The Journal of Systems and Software 208 (2024) 111881

Contents lists available at ScienceDirect

The Journal of Systems & Software

journal homepage: www.elsevier.com/locate/jss

Potential effectiveness and efficiency issues in usability evaluation within digital health: A systematic literature review☆

Bilal Maqbool *, Sebastian Herold

Department of Mathematics and Computer Science, Faculty of Health, Science and Technology, Karlstad University, Universitetsgatan 2, Karlstad, 651 88, Värmland, Sweden

ARTICLE INFO

Dataset link: https://doi.org/10.5281/zenodo.8218076

Keywords:
Software quality
Usability evaluation method
Digital healthcare (DH)
eHealth
Systematic literature review (SLR)

ABSTRACT

Context: Digital Health (DH) is widely considered essential for sustainable future healthcare systems. Software quality, particularly usability, is crucial for the success and adoption of most DH products. However, concerns about the effectiveness and efficiency of usability evaluation of DH products have been raised.

Objective: This article aims to analyse the prevalence and application contexts of usability evaluation methods in DH and to highlight potential issues related to their effectiveness and efficiency.

Method: A systematic literature review of usability evaluation studies, published by (academic) practitioners between 2016 and April 2023, was conducted. 610 primary articles were identified and analysed, utilising five major scientific databases.

Results: Our findings show a preference for inquiry (85%) and testing (63%) methods, with inspection used less frequently (17%). The published studies employed methods like questionnaires (75%); notably the SUS (49%), semi-structured interviews (25%), and heuristic evaluations (73%), with percentages based on their group. Data collection mainly involved the use of participant feedback (45%), audio/video recordings (44%), and system logs (20%), with both qualitative and quantitative data analyses prevalent in studies. However, several usability characteristics such as accessibility, memorability, and operability were found to be largely overlooked, and automation tools or platforms were not widely used. Among the systems evaluated were mHealth applications (70%), telehealth platforms (36%), health information technology (HIT) solutions (29%), personalized medicine (Per. Med.) (17%), wearable devices (12%), and digital therapeutics (DTx) interventions (6%), with the participation of general users, patients, healthcare providers, and informal caregivers varying based on the health condition studied. Furthermore, insights and experiences gathered from 24 articles underscored the importance of a mixed-method approach in usability evaluations, the limitations of traditional methods, the necessity for sector-specific customisation, and the potential benefits of remote usability studies. Moreover, while eye-tracking emerged as a promising evaluation technique, careful execution and interpretation are crucial to avoid data misinterpretation.

Conclusion: The study's findings showed that employing a combination of inquiry and testing-based methods is prevalent for evaluating DH platforms. Despite an array of DH systems, method distribution remained consistent across platforms and targeted user groups. The study also underlines the importance of involving target user groups in the process. Potentially affected cognitive abilities of participants and potential user groups of interest have to be taken into account when choosing evaluation methods, and methods might therefore need to be tailored. Complementary inspection methods might be particularly useful when recruiting representative participants is difficult. Several potential paths for future research are outlined, such as exploring novel technologies like artificial intelligence, for improved automation tool support in the usability evaluation process.

☆ Editor: Antonia Bertolino.

* Corresponding author.
E-mail addresses: bilal.maqbool@kau.se (B. Maqbool), sebastian.herold@kau.se (S. Herold).

https://doi.org/10.1016/j.jss.2023.111881
Received 20 September 2022; Received in revised form 16 October 2023; Accepted 19 October 2023
Available online 3 November 2023

1. Introduction

1.1. Motivation

As the World Health Organisation (WHO) outlines, healthcare systems worldwide experience many challenges (World Health Organization et al., 2018). A growing and ageing population, with an estimated 16% of the global population being 65 years or older by 2050, will require these systems to provide high quality services in a cost-efficient way (Thomson et al., 2009b,a). Easily accessible healthcare services and careful analysis of the needs of people particularly dependent on these services, such as the elderly or people living with disabilities, are crucial for effective and efficient healthcare (Mulumba et al., 2014; Salvador-Carulla and Saxena, 2009).

The rise of digital technologies has enabled healthcare providers to improve their services and systems (Della Mea, 2001) and to manage health and well-being more effectively (Maramba et al., 2019). The digital healthcare (DH) market has grown exponentially, with software being a crucial part of almost all relevant systems (Maramba et al., 2019; PwC, 2016). Consequently, software quality plays a key role in the success or failure of these systems, just as in any software-based product (Mack and Sharples, 2009). The European Commission and WHO have identified high levels of software quality as critical for DH solutions (Commission, 2002, 2006; Haas and Sembritzki, 2006). A lack of software quality in DH has been reported to be the cause for, e.g., drug reporting failures, software errors in critical care, failure in data quality, loss of data integrity, and improper system use (Bowman, 2013; Hoffman, 2014; Sittig et al., 2018; Ebad, 2020). Delays in diagnosis, longer hospital stays, increased treatment costs, harm to patient's health and life, and other consequences can result from such failures.

Usability is considered one important aspect of software quality (Househ et al., 2015; Paz and Pow-Sang, 2016; ISO, 2011). Better usability can improve well-being and productivity, while also reducing stress and the risk of harm to people using a system or being affected by its actions (DIS, 2009). It furthermore can lead to greater user acceptance (Jaspers et al., 2008; Huryk, 2010). Hence, the attainment of satisfactory levels of usability is regarded as crucial for the organisational success of companies producing software (Suvanmanee et al., 2020). Consequently, usability has been recognised as a critical success factor for DH technologies (Baharuddin et al., 2013).

Achieving high levels of usability is a challenging task (Zahra et al., 2018; Paz and Pow-Sang, 2016). Usability evaluation can help reach the degree of usability desired in software products (Zapata et al., 2015; Maramba et al., 2019). An earlier exploratory study, however, discovered that software engineers working in the field of DH struggle to spend sufficient time on adequate software testing in general, and usability evaluation in particular (Maqbool and Herold, 2021). An observations concluded that this is not because adequate techniques are missing (Delikostidis, 2007), but more likely by the difficulty of applying a technique fitting the specific context and purpose (Zahra et al., 2018; Paz and Pow-Sang, 2016).

1.2. Goals and contributions of this study

This study is motivated by the obvious gap described above. On the one hand, there is a rich corpus of tried-and-tested usability evaluation methods; on the other hand, evidence suggests that applying those methods to DH systems is not effective or efficient enough.

The goal of this study is to investigate this gap, to better understand how usability evaluation methods are applied in the field of DH applications, and to identify potential points for improving the effectiveness and/or efficiency of usability evaluation in this domain.

To this purpose, we conducted a systematic literature review (SLR) of 610 primary studies published between 2016 and April 2023, describing usability evaluation of DH applications [S1-S610][1] The studies were analysed and categorised along several dimensions related to the methods used and relevant contextual aspects, such as the DH technologies (system types) investigated, or targeted user groups. Based on the resulting picture, we synthesised current problems and potential areas of usability evaluation in which effectiveness and/or efficiency could be improved.

The main contributions of the paper are summarised here:

- An overview of the methods and practices used in DH usability evaluation across diverse contexts.
- An analysis of potential effectiveness and/or efficiency issues based on that overview.
- Recommendations for software practitioners to evaluate the usability of DH applications and direction for researchers to investigate potential research gaps.

1.3. Outline

The remaining article is structured as follows. The following section introduces concepts and terminology related to DH, software quality, and usability. Section 3 provides an overview of previous research. In Section 4, the research methodology and implemented protocol are explained. The results are presented and discussed in Sections 5 and 6, respectively. Section 7 presents potential threats to the validity and reliability of the study, and Section 8 concludes the article.

2. Background and terminology

2.1. Digital health (DH)

Several terms have been introduced and applied over the years to describe the progressing digitalisation of the healthcare sector. During the late 1990's and early 2000's, *eHealth* or *e-Health*, emerged as a term mainly used by leading IT businesses in their marketing efforts to describe the opportunities that the Internet offered to improve health care (Eysenbach, 2001). As common for terms invented that way, eHealth lacked a clear definition at that time. Nowadays, the WHO defines eHealth as "the cost-effective and secure use of information and communications technologies in support of health and health-related fields, including health care services, health surveillance, health literature, and health education, knowledge and research" (WHO, 2021).

The term *digital health (DH)* started to appear more often around 2010 and is often used and defined synonymously with eHealth (see for example Koskova (Kostkova, 2015) and Hsu (Hsu, 2015)). This understanding is contrasted with the definitions of the WHO that characterise DH as the "field of knowledge and practice associated with the development and use of digital technologies to improve health" that expands eHealth "to include digital consumers with a wider range of smart devices and connected equipment" (WHO, 2021). While eHealth according to the WHO definition sees Information and Communication Technologies (ICT) more as a mean to the end of providing "classical" healthcare, DH appears as a distinct field, considering advanced technologies like Internet-of-Things (IoT), artificial intelligence (AI), etc. as ways to deliver novel types of healthcare.

The US Food and Drug Administration (FDA) distinguishes five different subfields or technologies of DH, which are health information technology (HIT), mobile health (mHealth), telehealth (or telemedicine), wearable devices, and personalised medicine (Per. Med.) (Petersen, 2018).

[1] The repository has a file containing titles of the primary selected studies: https://doi.org/10.5281/zenodo.8218076.

1. *Health Information Technology* (HIT) is defined as ICT-based technology for monitoring, managing and communicating health information between DH systems, consumers, and health professionals (Fadahunsi et al., 2019). Examples of HIT are Electronic Health Record (EHR) or electronic prescription services;
2. *Mobile health (mHealth)* is based on mobile technology to deliver health services (mobile self-care) (Adibi, 2015). Commonly used mobile communication technologies are smartphones, tablets, and personal digital assistants (PDAs) (Cipresso et al., 2012);
3. In *telehealth*, or *telemedicine*, ICT is used to deliver healthcare services and connect patients and healthcare professionals over long distances (Shaw, 2009). Examples of applications are telerehabilitation, remote patient monitoring, and virtual doctor appointments;
4. *Wearable devices* in the context of DH are smart electronic devices that can be worn by people to monitor, analyse, and transfer health information (Düking et al., 2018). Fitness trackers and several types of body function sensors, like heartbeat or glucose sensors, are some examples of this category;
5. *Personalised medicine (Per. Med.)* deals with the use of digital technologies to deliver personalised healthcare, which, in contrast to dominant disease-focused approaches for providing healthcare, is largely based on knowledge of an individual's unique genetically inherited biological traits and exposure to environmental factors (Snyderman, 2012). Examples include genetic health risk assessment, but also individualised disease treatment and health planning;
6. In addition, the Digital Therapeutics Alliance (DTA) has introduced a sixth sub-field, *Digital Therapeutics (DTx)* (Hong et al., 2021). Digital therapeutics are software solutions that intervene in evidence-based therapies to monitor, manage, and prevent medical conditions (Patel and Butte, 2020). Applications include, for example, AI-supported diabetes monitoring or cognitive behavioural therapy for treating depression.

2.2. Software quality

The International Standardisation Organisation (ISO) defines software quality as the "degree to which a software system satisfies the stated and implied needs of its various stakeholders, and therefore provides value" (ISO, 2011). Maintaining the quality of a software system has been associated with the software industrial growth (Thota et al., 2020), customer satisfaction (Hossain et al., 2013) and a software companies competitiveness and profitability (Thota et al., 2020; Angelova and Zekiri, 2011).

Quality models attempt to capture the different quality characteristics of software in a hierarchical scheme. The ISO-9126 quality model was established in 2001 and received input from prior quality models to define the requirements for assessing software quality (Miguel et al., 2014). The ISO/IEC 25010 model, an improved version of the ISO-9126 model, was established in 2007 (ISO, 2011). ISO/IEC 25010 is categorised into eight top-level characteristics: functional suitability (functional completeness, correctness, appropriateness), performance efficiency, compatibility, *usability*, reliability, security, maintainability, and portability (ISO, 2011).

Usability has been frequently named as an essential quality factor for DH applications, such as appropriate user interfaces (UI) or user experience (UX), system integration, medical device compatibility, and performance testing (World Health Organization et al., 2018; Commission, 2002, 2006; Haas and Sembritzki, 2006). Studies have shown that software usability, reliability, and performance are the most influential factors on user satisfaction in DH (Aghazadeh et al., 2014, 2015). Among the reasons for the repeatedly emphasised importance of usability is its influence on other critical quality aspects, as the American Medical Informatics Association (AMIA) concluded: appropriate user interface design (usability) can, for example, improve software efficiency and patient safety (Middleton et al., 2013; Sagar and Saha, 2017).

2.3. Usability and usability evaluation

The term usability has been defined in various ways by different standards and researchers. Nielsen, a pioneer in the field of software usability, defines the five characteristics of good software usability as efficiency, errors-free, learnability, satisfaction, and memorability (Nielsen, 1994). The concept of usability emerged from the term user-friendliness and is defined by ISO 9241–11 as, "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" (ISO, 2018; Dix et al., 2003). The usability according to Nielsen, ISO 9241–11, and ISO/IEC 25010 usability characteristics are unified and described in the following list:

1. *Accessibility:* "The extent to which a system or product can be used by people with a diverse set of competencies and characteristics to achieve a specific objective in a particular context".
2. *Aesthetics:* "The degree to which a user interface allows for pleasant and satisfactory interaction".
3. *Appropriateness/Usefulness:* "The ability of users to determine if a product or system is suitable for their needs".
4. *Effectiveness:* "The accuracy and completeness with which users achieve specified goals".
5. *Efficiency:* "Systems should be efficient to use. When a user has fully learned the system, productivity will be possible on a high level".
6. *User error protection:* "The degree to which a system protects users against making errors".
7. *Learnability:* "Systems should be easy to learn. Users can rapidly start getting some work done with the system".
8. *Memorability:* "Systems should be easy to remember, making it possible for casual users to return to the system after some period of not using the system, without having to learn everything all over again".
9. *Satisfaction:* "The system should be pleasant to use; which makes users subjectively satisfied (comfort and acceptability of use) when using it".
10. *Operability:* "The capability of the software product to enable the user to operate and control it".

Usability evaluation comprises methods and techniques for evaluating a software product regarding its usability properties through systematic testing by engaging representative users, usability experts, or employing alternative means (automated review analysis, etc.). Evaluating software usability should be performed repeatedly throughout its development cycle (Genise, 2002). Delikostidis summarises that usability evaluation methods are generally divided into three categories: inquiry, inspection, and testing (Delikostidis, 2007).

In **inquiry methods**, usability analysts interact with users, observe them using the system, or allow them to respond verbally or in writing to questions to learn more about their likes, dislikes, requirements, and understanding of the system (Ivory, 2013). Some examples of inquiry techniques are:

- *Questionnaires or Surveys*: A questionnaire is a type of text-based survey in which participants are asked questions about the system of interest.
- *Observations*: This procedure aims to observe and study the participant's interaction with the system.
- *Interviews*: One-on-one (un/semi)-structured conversation, usually recorded and transcribed for later reference.
- *Focus group discussions (FGD)*: These are verbal interactions about the use of the system with a group of participants with similar backgrounds or experiences.

In the **inspection method**, experts and professionals review systems based on experience, knowledge, and standards (Ivory, 2013). Some examples of inspection techniques are:

- *Heuristic evaluation*: Experts assess the application usability based on rules of thumb and judge its conformity with recognised usability standards (the "heuristics") (Nielsen and Molich, 1990; Nielsen and Mack, 1994).
- *Cognitive walkthroughs*: Experts evaluate the usability of an application based on user goals and usage scenarios. It is a way of reviewing the system that focuses on how it effectively enables "exploratory learning" (Rieman et al., 1995).
- *Guideline checklists*: A list of rules and policies to check if they are followed in the review system (Dumas and Salzman, 2006).

Methods of the **testing** category have in common that the testing is performed through observing real or representative users executing a set of well-defined tasks with the system of interest (Delikostidis, 2007; Ivory, 2013; Paz and Pow-Sang, 2016). In task- and scenario-based testing as a subcategory, the focus is on obtaining quantifiable information about the performance of the participants who are instructed with the tasks, such as task completion rates, execution times, etc. In think-aloud protocols, as the second subcategory, participants are asked to verbalise their thoughts, feelings, and opinions while performing a task (Güss, 2018). This aids in learning how users engage with the system and the decisions they make. Thinking aloud can therefore provide insights into the cognitive processes of users.

2.4. A glimpse of usability evaluation in practice

In a hospital system facing high levels of nurse burnout, researchers conducted an investigation into the relationship between EHR usability and professional dissatisfaction (Melnick et al., 2021). Utilising the SUS, tailored to the EHR system, a strong association was found between the system's low usability score and increased nurse burnout. This finding prompted a need for user-centred design interventions and a re-evaluation of the interface to enhance user experience. Thus revealing the essential connection between technology usability and professional satisfaction in healthcare settings.

Another study used customised questionnaires to assess the usability of a decade-old Computerized Physician Order Entry (CPOE) system. The evaluation revealed multiple usability problems that could lead to user errors and inefficiencies (Khajouei et al., 2011).

Usability evaluation methods like the traditional heuristic evaluation method, often used for general software, can be modified or adapted based on system or goals, such as evaluating patient safety in medical devices (Zhang et al., 2003). This modification was proven successful in identifying usability problems. It can also quickly identify major usability issues linked to human errors in medical device use, mainly due to interface design problems.

A study by Schnall et al. investigated the usability of a mhealth app for HIV-infected adults using a mixed-method (focus groups, think-aloud protocols, and a heuristic evaluation checklist) and iterative approach (Schnall et al., 2016). The study provided insight into areas for improvement, particularly related to the need for clear instructions and an easy-to-use navigation system.

Furthermore, usability evaluation has evolved to include a combination of traditional and more objective methods (Koester et al., 2017). For instance, eye tracking and remote monitoring have been recommended for use alongside conventional approaches (Maramba et al., 2019), allowing researchers to pinpoint the origins of a problem and investigate the stages at which an issue arises, such as in perception or comprehension (Schiessl et al., 2003). However, without proper planning, resources, and understanding of eye-tracking data, this can lead to poor design decisions and negative effects on users and businesses (Pernice, 2017; Koester et al., 2017).

These are a few examples, demonstrating the diverse applications and adaptability of various usability evaluation methods within the field of DH. They also briefly highlight how these methods can lead to improvements within DH.

3. Related work

Usability evaluation in DH has been investigated in a couple of secondary studies like systematic literature reviews (SLR) in recent years.

Zapata et al. reviewed empirical studies on the usability of mHealth applications (Zapata et al., 2015). They analysed 22 articles published between 2000 and April 2014 and found that surveys/questionnaires, interviews and 'think-aloud' were the most common data collection methods. In 73% of the studies, the evaluation of applications was conducted only through the use of questionnaires and interviews. Operability was the most examined usability characteristic, since studies focused mainly on the user interface's 'ease of use'. The researchers recommend exploring further automation of various aspects of usability evaluation.

A systematic mapping study published in 2016 discovered that questionnaires, testing, heuristic evaluation, and interviews were among the most commonly used methods to assess usability (Paz and Pow-Sang, 2016). Among other findings, the study noted that reported automation/tool support was mainly related to user activity tracking and metric-based measurements. The researchers recommended exploring variants of usability tools and techniques that can be utilised for distinct or common scenarios/contexts.

A systematic review published in 2017 and covering nine papers focused on serious games in healthcare contexts. It was discovered that Likert-scale questionnaire (survey), interview, think-aloud, recording (logs or performance metrics), and observation were the most commonly used usability evaluation methods (Schmidt and De Marchi, 2017). The author stated that the major factors influencing the acceptance and usability of serious games appear to be unclear and, hence, evaluating their usability proves to be difficult. Another, wider literature review on serious games came to similar conclusions and showed that questionnaires, interviews, observation, heuristic evaluation, FGD, and logs or performance metrics were the most frequently used usability evaluation methods (Yanez-Gomez et al., 2017).

Ye et al. conducted an SLR related to the usability evaluation of diabetes mobile apps (Ye et al., 2017). Surveys, interviews, laboratory tests, user tests, questionnaires, and expert and heuristic evaluations were used in the investigated studies. The authors state that although evidence-based guidelines have proven to be effective in the development of applications, their utilisation in the context of mobile diabetic apps remains limited. They also advocated for a systematic and consistent approach to report guidelines, which would assist in educating the research community about the adherence to evidence-based guidelines for diabetes mobile app development.

Zahra et al. conducted an SLR to explore usability dimensions and suggest a commonly used usability model for chronic disease mobile applications (Zahra et al., 2018). Among the key findings was the observation that most of the applications mentioned in the studied articles were not evaluated for usability. Moreover, the authors judged existing usability evaluation models as unable to capture the complexities of this type of application.

Maramba and co authors reviewed research articles on usability evaluation in eHealth published between 2014 and 2017 and found that "questionnaire, task completion, 'think-aloud', interviews, heuristic evaluation, and focus groups" were the most frequently used usability evaluation methods (Maramba et al., 2019). Furthermore, the 10 most frequently reported health conditions or diseases for which usability of eHealth apps was evaluated were: "mental health, cancer, nutrition, child health, diabetes, telemedicine, cardiovascular disease, HIV, health information systems, and smoking". The authors also note that automated methods of usability evaluation common in other domains were not widely used in DH.

Ansaar et al. conducted an SLR on usability evaluation processes for mHealth applications (Ansaar et al., 2020). Questionnaires and interviews were found to be the most frequently used usability evaluation

Table 1
Comparative overview of key criteria between related work and the proposed study in the context of DH.

