health Information Technology Design. *JMIR Human Factors*. 2014;1(1):e3524. doi:10.2196/humanfactors.3524
12. Rutter S, Zamani E, McKenna-Aspell J, Wang Y. Embedding equality, diversity and inclusion in usability testing: Recommendations and a research agenda. *International Journal of Human-Computer Studies*.
13. Benkhalti M, Espinoza M, Cookson R, Welch V, Tugwell P, Dagenais P. Development of a checklist to guide equity considerations in health technology assessment. *International Journal of Technology Assessment in Health Care*. 2021;37(1):e17. doi:10.1017/S0266462320002275
14. Goli S. Using AI to analyze how your users think & feel. https://trymata.com/blog/2023/06/14/ai-user-testing-analyze-how-users-think-feel/
15. UserTesting. UserTesting AI. https://www.usertesting.com/platform/AI
16. Bhargava A. Revolutionising User Testing: How AI will transform the way we evaluate user experience. https://medium.com/design-bootcamp/revolutionising-user-testing-how-ai-will-transform-the-way-we-evaluate-user-experience-f441ffff0ccb

17. MacMillan A. Usertesting's vision for Artificial Intelligence: AI and ML. July 27, 2023. https://www.usertesting.com/blog/UserTesting-AI-vision
18. SAE International. SAE levels of Driving AutomationTM refined for clarity and international audience. https://www.sae.org/blog/sae-j3016-update
19. Specht A, Obaidi M, Nagel L, Stess M, Klünder J. What is Needed to Apply Sentiment Analysis in Real Software Projects: A Feasibility Study in Industry. In: Lárusdóttir MK, Naqvi B, Bernhaupt R, Ardito C, Sauer S, eds. *Human-Centered Software Engineering*. Springer Nature Switzerland; 2024:105-129. doi:10.1007/978-3-031-64576-1 6
20. Peres S, Phillips R. Validation of the System Usability Scale (SUS). *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 2013;57:192-196. doi:10.1177/1541931213571043
21. Chaudhry BM. Concerns and Challenges of AI Tools in the UI/UX Design Process: A Cross-Sectional Survey. In: *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM; 2024:1-6. doi:10.1145/3613905.3650878
22. Knearem T, Khwaja M, Gao Y, Bentley F, Kliman-Silver CE. Exploring the future of design tooling: The role of artificial intelligence in tools for user experience professionals. In: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM; 2023:1-6. doi:10.1145/3544549.3573874
23. Schwartz R, Vassilev A, Greene K, Perine L, Burt A, Hall P. *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*. National Institute of Standards and Technology (U.S.); 2022:NIST SP 1270. doi:10.6028/NIST.SP.1270
24. Jakob Nielsen. Usability 101: Introduction to Usability. January 3, 2012. https://www.nngroup.com/articles/usability-101-introduction-to-usability/
25. 18F Methods. A collection of tools to bring human-centered design into your project. https://guides.18f.gov/methods/
26. Alita Joyce. Formative vs. Summative Evaluations. July 28, 2019. https://www.nngroup.com/articles/formative-vs-summative-evaluations/
27. 18F Methods. Validate Test a design hypothesis. https://guides.18f.gov/methods/validate/usability-testing/#what
28. UserTesting. Moderated vs. unmoderated usability testing: the pros and cons. June 2, 2022. https://www.usertesting.com/resources/topics/moderated-vs-unmoderated-usability-testing
29. hotjar. The different types of usability testing methods for your projects. October 9, 2023. https://www.hotjar.com/usability-testing/methods/
30. Cole Stryker, Eda Kavlakoglu. What is artificial intelligence (AI)? August 16, 2024. https://www.ibm.com/topics/artificial-intelligence
31. IBM. What is a machine learning algorithm? https://www.ibm.com/topics/machine-learning-algorithms
32. Levy JJ, O'Malley AJ. Don't dismiss logistic regression: the case for sensible extraction of interactions in the era of machine learning. *BMC Medical Research Methodology*. 2020;20(1):171. doi:10.1186/s12874-020-01046-3
33. Castro HM, Ferreira JC. Linear and logistic regression models: when to use and how to interpret them? *J Bras Pneumol*. 48(6):e20220439. doi:10.36416/1806-3756/e20220439
34. Breiman L. Random Forests. *Machine Learning*. 2001;45(1):5-32. doi:10.1023/A:1010933404324
35. Taunk K, De S, Verma S, Swetapadma A. A Brief Review of Nearest Neighbor Algorithm for Learning and Classification. In: *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*. ; 2019:1255-1260.