Review type	Timeline	Articles	Databases	Evaluated usability methods?	Evaluated usability characteristics?	Evaluated health conditions?	Evaluated user groups?	DH systems	Evaluated automated methods?	Evaluated researchers' experience?	Pub. year
SLR	2000–2014	22	1, 2, 3, 4, 5	✓	✓	✓	✓	mHealth	×	×	2015 (Zapata et al., 2015)
SMS	2012–2015	215	2, 6, 7	✓	×	×	×	General Domains	×	×	2016 (Paz and Pow-Sang, 2016)
SLR	2015–2016	9	1, 2, 3, 8	✓	×	×	Partial	Mobile Serious Games	×	×	2017 (Schmidt and De Marchi, 2017)
SLR	2003–2015	187	2, 3, 7	✓	✓	×	✓	Serious Games	×	×	2017 (Yanez-Gomez et al., 2017)
SLR	2006–2016	20	4, 6, 15	✓	×	×	×	Diabetes Mobile Apps	×	×	2017 (Ye et al., 2017)
SLR	2010–2017	60	1, 2, 8, 10, 11, 12	✓	✓	×	Partial	Chronic Disease Mobile Apps	×	×	2018 (Zahra et al., 2018)
Scoping review	2014–2017	133	2, 3, 4, 9	✓	×	✓	✓	Various DH	Partial	×	2019 (Maramba et al., 2019)
SLR	2010–2017	19	1, 2, 3, 4, 5	✓	✓	×	×	mHealth	×	×	2020 (Ansaar et al., 2020)
SLR	1986–2019	51	9, 13, 14, 15, 16	✓	✓	×	×	EHR	×	×	2021 (Wronikowska et al., 2021)
SLR	2017–2021	55	1, 2, 3	✓	✓	×	×	Healthcare Websites	×	×	2022 (Saad et al., 2022)
Our SLR	2016–April 2023	610	2, 3, 6, 8, 17	✓	✓	✓	✓	All DH	✓	✓	2023

1 = ScienceDirect, 2 = ACM, 3 = IEEE, 4 = PubMed, 5 = Wiley, 6 = Scopus, 7 = ISI (WoS & WoK), 8 = Springer
9 = CINAHL, 10 = Taylor and Francis, 11 = JMIR, 12 = UPA, 13 = Medline, 14 = EMBASE, 15 = Cochrane, 16 = Open Grey, 17 = WoS.

methods. The majority of usability evaluation studies have relied on descriptive statistics analysis, such as for the SUS rating, task competition time, task performance, etc. Attractiveness of applications as a usability characteristic was neglected in primary studies, whereas operability was most frequently assessed. Usability for older and novice users was, in general, not well assessed. The authors argue based on their observations for mixed-method approaches for more reliable usability evaluation results.

A study, published in 2021, presented a review of 51 studies assessing the usability of EHR systems for hospital staff (Wronikowska et al., 2021). The study identified 11 different usability evaluation methods, with testing with users being the most prevalent, followed by questionnaires, heuristic evaluation and interviews. Seven different usability metrics, with ISO9241-11 and Nielsen's components being the most common. The study underscores a notable lack of consistency in the evaluation methods and metrics used across the field, potentially impacting healthcare work efficiencies. The authors advocate for the use of the ISO 9241–1 usability metrics (efficiency, effectiveness, satisfaction) along with a "usefulness" metric as part of a more standardised evaluation process for EHR systems, with the aim to enhance their development and implementation.

The usability of healthcare websites, often used by doctors and patients, has emerged as a key concern in recent healthcare technology advancements (Saad et al., 2022). An SLR explores this issue using studies published between 2017 and 2021, which identified and analysed 55 primary papers, aimed to aid designers in developing easy-to-use healthcare websites. Several evaluation methods, such as questionnaires and task-based evaluations, have been employed to examine usability features, including feedback and accuracy. At the same time, it brought attention to usability concerns within these systems, specifically related to challenges in navigation. The findings support the importance of iterative design testing, prioritising user-centred design, and conducting research on tool-based evaluation in order to enhance the accessibility of healthcare websites.

Current studies on usability evaluation in DH have gaps, as evidenced in our literature review (also see Table 1). Most do not cover the wide range of DH systems or evaluate the breadth of usability evaluation practices, with respect to their effectiveness and/or efficiency, applied in the DH field. Additionally, there is a lack of investigation into health conditions, user groups, and the use of automated tools or platforms. Also, researchers' practical experience with DH usability evaluation has not been explored. Therefore, we believe that conducting a comprehensive literature review on current usability evaluation practices in Digital Health is essential to understand potential issues that may affect their effectiveness or efficiency. This review will provide a broad, updated view of current usability practices, helping shape future design and evaluation strategies.

4. Research methodology: Systematic literature review

The study protocol applied for this SLR follows the guidelines developed by Kitchenham and Charters (Kitchenham and Charters, 2007). In the following subsections, we present the motivating research goals and questions for this study, the applied article search and selection strategies as well as the used data extraction and analysis strategies.

4.1. Research questions

As described in Section 1, this study is motivated by the previously acknowledged challenge of efficient usability evaluation in the DH domain. We, therefore, defined our research goals as listed below:

- **RG 1**: Characterise the use of usability evaluation methods in DH.
- **RG 2**: Identify areas of improvement in the use of these methods in terms of effectiveness and efficiency.

Research goals are used to characterise the use of usability evaluation techniques in DH in detail (RG 1) and to explore potential areas for effectiveness and efficiency improvement (RG 2). Based on the research goals, the following research questions were defined:

- **RQ 1**: How are various usability evaluation methods characterised and applied in the field of DH?
- **RQ 2**: What potential challenges and limitations affect the effectiveness and efficiency of usability evaluation in DH, and what could be the potential areas to explore further to address such issues?

RG 1 is refined into one major research question (RQ 1). RQ 1 broadly aimed to investigate the prevalence of the use of existing usability evaluation methods as well as circumstantial information about its use, such as the type of systems that evaluation methods are applied to. We further shift focus on, to which degree different characteristics of usability are covered in the reported usability evaluation studies (see Section 2.3). We also investigate to which degree automation is already employed in the domain of interest to reduce efforts.

RG 2 is refined by the research questions RQ 2. RQ 2 focuses on the perceptions as expressed in the selected primary articles. It aims to understand the challenges and limitations associated with various usability evaluation methods, as perceived by researchers and professionals engaged in usability evaluations. Further, RQ 2 broadly aims to examine, discuss and draw conclusions from the identified results and literature on the efficiency and effectivity of the identified usability methods.

Table 2
Inclusion and exclusion criteria.

ID	Criteria
Inclusion	
IC1	The article is related to usability evaluation in digital health.
IC2	The article is a primary study.
IC3	The article was published between 2016–April 2023.
IC4	The article is written in English.
IC5	The article was published as a peer-reviewed conference or journal article.
IC6	The article contains a significant amount of original work (no duplication of previous works).
IC7	The full text of the article is accessible via the available subscriptions or provided by the article authors.
Exclusion	
EC1	The article focuses on evaluating the usability of hardware in a digital health context.
EC2	The article only discusses usability evaluation in theory without practically evaluating a DH system/technology or usability evaluation methods in a DH context.
EC3	The article is a duplicate of an article retrieved from another database.

4.2. Search process

Databases: We selected five major electronic databases of scientific literature for this SLR: Scopus, ACM, IEEE Xplore, Springer, and Web of Science (WoS). The first four of these represent four of the most relevant scientific publishers in the field of computer science and software engineering (Scopus, ACM, IEEE, Springer). Web of Science is owned by Clarivate which is an analytics business in the area of bibliometrics and scientometrics. Together, these databases cover all relevant conferences and journal publications in the field of software engineering and computer science (Dyba et al., 2007).

Inclusion and exclusion criteria: We defined seven inclusion and three exclusion criteria in order to characterise articles that were potentially relevant to the topic of interest (see Table 2). All inclusion criteria were mandatory for articles to proceed to the data extraction phase of the protocol, while articles meeting any exclusion criteria were removed.

IC1 includes articles that either are explicitly about usability evaluation in a DH context, or discuss a system, service, or product including the process and results of evaluating its usability. IC2 limits the scope to primary studies and therefore excludes other literature surveys; these are instead discussed as related works in Section 3. We limited the publication period for relevant articles to 2016–April 2023 (see IC3) as most previous studies covered the body of work up to then (Schmidt and De Marchi, 2017; Ye et al., 2017; Maramba et al., 2019; Ansaar et al., 2020). The remaining criteria restrict the set of articles to those accessible to us as authors in terms of language and availability (IC4 and IC7) and to articles meeting minimal academic quality criteria such as having been peer-reviewed and being sufficiently novel (IC5 and IC6).

EC1 excludes articles mainly concerned with scenarios in which hardware components are the main subject of usability evaluation. As our research questions refer to usability evaluation in practice, we also explicitly exclude purely philosophical papers or articles outlining solutions without empirical evaluation (EC2). With EC3, we make sure to eliminate duplicate articles identified in more than one database.

Search terms and string: In order to construct suitable search strings to retrieve articles from the electronic literature databases, we first identified the relevant concepts for the search. Three concepts, or terms, were selected as bases for the search string construction: "Usability", "Test", and "eHealth". After that, we listed synonyms for these terms or alternative terms for the same or related concepts, e.g., "evaluation" as closely related and often synonymously used term for "testing".

With "Usability", we also incorporated "Ease of Use" as a synonym in our search string. While usability is a broader concept, we found that "Ease of Use" is often employed synonymously in various studies. Including it as a search term ensured that we did not exclude any relevant literature.

The concept of "eHealth" was also broadened to include its variations as well as related terms such as "welfare", and "telehealth". We incorporated the term "welfare" to expand the search beyond traditional healthcare and into the broader context of overall well-being, which is often the ultimate goal of many eHealth applications.

We selected "Telehealth", as it represents a wider perspective than "telemedicine", capturing a wide range of remote healthcare services, not limited to remote clinical services.

Common techniques for search string construction are then used to formulate the overall query (Rumsey, 2008; Spezi, 2016). First, concepts and their synonyms form disjunctive clauses. The resulting clauses are then connected via the NEAR proximity operator (to express required textual proximity between usability and testing concepts) and the AND operator. Finally, words in the search string were truncated and wildcards were applied to reflect stemming/lemmatisation to be able to find usages of inflected forms of the search terms. These steps resulted in the following search string: (Usability OR "Ease of use") NEAR/10 (Testing OR Evaluati* OR Analysis OR Examin* OR Assess*) AND (eHealth* OR e-health* OR Health* OR Welfare OR Telehealth OR mHealth)

For *Scopus*, the same search string was used, but *NEAR* proximity operator syntax was replaced by the equivalent *W* proximity operator. After finalising the search concepts and the string, the search query was executed on the selected databases. Available filters reflecting our inclusion criteria, e.g., regarding the period of interest or language of articles, were set accordingly. Search results were limited thematically to the subject of computer science where applicable. For *Web of Science*, *ACM* and *Scopus*, article titles, abstracts, and keywords were queried. While the default settings were used for *IEEE* (all meta information) and *Springer* (default searched content not documented).

4.3. Paper selection

Study selection is a multistage process, and this section explains the process of selecting research papers for the study. The PRISMA flow diagram (Moher et al., 2009) (see Fig. 1) shows the detailed process for selecting the final set of review articles.

In the identification stage, the search query was last executed on the selected databases on May 4, 2023, as described in the previous section. Duplicate articles were identified using the JabRef tool (Basak, 2014) and a Microsoft Excel script that we designed. After deduplication, 1021 articles were removed, leaving 4724 articles for a comprehensive title and abstract screening to determine if they met the remaining inclusion or exclusion criteria, which primarily included IC1, IC2, IC6, IC7, EC1, and EC2. Screening rules were established, and the authors assessed each article's relevance on a scale of 0 for irrelevant, 1 for unclear/ambiguous, and 2 for meeting the selection criteria. After independent reviews, both authors discussed and reviewed their individual scores frequently. In cases of diverging scores, we either reached an agreement about the final score during the discussion, postponed a decision and both reviewers read the paper in more detail, or, as the ultimate conclusion, included the paper to err on the safe side. In the end, we finalised 610 research articles for data extraction, analysis, and synthesis.[2] We eliminated 24 articles that could not be accessed according to IC7 and eight articles with non-English main texts despite having an English abstract.

[2] See the list of 763 fully reviewed articles in SLR repository: https://doi.org/10.5281/zenodo.8218076.

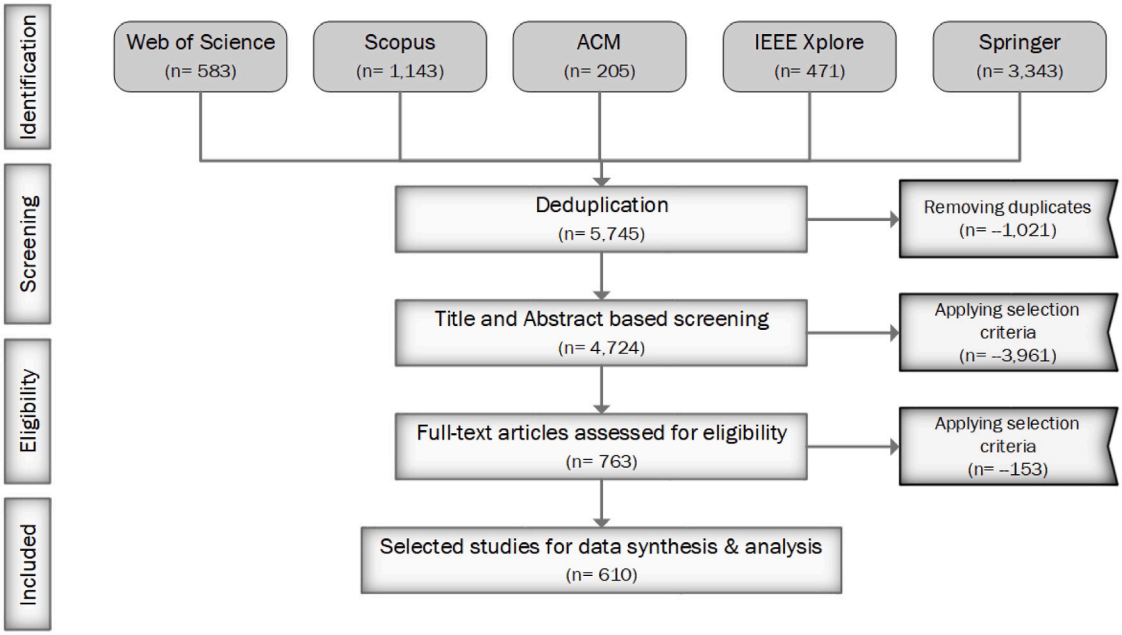

Fig. 1. PRISMA flow diagram.

Table 3
Data extraction form.

ID	Concern mapping to research question	Data item	Explanation
I1	General	Bibliographic information	This item included title, publication year, authors, etc.
I2		Health condition or disease	A condition merely denotes a person's state of health; a condition that causes illness is classified as a disease or disorder.
I3	Context	System targeted users	The key user group to which the digital health system is aimed at in order to solve their issues.
I4		System type	Domains of digital health technologies, e.g., mHealth.
I5		System maturity	Stage of the digital health system life cycle (mainly, prototype, incremental or fully developed system).
I6	RQ 1	Data collection methods	Primary methods used to collect usability evaluation data.
I7	RQ 1, RQ 2	Data recording methods/type	Methods/type used to record events, actions, and user interactions with the system.
I8	RQ 1, RQ 2	Data analysis methods	Methods to discover relevant information from recorded and gathered data, report results, and assist decision-making.
I9	RQ 1, RQ 2	Participant type	User group participated in usability evaluations.
I10	RQ 1, RQ 2	No. of participants	No. of participants involved during usability evaluations.
I11	RQ 1, RQ 2	System evaluated usability characteristics	Characteristics used to determine how a digital healthcare system will best serve consumers.
I12	RQ 1, RQ 2	Usability evaluation tools	Automated tools or methods used to assist usability evaluations.
I13	RQ 2	Usability method perception	Researchers' perception about employed usability method(s) in practice.

4.4. Data extraction

A data extraction form was developed to answer the research questions. The covered data items, along with an explanation and the association with the research questions that each data item supports to address, are depicted in Table 3. The form was used to elicit data relevant to answer the postulated research questions. Each paper in the set that resulted from the previous phase was thoroughly examined, with one of the authors filling in the data extraction form and the other author verifying its accuracy. The data extracted from each article was recorded in a worksheet (Excel) and then coded in NVivo 12 to categorise and compare extracted facets. Three types of information were collected from each study: general information such as the paper title, publication year, and the paper's primary goal(s); contextual information about usability evaluation; and the information that can be directly used to answer research questions, such as the usability evaluation methods.

As a pilot run and to minimise errors and mitigate researcher bias during the data extraction phase, both authors independently extracted data from the initial five sample articles. Subsequently, the first author reviewed and extracted the other selected articles, while the second author reviewed and validated the data extraction form continuously but randomly. All differences and disagreements were raised and resolved in regular weekly consensus meetings.

4.5. Data analysis

The data extracted from the primary studies was analysed so that the answers to the research questions could be made both qualitatively and quantitatively. Frequencies/distributions for most facets were computed as well as for combinations of facets to find interesting associations. Usability evaluation tools were explored, analysed and compared with each other to explore the potential for using automated usability tools or solutions to facilitate usability evaluation studies. The retrieved articles were analysed to identify articles that focused on the feedback, researchers and professionals have expressed about the effectiveness, efficiency, and satisfaction of usability evaluation methods used in practice. The perceptions were categorised under the

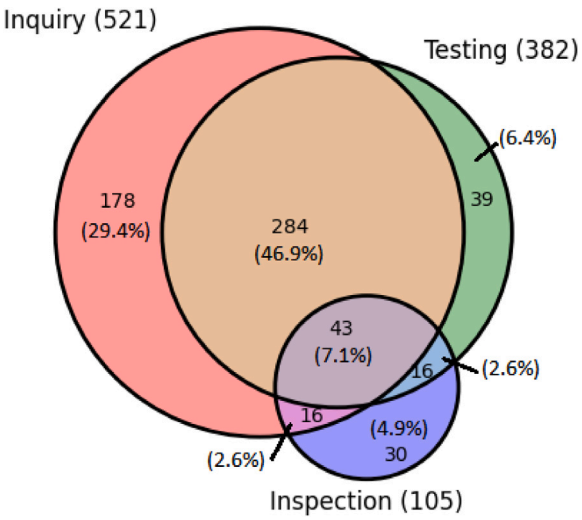

Fig. 2. Frequencies of (combinations of) usability evaluation methods.

Table 4
The seven most frequently applied (combinations of) data collection methods for inquiries.

Methods	Frequency (relative)
Questionnaires	295 (56.7%)
Questionnaires + Interviews	73 (14.0%)
Questionnaires + Observation	36 (6.9%)
Questionnaires + Interviews + Observations	32 (6.2%)
Interviews	32 (6.2%)
Observations	15 (2.9%)
FGD	12 (2.3%)

respective groups and analysed to identify shortcomings or merits of usability evaluation methods used in different contexts and scenarios.

5. Results

A total of 610 articles published between 2016 and April 2023 were identified and analysed in this SLR. The following subsections present the results per relevant aspect as defined in the data extraction form (see Table 3). Please note that for the sake of brevity, this section mainly reports aggregated frequencies of articles instead of detailed lists of articles per category. An overview of a usability evaluation taxonomy designed based on results in the context of DH is presented in Fig. 3. Furthermore, a replication package can be assessed to find the list of primary selected and rejected studies after a full paper review, and an NVivo project used to categorise (classify) and compare extracted facets.[3]

5.1. Usability evaluation methods

Two main types of usability evaluation were utilised in the selected papers: inquiry (n = 521) and testing (n = 382) methods. Inspection methods appear to play a minor role (n = 105). Fig. 2 breaks these figures down into figures for unique combinations of methods. The results show that using inquiry methods only (29.4%), combining them with testing (46.9%), or combining them with inspection and testing methods (7.1%) account for over 83% of the studied articles.

As explained in Section 2.3 and as manifested in the developed data extraction form, data collection, recording, and analysis can be seen as conceptual building blocks of any usability evaluation method. The following subsections cover these building blocks.

Questionnaires (n = 452, 74.6%)[4] were identified as the most widely used **data collection method** in *inquiry* approaches, followed by interviews (n = 152, 25.1%), observations (n = 96, 15.8%), and FGD (n = 32, 5.3%).

Table 4 shows the seven most frequent combinations of techniques, which account in sum for over 90% of papers that describe usability evaluation involving inquiries. Questionnaires are part of the first four of these combinations, with almost accounting for 84% of the papers. In almost 57% of the papers, questionnaires were even used as the sole inquiry method.

Standardised questionnaires were the most commonly applied types of questionnaires, with the *System Usability Scale (SUS)* being predominant (n = 221, 49.0% of studies employing questionnaires). Other standardised questionnaires include the *Post-Study System Usability Questionnaire (PSSUQ) (n = 25, 5.3%)*, the *NASA Task Load Index (NASA-TLX)* (n = 19, 4.2%), the *User Experience Questionnaire (UEQ)* (n = 18, 3.9%), the *Questionnaire for User Interaction Satisfaction (QUIS)* (n = 14, 3.1%), the *Usefulness, Satisfaction, and Ease of use (USE)* (n = 10, 2.2%), and the *Health Information Technology Usability Evaluation Scale (Health-ITUES)* (n = 9, 2.0%). Customisation or modification of generic/standardised questionnaires was observed in 27 studies. In four studies, new usability questionnaires were developed.

Among the studies employing *interviews* as a data collection method, semi-structured interviews were frequently used (n = 104, 68.4%). Also, 42 articles (27.6%) did not explicitly mention the type of interview applied. In only 32 instances (6.2%), interviews were applied as the only mean of data collection.

Among *inspection* methods, heuristic evaluations were the most prominent data collection techniques in studies (n = 77, 73.3% of studies applying inspections). In particular, Jakob Nielsen's usability heuristics for user interface design (Molich and Nielsen, 1990; Nielsen and Mack, 1994) were used most frequently (n = 49, 63.6%). Six studies were reported to have customised or used modified versions of heuristics. Cognitive walkthroughs were identified as an applied inspection method identified in 24 studies. Furthermore, generic design guideline checklists were used in nine studies.

Studies applying *testing* methods apply, almost without exception, task and scenario-based techniques to collect data (344 of 382). In 258 of these papers (75%), those techniques were described as the only method applied, while in 80 papers they were used in combination with think-aloud protocols (21%). Only three papers have employed think-aloud protocols exclusively. Additionally, in a total of 32 articles, evaluators carried out randomised testing, which involved test participants without predetermined tasks. Moreover, four articles also utilised the Wizard-of-Oz technique, while two employed card sorting techniques.

In the descriptions of their studies, a total of 163 articles explicitly mentioned the **type of data recorded** during the usability evaluation process. The most frequently recorded data included notes or participants' written or verbal feedback (n = 74, 45.4%), audio or video recordings (n = 72, 44.1%), system logs such as interaction, event, or error logs (n = 33, 20.2%), eye tracking (n = 16, 9.8%), screen recording (n = 11, 6.7%), and emotion recognition (n = 2, 1.2%).

For 276 (45.2%) of the analysed articles, the authors explicitly referred to the applied **data analysis methods** whereby multiple mentions occurred. In 167 papers (60.5%), we could identify the use of qualitative data analysis methods. These included, among others, coding analysis (n = 79, 47.3%), thematic analysis (n = 79, 47.3%), general inductive analysis (n = 31, 18.6%), grounded theory (n = 4, 2.4%),

[3] The SLR repository (replication package): https://doi.org/10.5281/zenodo.8218076.

[4] Of 606 studies employing usability methods (inquiry, inspection, or testing).

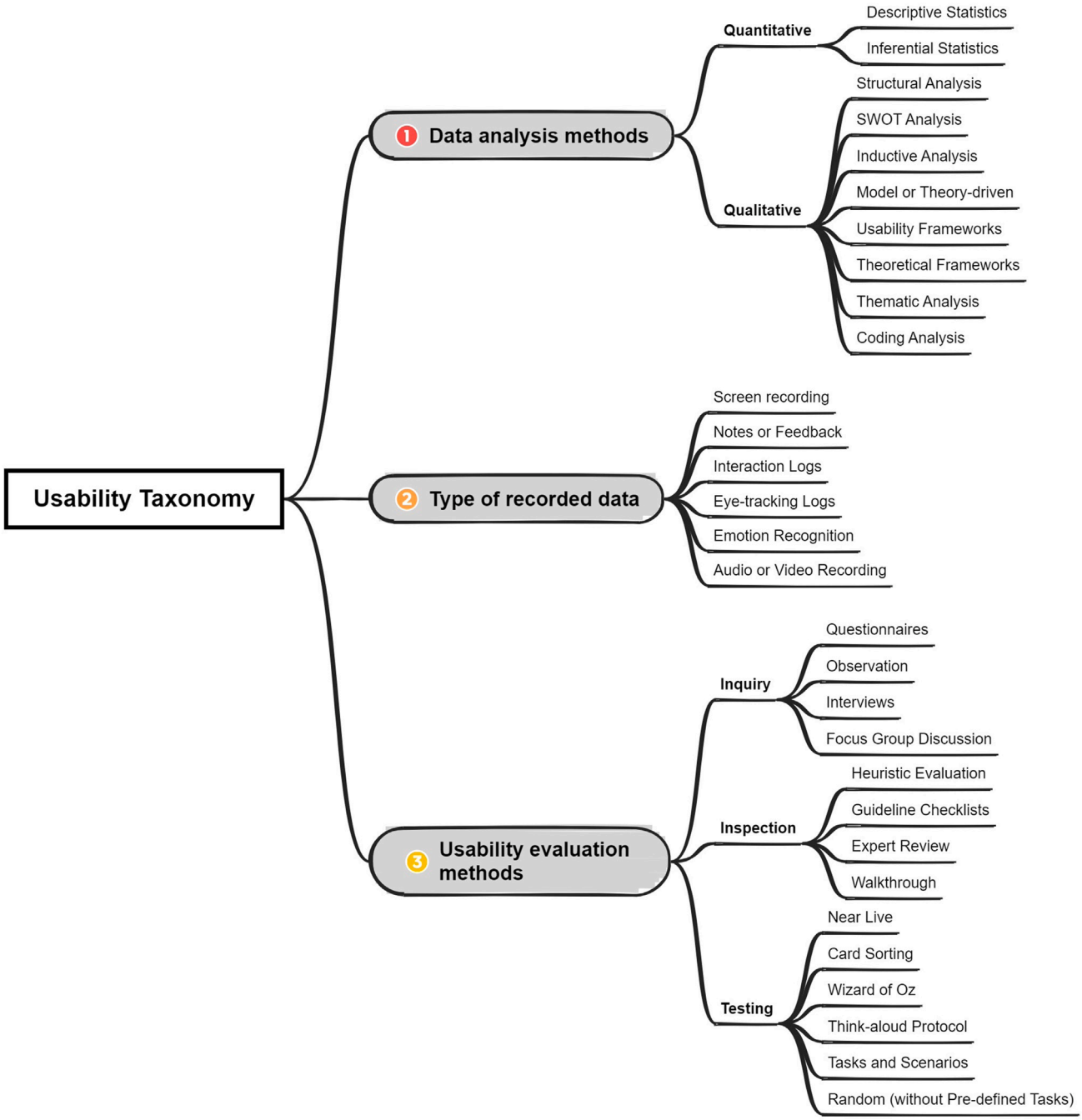

Fig. 3. A usability evaluation taxonomy — based on findings.

and Framework Analysis method and Usability Problem Taxonomy (FA-UPT) (n = 4, 2.4%).

Quantitative analysis methods were performed in 177 papers (64.1%) of which 123 (69.5%) included descriptive statistics and 89 (50.3%) inferential statistics.

5.2. Evaluated usability characteristics

The usability characteristics as listed in Section 2.3 were applied to categorise according to which characteristics they were investigated in the applied usability evaluation studies. The frequencies are listed in Table 5. Satisfaction, effectivity, efficiency, and learnability appeared in more than half of the papers, respectively, while the characteristics of aesthetics, operability, usefulness, memorability, and accessibility were less assessed. Table 6 shows a breakdown of the total frequencies of usability characteristics according to subfields of DH. It shows that the four characteristics first mentioned above are prevalent across all the subfields.

5.3. Participants

The systems tested in the analysed studies targeted different user groups. We classified these user groups as general users (n = 152),

Table 5
Evaluated usability characteristics.

Characteristics	Frequency (relative)
Satisfaction	551 (90.3%)
Effectiveness	424 (69.5%)
Efficiency	359 (58.9%)
Learnability	318 (52.1%)
User error protection	200 (32.8%)
Accessibility	85 (13.9%)
Memorability	79 (13.0%)
Usefulness	69 (11.3%)
Operability	49 (8.0%)
Aesthetics	31 (5.1%)

healthcare personnel (e.g., doctors, nurses; n = 245), informal caregivers (e.g., family members providing care of ageing relatives; n = 29), and patients (n = 400).

Table 7 lists percentages of usability evaluation studies comparing targeted user groups and groups of users actually participating in the usability evaluation.

IT experts, such as programmers or usability evaluation experts, are added as an additional group that is sometimes used as a substitute for user groups or as a necessary group for example in inspections. Averaged over the user groups, the participating user group was also the target audience of the system under investigation in ca. 49% of studies. This is true for patients (42% of patient-targeted studies), healthcare personnel (HCP) (52%), general users (64%), and informal caregivers (37%).

We also investigated the distribution of data collection methods across the different participating user groups, which turned out to reflect the overall distribution of methods as identified and described in data collection methods (see Section 5.1). This is also true for inquiry methods, which means that questionnaires are the dominant method across all types of participating user groups.

5.4. Usability evaluation tools

Only a few studies (n = 25, 4.1%) reported the use of various tools and methods to automate different parts of the usability evaluation process. In 15 instances, user activity tracking tools were used during *testing*. These included *Morae*[5] (n = 8) (Au et al., 2008), *Google Analytics*[6] (n = 3), *TURF*[7] (n = 2), and a customised data capture system that was later validated with Google Analytics analysis (n = 1).

Other tools applied in studies are mainly concerned with accessibility. *AChecker*[8] (n = 3), an online accessibility testing tool. Three studies employed Natural Language Processing (NLP) techniques to analyse user reviews or comments on DH applications. *TAW*[9] (n = 2) was used for website accessibility analysis based on W3C. One study used *CogTool*[10] (n = 1), which is a user interface (UI) prototyping tool used for predictive performance measures. *Google Chrome SEE application*[11] (n = 1) allowed experiencing the web as a person with a vision impairment, for example as being colour-blind would see Sik-Lányi et al. (2017). One study used different readability level methods and tools to assess the readability of English content on DH websites. Another research employed a readability analysis tool (*Readable*,[12] n = 1) to improve writing by measuring text readability density using Flesch-Kincaid readability score. Among others, *Webpage analyser (n = 1)*,[13] a free tool was used to test website speed to report website performance as well as provide suggestions on how to optimise website performance. Furthermore, a study (n = 1) proposed a new methodology or tool for emotion recognition to objectively, autonomously, and remotely evaluate the usability of an eHealth website and identify mutually exclusive emotions when users interacted with the website.

5.5. Evaluated system types

The usability of different types of systems was assessed across various stages of system maturity, including fully developed or deployed systems (n = 314, 51.5%), prototypes (n = 214, 35.1%), and incrementally developed systems (n = 87, 12.3%). These software systems were classified according to the different subfields of DH systems or technologies, as described in Section 2. The systems evaluated in the different studies included mHealth applications (n = 430, 70.5%), telehealth platforms (n = 222, 36.4%), Health Information Technology (HIT) solutions (n = 175, 28.7%), Personalized Medicine (Per. Med.) solutions (n = 103, 16.9%), wearable devices (n = 72, 11.8%), and Digital Therapeutics (DTx) interventions (n = 39, 6.4%).

The study found a range of mHealth interventions assessed for usability in primary studies, encompassing diverse technological platforms. The predominant platform utilised was smartphones (n = 244), followed by gaming applications (n = 82), including exergames and serious games (SG). Websites (n = 79), tablets-based apps (n = 40), health communication chatbots or platforms (n = 16), desktop applications (n = 15), Augmented Reality (AR)-based interventions (n = 9), and smart TV applications (n = 8) were also found in the study.

The study also found a variety of telehealth or telemedicine systems, including virtual care and telerehabilitation services (n = 57), health monitoring and management solutions (n = 38), medical appointment or management systems (n = 5), as well as general smartphone, AR/VR, IoT-based, and chatbot applications (n = 129).

The study found various types of HIT systems, which were categorised as follows: health records and information management systems (n = 86), such as electronic health records (EHR), electronic patient records (EPR), personal health record (PHR), and health information systems (HIS); clinical decision support and advisory systems (n = 59); patient interaction and engagement systems (n = 11); and healthcare operations and management systems (n = 9).

Key personalised healthcare included digital rehabilitation tools (n = 28), health and wellness management (n = 22), assistive technologies (n = 15), clinical decision support systems (CDSS) (n = 13), and personalised medication management (n = 4).

Wearable devices included virtual and augmented reality devices (n = 49), fitness and activity tracking (smartwatch) (n = 15), and general health monitoring devices (n = 5), like electronic sphygmomanometers and wearable fall risk assessment systems (FRAS).

DTx included assistive technologies (n = 17), physical health therapeutics or rehabilitation (n = 14), nutritional therapeutics (n = 3), and general smartphone and Alexa-based therapeutics (n = 5).

Tables 8 and 9 show a breakdown of the figures for applied data collection and inquiry methods, respectively. They show that the overall prevalence of the inquiry method, and within these the widespread use of questionnaires, noted before is independent of the technology investigated.

[5] Discontinued now.
[6] https://analytics.google.com/analytics/web/
[7] https://sbmi.uth.edu/nccd/turf/
[8] https://achecker.achecks.ca/checker/index.php/
[9] https://www.tawdis.net/index/
[10] https://www.cogtool.org/
[11] Not accessible now, last accessed: 07 Aug 2023. https://chrome.google.com/webstore/detail/see/dkihcccbkkakkbpikjmpnbamkgbjfdcn/.
[12] https://readable.com/
[13] http://www.websiteoptimization.com/services/analyse/

Table 6
System types and their evaluated usability characteristics.

	DTX *Col %*	HIT *Col %*	mHealth *Col %*	Per. Med. *Col %*	Telehealth *Col %*	Wearable devices *Col %*
Accessibility	4.4%	2.8%	4.2%	5.5%	5.7%	4.0%
Aesthetics	2.2%	1.2%	1.6%	1.4%	1.2%	2.0%
Effectiveness	22.8%	19.3%	19.4%	20.1%	20.2%	20.0%
Efficiency	16.2%	18.8%	16.2%	15.6%	14.9%	14.4%
Learnability	14.7%	13.3%	15.5%	14.6%	13.7%	14.8%
Memorability	1.5%	3.3%	4.0%	3.1%	3.0%	2.0%
Operability	2.9%	1.8%	1.9%	5.0%	3.0%	4.4%
Satisfaction	27.9%	25.4%	25.3%	22.3%	25.4%	27.6%
Usefulness	2.9%	2.8%	3.3%	4.3%	4.6%	4.0%
User error protection	4.4%	11.4%	8.6%	7.9%	8.4%	6.8%

Table 7
Comparison of targeted vs participating user groups (relative frequencies).

Participated / Targeted	IT experts	General users	Healthcare personnel	Informal caregivers	Patients
General users	14.4%	63.8%	9.6%	0.5%	11.7%
Healthcare personnel	12.9%	15.5%	51.9%	2.1%	17.6%
Informal caregivers	16.7%	14.8%	16.7%	37.0%	14.8%
Patients	13.3%	26.3%	16.6%	2.2%	41.7%

Table 8
Evaluated system types and usability evaluation methods.

	Inquiry	Inspection	Testing
DTx	56.7%	8.3%	35.0%
HIT	51.5%	11.0%	37.6%
mHealth	51.4%	10.4%	38.2%
Per. Med.	50.6%	7.8%	41.7%
Telehealth	52.3%	9.2%	38.5%
Wearable devices	54.0%	5.7%	40.3%

Table 9
Evaluated system types and employed inquiry methods.

	Questionnaires	Interviews	Observation	FGD
DTx	70.0%	22.5%	5.0%	2.5%
HIT	59.5%	20.5%	15.5%	4.5%
mHealth	60.4%	21.2%	13.6%	4.8%
Per. Med.	63.3%	21.1%	9.4%	6.3%
Telehealth	62.3%	20.9%	12.8%	4.0%
Wearable devices	69.7%	20.2%	9.0%	1.1%

5.6. Health condition or disease

The UKCRC Health Research Classification System (HRCS) is a system to categorise diseases and conditions (UKCRC, 2017). The health categories are based on the International Classification of Diseases (ICD) codes of the World Health Organisation (WHO), which encompass all areas of disease and health conditions. From each article, we extracted information about the condition of the disease of interest, if applicable, and categorised it according to the HRCS.

Table 10 lists the identified categories in descending order of their frequencies; categories missing in the table were not populated by any of the analysed papers.

A more detailed look at the association between the condition and disease category of interest and applied data collection techniques confirmed the previous results regarding the prevalence of techniques.

Table 11 shows the percentage of user groups who participate in usability evaluations of DH technologies. These technologies were specifically targeted and developed for those user groups and the top 10 prevalent health conditions or diseases. In the usability evaluation studies, a mixed level of participation was observed across the different user groups, depending on the specific health condition or disease category. While these systems are designed for a wide range of health conditions and user groups, the data unveils a relatively low participation rate of patients across all health condition categories. The average participation rate for patients, in general, was found to be 43.4%, with a high of 56.3% for eye conditions and a low of 32.4% for respiratory conditions. General users were seen to participate more frequently, averaging 60.5%, with the maximum participation in musculoskeletal (72.7%) and metabolic and endocrine (71.4%) conditions. However, their participation dropped to 40% in neurological conditions.

Healthcare providers (HCPs) exhibited an average participation rate of 46.8%, but their engagement varied widely. They were most involved in the cases of metabolic and endocrine conditions at 60.0%, but were entirely absent in eye condition studies. Informal caregivers, on the other hand, had an average participation rate of 47.9%. The participants showed complete involvement (100%) in studies related to respiratory, metabolic, and endocrine disorders, but did not participate at all in studies related to musculoskeletal conditions.

Table 10
Health condition and disease.

Category	Frequency (relative)
CAT 1. Generic health relevance e.g., well-being, ageing, chronic diseases	280 (38.1%)
CAT 2. Mental health e.g., cognitive impairment, dementia, autism, depression	103 (14.0%)
CAT 3. Musculoskeletal e.g., physical rehabilitation, fall prevention	79 (10.8%)
CAT 4. Neurological e.g., dementia, autism, parkinson's, alzheimer	56 (7.6%)
CAT 5. Respiratory e.g., asthma, COVID-19, tuberculosis	46 (6.3%)
CAT 6. Metabolic and endocrine e.g., diabetes	42 (5.7%)
CAT 7. Reproductive health and childbirth e.g., pregnancy, childcare	27 (3.7%)
CAT 8. Cancer and neoplasms e.g., cancer, HIV	24 (3.3%)
CAT 9. Cardiovascular e.g., heart attack	19 (2.6%)
CAT 10. Eye e.g., visual impairment, blindness	14 (1.9%)
CAT 11. Renal and urogenital e.g., kidney disease, pelvic inflammatory disease	13 (1.8%)
CAT 12. Stroke	8 (1.1%)
CAT 13. Blood, CAT 14. Congenital disorders CAT 15. Infection, CAT 16. Oral and gastrointestinal	4 (0.5%) (*each*)
CAT 17. Ear, CAT 18 Inflammatory and immune system	3 (0.4%) (*each*)
CAT 19 Injuries and accidents	2 (0.3%)

Table 11
User participation in usability evaluation studies for DH systems designed for specific health conditions or diseases.

Category	Patients	General users	HCP	Informal caregivers
Generic health relevance	39.6%	64.2%	55.0%	29.2%
Mental health	41.8%	66.7%	45.7%	40.0%
Musculoskeletal	44.2%	72.7%	39.3%	0.0%
Neurological	45.0%	40.0%	55.0%	47.1%
Respiratory	32.4%	63.0%	53.3%	100.0%
Metabolic and endocrine	51.0%	71.4%	60.0%	100.0%
Reproductive health and childbirth	43.5%	66.7%	58.8%	33.3%
Cancer and neoplasms	42.9%	–	46.7%	–
Cardiovascular	37.0%	50.0%	54.6%	–
Eye	56.3%	50.0%	0.0%	33.3%
Avg.	43.4%	60.5%	46.8%	47.9%

Overall, this data indicates that while DH systems are being developed across various health conditions, the user groups targeted by these systems are not always adequately represented in usability evaluation studies, potentially impacting the systems' overall effectiveness.

5.7. Expressed issues with usability evaluation

A total of 24 articles were identified that conducted usability evaluations of DH technologies as well as discussions on opinions, issues, and concerns with regard to the practical application of usability evaluation methods. Valuable insights were shared, highlighting the experiences faced during the implementation of these methods and offering diverse perspectives on their effectiveness and efficiency.

Several of them were concerned with the use of questionnaires. Georgsson and Staggers noticed in their study that their open-ended questionnaires uncovered fewer usability issues but more high-severity ones than the previously performed scenario-based testing [S13]. They suspect that participants were exhausted after testing and answered questionnaires less accurately. Lei-Juan et al. noticed in a study with elderly participants that questionnaires indicated more positive answers regarding usability than observations would suggest [S31]. They assume that the elderly participants might have held back too negative opinions about the tested product out of politeness or of being negatively judged as incapable or "unwise". In a study applying the SUS protocol, the authors expressed the need for complementary methods, for example, think-aloud protocols [S170]. Holmes et al. in the same vein, noticed that they were able to assess the difficulty of using the system under test but not if the participants enjoyed using the system [S199].

Chumkasian et al. examine the deficiencies in traditional usability evaluation techniques like SUS and Health-ITUES when applied to mobile health apps [S337]. They argue for the relevance of MAUQ, a specialised questionnaire for mHealth app evaluation, but note that it requires customisation for their specific healthcare application context. Through validation of a modified MAUQ in their research, they endorse the necessity for adaptability in these evaluation methods. The authors emphasise the need for domain-specific strategies to address unique usability challenges within mhealth apps. However, they also recognise that such methods may not be universally applicable and often need tailoring to particular application contexts or user demographics. This underscores the importance of a broad spectrum of specific methodologies for comprehensive usability assessments, which may require additional effort and resources.