doi:10.1109/ICCS45141.2019.9065747

36. Cheng-Jin Du, Da-Wen Sun. Support Vector Machine. In: *Computer Vision Technology for Food Quality Evaluation*. ; 2008.
37. Steven Walczak, Narciso Cerpa. Artificial Neural Networks. In: *Encyclopedia of Physical Science and Technology (Third Edition)*. ; 2003.
38. Khurana D, Koli A, Khatter K, Singh S. Natural language processing: state of the art, current trends and challenges. *Multimed Tools Appl*. 2023;82(3):3713-3744. doi:10.1007/s11042-022-13428-4
39. United Nations Principles for Responsible Investment. Diversity equity and inclusion. 2020. https://www.unpri.org/sustainability-issues/environmental-social-and-governance-issues/social-issues/diversity-equity-and-inclusion#:~:text=Equity%20means%20people%20have%20fair,resources%20and%20power%20to%20thrive
40. Milken Institute School of Public Health. Equity vs. Equality: What’s the Difference? November 5, 2020. https://onlinepublichealth.gwu.edu/resources/equity-vs-equality/
41. Melissa Dudek, Laura Eisenbeis, Naomi Alem, Jerry Ronaghan, Dutch MacDonald, Kedra Newsom Reeves. The Importance of Being Equitable in Product Design. February 9, 2022. https://www.bcg.com/publications/2022/the-importance-of-equitable-products
42. Government of Canada. Inclusion of marginalized people. June 5, 2017. https://www.international.gc.ca/world-monde/issues_development-enjeux_developpement/human_rights-droits_homme/inclusion.aspx?lang=eng
43. United Nations. Vulnerable Groups: who are they? https://www.un.org/en/fight-racism/vulnerable-groups

P1. Souza KES, Seruffo MCR, De Mello HD, Souza DDS, Vellasco MMBR. User Experience Evaluation Using Mouse Tracking and Artificial Intelligence. *IEEE Access*. 2019;7:96506-96515. doi:10.1109/ACCESS.2019.2927860

P2. Fernandes PM, Lopes M, Prada R. Agents for Automated User Experience Testing. In: *2021 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*. IEEE; 2021:247-253. doi:10.1109/ICSTW52544.2021.00049

P3. Dingli A, Cassar S. An Intelligent Framework for Website Usability. *Advances in Human-Computer Interaction*. 2014;2014:1-13. doi:10.1155/2014/479286

P4. Fan M, Li Y, Truong KN. Automatic Detection of Usability Problem Encounters in Think-aloud Sessions. *ACM Trans Interact Intell Syst*. 2020;10(2):1-24. doi:10.1145/3385732

P5. Liyanage NL, Vidanage K. Site-ability: A website usability measurement tool. In: *2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE; 2016:257-265. doi:10.1109/ICTER.2016.7829929

P6. Harms P. Automated Usability Evaluation of Virtual Reality Applications. *ACM Trans Comput-Hum Interact*. 2019;26(3):1-36. doi:10.1145/3301423

P7. Kuang E. Crafting Human-AI Collaborative Analysis for User Experience Evaluation. In: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM; 2023:1-6. doi:10.1145/3544549.3577042

P8. Soure EJ, Kuang E, Fan M, Zhao J. CoUX: Collaborative Visual Analysis of Think-Aloud Usability Test Videos for Digital Interfaces. *IEEE Trans Visual Comput Graphics*. 2022;28(1):643-653. doi:10.1109/TVCG.2021.3114822

P9. Ponce P, Balderas D, Peffer T, Molina A. Deep learning for automatic usability evaluations based on images: A case study of the usability heuristics of thermostats. *Energy and Buildings*. 2018;163:111-120. doi:10.1016/j.enbuild.2017.12.043

P10. Bakaev M, Gaedke M, Khvorostov V, Heil S. Extending Kansei Engineering for Requirements Consideration in Web Interaction Design. In: Bozzon A, Cudre-Maroux P, Pautasso C, eds. *Web Engineering*. Vol 9671. Lecture Notes in Computer Science. Springer International Publishing; 2016:513-518. doi:10.1007/978-3-319-38791-8_39

P11. Lin T, Xie T, Chen Y, Tang N. Automatic cognitive load evaluation using writing features: An exploratory study. *International Journal of Industrial Ergonomics*. 2013;43(3):210-217. doi:10.1016/j.ergon.2013.02.002

P12. Oztekin A. A decision support system for usability evaluation of web-based information systems. *Expert Systems with Applications*. 2011;38(3):2110-2118. doi:10.1016/j.eswa.2010.07.151