Khajouei et al. [S77] and Farzandipour et al. [S411] compared properties of heuristic evaluations (HE) and cognitive walkthroughs (CW) for usability evaluation and attested comparable effectiveness. While HE proved more beneficial for overall user satisfaction, the researchers observed that CW was superior in identifying high severity usability problems. CW seems more appropriate when learnability is the focus of the evaluation, especially if the main user group of interest consists of novice users. Beauchemin et al. used both techniques and think-aloud protocols to iteratively improve a self-management application [S167]. They experienced that cognitive walkthroughs were efficient in validating that novel features were usable while heuristic evaluations were rather used to verify that updates or features were really needed, as expressed during think-aloud. Capeleti et al. further found that heuristic methods might not address all wearable technology challenges, hinting at the need for more specific testing strategies and considering users' group mental model differences [S269]. The complementary roles of user testing and heuristic evaluations were found important in addressing various usability and design issues [S269, S271]

In a different study, Khajouei et al. compared checklists with other usability evaluation methods [S135]. They concluded that checklists were easier to use for novice usability evaluators but less suited to identify high-severity issues than other heuristic techniques, and therefore suggest using them with caution if usability is a major concern. Other researchers recommend combining heuristic evaluation with other approaches [S35] and to involve end-users as well as system experts [S121].

Petrie and Wakefield [S299] and Wahyuningrum et al. [S500] conducted separate studies and found that remote studies offer more benefits compared to face-to-face studies, despite the technical challenges involved. They found several advantages of remote evaluations, including speed, cost savings, increased security and privacy for participants, discrete observation and support from the research team, and the opportunity to conduct tests in a natural setting. However, they also highlighted the limitations of remote testing, specifically reduced experimental control and emphasised that remote evaluations require careful planning and efforts but present promising opportunities, especially during disruptive situations like a pandemic.

Regarding methods categorised as testing, Broekhuis et al. stressed the threat to external validity if usability is performed in a lab-like setting [S170]. Capeleti et al. further indicate the importance of context in usability evaluation, showing results can differ between laboratory and real-world settings [S269]. Particularly in DH, the actual usage context might simply differ fundamentally from such a synthetic setting. They, in line with other studies, therefore stress the relevance of remote usability evaluation [S121, S233, S299, S500].

Furthermore, Schöne et al. researched adapting usability tests for cognitively impaired individuals, highlighting caregivers' crucial role [S351]. The study found that tailoring tests with simplified language, modified rating systems, and familiar environments significantly aids the process. The test leader, requiring an understanding of the testing process and the ability to adapt to cognitive limitations, played a vital role, with the caregivers' presence enhancing communication and motivation, despite potential interruptions, which were mitigated by pre-set signals. The importance of observation and collaboration with caregivers was emphasised, given the limited feedback from cognitively impaired individuals and considering caregivers' insights about the users' abilities. Finally, the study acknowledged the willingness of cognitively impaired individuals to use technology and take part in such studies, advocating for more inclusive and user-centred design approaches.

One study provided evidence for task completion as a metric, showing a greater association with usability benchmark results than SUS questionnaires [S170]. However, Le et al. stressed the relevance of defining representative tasks as the performance of participants depends strongly on the defined tasks [S30]. Tasks unlikely to be performed in reality may not reflect a system's usability in practice.

Few authors commented on the use of think-aloud protocols. Le et al. investigate "insight evaluation", which consists of testing by letting users perform open-ended tasks while thinking aloud [S30]. They noted that the results of insight evaluations corresponded closely to the results of usability benchmarks. Moreover, participants expressed feeling less like the actual subject of the tests than in other methods, due to the open nature of insight evaluations. However, possibly for the same reason, participants appeared to feel insecure about performing the task "right" and therefore often sought guidance from the moderator. The challenge for the moderator is to provide that guidance without influencing the direction of the participants' exploration of the system. Richardson et al. compared a think-aloud protocol with "near-live" usability testing in which healthcare professionals performed predefined tasks with a patient actor [S95]. They found that think-aloud results were mainly useful in improving ease-of-use aspects, while near-live testing provided helpful insights regarding the integration of the tested tool in the workflow. Cho et al. stress that combining think-aloud with eye-tracking techniques can provide valuable insights [S121]. They mentioned common obstacles, like losing eye movements due to user fatigue, lack of clear criteria of what constitutes a good eye movement pattern, and cost of equipment [S179].

Following the think-aloud protocol, in-depth interviews revealed the majority of problems [S13]. Furthermore, open-ended interview questions were found important and useful since they allowed users to freely express their opinions and provide insights on issues that may not have emerged during testing. The card sorting technique has also proven to be useful for identifying targeted end-user information preferences and establishing application components before developing DH prototypes [S121]. Other researchers discovered that feedback from end users and eLearning experts was helpful in analysing and improving the prototype [S233].

Overall, most authors presenting results about or analysing usability evaluation methods advocate mixed/multi-method approaches [S13, S19, S33, S56, S233, S261, S269]. Several of them mentioned that using mixed-method approaches led to improvements that would not have been detected as required or desired using a single method [S19, S233, S261, S269].

6. Discussion

Section 6.1 presents and summarises the key findings of the study. These findings shed light on the prevalence of usability evaluation in DH. Further subsequent sections are focused on exploring emerging themes around the effectiveness and efficiency of usability evaluation in DH. By exploring these aspects, we gain insights into the strengths, weaknesses, and potential areas for further exploration and improvement in usability evaluation practices, both in DH and in general. Such insights can be invaluable for shaping future research and advancements in DH usability evaluation and improving the effectiveness of DH systems.

6.1. Summary of key findings

The results showed that a combination of both **inquiry and testing methods** was the most frequently employed, evident in 47% of the studies. While, inspection methods were not as widely used, being employed as a single method in 5% of the studies.

Questionnaires, especially standardised ones such as SUS, emerged as the most common data collection method among the inquiry methods, accounting for 75% of their use. *Interviews*, on the other hand, were applied in about a quarter of inquiry-based studies.

Regarding **testing methods**, *task and scenario-based* usability evaluation was applied in approximately 75% of the studies. Further, about 20% of these studies combined think-aloud protocols, providing a more comprehensive understanding of the participants' thought processes.

The distribution of predominantly used methods and techniques was similar across all *DH systems types* and *targeted user or patient groups*. The results also showed that the majority of the studies (41%) were focused on investigating the usability of the mHealth systems compared to other DH systems, such as Telehealth (21%), HIT (16%), Per. Med (10%), Wearable (7%), and DTx (4%).

On average, 49% of the studies engaged **participants** from their targeted user groups. Often, the participants' insights were supplemented with those of healthcare professionals or IT experts. Particularly, IT experts were mostly involved in studies implementing inspection methods. In contrast, the inclusion of informal caregivers was less.

The top five **usability characteristics** evaluated in DH studies were satisfaction, effectiveness, efficiency, learnability, and user error protection. However, aesthetics, operability, usefulness, memorability, and accessibility were less frequently evaluated, pointing to a potential area for further research.

A low percentage of studies (only 4%) reported the use of **tools or platforms** to automate or semi-automate the usability evaluation process. This suggests a deficit in tool utilisation or reporting. These tools or platforms were often used to track user activity or to validate the overall accessibility of DH systems.

6.2. Dominance of inquiry methods

Our study unveils the predominance of inquiry methods in evaluating DH solutions' usability, as employed in 29% of the studies exclusively. These methods, such as interviews and focus groups, facilitate direct user engagement and valuable insights (Lyzara et al., 2019). Providing qualitative understanding unobtainable elsewhere, enable in-depth exploration of user needs, motivations, and preferences (Anuar et al., 2014; Rocha et al., 2014; Bolívar, 2014). Adapt to diverse research goals and contexts, enhancing understanding of users' backgrounds, preferences, and challenges. Additionally, they assist with the early detection of usability issues and timely adjustments are feasible.

Despite the obvious benefits, inquiry methods are not free of limitations (Anuar et al., 2014; Rocha et al., 2014; Bolívar, 2014). The need for extensive planning and coordination can make these methods resource-intensive and time-consuming, potentially affecting project timelines (Lyzara et al., 2019). The quality of data can be compromised due to question-wording, order, interviewer bias or inexperience, thus affecting reliability and validity (dos Santos Grecco et al., 2021; Ouellet et al., 2021; Alcaraz-Quiles et al., 2018; Verkijika and De Wet, 2018). However, skilled researchers can mitigate these issues by designing comprehensive protocols and careful evaluations (Brédart et al., 2014). Furthermore, sample representativeness poses a significant challenge, especially when dealing with large samples (Anuar et al., 2014; Bolívar, 2014; Koopman and Zuccon, 2016). Researchers might have to limit sample size due to associated costs and effort, potentially resulting in under-representation and reduced generalisability of findings.

Approximately 41% of the studies report using just one method for usability evaluation, which might overlook non-obvious usability problems. In particular, inquiry methods may fail to detect non-obvious usability problems, as users with cognitive impairment might struggle to articulate or recognise them during interviews. The consensus is that combining different methods improves usability issue identification due to their varied strengths and weaknesses [S35, S121, S512, (Lyzara et al., 2019)]. This combination ranges from detecting more issues to identifying fewer but high-impact issues, and from being more suitable for beginners or experienced users to be ideal for driving new

features or validating the usability of new features (see Section 5.7). Complementary strategies, like combining inquiry methods with testing, further enrich the evaluation strategy for DH products. Riihiaho (2018). Such a combination can enhance the inquiry process' effectiveness by providing empirical data from testing to supplement the insights [S512]. This approach can facilitate data validation and result in cross-referencing for improved reliability.

The studies also underline the vital role of ensuring data privacy in healthcare, where usability evaluation methods often collect sensitive data (Fadahunsi et al., 2021; Buie and Murray, 2012; Genov et al., 2009). Any breach could undermine user trust and legality. Furthermore, potential security risks should be considered, as insecure platforms can expose collected data to threats. Adherence to privacy, security, and ethical considerations, such as informed consent and data privacy and management planning, directly can influence trust, legality, and overall effectiveness and efficiency of the usability evaluation process.

Socio-cultural aspects add another layer of complexity to the usability evaluation. For instance, a study on a telehealth application for diabetes management before or during Covid-19 might yield different usability insights based on social limitations and cultural contexts (Karim et al., 2021; Egle et al., 2022). Users' priorities may differ, with some focusing on cost-effectiveness while others prioritise data privacy and security of DH products (Khoja et al., 2013). Cultural and linguistic differences, influencing user feedback and perceptions (Ouellet et al., 2021), require inquiry methods to adapt for cultural sensitivity and contextual relevance (Riihiaho, 2018).

Emerging technologies such as AI chatbots and virtual reality (VR) interfaces present both new challenges and opportunities (Haverinen et al., 2019). Augmented reality (AR) and VR offer immersive experiences that differ from traditional interfaces, raising unique usability concerns like physical comfort, motion sickness, and depth perception difficulties. These interaction modes can raise the bar not only for inquiry methods, but also for other methods to effectively capture, analyse, and evaluate usability issues. Adapting methods to measure the usability of such technologies is crucial for improving the overall user experience.

Usability evaluation, generally costly and time-consuming, often leaves developers in uncertainty for efficient and context-specific testing, while managers face delays due to repeated revisions needed to correct usability issues (Tamir et al., 2008). The task of estimating the resources needed for usability evaluations in DH, which are often critical and sensitive in nature, further complicates the matter. The task of estimating resources such as test size, effort (person-hours), cost, and schedule for evaluations, a process that itself adds time and expense (Tamir et al., 2008; Kaur and Kaur, 2022). In terms of estimating usability evaluation efforts, further research is required to establish the applicability of existing or customised usability evaluation methods for DH technologies.

Improving methods for prevalent DH user groups remains a question for further research. This includes designing tailored questionnaires or interviews for evaluations with the elderly and adjusting them for people with disabilities who may interact uniquely with DH apps. Tailored testing may involve assistive technologies like screen readers or voice commands, and adapting interview methods to participants' cognitive abilities ensures inclusiveness and focused design. By making the language clear, using visual aids, changing the pace, and thoughtfully including support like caregivers, the process becomes more empathetic and efficient. Such approaches offer deeper insights into unique user needs, helping developers create more inclusive DH solutions.

In conclusion, inquiry methods are essential for usability evaluation. When integrated with other methods, they provide a comprehensive usability evaluation, harmonising qualitative and quantitative data. It is further recommended to clearly define research objectives, develop structured protocols, carefully plan participant recruitment, maintain neutrality, systematically analyse data, and triangulate with other methods to enhance the effectiveness and efficiency of inquiry methods (Nielsen, 1993; Kushniruk and Patel, 2004; Dillman et al., 2014).

6.3. Dominance of questionnaire

A majority of the literature, including our findings, evident in 57% of our studied cases uses questionnaires as the sole method, suggesting the dominance of questionnaires as a primary method in assessing the usability of DH technologies (Zapata et al., 2015; Paz and Pow-Sang, 2016; Maramba et al., 2019; Wang and Huang, 2022). The simplicity, efficiency, and cost-effectiveness of questionnaires have contributed to their widespread usage (Brooke, 2013; Lyzara et al., 2019; Wang and Huang, 2022). They offer a practical and scalable approach to collecting data and analysis from diverse user groups (Barnum, 2010; Riihiaho, 2018). Closed-ended questionnaires offer a quantifiable format that facilitates the generation of ample data, which can be easily interpreted and compared (Connelly et al., 2022), which is easier to fill in than long and open-ended questionnaires (Milne, 1999). Open-ended questionnaires can provide deeper insights (Walji et al., 2014) but also require more resources and time and often are difficult to analyse (Varela and Ares, 2012).

Moreover, the advent of new technologies has notably improved the administration and analysis of questionnaires. For instance, the ability to distribute and complete questionnaires online enhances both accessibility and efficiency (Couper, 2000). Furthermore, advancements in AI and ML have opened avenues for rapid and precise interpretation of results, particularly beneficial for large-scale studies (Davenport and Kalakota, 2019).

Despite their popularity, questionnaires are not without challenges. Planning, designing, and administering effective questionnaires are not easy, and it becomes increasingly challenging if the tasks they relate to are complex and/or critical, especially in unmoderated testing (Sauro and Lewis, 2011; Milne, 1999). The precision, clarity, and unbiased nature of questions significantly impact data quality (Larsen, 2003; Riihiaho, 2018). Furthermore, self-reported questionnaires also can introduce potential response bias, as respondents may not provide truthful answers due to social desirability bias or may fail to recall past events accurately (Riihiaho, 2018). Therefore, to reduce bias, potentially sensitive questions should be designed carefully and sent electronically or by mail rather than personally.

Given that questionnaires are typically administered after or even without user interaction with a product, feedback collection is often retrospective. This potentially delays the identification or resolution of usability issues, and issues with recall remain a thread with people having cognitive issues. Moreover, questionnaires, which primarily gather quantitative data, may not capture non-verbal cues or perceptive information, limiting their capacity to fully comprehend the emotional and other affective aspects of usability.

Among standardised questionnaires, the SUS was the most popular, with attributes of quick administration, reliability, and wide applicability for both small and large sample sizes (Tullis and Stetson, 2004; Brooke, 2013). It is also relevant and commonly used for comparing two different versions of a system based on how each is perceived in terms of usability (Brooke, 2013). However, we noticed that in the investigated studies, SUS was generally utilised to assess only one version of a system. The information value of such cases is very limited, as there is no baseline for comparison. It has also been noted that the determined usability score has little to do with task performance (Brooke, 2013). Furthermore, these methods do not always yield a detailed understanding of specific usability issues, and thus can sometimes lead to a less thorough understanding of a product's unique usability challenges. Furthermore, both SUS and PSSUQ are free to use and technology independent questionnaires (Lewis, 2002; Brooke, 2013) but can be used depending on the situation and the case. For example, SUS is used when it is necessary to measure the user's perceived usability of the system/product (Brooke, 2013) while PSSUQ is important for measuring perceived customer satisfaction (Lewis, 1992). Furthermore, the Single Ease Question (SEQ) can be used to

measure task level usability instead of SUS. QUIS has undergone several revisions and has the ability to provide general feedback on the system (Schneiderman, 1987; Hodrien et al., 2021). But in comparison to SUS, QUIS is more lengthy, time-consuming and costly to use because it requires a licensing fee for paper/online versions.

NASA-TLX measures perceived workload, including mental and physical effort, and is the standard in many Human Factors and Ergonomics studies (Hart and Staveland, 1988). Unlike SUS, PSSUQ, and UEQ, NASA-TLX lacks clear benchmark scores for acceptable workload and is complex and lengthy, thus less commonly used in UX research (Miller, 2001; Andersson, 2017). It indicates whether interface changes improve workload without specifying fixes. However, it can disrupt the study's flow and cause fatigue due to being a lengthy questionnaire that must be completed after each important activity. NASA-TLX can be helpful in scenarios where the human error must be minimised, such as healthcare (Hart, 2006; Nur et al., 2020).

Moreover, cultural context, language nuances, and user familiarity with certain concepts greatly influence the effectiveness of questionnaire design and data interpretation (Sonderegger and Sauer, 2013; Riihiaho, 2018). Questions must be culturally sensitive and avoid presumptions about user familiarity with certain concepts. Language nuances and cultural perspectives can influence question understanding and responses, potentially leading to bias or misunderstanding.

The European Union's medical device directive (MDR) guides the design and development of medical software, yet lacks explicit guidelines for validating aspects like "performance, safety, correctness, and usability" (Bochicchio et al., 2021). This shortcoming becomes particularly crucial with digital therapeutics (DTx), where quality bears direct implications on human security and safety. DTx demands distinct evaluation techniques to meet its specific needs and the constraints of clinical trials, validating new therapies' effectiveness and safety. So, the application of usability methods, such as questionnaires, also needs to be adapted to DH system types. However, in general, due to higher reliability, it is recommended to use standardised questionnaires rather than constructing new questionnaires (Hornbæk and Law, 2007). Instead of designing new assessment methods, researchers also employed established ones for DTx (Bochicchio et al., 2021), notably the International Measurement System (IMS) (Aitken and Lye, 2015) and the Mobile Application Rating Scale (MARS) (Stoyanov et al., 2015). Both are widely recognised and leveraged in the DH field, with IMS providing a more objective evaluation of DTx and MARS offering a broader, multi-dimensional approach to assessing the quality of health apps. A study further showed that implementing a standardised collaborative method to evaluate DTx apps' usability and quality in clinical trials can enhance the validation process and optimise resource allocation in developing effective digital therapeutics (Aitken and Lye, 2015).

Meanwhile, results also showed that 4% of articles customised or modified standardised questionnaires, or 20% of articles used general self-designed close-ended questionnaires. Customised questionnaires can capture unique product features or usability factors specific to the context of use. This shows practice and a need for customised questionnaires, which can target specific areas of interest inadequately covered by standardised methods (Nielsen et al., 2014), but with proven reliability and trials (Riihiaho, 2018). It is also worth noting that the standardised Health-ITUES questionnaire for the healthcare domain is not widely used, although it allows for adaptability, customisation, and deeper analysis (Yen et al., 2010; Schnall et al., 2018), which we assume to be due to the higher degree of familiarity with SUS among practitioners. However, the design of a well-structured, unbiased, and effective customised questionnaire requires considerable expertise, and it can be time-consuming. Still, their validity and reliability might not match those of standardised counterparts (Hornbæk and Law, 2007).

This leads us back to the critical point of balancing effectiveness and efficiency in usability evaluation. While questionnaires, especially standardised ones, offer efficiency, their effectiveness in examining context-specific usability aspects may be limited. This might explain the common recommendation of the use of mixed-method approaches, which balance the depth of insight (effectiveness), like from interviews and observations, with the extensive coverage (efficiency) of questionnaires (Batsis et al., 2018). Furthermore, the inclusion of open-ended questions in questionnaires can help users provide detailed explanations or insights into their experiences, thereby contributing to a better understanding of usability issues (Harland and Holey, 2011). Additionally, piloting questionnaires can help identify any ambiguities or issues before the actual evaluation (Reeves et al., 2002).

In summary, the prevailing use of questionnaires, particularly the SUS, in DH usability evaluations could be complemented with other questionnaires for improved effectiveness. They should not be used in isolation, and a comprehensive understanding of usability can be achieved with a mixed-method approach tailored to the context and objectives. Furthermore, a critical review of various questionnaires is needed to discover the appropriate selection of standardised or customised questionnaires for specific user groups or health systems in DH.

6.4. Dominance of task and scenario-based evaluation

The task and scenario-based methods were found to be the dominant method in usability evaluations, accounting for 75% of the studies. These methods involve the design of realistic tasks, representing actual anticipated use, and reflecting the key components of the system being evaluated. Task and scenario-based usability evaluations are robust and effective (Earnshaw et al., 2001), but it has to be carefully crafted and formulation is very important (Fritz and Berger, 2015). This is also critical considering the time and effort generally spent recruiting test users.

In general, it is said, that the more tasks assigned to each user, the higher the probability of detecting potential issues (Lindgaard and Chattratichart, 2007). However, the tasks should be concise, typically not exceeding an hour to maintain user concentration, yet sufficiently diverse and comprehensive to allow users to explore the system (Riihiaho, 2018). Optimal and well-crafted tasks and scenarios aid in the design of the test by removing artificiality and allowing to focus on the real needs of the user and match test objectives/goals (Rubin and Chisnell, 2008; Russ and Saleem, 2018; Barnum, 2020). These factors determine the accuracy and usefulness of a test, and evaluators must have prior experience, a clear grasp of the objective, and a consistent effort to improve task writing skills. However, there are not enough papers that give an overall direction for developing usability scenarios in the healthcare sector, which makes it more difficult to develop and use tasks and scenarios (Russ and Saleem, 2018).

One crucial element of tasks and scenario-based methods is ensuring task relevance and quality, which has a direct impact on the number and types of issues uncovered (Skov and Stage, 2012). The tasks should be clear, meaningful, and unambiguous. This can elicit natural interaction with the product, allowing for an authentic assessment of its usability. Even the wording of the tasks must be carefully considered, so as not to provide distinct clues to the required actions. Moreover, these usability evaluations often start with simpler tasks (Riihiaho, 2018). This eases the user into the test setting, providing them with an opportunity to familiarise themselves with the environment. Some tasks should be independent of one another and presented one at a time. If time constraints arise, this design permits the exclusion of some tasks without significantly impacting the overall test.

However, while task and scenario-based testing prove to be influential in usability evaluation, several potential issues must be acknowledged. Incorporating scenarios provides a socio-organisational context for the tasks, essentially connecting them. This context becomes particularly important when testing with specific cultural groups, as some users may find isolated tasks hard to interpret without relevant background scenarios (Clemmensen et al., 2009). Scenarios might take

the form of a specific shift in a nurse's workload, providing relatable contexts for the user. For instance, consider the case of a nurse working a night shift in an emergency department, facing a surge of critically ill patients needing immediate care. Amid the chaos, the nurse must prioritise tasks, perform life-saving procedures, and cooperate with the medical team for optimal patient care. This intense, relatable scenario centres users, enabling them to empathise with the high-pressure decisions nurses make and can highlight the complexities of nursing and the system in use.

Mentions of potentially inaccurate results due to the exhaustion of participants or to their perception/understanding of their role in testing (see Section 5.7), point at the necessity to carefully select usability evaluation approaches with typical user groups in DH in mind. The cognitive load implied by usability evaluation methods, such as task and scenario-based testing combined with think-aloud, or lengthy questionnaires and interviews needs to be considered for obtaining reliable results from any users, particularly from elderly or patients with conditions affecting cognitive abilities. Guidelines for performing usability evaluation for specific groups, e.g., the elderly, exist (e.g., by Silva and Nunes (2010) or Loring et al. (Engdahl et al., 2009)). Our results, however, do not indicate any significantly different use of methods for specific user groups, like a widely adopted application of such guidelines. Furthermore, the prevalence of task and scenario-based methods underline the importance of adequate task/scenarios (Russ and Saleem, 2018). Following guidelines for task writing might hence improve the effectiveness and efficiency; overall direction for developing them specifically for the healthcare sector is scarce though.

Furthermore, task and scenario-based methods involve participants, which can be challenging to recruit, particularly true for people with serious health or mobility concerns. In such cases, task and scenario-based methods can also be used with complementary methods. When preparing usability evaluations, it is advisable to conduct one or two preliminary usability inspections to identify potential usability issues early on Riihiaho (2018). If significant problems are identified during these assessments, they can be addressed before involving actual users. Alternatively, the identified issues can be validated through test tasks with real users.

Inspection methods are often considered easy and cost-effective methods that can be performed by small numbers of inspectors (3–5) (Ghasemifard et al., 2015; Wijesinghe et al., 2016; Riihiaho, 2018). These methods are also suitable for evaluating usability at early stages of system maturity and may be used to investigate a wide range of usability problems, in particular when used in combination (Riihiaho, 2018; Aziz et al., 2021; Hollingsed and Novick, 2007). Given the low number of studies identified to apply inspection methods in DH, the full potential to increase the effectiveness of usability evaluation through inspections is not exploited. Although inspection methods can aid in obtaining expert opinion and recognising the user's goals (Aziz et al., 2021), they should not be seen as a substitute for testing with real and representative users (Hollingsed and Novick, 2007). They may, however, detect some usability issues early on and hence reduce the cost involved in performing and acting upon later user-based testing.

Complementing testing with user interviews in real-use contexts is also seen as a valuable approach for several reasons (Riihiaho, 2018). Firstly, it provides valuable insights and ideas for developing appropriate scenarios and test tasks. Secondly, these interviews offer a more profound understanding of the users' objectives, enabling us to focus the tests on specific areas. Lastly, gathering information about the users' goals and priorities helps in creating effective redesign proposals to address the issues identified during the tests. But, according to Jakob Nielsen, "What users say and what they do are different" (Nielsen, 2001). Combining user observations with testing can also be a good practice to look further, as it will provide insight into how people do things because sometimes it can be difficult for them to explain how they use the system or perform activities (Mortensen, 2020).

The findings showed that think-aloud protocols were frequently employed in combination with task and scenario-based techniques for conducting usability evaluations in DH. Despite their effectiveness and high value, obtaining and analysing data from think-aloud sessions can be time-consuming and resource-intensive due to the fact that it generally includes gathering and evaluating all of a user's verbalisation (Nielsen et al., 2002; Güss, 2018; Riihiaho, 2018). When considering think-aloud for usability evaluation, one should be aware of the different types. *Concurrent* and *retrospective* think-aloud methods both highlight a similar set of usability issues and task performance, but they differ in certain ways (Ohnemus and Biers, 1993; Van Den Haak et al., 2003; Van den Haak et al., 2004). The concurrent think-aloud approach is a more accurate representation of a task-oriented usability evaluation, while the retrospective think-aloud approach is more likely to provide a wider range of user reactions (Van Den Haak et al., 2003). One consequence of the concurrent think-aloud protocols might be a negative effect on task performance through the increased cognitive load to think-aloud while performing tasks (which is a general issue in think-aloud, though) (Ericsson and Simon, 1984; Branch, 2000b). Concurrent think-aloud protocols may therefore be inefficient if task complexity is high, and this raises questions about their effectiveness in such situations. Thinking aloud, in general, may not be possible for each user group to participate effectively (Branch, 2000a; Nielsen et al., 2002; Fan et al., 2020b). The findings of Franz et al. in using think-aloud with older adults, for example, suggest that whether or not to use any of the two techniques depends on the characteristics of the participating individual and may be related to other factors (Franz et al., 2019). This finding is closely connected to our finding, in general, regarding the selection of the appropriate methods for the targeted user group. Therefore, the use of task and scenario methods and think-aloud protocols should carefully consider the cognitive characteristics of the prospective participants in the DH usability evaluation. Some efforts have been made to address some of the issues associated with think-aloud methods, and some possibilities and directions have been considered and validated to automate some of its aspects for health information technology (HIT) (Russ and Saleem, 2018) and other domains (Johanssen et al., 2019; Fan et al., 2020a).

To summarise, the importance of task and scenario-based usability evaluation, both separately and in combination with other methods, is well-known. However, careful planning is needed to avoid common pitfalls and maximise the effectiveness of the evaluation process. By understanding the users' goals and priorities, appropriate tasks and scenarios can be designed, and a meaningful evaluation process can be conducted. Properly planned and executed, task and scenario-based usability evaluation can offer deep insights and actionable data to guide product or system improvement.

6.5. Dominance of quantitative data collection and analysis methods

The results identify the prevalence and use of quantitative data collection and analysis methods in usability evaluations of DH products, highlighting the reliance on quantitative data in identifying usability issues, such as satisfaction ratings, task completion percentage and times, error rate, etc. The relative effectiveness and efficiency of quantitative and qualitative approaches vary, depending on the usability evaluation goals and the specific characteristics of the DH products under investigation.

Quantitative analysis methods, employed in 51% of the studies we analysed, appear to have gained prominence due to their inherent objectivity and the potential for generalisability (Simpson, 2015). Such methods, including descriptive and inferential statistics, can facilitate a systematic and standardised evaluation procedure. Descriptive statistics provide the basic features of the dataset and offer simple summaries about the sample and the measures. The use of inferential statistics can allow for hypothesis testing and the drawing of conclusions about populations based on sample data. Such a framework not only strengthens

the reliability of gathered data but also strengthens the reproducibility of such studies. Assisting in uncovering and analysing issues like design or functionality flaws quickly, thereby enhancing the overall efficiency of the evaluation process. However, an over-reliance on quantitative methods could prevent the comprehensive assessment of user experience (Goertzen, 2017). While these methods offer efficiencies, they can potentially miss context-specific or user-focused issues, areas where qualitative methods excel.

Quantitative methods, despite their extensive use, do come with further challenges. These include the risks of reductionism, where complex human behaviours and responses are overly simplified into numerical values. In addition, the complex nature and diversity of DH products could necessitate unique, context-specific usability metrics and analysis methods. Furthermore, conducting inferential statistics requires a sound understanding of statistical concepts and the competence to choose and apply the appropriate statistical tests. The collection of quantitative data can be resource-intensive, especially in cases where large sample sizes are needed. Additionally, designing and validating quantitative instruments like surveys or scales can be a demanding process in such cases.

The results show a comparatively less use of qualitative data collection methods, such as interviews, think-aloud protocols, and observation. However, qualitative data analysis methods were applied in 61% of the articles. Techniques such as coding and thematic analysis empower researchers to look into user behaviour and feedback more intensely, fostering a holistic view of product usability. Coding analysis and thematic analysis are both well-recognised for their effectiveness in unveiling patterns and themes in large datasets (Braun and Clarke, 2006). While grounded theory, despite its powerful ability to generate new theory from data, is only represented in a mere 2.4% of the articles. This underutilisation may stem from the method's demand for rigour, time, and a high level of expertise from the researcher (Holton, 2007). In general, qualitative methods, such as coding analysis and thematic analysis, also require a significant investment of time and resources as compared to quantitative analysis methods (Coates et al., 2021; Soni et al., 2022). Coding and categorising qualitative data is a labour-intensive process that demands fine attention to detail. The iterative process of refining codes and themes to accurately reflect the data can be time-consuming. Moreover, achieving reliability in these methods necessitates a deep understanding of the subject matter, an ability to interpret patterns in data, and, in some cases, the involvement of multiple analysts to ensure the reliability of the analysis. Furthermore, qualitative data collection can often be lengthy and complex, involving in-depth interviews, prolonged observations, or extensive document analysis.

In the era of rapidly evolving technologies like AI, the landscape of usability evaluation studies for DH products is set to change. Such technologies can considerably improve the effectiveness and efficiency of data analysis generated by healthcare applications for improved functionality as well as for other reasons such as usability analysis (De Choudhury et al., 2016). AI-powered tools can process vast data volumes with improved speed and accuracy, augmenting the efficiency of quantitative methods (Raghupathi and Raghupathi, 2014). Similarly, such tools can assist with the analysis of qualitative data, identifying patterns and trends that may otherwise go unnoticed. However, the use of these technologies needs careful consideration to ensure that they neither undermine the richness of qualitative data nor compromise data privacy (Price and Cohen, 2019). Furthermore, such technologies can provide new opportunities and challenges for the collection, processing, and interpretation of usability data, such as by enabling the automation of usability metrics or the analysis of unstructured data (Chen et al., 2014).

In order to improve the effectiveness and efficiency of DH products' usability evaluations, we recommend a mixed-methods approach integrating both quantitative and qualitative analysis methods, as supported by our results. Such a strategy can offer a more comprehensive and multifaceted understanding of usability while addressing the inherent limitations of each method. For instance, triangulation can be used to validate findings, employing multiple data sources and methods. However, the mixed-methods approach, though beneficial, can escalate issues if not properly managed. This approach necessitates expertise in both qualitative and quantitative research techniques (Bryman, 2006). It might also be difficult and time-consuming to integrate and compare results derived from various methods.

Given the inherent complexities associated with qualitative and quantitative data analysis methods, depending upon evaluation goals and available resources, researchers might consider alternative complementary evaluation techniques that may offer similar insights with less resource intensity. These methods, varying in their approach and application, carry distinct advantages relevant to different research contexts. For example, Narrative Analysis may provide a less resource-intensive analysis while offering contextually rich insights (Riessman, 2008; Braun and Clarke, 2006). Simultaneously, Content Analysis stands as a flexible alternative, enabling the examination of various communication modes and finding application in both qualitative and quantitative contexts (Krippendorff, 2018). Meta-analysis consolidates findings from multiple studies into a comprehensive statistical review, thereby offering robust insights without necessitating the resources needed for new quantitative investigations (Borenstein et al., 2021). While, in certain circumstances, a Case Study approach could be preferred, focusing on in-depth analysis of a single or a small group of units, and offering comprehensive insights with less demand on resources and time compared to broader qualitative methods (Yin, 2017). Emerging technological advancements have shown Computational Methods, particularly useful when analysing large datasets, with the potential to unveil unique insights that may not be covered by some traditional methods (Lazer et al., 2009). Bayesian Methods, meanwhile, provide an alternative in inferential statistics, integrating prior knowledge into statistical modelling, thereby proving more efficient and adaptable, especially when dealing with complex or small datasets (Gelman and Shalizi, 2013).

In conclusion, while complementary or alternatives present potential for effectiveness or efficiency, their selection should align with the research question, data nature, and study objectives. Researchers and usability evaluators must remain mindful of the unique assumptions, requirements, and potential challenges accompanying each method before deciding on an approach.

6.6. Distribution of and under-represented usability characteristics

Assessing the effectiveness of DH products is a complex task, demanding not only the participation of diverse users but also the careful design of tests and interpretation of results (Riihiaho, 2018). This process is further complicated by the need for predetermined measures specific to each system's attribute, and the rapid pace of technological change makes iterative evaluations increasingly resource-intensive. Our analysis of various studies unveils a mixed situation in evaluating usability characteristics. Satisfaction, effectiveness, efficiency, and learnability are frequently studied, highlighting their importance in ensuring that users are able to achieve the goals they want to achieve (effectiveness) productively (efficiency) with the help of the system, learning to do so (learnability), and finding the system pleasant to use (satisfaction). However, on the other hand, it unveils potential areas of concern, particularly under-represented or under-reported usability characteristics. Accessibility, memorability, usefulness, operability, and aesthetics are less emphasised, although several authors argue for their importance (Coursaris and Kim, 2011; Sivaji et al., 2014; Shamsujjoha et al., 2021), could potentially impact the effectiveness and efficiency of usability evaluations. Moreover, the findings on operability contradict previous literature. Zapata et al. noticed that operability was the most frequently investigated characteristic, however, covered a much smaller sample of studies (n = 22) (Zapata et al., 2015). Ansaar et al.

made the same observation in their study of 19 papers (Ansaar et al., 2020).

Concerns arise due to the heavy emphasis on immediate user satisfaction and task completion in usability evaluations. The neglect of under-represented characteristics might overlook key usability issues that manifest over time or affect specific user groups. Such imbalance could lead to reduced user satisfaction and eventual system abandonment. These, along with other characteristics, can be significant in usability evaluation, and failing to evaluate them can lead to poor evaluation and misleading interpretations. For example, if a DH application's usability evaluation is perceived as ineffective, it may indicate that these characteristics were not properly tested with the methods employed. Similarly, work context and content must be considered for effective evaluation of DH applications (Edwards et al., 2008; ISO, 2018; Dix et al., 2003). Furthermore, if relevant user groups are not involved in the testing process, issues relating to particular usability characteristics may go undetected. For example, the elderly may have difficulty understanding and learning to use applications, but younger ones may not (Leung, 2009).

In general, an effective usability evaluation should encompass a wide range of characteristics while considering resources and test objectives, particularly those that are important for the specific system under review (Nielsen, 2005). For instance, in personalised healthcare or digital therapeutics, user error protection or operability could be crucial. Ignoring these elements during the early stages might result in an incomplete understanding of a system's usability, leading to subsequent corrections that could have been avoided. For Personal Health Records (PHRs), learnability, operability, and error prevention are essential due to the self-managing nature of patients. Telemedicine platforms may focus on effectiveness and efficiency due to their role in timely patient-doctor communication. Likewise, mobile health apps require high levels of efficiency, operability, and accessibility for effective utilisation. Accessibility is especially important for DH systems with a broad target group, like national health information websites or EHR systems. For wearables and IoT devices, operability and aesthetics are paramount given their function in data collection and active use.

Focusing excessively on a few characteristics could compromise the efficiency or effectiveness of the evaluation process. For a holistic usability evaluation, all relevant characteristics must be given due consideration. Overlooking certain aspects, such as accessibility or operability, could result in missed opportunities for improvement, thereby reducing the evaluation's efficiency and necessitating additional work. It is also important to note that the chosen testing method can influence which characteristics are most appropriate (Riihiaho, 2018). For example, methods that may slow user performance, like the thinking-aloud method, may benefit from emphasising characteristics other than performance. Furthermore, consider suitable usability methods and characteristics while keeping test users' capabilities or characteristics in mind. For example, thinking aloud methods, will not be appropriate for participants with limited cognitive abilities.

The diversity of the user base is a critical factor that can impact usability evaluations. Evaluations designed for generic users risk not addressing the specific needs of minority groups. This is particularly relevant for mhealth, wearables, or telehealth services, where accessibility for elderly or disabled users is essential. Different user groups may have varying preferences for usability characteristics, depending on their unique needs and experiences. For example, healthcare professionals, who often face heavy workloads and time-sensitive decisions, might tend to prioritise efficiency and effectiveness. This focus could help them navigate systems more quickly and accurately. On the other hand, patients managing chronic conditions, who may use systems frequently but could be unfamiliar with advanced technology, might find learnability and memorability more essential. Although these preferences are not universally true but could provide insight into how different groups approach and interact with technology in specific contexts or needs. In societies where the elderly are less familiar with technology due to limited exposure, the usability of DH platforms should emphasise characteristics like accessibility, learnability, and memorability. These diverse user requirements underscore the need for a user-centric approach to usability characteristics assessment.

Supplementary evaluation parameters, like privacy and ethical considerations, are also important, especially when dealing with vulnerable populations or health-related data (Pramesha Chandrasiri et al., 2019; Nittari et al., 2020). The socio-cultural context of the end users can significantly influence the usability of a system, emphasising the need for a comprehensive approach to usability evaluation. For example, users from different socio-cultural backgrounds might interpret medical information differently. A usability attribute that might be important in this context is "localization", which involves customising the application according to the cultural norms and language of the user, leading to improved understanding and adoption of the application.

Therefore, a more comprehensive and inclusive approach is required for usability evaluation, one that incorporates a wide spectrum of relevant characteristics, caters to a diverse range of user groups and balances the cost of resources and efforts. Furthermore, it must be carefully considered if the chosen methods reflect/assess the required characteristics or not. Such a balanced, all-inclusive strategy is also necessary to navigate and meet the opportunities and challenges posed by the rapidly progressing technological environment.

6.7. Limited focus on automated usability evaluation tools

In general, automation tools can minimise human error, thereby increasing the reliability of software testing (Khaliq et al., 2022). Despite these benefits, automated usability evaluation tools adoption in DH is still low. Understanding their use benefits and limitations, as well as examining potential barriers to adoption, can provide insights and potential future directions to enhance the effectiveness and efficiency of usability evaluations in DH.

A lack of methods/tools to evaluate usability in DH has been reported [S11, S31, S40, (Maramba et al., 2019)]. Our findings are inline, since around 4% of studies reported that any tools were used to (semi-) automate the usability evaluation process or specific parts. This does not necessarily mean that no tools are used, but that the investigated reports of usability studies put a stronger emphasis on methods and results. General purpose tools used, e.g., to perform quantitative or qualitative data analysis or to collect data, might hence be not included in the description of the study. However, even considering this aspect, the number of studies using tools and the variety of tools reported in the literature are relatively low. Our findings also show that tool support in usability evaluation is predominantly limited to user activity tracking and website accessibility checking tools. It is also worth noting that the most frequently mentioned tool (Morae) has been discontinued.

Automation tools can reduce the time required for data collection, processing, and analysis, speeding up the usability evaluation process. For instance, Morae type tools can record and annotate user interactions automatically, allowing for consistent measurement across different sessions or participants, eliminating the need for researchers to do it manually. Google Analytics can precisely track a range of user behaviours, from time spent on a page to click paths. Automated tools can provide an objective view of user interactions, avoiding observer bias. They record user activity in real-time, offering an unfiltered perspective of user behaviours. The emergence and adoption of Natural Language Processing (NLP) in studies represent a promising advancement towards more advanced and automated tools (Haggag, 2022; Haque and Rubya, 2022; Silva and Santos, 2023). NLP's ability to analyse user reviews and comments on DH applications can significantly enhance the effectiveness of usability evaluations by offering valuable qualitative feedback. Moreover, the inclusion of prototyping tools like CogTool and readability testing with Readable and other methods demonstrates the expanding landscape of usability evaluation. CogTool's predictive performance measure and readability scoring systems

can boost evaluation efficiency by identifying potential usability issues before conducting actual user testing. Early detection of these issues enables quicker iterations and improvements, ultimately enhancing the overall efficiency of the usability evaluation process.

The adoption of automation tools for usability evaluation in healthcare faces various challenges. Currently, Morae, Google Analytics, and TURF are widely used for user activity tracking during testing phases. However, these tools have a general focus and lack healthcare-specific adaptations. This limitation can become evident when applying them to specialised DH systems designed for specific user groups, like the elderly with Parkinson's disease, and cognitive or visual impairments. Moreover, while the Google Chrome SEE application can simulate the experience of colour-blind users, it fails to encompass the full range of visual impairments that elderly users may encounter. Healthcare regulations and patient privacy concerns may also demand additional features not present in many tools. Being mindful of these limitations enables researchers and practitioners to select the appropriate tools and avoid overlooking crucial elements unique to DH.

Interpreting large quantities of data from automation tools can also be challenging, often diverting attention from nuanced qualitative insights into user behaviours. This can result in analysis focusing primarily on numbers, neglecting valuable qualitative aspects (Hendren et al., 2023). It is vital to balance quantitative and qualitative analysis, incorporating user feedback, observations, and contextual information for a holistic interpretation of user behaviours. Furthermore, complementing different tools can allow for a mixed-method approach to usability evaluation. For example, natural language processing can be employed to analyse qualitative user feedback on DH applications, providing in-depth insights into user attitudes and experiences. Simultaneously, tools like Google Analytics can offer quantitative user engagement data, allowing for a more holistic evaluation of usability.