P13. Moutinho Da Ponte MJ, Morais Da Silveira A. A Methodology for Evaluation the Usability of Software for Industrial Automation Using Artificial Neural Networks: Case Study--Eletrobr&#x0E1;s. In: *2008 International Conference on Computational Intelligence for Modelling Control & Automation*. IEEE; 2008:430-435. doi:10.1109/CIMCA.2008.18

P14. Robal T, Marenkov J, Kalja A. Ontology Design for Automatic Evaluation of Web User Interface Usability. In: *2017 Portland International Conference on Management of Engineering and Technology (PICMET)*. IEEE; 2017:1-8. doi:10.23919/PICMET.2017.8125425

P15. Kuang E, Jahangirzadeh Soure E, Fan M, Zhao J, Shinohara K. Collaboration with Conversational AI Assistants for UX Evaluation: Questions and How to Ask them (Voice vs. Text). In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM; 2023:1-15. doi:10.1145/3544548.3581247

P16. Fan M, Wu K, Zhao J, Li Y, Wei W, Truong KN. VisTA: Integrating Machine Intelligence with Visualization to Support the Investigation of Think-Aloud Sessions. *IEEE Trans Visual Comput Graphics*. Published online 2019:1-1. doi:10.1109/TVCG.2019.2934797

P17. Fan M, Yang X, Yu T, Liao QV, Zhao J. Human-AI Collaboration for UX Evaluation: Effects of Explanation and Synchronization. *Proc ACM Hum-Comput Interact*. 2022;6(CSCW1):1-32. doi:10.1145/3512943

P18. Gupta S, Epiphaniou G, Maple C. AI-augmented usability evaluation framework for software requirements specification in cyber physical human systems. *Internet of Things*. 2023;23:100841. doi:10.1016/j.iot.2023.100841

P19. Fan M, Zhu L. Think-Aloud Verbalizations for Identifying User Experience Problems: Effects of Language Proficiency with Chinese Non-Native English Speakers. In: *The Ninth International Symposium of Chinese CHI*. ACM; 2021:22-32. doi:10.1145/3490355.3490358

P20. Fan M, Zhao Q, Tibdewal V. Older Adults' Think-Aloud Verbalizations and Speech Features for Identifying User Experience Problems. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM; 2021:1-13. doi:10.1145/3411764.3445680

P21. Yang B, Wei L, Pu Z. Measuring and Improving User Experience Through Artificial Intelligence-Aided Design. *Front Psychol*. 2020;11:595374. doi:10.3389/fpsyg.2020.595374

P22. Oztekin A, Delen D, Turkyilmaz A, Zaim S. A machine learning-based usability evaluation method for eLearning systems. *Decision Support Systems*. 2013;56:63-73. doi:10.1016/j.dss.2013.05.003

P23. Bakaev M, Mamysheva T, Gaedke M. Current trends in automating usability evaluation of websites: Can you manage what you can't measure? In: *2016 11th International Forum on Strategic Technology (IFOST)*. IEEE; 2016:510-514. doi:10.1109/IFOST.2016.7884307

P24. Stahlke S., Nova A, Mirza-Babaei P. Artificial Playfulness: A Tool for Automated Agent-Based Playtesting. In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM; 2019:1-6. doi:10.1145/3290607.3313039

P25. Taati B, Snoek J, Mihailidis A. Towards Aging-in-Place: Automatic Assessment of Product Usability for Older Adults with Dementia. In: *2011 IEEE First International Conference on Healthcare Informatics, Imaging and Systems Biology*. IEEE; 2011:205-212. doi:10.1109/HISB.2011.43

P26. De Souza Santos F, Vinícius Treviso M, Gama SP, De Mattos Fortes RP. A Framework to Semi-automated Usability Evaluations Processing Considering Users' Emotional Aspects. In: Kurosu M, ed. *Human-Computer Interaction. Theoretical Approaches and Design Methods*. Vol 13302. Lecture Notes in Computer Science. Springer International Publishing; 2022:419-438. doi:10.1007/978-3-031-05311-5_29

P27. Gannouni S, Belwafi K, Aledaily A, Aboalsamh H, Belghith A. Software Usability Testing Using EEG-Based Emotion Detection and Deep Learning. *Sensors*. 2023;23(11):5147. doi:10.3390/s23115147

P28. Yurkin VA, Saradgishvili SE, Voinov NV, Molodyakov SA. Applying Neural Network to Assess Application User Experience. In: *2023 IV International Conference on Neural Networks and Neurotechnologies (NeuroNT)*. IEEE; 2023:39-42. doi:10.1109/NeuroNT58640.2023.10175849

P29. Čejka M, Masner J, Jarolímek J, et al. UX and Machine Learning – Preprocessing of Audiovisual Data Using Computer Vision to Recognize UI Elements. *AOL*. 2023;15(3):35-44. doi:10.7160/aol.2023.150304