Researchers and usability experts may not be using many automation tools due to their unfamiliarity or lack of training with them. The ever-changing technological landscape presents a challenge, with the constant emergence of new tools and updates. Without adequate training and exposure, researchers may struggle to grasp and utilise such tools effectively in their studies. Advanced automation tools often come with substantial financial costs, which may also be a barrier for researchers running on limited budgets. The incorporation of automation tools into existing research processes can be a complex and disruptive task. Such integration requires careful planning and management, which could pose significant challenges, particularly in multidisciplinary research teams or for those with established or already running workflows.

The lack of peer support or an established precedent in the research community or organisation for using such usability tools can discourage researchers from adopting them. Many researchers and usability evaluators, especially new ones, may prefer to stick to tried-and-tested methods and may be reluctant to venture into unfamiliar technological territory without the backing of their peers or organisations. Other researchers reviewed more holistic tool support for usability evaluations and discovered several challenges, among others, the difficulty of interpreting test results, lack of actionable recommendations, or failure of tools to point out high-priority usability issues (Mvungi and Tossy, 2015; Namoun et al., 2021). These complications may also pose significant barriers to the integration of more advanced tool support.

Apart from the tools and approaches mentioned in the results section (Section 5.4), we also investigated additional tools and platforms available in the market to explore their potential use in specific situations. Notably, there are several tools or platforms for conducting remote moderated or unmoderated user research and usability evaluations. For example, *Hotjar*,[14] *Usabilla*,[15] *Qualaroo*,[16] *UserZoom (now part of UserTesting)*, *Loop11*,[17] and *TryMyUI*[18] are popular tools for conducting online surveys with real users, capturing user feedback for data collection and analysis, and delivering real-time results. *Lookback*[19] and *UserTesting*[20] provide a platform for moderated and unmoderated user research and usability evaluation interviews. *UserTesting* is supported with real-time video interviews and machine learning-driven dashboards and visualisations, which help in the discovery of insights to effectively and efficiently highlight usability concerns. *Hotjar* and *UserTesting* are two of the most popular tools to analyse and empathise with users.

Tools such as *Qualaroo*, *CrazyEgg*,[21] *Loop11*, *Optimizely*,[22] *TryMyUI*, *Lookback*, and *Inspectlet*[23] facilitate to conduct, monitor, and analyse usability evaluation using real-world tasks. These tools provide effective and efficient tracking and analysis of user activities, analytics, insights, and behaviours. These tools help to understand general usage patterns and identify areas of customer frustration and the most interacted content. *CrazyEgg* and *Inspectlet* can further add new dimensions to the Google Analytics insights. Furthermore, *Inspectlet* can provide funnel analytics to detect critical events across the user journey to effectively manage testing and enhance user experience. Additionally, *Qualaroo* offers AI-powered analytics and reporting that help anticipate what works and what does not. *CrazyEgg*, *Loop11*, *Optimizely*, and *Inspectlet* are suited for A/B testing to facilitate conversion optimisation. *Hotjar*, *CrazyEgg*, *Loop11*, *TryMyUI*, and *Inspectlet* record user sessions and provide various forms of heatmaps, scroll-maps, and click reports that assist in visualising the density of user engagement and placing content in the order of preference to consumers so that they can effectively and efficiently achieve their goals.

We were unable to discover many tools that perform automated usability evaluations. Few tools are available to access the readability and colour contrast of the website, such as *Juicy Studio*[24] and *Webfx's Readability Test Tool*.[25] *Dead-Link checker*[26] tool is used for checking dead/broken links. Furthermore, some accessibility tools are no longer available, such as *Check My Colours* and *Browsershots* tools for checking website colour choices and accessibility, respectively. It is important to note that a complete focus on accessibility in usability tools is ineffective. We agree with Jakob Nielsen that tools must adopt various usability perspectives to make (automated) usability evaluation more effective and efficient (Nielsen, 2005). Other tools that evaluate other perspectives of usability or assist other activities can be quite beneficial. Such as *happyscribe*[27] is a transcription service that uses AI technology and is 85% accurate, which can be quite efficient in transcribing usability interviews. Furthermore, a variety of usability evaluation tools were researched and developed, but were rarely made public (Dingli and Mifsud, 2011; Kaasila et al., 2012; Geng and Tian, 2014; Grigera et al., 2017).

Using tools to automatically track user actions or system events and analyse usability data can also help increase test effectiveness and efficiency, and deliver more reliable results (Bastien, 2010). Although some tools are available, there are still some research difficulties and technological challenges (Mvungi and Tossy, 2015; Namoun et al., 2021). Usability evaluation involves unpredictable and gestural inputs, which can only be given by humans. So, developing an automated

[14] https://www.hotjar.com/
[15] https://www.getfeedback.com/resources/cx/usabilla/
[16] https://qualaroo.com/
[17] https://www.loop11.com/
[18] https://www.trymyui.com/
[19] https://www.lookback.com/
[20] https://www.usertesting.com/
[21] https://www.crazyegg.com
[22] https://www.optimizely.com/
[23] https://www.inspectlet.com/
[24] https://juicystudio.com/services.php
[25] https://www.webfx.com/tools/read-able/
[26] https://www.deadlinkchecker.com/
[27] https://www.happyscribe.com/transcription

solution for such kinds of evaluations is challenging. Human interaction, psychology, and behaviour can be difficult to simulate with scripts. However, some activities can be automated in the design and analysis stages. Companies may also benefit from automated usability evaluation frameworks since they conduct such tests on a regular and large scale for a variety of products at various development stages. Automated usability evaluation can benefit in resource savings, cost reduction, increased overall test coverage, performance testing, etc. Despite their benefits, automated usability evaluation tools must be used alongside traditional usability evaluation methods.

Finding or developing appropriate (automated) usability evaluation tools can be challenging, but it can improve test effectiveness and efficiency, and deliver more reliable results. Furthermore, automating other aspects of usability evaluation can be beneficial, such as test case or script generation, interview or think-aloud transcription and analysis, etc. There seems potential for tools that can go beyond conventional metrics and assess more nuanced aspects of usability such as user group-specific interactions with systems, which can significantly impact a user's experience with DH systems.

Meanwhile, while automation tools have evident benefits in usability evaluation, their adoption is restricted by several limitations and barriers, which may explain their infrequent use in DH usability research. Increased awareness and understanding of such tools, and the development of cost-effective, user-centric, and healthcare-specific tools may promote their adoption. Ultimately, improving the effectiveness and efficiency of usability evaluations in the DH sector. Moreover, the research community should foster an environment that encourages innovation and the exploration of novel methodologies. Future studies could further investigate such barriers and develop strategies to promote the adoption of automation tools in DH usability evaluations. The use of tools for automatically tracking and analysing user behaviour, designing and analysing other aspects of usability evaluation (such as transcription), and automatic usability evaluation (heuristics- or guidelines-based UI and accessibility testing) has shown to be effective. Further research is required to investigate the effectiveness and efficiency of using AI or ML approaches to automate (some aspects of) usability evaluation.

6.8. Low participation of patients

The efficacy of usability evaluation depends largely on the chosen inquiry and testing methods, with each offering unique advantages, such as gaining customer feedback, fostering inclusivity, ensuring precision, and facilitating in-depth insights and discussions (Lyzara et al., 2019; Aziz et al., 2021). However, their effectiveness and efficiency can be undermined by difficulties in participant recruitment, especially when target populations are characterised by specific traits such as age group, health conditions, and geographical location. The challenge is even worse when DH apps are tailored for niche user groups. For example, it is challenging to recruit a large number of cancer patients for usability evaluation (Sygna et al., 2015; Jacobs et al., 2019; Mudaranthakam et al., 2022). The same is true for patients with mental health (Patel et al., 2003; Liu et al., 2018), diabetes (Nguyen et al., 2014; Kanuch et al., 2016), disability, ageing and well-being problems (Van der Geest, 2006). Further key challenges include patient availability, comfortability, capability, well-being, awareness, social, cultural and ethical issues (De Bleecker and Okoroji, 2018; Kadam et al., 2016).

There is a widely accepted principle in healthcare, "Nothing about me, without me", which emphasises the central role of patients in all decisions related to their health (Delbanco et al., 2001). Their inputs, often ignored, are crucial in shaping the acceptability, usability, and effectiveness of healthcare systems (Katsulis et al., 2016). Despite the crucial role of these studies, results indicate that only 42% of these studies include patients as participants, despite being the primary users of such systems. Meanwhile, HCPs and general users, who provide clinical relevance and lay perspectives respectively, are disproportionately represented. Healthy participants ("general users") are however also used as substitutes for actual patients; this, however, seems to happen significantly less often than previous research on usability evaluation for serious games in healthcare suggests (Yanez-Gomez et al., 2017). Although valuable, the insights of HCPs and general users cannot supplant the input of patients, who offer an indispensable perspective due to their direct experience with the healthcare condition in question. The under-representation of informal caregivers, vital for managing chronic health conditions, signals a missed opportunity in creating comprehensive DH systems (Schaller et al., 2016). Ignoring caregiver input may lead to potential gaps in addressing specific needs and challenges that only they, given their hands-on care experience, can identify and articulate.

In general, several barriers to patient participation in usability studies exist. Logistical challenges and rigorous usability evaluation sessions can deter older individuals or those with serious health conditions. Conventional inquiry methods may prove unsuitable for some demographics, resulting in incomplete or misleading usability assessments (Choi et al., 2022). For instance, patients with Alzheimer's or dementia may struggle to communicate their experiences effectively due to cognitive impairments. Patients with serious illnesses, like cancer, also face distinctive challenges in participating in inquiry-based evaluations. The stress of intensive treatments or the weight of emotional or ethical factors can limit their willingness or capability to engage in long interviews or detailed surveys. As their healthcare demands often take follow-ups and medical checkups, it becomes challenging to recruit an adequate sample size or obtain thorough feedback using traditional inquiry methods. Additionally, the digital divide further impedes participation, especially for those in rural areas or from lower socio-economic backgrounds. The low participation of patients in usability studies is a pervasive issue, and the context of DH platforms is no exception. Some individuals with anxiety disorders or depression may feel hesitant to use some platforms for self-managing their mental health (Santoso et al., 2021) and so be less interested to participate in evaluations. External factors like the COVID-19 pandemic might also impose restrictions on the participant's ability to partake in studies (Egle et al., 2022). Patients may hesitate to engage with DH systems due to concerns like data breaches, time constraints, fear of stigmatisation, privacy issues, ethical considerations, sociocultural factors, or doubts about the usefulness of DH systems or apps (Gasteiger et al., 2021; Lei-Juan et al., 2016; Egle et al., 2022).

Obtaining informed consent poses another significant challenge. Especially in the case of cognitively impaired patients, ensuring that they fully understand the implications, benefits, and potential risks of the study is crucial. Without this understanding, they might agree to participate without truly comprehending the consequences. Due to some personal instances of research misconduct or systematic prejudice, patients may lack faith in such research (Gasteiger et al., 2021). Additionally, some people may limit or hide opinions on the system's usability in order to avoid being perceived as technologically illiterate or inexperienced (Lei-Juan et al., 2016). Socio-cultural differences can also influence social interactions and communication methods, with some cultures placing a high value on hierarchy and status between test users and moderators, while others emphasise individualistic values and encourage direct criticism, reflecting a culture that fosters independent judgment (Schrier, 1992; Yeo, 2000). Furthermore, communication approaches vary widely across cultures, with low-context societies favouring explicit, direct communication and readily offering critique even to unfamiliar observers or moderators. Conversely, high-context cultures lean towards nuanced, indirect communication, where comfort in critique typically arises within established relationships. It is critical to consider alternative usability evaluation methods that better address such user groups' limitations. Remote usability evaluation or observational methods could provide valuable insights into real-time

behaviours in their natural environments (Ahmad et al., 2021; Breeman et al., 2021; Riihiaho, 2018). Early design decisions can also benefit from prototyping, especially with older adults and diverse collaboration incorporating various stakeholders (Ahmad et al., 2021; Breeman et al., 2021).

The under-representation of intended user groups can lead to the development of DH systems that inadequately cater to their needs, preferences, and abilities. The limited sample size also brings into question the reliability of such studies and the efficient allocation of resources spent on refining the systems. Further solutions need to be explored to increase patient participation. These strategies should consider patients' physical limitations, cognitive capacities, and privacy concerns. Despite the challenges, it is critical to remember that the voices of patients remain a fundamental aspect in designing inclusive, effective, efficient, and truly user-centred DH systems.

6.9. Recommendations

Our SLR findings have recommendations for both researchers planning research on usability evaluation of DH apps and practitioners working in DH app development and/or specialised testing companies that are experiencing such issues and want to effectively integrate usability evaluation methods into the project lifecycle. This section presents a summary of the main recommendations.

The identified findings lead us to the following recommendations for usability evaluation practitioners:

- *The need for a complementary approach:* Since inquiry methods are used in 29% of studies exclusively, and questionnaires based evaluation dominates with 75% of the overall inquiry-based studies, usability experts should consider diversifying their approach. Combining different methods could offer a more comprehensive evaluation and avoid the discussed limitations.
- *The importance of automation tools:* With only 4% of studies using automation or semi-automation tools, there is a need for usability experts to develop, explore, implement, and promote tools or methods in DH usability evaluations. Automation tools may improve efficiency and offset some resource-intensiveness of the usability evaluation process.
- *Addressing the low participation of target users:* With an average of 49% of studies engaging participants from their targeted user groups, usability experts should explore strategies to increase their participation.
- *Emphasis on under-evaluated usability characteristics:* Areas like operability, usefulness, memorability, and accessibility are less evaluated. Usability experts should work towards incorporating these aspects into their usability evaluations, aligning the focus with the actual needs and preferences of different user groups and DH systems.

From a research perspective, we conclude that the following directions might be important to better support usability evaluation in DH:

- *Tailoring methods:* Investigate how to improve methods for prevalent DH user groups. For example, designing tailored questionnaires, tasks and scenarios, or interviews to conduct usability evaluation with the elderly. Further research is required to discover domain and user group specific guidelines for writing usability tasks and scenarios.
- *Empirical investigation:* Furthermore, substantial empirical evidence concerning the effectiveness and efficiency of tailored and already applied methods in practice is required. For example, evaluate the efforts spent (e.g., number of hours) in usability evaluation of a specific mhealth app with elderly participants, comparing customised usability methods to traditional approaches to determine efficiency and effectiveness.
- *Investigating barriers to the adoption of automation tools:* The low adoption rate of automated usability evaluation tools in DH presents an area for further research. Understanding the barriers to adoption, including healthcare-specific adaptations and costs, could foster innovation and improvement in this area.
- *Exploring AI potential in usability evaluation:* Further research is required to investigate the feasibility, effectiveness, and efficiency of using artificial intelligence or machine learning approaches to automate (some aspects of) usability evaluation.
- *Understanding and enhancing patient participation:* The low participation of patients is a significant challenge. Researchers should investigate the barriers, including health condition-specific factors, and develop strategies to enhance patient engagement. This can lead to more relevant and effective DH platforms.

7. Validity and reliability

Although we designed and structured the study to ensure accurate and unbiased findings, there are potential threats to the validity and reliability of the study.

7.1. External validity

The degree to which the results of a study may be generalised to a larger population is referred to as external validity (Mitchell and Jolley, 2010). In our context, this is the question to which degree the findings distilled from the investigated studies can be generalised to usability evaluation in DH in general.

Inaccurate publication selection: One threat to the external validity of every literature review is the possibility of including irrelevant or missing relevant articles. In Section 4, we describe a systematic approach to study selection. All articles were screened independently by both authors, and disagreements related to the relevance of articles were discussed and decided upon jointly based on more detailed reading if necessary. We are confident to have excluded irrelevant articles as far as possible by doing so.

The risk of missing relevant articles was mitigated by choosing five highly relevant databases for article retrieval, covering all major relevant publication venues. Still, there are chances that we may have missed some relevant literature. This risk could have been reduced by applying snowballing techniques (Wohlin, 2014). We considered this too time-consuming for this SLR given the number of articles retrieved from the databases. Following a forward snowballing pilot run, we discovered that the test articles' dataset, published between 2021 and April 2022, was already present in the sample of forward snowballing articles published between 2019 and 2020. Therefore, we decided not to continue exhaustive snowballing, especially considering the amount of effort required versus the potential outcome. It, therefore, seemed tolerable to us to omit snowballing.

We excluded grey literature and unpublished results since determining quality was challenging. However, there is a risk of overlooking insights from this type of articles, for example, usability evaluation studies of DH products evaluated and reported by companies. In order to mitigate this risk, we also plan to survey practitioners and organisations who develop and/or evaluate such systems in the near future, as this might provide more detailed insights into industrial practices than a general grey literature review.

There is a chance of having faults in the Microsoft Excel script that we used for automating the article selection process, for example, to remove duplicate articles contained in multiple databases. We mitigated this risk by thoroughly testing the script and cross-checking results with those obtained from similar functions in the JabRef tool (Basak, 2014).

Sampling: A SLR with a lower sample size will likely produce fewer generalisable conclusions. We believe that the comprehensive analysis of 610 publications produced more generalised results, which will increase the work's external validity. Furthermore, there seems a risk of missing publications before 2016. However, we attempted to limit this risk by exploring and reporting knowledge from the literature studies (see Section 3) that covered past years and key insights.

7.2. Construct validity

Construct validity is the extent to which the measurements taken are related to the phenomena being examined (Mitchell and Jolley, 2010). The phenomenon studied were characteristics of usability evaluation studies in DH and possible issues related to effectiveness and efficiency. The measurement taken is the categorising of studied articles into several dimensions considered relevant, based on which we draw conclusions about potential issues. There is a risk of incorrectly categorising them, but we tried to minimise this risk by employing multi-stage extraction, analysis, reviewing, and refining activities (also explained in Section 4 and Section 7.1).

7.3. Internal validity

Internal validity is defined as the extent to which a study establishes that an effect is caused by a certain factor, and by this factor alone (Mitchell and Jolley, 2010). We do not discuss any risk to internal validity because this SLR does not aim to establish a cause–effect relationship of any kind.

7.4. Reliability

Reliability refers to consistency in the data gathering and analysis in the sense that being consistent, means that repeating both should lead to the same results.

Data extraction inaccuracy and research bias: We may not have obtained accurate results due to the subjective nature of the data extraction process. We tried to mitigate this risk by developing a comprehensive data extraction form. Moreover, during this process, mutual validation of extracted data among the researchers extracting the data was performed to minimise this risk.

Manual data extraction might have resulted in unintentional bias. More importantly, personal experience and knowledge of usability evaluation had the potential to lead to biased results. Data extraction was done iteratively to address this threat. First, all possible data relevant to the data items were elicited. Then, the extracted data was categorised according to the common themes, and finally, the data was refined and classified according to well-defined standards, guidelines, literature, and information related to the topics.

Lastly, the tendency of a scientific study to favour the interests of the funding partner is referred to as *funding bias*. However, in our study, the sponsor had no personal interest or any direct involvement in the design, analysis, or publishing. We had complete autonomy to disclose both positive and negative results.

We furthermore provide a comprehensive replication package[3] that enables other researchers to validate our results and as a basis to improve the validity and reliability of results in the future.

8. Conclusion

Usability is one of the key factors for the adoption and success of DH products. Only with accessibility, usefulness, learnability, etc. for people in mind who will need those products the most, DH will continue to contribute to sustainable healthcare systems. Usability evaluation is an important piece in the puzzle of ensuring usability. However, concerns have been raised about the effectiveness and efficiency of how this evaluation is done in DH.

This systematic literature review provides a comprehensive study of the usability evaluation of DH technologies in recent years. We reviewed 610 articles and extracted information related to the usability evaluation methods applied and the contexts they were applied in. The results paint a pretty homogeneous picture of the landscape of DH usability evaluation. Inquiries and task and scenario-based testing are prevalent methods, as well as questionnaires for data collection, across all types of applications and targeted user groups, including healthcare professionals and patients with diverse conditions and diseases. A lack of usability evaluation for targeting certain usability characteristics was noted such as accessibility, memorability, and operability. These overlooked characteristics may hinder the full potential of DH products. The use of tools for automating (steps of) usability evaluation was hardly used or documented. This represents a significant area for growth and efficiency improvement.

Through a detailed examination of the prevalent themes that emerged from our study, we have deepened our understanding of the effectiveness and efficiency of usability evaluation in DH. This exploration has uncovered the strengths and weaknesses inherent in current practices, highlighting potential areas that need further investigation and improvement. The knowledge and perspectives gained from this study provide a foundation upon which researchers, practitioners, and other stakeholders can build to drive meaningful advancements in DH.

As user groups of DH products are common in which reduced cognitive abilities might be common, we recommend carefully selecting evaluation methods adapted to specific, relevant user groups, e.g., older adults or people with dementia. Methods like think-aloud in these scenarios might not only be ineffective but also frustrating or even humiliating for participants. Methods should furthermore be selected with the specific usability characteristic of interest in mind and the use of inspections and other methods, in particular when recruiting participants from specific user groups is difficult, should be considered as complementary.

The results also suggest possible routes for future research. Refining existing methods, such as questionnaires, interviews, and tasks and scenarios, etc. for DH or/and typical user groups and technologies appears desirable to increase usability evaluation effectiveness and efficiency. More sophisticated tool support, for example, based on artificial intelligence, seems further desirable to support this.