P30. Batliner M, Hess S, Ehrlich-Adám C, Lohmeyer Q, Meboldt M. Automated areas of interest analysis for usability studies of tangible screen-based user interfaces using mobile eye tracking. *AIEDAM*. 2020;34(4):505-514. doi:10.1017/S0890060420000372

P31. Duarte RP, Cunha CA, Cardoso JC. Automatic User Testing and Emotion Detection in Interactive Urban Devices. In: Gervasi O, Murgante B, Taniar D, et al., eds. *Computational Science and Its Applications – ICCSA 2023*. Vol 13957. Lecture Notes in Computer Science. Springer Nature Switzerland; 2023:3-18. doi:10.1007/978-3-031-36808-0_1

P32. Batch A, Ji Y, Fan M, Zhao J, Elmqvist N. uxSense: Supporting User Experience Analysis with Visualization and Computer Vision. *IEEE Trans Visual Comput Graphics*. 2024;30(7):3841-3856. doi:10.1109/TVCG.2023.3241581

P33. Kamińska D, Zwoliński G, Laska-Leśniewicz A. Usability Testing of Virtual Reality Applications—The Pilot Study. *Sensors*. 2022;22(4):1342. doi:10.3390/s22041342

P34. Asghar M, Bajwa IS, Ramzan S, Afreen H, Abdullah S. A Genetic Algorithm-Based Support Vector Machine Approach for Intelligent Usability Assessment of m-Learning Applications. Kumar M, ed. *Mobile Information Systems*. 2022;2022:1-20. doi:10.1155/2022/1609757

P35. Golondrino GEC, Sanabria LFM, Muñoz WYC. Proposal of an Automated Tool for Conducting Usability Inspections based on Nielsen Heuristics. 2021;14(9).

P36. Liu Z, Chen C, Wang J, Huang Y, Hu J, Wang Q. Owl eyes: spotting UI display issues via visual understanding. In: *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering*. ACM; 2020:398-409. doi:10.1145/3324884.3416547

P37. Novák JŠ, Masner J, Benda P, Šimek P, Merunka V. Eye Tracking, Usability, and User Experience: A Systematic Review. *International Journal of Human–Computer Interaction*. 2024;40(17):4484-4500. doi:10.1080/10447318.2023.2221600

P38. Speicher M, Both A, Gaedke M. WaPPU: Usability-Based A/B Testing. In: Casteleyn S, Rossi G, Winckler M, eds. *Web Engineering*. Vol 8541. Lecture Notes in Computer Science. Springer International Publishing; 2014:545-549. doi:10.1007/978-3-319-08245-5_47

P39. Hwang H, Lee Y. Usability Problem Identification Based on Explainable Neural Network in Asynchronous Testing Environment. *Interacting with Computers*. 2021;33(2):155-166. doi:10.1093/iwc/iwab018

P40. Yisi Liu, Olga Sourina, Hui Ping Liew, et al. Human Factors Evaluation in Maritime Virtual Simulators Using Mobile EEG-Based Neuroimaging. In: *Human Factors Evaluation in Maritime Virtual Simulators Using Mobile EEG-Based Neuroimaging*. Vol Volume 5: Transdisciplinary Engineering: A Paradigm Shift. Advances in Transdisciplinary Engineering. ; :261-268.

P41. Hedegaard S, Simonsen JG. Mining until it hurts: automatic extraction of usability issues from online reviews compared to traditional usability evaluation. In: *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*. ACM; 2014:157-166. doi:10.1145/2639189.2639211

P42. El-Halees AM. Software Usability Evaluation Using Opinion Mining. *JSW*. 2014;9(2):343-349. doi:10.4304/jsw.9.2.343-349

P43. Dolunay B, Akgunduz A. Automated end-user behaviour assessment tool for remote product and system testing. *Expert Systems with Applications*. 2008;34(4):2511-2523. doi:10.1016/j.eswa.2007.04.011

P44. Norman KL, Panizzi E. Levels of automation and user participation in usability testing. *Interacting with Computers*. 2006;18(2):246-264. doi:10.1016/j.intcom.2005.06.002

P45. Jorritsma W, Cnossen F, Dierckx RA, Oudkerk M, Van Ooijen PMA. Pattern mining of user interaction logs for a post-deployment usability evaluation of a radiology PACS client. *International Journal of Medical Informatics*. 2016;85(1):36-42. doi:10.1016/j.ijmedinf.2015.10.007

P46. Christoffer Korvald, Eunjin Kim, H. Reza. Evaluation and implementation of machine learning techniques in usability testing for web sites. Published online 2014. https://api.semanticscholar.org/CorpusID:59066735