CRediT authorship contribution statement

Bilal Maqbool: Conceptualization, Methodology, Validation, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing. **Sebastian Herold:** Conceptualization, Methodology, Validation, Investigation, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data is in replication package and linked/cited in paper. SLR repository: https://doi.org/10.5281/zenodo.8218076.

Acknowledgements

This work was funded, in parts, by Region Värmland through the DHINO project, Sweden (Grant: RUN/220266) and Vinnova through the DigitalWell Arena (DWA) project, Sweden (Grant: 2018-03025).

References

Adibi, S., 2015. Mobile Health: A Technology Road Map. Vol. 5, Springer.

Aghazadeh, S., Pirnejad, H., Aliev, A., Moradkhani, A., 2015. Evaluating the effect of software quality characteristics on health care quality indicators.

Aghazadeh, S., Pirnejad, H., Moradkhani, A., Aliev, A., 2014. The impact of software quality characteristics on healthcare outcome: a literature review. Integr. Inf. Technol. Manag. Qual. Care 201–204.

Ahmad, B., Beecham, S., Richardson, I., 2021. The case of golden jubilants: using a prototype to support healthcare technology research. In: 2021 IEEE/ACM 3rd International Workshop on Software Engineering for Healthcare. SEH, IEEE, pp. 68–71.

Aitken, M., Lye, J., 2015. Patient adoption of mHealth, report by the IMS institute for healthcare informatics, 11. pp. 1–57, Retrieved May 2023.
Alcaraz-Quiles, F.J., Urquia-Grande, E., Muñoz-Colomina, C.I., Rautiainen, A., 2018. E-government implementation: transparency, accessibility and usability of government websites. In: International E-government development: Policy, implementation and best practice. Springer, pp. 291–306.
Andersson, E., 2017. Cognitive Workload, Game Experience, and Intrinsic Motivation (Dissertation).
Angelova, B., Zekiri, J., 2011. Measuring customer satisfaction with service quality using American customer satisfaction model (ACSI model). Int. J. Acad. Res. Bus. Soc. Sci. 1 (3), 232.
Ansaar, M.Z., Hussain, J., Bang, J., Lee, S., Shin, K.Y., Woo, K.Y., 2020. The mHealth applications usability evaluation review. In: 2020 International Conference on Information Networking. ICOIN, IEEE, pp. 70–73.
Anuar, N., Idris, A.S., Misron, M.M., Norbaya, S., 2014. Evaluation of e-government (PBT) portal based on usability design. In: 2014 International Conference on Computational Science and Technology. ICCST, IEEE, pp. 1–7.
Au, F.T., Baker, S., Warren, I., Dobbie, G., 2008. Automated usability testing framework. In: Proceedings of the Ninth Conference on Australasian User Interface, Vol. 76. pp. 55–64.
Aziz, N.S., Sulaiman, N.S., Hassan, W.N.I.T.M., Zakaria, N.L., Yaacob, A., 2021. A review of website measurement for website usability evaluation. In: Journal of Physics: Conference Series. Vol. 1, IOP Publishing, pp. 1–10.
Baharuddin, R., Singh, D., Razali, R., 2013. Usability dimensions for mobile applications-a review. Res. J. Appl. Sci. Eng. Technol. 5 (6), 2225–2231.
Barnum, C.M., 2010. Usability Testing Essentials. Elsevier.
Barnum, C.M., 2020. Usability Testing Essentials: Ready, Set... Test!. Morgan Kaufmann.
Basak, S.K., 2014. Reference management software: a comparative analysis of JabRef and RefWorks. Language 1.
Bastien, J.C., 2010. Usability testing: a review of some methodological and technical aspects of the method. Int. J. Med. Inf. 79 (4), e18–e23.
Batsis, J., Zagaria, A., Kotz, D., Bartels, S., Boateng, G., Proctor, P., Carpenter-Song, E., 2018. Usability evaluation for the Amulet Wearable Device in rural older adults with obesity. Gerontechnology 17 (3), 151–159.
Bochicchio, M.A., Vaira, L., Mortara, A., De Maria, R., 2021. Which usability assessment for digital therapeutics and patient support programs? In: 2021 IEEE International Conference on Digital Health. ICDH, IEEE, pp. 276–282.
Bolívar, M.P.R., 2014. Measuring E-Government Efficiency: The Opinions of Public Administrators and Other Stakeholders. Springer.
Borenstein, M., Hedges, L.V., Higgins, J.P., Rothstein, H.R., 2021. Introduction to Meta-Analysis. John Wiley & Sons.
Bowman, S., 2013. Impact of electronic health record systems on information integrity: quality and safety implications. Perspect. Health Inf. Manag. 10 (Fall).
Branch, J.L., 2000a. Investigating the information-seeking processes of adolescents: The value of using think alouds and think afters. Libr. Inf. Sci. Res. 22 (4), 371–392.
Branch, J.L., 2000b. The trouble with think alouds: Generating data using concurrent verbal protocols. In: Proceedings of the Annual Conference of CAIS/Actes du congrès annuel de l'ACSI.
Braun, V., Clarke, V., 2006. Using thematic analysis in psychology. Qual. Res. Psychol. 3 (2), 77–101.
Brédart, A., Marrel, A., Abetz-Webb, L., Lasch, K., Acquadro, C., 2014. Interviewing to develop Patient-Reported Outcome (PRO) measures for clinical research: eliciting patients' experience. Health Qual. Life Outcomes 12 (1), 1–10.
Breeman, L.D., Keesman, M., Atsma, D.E., Chavannes, N.H., Janssen, V., van Gemert-Pijnen, L., Kemps, H., Kraaij, W., Rauwers, F., Reijnders, T., et al., 2021. A multi-stakeholder approach to ehealth development: Promoting sustained healthy living among cardiovascular patients. Int. J. Med. Inf. 147, 104364.
Brooke, J., 2013. SUS: a retrospective. J. Usability Stud. 8 (2), 29–40.
Bryman, A., 2006. Integrating quantitative and qualitative research: how is it done? Qual. Res. 6 (1), 97–113.
Buie, E., Murray, D., 2012. Usability in Government Systems: User Experience Design for Citizens and Public Servants. Elsevier.
Chen, X., Vorvoreanu, M., Madhavan, K., 2014. Mining social media data for understanding students' learning experiences. IEEE Trans. Learn. Technol. 7 (3), 246–259.
Choi, K.-S., Chan, S.-H., Ho, C.-L., Matejak, M., 2022. Development of a healthcare information system for community care of older adults and evaluation of its acceptance and usability. Digit. Health 8.
Cipresso, P., Serino, S., Villani, D., Repetto, C., Sellitti, L., Albani, G., Mauro, A., Gaggioli, A., Riva, G., 2012. Is your phone so smart to affect your state? An exploratory study based on psychophysiological measures. Neurocomputing 84, 23–30.
Clemmensen, T., Hertzum, M., Hornbæk, K., Shi, Q., Yammiyavar, P., 2009. Cultural cognition in usability evaluation. Interact. Comput. 21 (3), 212–220.
Coates, W.C., Jordan, J., Clarke, S.O., 2021. A practical guide for conducting qualitative research in medical education: part 2—coding and thematic analysis. AEM Educ. Train. 5 (4), e10645.
Commission, E., 2002. Quality criteria for health related websites.
Commission, E., 2006. Connected health: Quality and safety for European citizens.
Connelly, Y., Lotan, R., Brzezinski Sinai, Y., Rolls, D., Beker, A., Abensour, E., Neudorfer, O., Stocki, D., 2022. Implementation of a personalized digital app for pediatric preanesthesia evaluation and education: ongoing usability analysis and dynamic improvement scheme. JMIR Form. Res. 6 (5), e34129.
Couper, M.P., 2000. Web surveys: A review of issues and approaches. Public Opin. Q. 64 (4), 464–494.
Coursaris, C.K., Kim, D.J., 2011. A meta-analytical review of empirical mobile usability studies. J. Usability Stud. 6 (3), 117–171.
Davenport, T., Kalakota, R., 2019. The potential for artificial intelligence in healthcare. Future Healthc. J. 6 (2), 94.
De Bleecker, I., Okoroji, R., 2018. Remote Usability Testing: Actionable Insights in User Behavior Across Geographies and Time Zones. Packt Publishing Ltd.
De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., Kumar, M., 2016. Discovering shifts to suicidal ideation from mental health content in social media. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. pp. 2098–2110.
Delbanco, T., Berwick, D.M., Boufford, J.I., Ollenschläger, G., Plamping, D., Rockefeller, R.G., 2001. Healthcare in a land called PeoplePower: nothing about me without me. Health Expect. 4 (3), 144–150.
Delikostidis, I., 2007. Methods and Techniques for Field-Based Usability Testing of Mobile Geo-Applications (thesis). International Institute for Geo-Information Science & Earth Observation (ITC), Enschede.
Della Mea, V., 2001. What is e-health (2): The death of telemedicine? J. Med. Internet Res. 3 (2), e22.
Dillman, D.A., Smyth, J.D., Christian, L.M., 2014. Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method. John Wiley & Sons.
Dingli, A., Mifsud, J., 2011. Useful: A framework to mainstream web site usability through automated evaluation.
DIS, I., 2009. 9241-210: 2010. Ergonomics of human system interaction-Part 210.
Dix, A., Dix, A.J., Finlay, J., Abowd, G.D., Beale, R., 2003. Human-Computer Interaction. Pearson Education.
dos Santos Grecco, C.H., dos Santos, I.J.A.L., Farias, M.S., de Farias, L.P., Vianna Filho, A.M., 2021. A fuzzy method for the usability evaluation of nuclear medical equipment. Braz. J. Radiat. Sci. 9 (2A (Suppl.)).
Düking, P., Achtzehn, S., Holmberg, H.-C., Sperlich, B., 2018. Integrated framework of load monitoring by a combination of smartphone applications, wearables and point-of-care testing provides feedback that allows individual responsive adjustments to activities of daily living. Sensors 18 (5), 1632.
Dumas, J.S., Salzman, M.C., 2006. Usability assessment methods. Rev. Hum. Factors Ergon. 2 (1), 109–140.
Dyba, T., Dingsoyr, T., Hanssen, G.K., 2007. Applying systematic reviews to diverse study types: An experience report. In: First International Symposium on Empirical Software Engineering and Measurement. ESEM 2007, IEEE, pp. 225–234.
Earnshaw, R., Guedj, R., Van Dam, A., Vince, J., 2001. Frontiers of Human-Centered Computing, Online Communities and Virtual Environments. Springer Science & Business Media.
Ebad, S.A., 2020. Healthcare software design and implementation—A project failure case. Softw. - Pract. Exp. 50 (7), 1258–1276.
Edwards, P.J., Moloney, K.P., Jacko, J.A., Sainfort, F., 2008. Evaluating usability of a commercial electronic health record: A case study. Int. J. Hum.-Comput. Stud. 66 (10), 718–728.
Egle, F., Kluge, F., Schoene, D., Becker, L., Koelewijn, A.D., 2022. Development of an inertial sensor-based exergame for combined cognitive and physical training. In: 2022 IEEE-EMBS International Conference on Wearable and Implantable Body Sensor Networks. BSN, IEEE, pp. 1–4.
Engdahl, K., Leclerc, K., Loring, B., 2009. Conducting field studies with older users: Lessons for recruiting and testing older users. User Exp. Mag. 8 (1).
Ericsson, K.A., Simon, H.A., 1984. Protocol Analysis: Verbal Reports As Data. The MIT Press.
Eysenbach, G., 2001. What is e-health? J. Med. Internet Res. 3 (2), e20.
Fadahunsi, K.P., Akinlua, J.T., O'Connor, S., Wark, P.A., Gallagher, J., Carroll, C., Majeed, A., O'Donoghue, J., 2019. Protocol for a systematic review and qualitative synthesis of information quality frameworks in eHealth. BMJ Open 9 (3), e024722.
Fadahunsi, K.P., O'Connor, S., Akinlua, J.T., Wark, P.A., Gallagher, J., Carroll, C., Car, J., Majeed, A., O'Donoghue, J., 2021. Information quality frameworks for digital health technologies: systematic review. J. Med. Internet Res. 23 (5), e23479.
Fan, M., Li, Y., Truong, K.N., 2020a. Automatic detection of usability problem encounters in think-aloud sessions. ACM Trans. Interact. Intell. Syst. 10 (2), 1–24.
Fan, M., Shi, S., Truong, K.N., 2020b. Practices and challenges of using think-aloud protocols in industry: An international survey. J. Usability Stud. 15 (2).
Franz, R.L., Neves, B.B., Epp, C.D., Baecker, R., Wobbrock, J.O., 2019. Why and how think-alouds with older adults fail: Recommendations from a study and expert interviews. In: Perspectives on Human-Computer Interaction Research with Older People. Springer, pp. 217–235.
Fritz, M., Berger, P.D., 2015. Improving the User Experience Through Practical Data Analytics: Gain Meaningful Insight and Increase Your Bottom Line. Morgan Kaufmann.
Gasteiger, N., Ahn, H.S., Gasteiger, C., Lee, C., Lim, J., Fok, C., Macdonald, B.A., Kim, G.H., Broadbent, E., 2021. Robot-delivered cognitive stimulation games for older adults: Usability and acceptability evaluation. ACM Trans. Hum.-Robot Interact. 10 (4), 1–18.

Gelman, A., Shalizi, C.R., 2013. Philosophy and the practice of Bayesian statistics. Br. J. Math. Stat. Psychol. 66 (1), 8–38.
Geng, R., Tian, J., 2014. Improving web navigation usability by comparing actual and anticipated usage. IEEE Trans. Hum.-Mach. Syst. 45 (1), 84–94.
Genise, P., 2002. Usability Evaluation: Methods and Techniques: Version 2.0. University of Texas.
Genov, A., Keavney, M., Zazelenchuk, T., 2009. Usability testing with real data. J. Usability Stud. 4 (2), 85–92.
Ghasemifard, N., Shamsi, M., Kenari, A.R.R., Ahmadi, V., 2015. A new view at usability test methods of interfaces for human computer interaction. Glob. J. Comput. Sci. Technol..
Goertzen, M.J., 2017. Introduction to quantitative research and data. Libr. Technol. Rep. 53 (4), 12–18.
Grigera, J., Garrido, A., Rivero, J.M., Rossi, G., 2017. Automatic detection of usability smells in web applications. Int. J. Hum.-Comput. Stud. 97, 129–148.
Güss, C.D., 2018. What is going through your mind? Thinking aloud as a method in cross-cultural psychology. Front. Psychol. 1292.
Haas, P., Sembritzki, J., 2006. The European eHealth initiative–objectives and solutions. In: European Conference on eHealth 2006. Gesellschaft für Informatik eV, pp. 13–24.
Haggag, O., 2022. Better identifying and addressing diverse issues in mhealth and emerging apps using user reviews. In: Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering. pp. 329–335.
Haque, M.R., Rubya, S., 2022. "For an app supposed to make its users feel better, it sure is a joke"-an analysis of user reviews of mobile mental health applications. Proc. ACM Hum.-Comput. Interact. 6 (CSCW2), 1–29.
Harland, N., Holey, E., 2011. Including open-ended questions in quantitative questionnaires—theory and practice. Int. J. Therapy Rehabil. 18 (9), 482–486.
Hart, S.G., 2006. NASA-task load index (NASA-TLX); 20 years later. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 50, No. 9. Sage publications Sage CA, Los Angeles, CA, pp. 904–908.
Hart, S.G., Staveland, L.E., 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: Advances in Psychology. Vol. 52, Elsevier, pp. 139–183.
Haverinen, J., Keränen, N., Falkenbach, P., Maijala, A., Kolehmainen, T., Reponen, J., 2019. Digi-HTA: Health technology assessment framework for digital healthcare services. Finn. J. eHealth eWelfare 11 (4), 326–341.
Hendren, K., Newcomer, K., Pandey, S.K., Smith, M., Sumner, N., 2023. How qualitative research methods can be leveraged to strengthen mixed methods research in public policy and public administration? Public Adm. Rev. 83 (3), 468–485.
Hodrien, A., Fernando, T., et al., 2021. A review of post-study and post-task subjective questionnaires to guide assessment of system usability. J. Usability Stud. 16 (3), 203–232.
Hoffman, S., 2014. Medical big data and big data quality problems. Conn. Ins. LJ 21, 289.
Hollingsed, T., Novick, D.G., 2007. Usability inspection methods after 15 years of research and practice. In: Proceedings of the 25th Annual ACM International Conference on Design of Communication. pp. 249–255.
Holton, J.A., 2007. The coding process and its challenges. Sage Handb. Grounded Theory 3, 265–289.
Hong, J.S., Wasden, C., Han, D.H., 2021. Introduction of digital therapeutics. Comput. Methods Programs Biomed. 106319.
Hornbæk, K., Law, E.L.-C., 2007. Meta-analysis of correlations among usability measures. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 617–626.
Hossain, A., Kashem, M.A., Sultana, S., 2013. Enhancing software quality using agile techniques. IOSR J. Comput. Eng. 10 (2), 87–93.
Househ, M.S., Shubair, M.M., Yunus, F., Jamal, A., Aldossari, B., 2015. The use of an adapted health IT usability evaluation model (health-ITUEM) for evaluating consumer reported ratings of diabetes mHealth applications: implications for diabetes care and management. Acta Inform. Med. 23 (5), 290.
Hsu, J.M., 2015. Digital health technology and trauma: development of an app to standardize care. Anz J. Surg. 85 (4), 235–239.
Huryk, L.A., 2010. Factors influencing nurses' attitudes towards healthcare information technology. J. Nurs. Manag. 18 (5), 606–612.
ISO, 2011. ISO/IEC 25010:2011. Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models. https://www.iso.org/obp/ui/#iso:std:iso-iec:25010:ed-1:v1:en. (Accessed 9 October 2023).
ISO, 2018. ISO 9241-11:2018. Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts. https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en. (Accessed 9 October 2023).
Ivory, M.Y., 2013. Automated Web Site Evaluation: Researchers' and Practioners' Perspectives. Vol. 4, Springer Science & Business Media.
Jacobs, M., Hopkins, J., Mumber, M., Mynatt, E., 2019. Usability evaluation of an adaptive information recommendation system for breast cancer patients. In: AMIA Annual Symposium Proceedings. Vol. 2019, American Medical Informatics Association, p. 494.
Jaspers, M.W., Peute, L.W., Lauteslager, A., Bakker, P.J., 2008. Pre-post evaluation of physicians' satisfaction with a redesigned electronic medical record system. Stud. Health Technol. Inf. 136, 303.
Johanssen, J.O., Reimer, L.M., Bruegge, B., 2019. Continuous thinking aloud. In: 2019 IEEE/ACM Joint 4th International Workshop on Rapid Continuous Software Engineering and 1st International Workshop on Data-Driven Decisions, Experimentation and Evolution. RCoSE/DDrEE, IEEE, pp. 12–15.
Kaasila, J., Ferreira, D., Kostakos, V., Ojala, T., 2012. Testdroid: automated remote UI testing on android. In: Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia. pp. 1–4.
Kadam, R.A., Borde, S.U., Madas, S.A., Salvi, S.S., Limaye, S.S., 2016. Challenges in recruitment and retention of clinical trial subjects. Perspect. Clin. Res. 7 (3), 137.
Kanuch, S.W., Cassidy, K.A., Dawson, N.V., Athey, M., Fuentes-Casiano, E., Sajatovic, M., 2016. Recruiting and retaining individuals with serious mental illness and diabetes in clinical research: lessons learned from a randomized, controlled trial. J. Health Dispar. Res. Pract. 9 (3), 115.
Karim, N., Rybarczyk, M.M., Jacquet, G.A., Pousson, A., Aluisio, A.R., Bilal, S., Moretti, K., Douglass, K.A., Henwood, P.C., Kharel, R., et al., 2021. COVID-19 pandemic prompts a paradigm shift in global emergency medicine: multidirectional education and remote collaboration. AEM Educ. Train. 5 (1), 79–90.
Katsulis, Z., Ergai, A., Leung, W.Y., Schenkel, L., Rai, A., Adelman, J., Benneyan, J., Bates, D.W., Dykes, P.C., 2016. Iterative user centered design for development of a patient-centered fall prevention toolkit. Appl. Ergon. 56, 117–126.
Kaur, A., Kaur, K., 2022. Systematic literature review of mobile application development and testing effort estimation. J. King Saud Univ.-Comput. Inf. Sci. 34 (2), 1–15.
Khajouei, R., Wierenga, P., Hasman, A., Jaspers, M.W., 2011. Clinicians satisfaction with CPOE ease of use and effect on clinicians' workflow, efficiency and medication safety. Int. J. Med. Inf. 80 (5), 297–309.
Khaliq, Z., Farooq, S.U., Khan, D.A., 2022. Artificial intelligence in software testing: Impact, problems, challenges and prospect. arXiv preprint arXiv:2201.05371.
Khoja, S., Durrani, H., Scott, R.E., Sajwani, A., Piryani, U., 2013. Conceptual framework for development of comprehensive e-health evaluation tool. Telemedicine e-Health 19 (1), 48–53.
Kitchenham, B., Charters, S., 2007. Guidelines for performing systematic literature reviews in software engineering.
Koester, T., Brøsted, J.E., Jakobsen, J.J., Malmros, H.P., Andreasen, N.K., 2017. The use of eye-tracking in usability testing of medical devices. In: Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care, Vol. 6, No. 1. SAGE Publications Sage CA, Los Angeles, CA, pp. 192–199.
Koopman, B., Zuccon, G., 2016. A test collection for matching patients to clinical trials. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 669–672.
Kostkova, P., 2015. Grand challenges in digital health. Front. Public Health 3, 134.
Krippendorff, K., 2018. Content Analysis: An Introduction to Its Methodology. Sage Publications.
Kushniruk, A.W., Patel, V.L., 2004. Cognitive and usability engineering methods for the evaluation of clinical information systems. J. Biomed. Inf. 37 (1), 56–76.
Larsen, L.B., 2003. Issues in the evaluation of spoken dialogue systems using objective and subjective measures. In: 2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721). IEEE, pp. 209–214.
Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al., 2009. Social science. Computational social science. Science 323 (5915), 721–723.
Lei-Juan, H., Jian-Bing, L., Xin-Zhu, X., 2016. Applying usability testing in the evaluation of products and services for elderly people. DEStech Trans. Eng. Technol. Res..
Leung, R., 2009. Improving the learnability of mobile device applications for older adults. In: CHI'09 Extended Abstracts on Human Factors in Computing Systems. pp. 3125–3128.
Lewis, J.R., 1992. Psychometric evaluation of the post-study system usability questionnaire: The PSSUQ. In: Proceedings of the Human Factors Society Annual Meeting. Vol. 36, Sage Publications Sage CA, Los Angeles, CA, pp. 1259–1260.
Lewis, J.R., 2002. Psychometric evaluation of the PSSUQ using data from five years of usability studies. Int. J. Hum.-Comput. Interact. 14 (3–4), 463–488.
Lindgaard, G., Chattratichart, J., 2007. Usability testing: what have we overlooked? In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1415–1424.
Liu, Y., Pencheon, E., Hunter, R.M., Moncrieff, J., Freemantle, N., 2018. Recruitment and retention strategies in mental health trials–a systematic review. PLoS One 13 (8), e0203127.
Lyzara, R., Purwandari, B., Zulfikar, M.F., Santoso, H.B., Solichah, I., 2019. E-government usability evaluation: Insights from a systematic literature review. In: Proceedings of the 2nd International Conference on Software Engineering and Information Management. pp. 249–253.
Mack, Z., Sharples, S., 2009. The importance of usability in product choice: A mobile phone case study. Ergonomics 52 (12), 1514–1528.
Maqbool, B., Herold, S., 2021. Challenges in developing software for the Swedish healthcare sector. In: HEALTHINF. pp. 175–187.
Maramba, I., Chatterjee, A., Newman, C., 2019. Methods of usability testing in the development of ehealth applications: a scoping review. Int. J. Med. Inf. 126, 95–104.

Melnick, E.R., West, C.P., Nath, B., Cipriano, P.F., Peterson, C., Satele, D.V., Shanafelt, T., Dyrbye, L.N., 2021. The association between perceived electronic health record usability and professional burnout among US nurses. J. Am. Med. Inf. Assoc. 28 (8), 1632–1641.
Middleton, B., Bloomrosen, M., Dente, M.A., Hashmat, B., Koppel, R., Overhage, J.M., Payne, T.H., Rosenbloom, S.T., Weaver, C., Zhang, J., 2013. Enhancing patient safety and quality of care by improving the usability of electronic health record systems: recommendations from AMIA. J. Am. Med. Inf. Assoc. 20 (e1), e2–e8.
Miguel, J.P., Mauricio, D., Rodríguez, G., 2014. A review of software quality models for the evaluation of software products. Int. J. Softw. Eng. Appl. 5, 31–53.
Miller, S., 2001. Workload Measures. National Advanced Driving Simulator. University of Iowa Press, Iowa City, IA, USA.
Milne, J., 1999. Questionnaires: Advantages and disadvantages. In: Evaluation Cookbook.
Mitchell, M.L., Jolley, J.M., 2010. Research Design Explained. Thomson Wadsworth.
Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., Group*, P., 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann. Internal Med. 151 (4), 264–269.
Molich, R., Nielsen, J., 1990. Improving a human-computer dialogue. Commun. ACM 33 (3), 338–348.
Mortensen, D.H., 2020. Pros and cons of conducting user interviews. https://www.interaction-design.org/literature/article/pros-and-cons-of-conducting-user-interviews. (Accessed 1 October 2021).
Mudaranthakam, D.P., Alsup, A.M., Murakonda, V., Lin, T., Thompson, J., Gajewski, B., Mayo, M.S., 2022. Accelerating cancer patient recruitment through a mobile application (clinical trial finder). Cancer Inform. 21, 11769351211073114.
Mulumba, M., Nantaba, J., Brolan, C.E., Ruano, A.L., Brooker, K., Hammonds, R., 2014. Perceptions and experiences of access to public healthcare by people with disabilities and older people in uganda. Int. J. Equity Health 13 (1), 1–9.
Mvungi, J., Tossy, T., 2015. Usability evaluation methods and principles for the web. Int. J. Comput. Sci. Inf. Secur. 13 (7), 86.
Namoun, A., Alrehaili, A., Tufail, A., 2021. A review of automated website usability evaluation tools: Research issues and challenges. In: International Conference on Human-Computer Interaction. Springer, pp. 292–311.
Nguyen, T.T., Jayadeva, V., Cizza, G., Brown, R.J., Nandagopal, R., Rodriguez, L.M., Rother, K.I., 2014. Challenging recruitment of youth with type 2 diabetes into clinical trials. J. Adolesc. Health 54 (3), 247–254.
Nielsen, J., 1993. Usability Engineering. Academic Press, London.
Nielsen, J., 1994. Usability Engineering. Morgan Kaufmann.
Nielsen, J., 2001. First rule of usability? Don't listen to users. https://www.nngroup.com/articles/first-rule-of-usability-dont-listen-to-users/. (Accessed 1 October 2021).
Nielsen, J., 2005. Accessibility is not enough. https://www.nngroup.com/articles/accessibility-is-not-enough/. (Accessed 11 July 2022).
Nielsen, K., Abildgaard, J.S., Daniels, K., 2014. Putting context into organizational intervention design: Using tailored questionnaires to measure initiatives for worker well-being. Hum. Relat. 67 (12), 1537–1560.
Nielsen, J., Clemmensen, T., Yssing, C., 2002. Getting access to what goes on in people's heads? Reflections on the think-aloud technique. In: Proceedings of the Second Nordic Conference on Human-Computer Interaction. pp. 101–110.
Nielsen, J., Mack, R.L. (Eds.), 1994. Usability inspection methods. John Wiley & Sons, Inc., USA.
Nielsen, J., Molich, R., 1990. Heuristic evaluation of user interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 249–256.
Nittari, G., Khuman, R., Baldoni, S., Pallotta, G., Battineni, G., Sirignano, A., Amenta, F., Ricci, G., 2020. Telemedicine practice: review of the current ethical and legal challenges. Telemed. e-Health 26 (12), 1427–1437.
Nur, I., Iskandar, H., Ade, R., 2020. The measurement of nurses'mental workload using nasa-tlx method (a case study). Malays. J. Public Health Med. 20 (Special1), 60–63.
Ohnemus, K.R., Biers, D.W., 1993. Retrospective versus concurrent thinking-out-loud in usability testing. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 37, No. 17. SAGE Publications Sage CA, Los Angeles, CA, pp. 1127–1131.
Ouellet, J., Kang, J., Girouard, A., 2021. Cultural and linguistic variables in usability testing: A Canadian evaluation of international students. J. Usability Stud. 16 (2).
Patel, N.A., Butte, A.J., 2020. Characteristics and challenges of the clinical pipeline of digital therapeutics. NPJ Digit. Med. 3 (1), 1–5.
Patel, M.X., Doku, V., Tennakoon, L., 2003. Challenges in recruitment of research participants. Adv. Psychiatr. Treat. 9 (3), 229–238.
Paz, F., Pow-Sang, J.A., 2016. A systematic mapping review of usability evaluation methods for software development process. Int. J. Softw. Eng. Appl. 10 (1), 165–178.
Pernice, K., 2017. F-shaped pattern of reading on the web: Misunderstood, but still relevant (even on mobile). https://www.nngroup.com/articles/f-shaped-pattern-reading-web-content/. (Accessed 1 June 2022).
Petersen, A., 2018. Digital Health and Technological Promise: A Sociological Inquiry. Routledge.
Pramesha Chandrasiri, G., Halgamuge, M.N., Subhashi Jayasekara, C., 2019. A comparative study in the application of IoT in health care: data security in telemedicine. Secur. Priv. Trust IoT Environ. 181–202.
Price, W.N., Cohen, I.G., 2019. Privacy in the age of medical big data. Nat. Med. 25 (1), 37–43.
PwC, 2016. The digital patient is here — but is healthcare ready?.
Raghupathi, W., Raghupathi, V., 2014. Big data analytics in healthcare: promise and potential. Health Inf. Sci. Syst. 2, 1–10.
Reeves, R., Coulter, A., Jenkinson, C., Cartwright, J., Bruster, S., Richards, N., 2002. Development and Pilot Testing of Questionnaires for Use in the Acute NHS Trust Inpatient Survey Programme. Picker Institute, Europe.
Rieman, J., Franzke, M., Redmiles, D., 1995. Usability evaluation with the cognitive walkthrough. In: Conference Companion on Human Factors in Computing Systems. pp. 387–388.
Riessman, C.K., 2008. Narrative Methods for the Human Sciences. Sage.
Riihiaho, S., 2018. Usability testing. Wiley Handb. Hum. Comput. Interact. 1, 255–275.
Rocha, M.A.M., Sandoval, C.A.M., Bautista, J.S.A., Ramírez, H.L., 2014. Developing a usability study for Mexican Government Sites: the case study of the Portal del Empleo. In: Proceedings of the 5th Mexican Conference on Human-Computer Interaction. pp. 1–6.
Rubin, J., Chisnell, D., 2008. Handbook of Usability Testing: How to Plan, Design and Conduct Effective Tests. John Wiley & Sons.
Rumsey, S., 2008. How to Find Information: A Guide for Researchers. McGraw-Hill Education (UK).
Russ, A.L., Saleem, J.J., 2018. Ten factors to consider when developing usability scenarios and tasks for health information technology. J. Biomed. Inf. 78, 123–133.
Saad, M., Zia, A., Raza, M., Kundi, M., Haleem, M., 2022. A comprehensive analysis of healthcare websites usability features, testing techniques and issues. IEEE Access.
Sagar, K., Saha, A., 2017. A systematic review of software usability studies. Int. J. Inf. Technol. 1–24.
Salvador-Carulla, L., Saxena, S., 2009. Intellectual disability: between disability and clinical nosology. Lancet 374 (9704), 1798–1799.
Santoso, I.S., Ferdinansyah, A., Sensuse, D.I., Suryono, R.R., Hidayanto, A.N., et al., 2021. Effectiveness of gamification in mHealth apps designed for mental illness. In: 2021 2nd International Conference on ICT for Rural Development. IC-ICTRuDev, IEEE, pp. 1–6.
Sauro, J., Lewis, J.R., 2011. When designing usability questionnaires, does it hurt to be positive? In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 2215–2224.
Schaller, S., Marinova-Schmidt, V., Setzer, M., Kondylakis, H., Griebel, L., Sedlmayr, M., Graessel, E., Maler, J.M., Kirn, S., Kolominsky-Rabas, P.L., et al., 2016. Usefulness of a tailored eHealth service for informal caregivers and professionals in the dementia treatment and care setting: the eHealthMonitor dementia portal. JMIR Res. Prot. 5 (2), e4354.
Schiessl, M., Duda, S., Thölke, A., Fischer, R., 2003. Eye tracking and its application in usability and media research. MMI-interaktiv J. 6 (2003), 41–50.
Schmidt, J.D.E., De Marchi, A.C.B., 2017. Usability evaluation methods for mobile serious games applied to health: a systematic review. Univ. Access Inf. Soc. 16 (4), 921–928.
Schnall, R., Cho, H., Liu, J., et al., 2018. Health information technology usability evaluation scale (health-ITUES) for usability assessment of mobile health technology: validation study. JMIR mHealth uHealth 6 (1), e8851.
Schnall, R., Rojas, M., Bakken, S., Brown, W., Carballo-Dieguez, A., Carry, M., Gelaude, D., Mosley, J.P., Travers, J., 2016. A user-centered model for designing consumer mobile health (mHealth) applications (apps). J. Biomed. Inf. 60, 243–251.
Schneiderman, B., 1987. Designing the user interface: Strategies for effective human-computer interaction.
Schrier, J.R., 1992. Reducing stress associated with participating in a usability test. In: Proceedings of the Human Factors Society Annual Meeting, Vol. 36, No. 16. SAGE Publications Sage CA, Los Angeles, CA, pp. 1210–1214.
Shamsujjoha, M., Grundy, J., Li, L., Khalajzadeh, H., Lu, Q., 2021. Human-centric issues in ehealth app development and usage: A preliminary assessment. In: 2021 IEEE International Conference on Software Analysis, Evolution and Reengineering. SANER, IEEE, pp. 506–510.
Shaw, D.K., 2009. Overview of telehealth and its application to cardiopulmonary physical therapy. Cardiopulm. Phys. Ther. J. 20 (2), 13.
Sik-Lányi, C., Szűcs, V., Guzsvinecz, T., 2017. Usability and colour-check of a healthcare WEB-site. In: 2017 IEEE 30th Neumann Colloquium. NC, IEEE, pp. 000111–000116.
Silva, P.A., Nunes, F., 2010. 3 x 7 usability testing guidelines for older adults.
Silva, P.A., Santos, R., 2023. An opinion mining methodology to analyse games for health. Multimedia Tools Appl. 82 (9), 12957–12976.
Simpson, S.H., 2015. Creating a data analysis plan: What to consider when choosing statistics for a study. Can. J. Hosp. Pharm. 68 (4), 311.
Sittig, D.F., Belmont, E., Singh, H., 2018. Improving the safety of health information technology requires shared responsibility: It is time we all step up. In: Healthcare. Vol. 6, Elsevier, pp. 7–12.
Sivaji, A., Abdollah, N., Tzuaan, S.S., Khean, C.N., Nor, Z.M., Rasidi, S.H., Wai, Y.S., 2014. Measuring public value UX-based on ISO/IEC 25010 quality attributes: Case study on e-Government website. In: 2014 3rd International Conference on User Science and Engineering. I-USEr, IEEE, pp. 56–61.
Skov, M.B., Stage, J., 2012. Training software developers and designers to conduct usability evaluations. Behav. Inf. Technol. 31 (4), 425–435.

Snyderman, R., 2012. Personalized health care: from theory to practice. Biotechnol. J. 7 (8), 973–979.
Sonderegger, A., Sauer, J., 2013. The influence of socio-cultural background and product value in usability testing. Appl. Ergon. 44 (3), 341–349.
Soni, H., Ivanova, J., Wilczewski, H., Bailey, A., Ong, T., Narma, A., Bunnell, B.E., Welch, B.M., 2022. Virtual conversational agents versus online forms: Patient experience and preferences for health data collection. Front. Digit. Health 4, 954069.
Spezi, V., 2016. Is information-seeking behavior of doctoral students changing?: a review of the literature (2010–2015). New Rev. Acad. Librariansh. 22 (1), 78–106.
Stoyanov, S.R., Hides, L., Kavanagh, D.J., Zelenko, O., Tjondronegoro, D., Mani, M., 2015. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR mHealth uHealth 3 (1), e3422.
Suvanmanee, W., Kee, D.M.H., Lee, Z.W., Low, C.M., Zakwan, M.A., Gupta, A., Pandey, R., Alfahad, F.F., Quttainah, M.A., 2020. The relationship between customer satisfaction and organizational success: A study of Panasonic. J. Community Dev. Asia 3 (3), 48–57.
Sygna, K., Johansen, S., Ruland, C.M., 2015. Recruitment challenges in clinical research including cancer patients and caregivers. Trials 16 (1), 1–9.
Tamir, D., Komogortsev, O.V., Mueller, C.J., 2008. An effort and time based measure of usability. In: Proceedings of the 6th International Workshop on Software Quality. pp. 47–52.
Thomson, S., Foubister, T., Figueras, J., Kutzin, J., Permanand, G., Bryndová, L., Organization, W.H., et al., 2009a. Addressing financial sustainability in health systems.
Thomson, S., Foubister, T., Mossialos, E., Organization, W.H., et al., 2009b. Financing Health Care in The European Union: Challenges and Policy Responses. World Health Organization. Regional Office for Europe.
Thota, M.K., Shajin, F.H., Rajesh, P., et al., 2020. Survey on software defect prediction techniques. Int. J. Appl. Sci. Eng. 17 (4), 331–344.
Tullis, T.S., Stetson, J.N., 2004. A comparison of questionnaires for assessing website usability. In: Usability Professional Association Conference. Vol. 1, Minneapolis, USA, pp. 1–12.
UKCRC, 2017. Health categories. https://hrcsonline.net/health-categories/. (Accessed 1 March 2022).
Van Den Haak, M., De Jong, M., Jan Schellens, P., 2003. Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. Behav. Inf. Technol. 22 (5), 339–351.
Van den Haak, M.J., de Jong, M.D., Schellens, P.J., 2004. Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: A methodological comparison. Interact. Comput. 16 (6), 1153–1170.
Van der Geest, T., 2006. Conducting usability studies with users who are elderly or have disabilities. Tech. Commun. 53 (1), 23–31.
Varela, P., Ares, G., 2012. Sensory profiling, the blurred line between sensory and consumer science. A review of novel methods for product characterization. Food Res. Int. 48 (2), 893–908.
Verkijika, S.F., De Wet, L., 2018. A usability assessment of e-government websites in Sub-Saharan Africa. Int. J. Inf. Manage. 39, 20–29.
Walji, M.F., Kalenderian, E., Piotrowski, M., Tran, D., Kookal, K.K., Tokede, O., White, J.M., Vaderhobli, R., Ramoni, R., Stark, P.C., et al., 2014. Are three methods better than one? A comparative assessment of usability evaluation methods in an EHR. Int. J. Med. Inf. 83 (5), 361–367.
Wang, C.-M., Huang, C.-S., 2022. Using digital technology to design a simple interactive system for nostalgic gaming to promote the health of slightly disabled elderly people. Int. J. Environ. Res. Public Health 20 (1), 128.
WHO, 2021. Global strategy on digital health 2020–2025.
Wijesinghe, N., Prasad, P., Alsadoon, A., Elchouemi, A., 2016. Usability evaluation of electronic health records using user integrated heuristic walkthrough method. In: 2016 International Conference on Information and Digital Technologies. IDT, IEEE, pp. 305–310.
Wohlin, C., 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering. pp. 1–10.
World Health Organization, et al., 2018. A vision for primary health care in the 21st century: towards universal health coverage and the sustainable development goals. Tech. Rep., World Health Organization (WHO).
Wronikowska, M.W., Malycha, J., Morgan, L.J., Westgate, V., Petrinic, T., Young, J.D., Watkinson, P.J., 2021. Systematic review of applied usability metrics within usability evaluation methods for hospital electronic healthcare record systems: Metrics and evaluation methods for eHealth systems. J. Eval. Clin. Pract. 27 (6), 1403–1416.
Yanez-Gomez, R., Cascado-Caballero, D., Sevillano, J.-L., 2017. Academic methods for usability evaluation of serious games: a systematic review. Multimedia Tools Appl. 76 (4), 5755–5784.
Ye, Q., Boren, S.A., Khan, U., Kim, M.S., 2017. Evaluation of functionality and usability on diabetes mobile applications: a systematic literature review. In: International Conference on Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Springer, pp. 108–116.
Yen, P.-Y., Wantland, D., Bakken, S., 2010. Development of a customizable health IT usability evaluation scale. In: AMIA Annual Symposium Proceedings. Vol. 2010, American Medical Informatics Association, p. 917.
Yeo, A.W., 2000. Are usability assessment techniques reliable in non-western cultures? Electron. J. Inf. Syst. Develop. Countries 3 (1), 1–21.
Yin, R., 2017. Case study research and applications: Design and methods. Oaks.
Zahra, F., Mohd, H., Hussain, A., Omar, M., 2018. Usability dimensions for chronic disease mobile applications: a systematics literature review. In: Knowledge Management International Conference. KMICe, KMICe, pp. 363–368.
Zapata, B.C., Fernández-Alemán, J.L., Idri, A., Toval, A., 2015. Empirical studies on usability of mHealth apps: a systematic literature review. J. Med. Syst. 39 (2), 1–19.
Zhang, J., Johnson, T.R., Patel, V.L., Paige, D.L., Kubose, T., 2003. Using usability heuristics to evaluate patient safety of medical devices. J. Biomed. Inf. 36 (1–2), 23–30.

Bilal Maqbool is currently pursuing his doctorate in Computer Science in the Software Quality and Digital Modernisation (SQuaD) Research Group at the Department of Mathematics and Computer Science at Karlstad University, Sweden. He holds an M.S. degree in Software Engineering from the Computer and Software Engineering Department of the National University of Sciences and Technology (NUST), Pakistan. He also served as a Senior Research Assistant at NUST, contributing to the field of software engineering. Bilal's research interests focus on software design, usability/UX evaluation, software quality, and business process automation through natural language processing (NLP). He can be contacted at: bilal.maqbool@kau.se.

Sebastian Herold is an associate professor in the Software Quality and Digital Modernisation (SQuaD) Research Group at the Department of Mathematics and Computer Science of Karlstad University, Sweden. He received his Ph.D. from Clausthal University of Technology, Germany, and worked at Lero - The Irish Software Research Centre in Limerick, Ireland. Sebastian's research interests focus on software architecture and design, software evolution, usability/UX evaluation, and software quality. He can be contacted at: sebastian.herold@kau.se.