scientific reports

Check for updates

OPEN

Developing a standardized framework for evaluating health apps using natural language processing

Julian Herpertz[1,2,6], Bridget Dwyer[2], Jacob Taylor[3], Nils Opel[1,4,5] & John Torous[2]✉

Despite regulatory efforts, many smartphone health applications remain unregulated, raising concerns about privacy, security, and evidence-based effectiveness. The lack of standardized regulation has led to the proliferation of over 130 frameworks, introducing new criteria and methodologies for app evaluation. The sheer number of frameworks, coupled with their varying approaches to app evaluation, create challenges for comparison. Our study aims to synthesize existing knowledge and propose a standardized app evaluation framework. We conducted a synthesis of reviews on health app evaluation frameworks. Using natural language processing (NLP), we analyzed evaluation domains and grouped them into clusters based on semantic similarities. Standardized definitions for these clusters were developed. We identified eight review articles that met the inclusion criteria, each proposing between six and 17 app evaluation domains. Using NLP, we identified five clusters of app evaluation: Effectiveness & Development, Technology & Functionality, Validity & Legal, Safety & Privacy, and Implementation & Ethics, each of which was assigned a standardized definition. The clusters align with but expand on the American Psychiatric Association's evaluation domains, incorporating critical aspects such as inclusivity, safety, engagement, and ethical principles. Temporal analysis revealed an increasing focus on Effectiveness & Development, while Safety & Privacy showed a stagnation in attention over time.

Keywords App evaluation framework, Information management, Mobile health apps, Telemedicine, Natural Language processing

There are over 2 million apps on the Google Play Store[1] and Apple's App Store[2] and more than 300,000 have been classified as mobile health applications (mHealth apps)[3]. MHealth utilizes mobile devices like smartphones and tablets to provide healthcare services and facilitate health management. More than 300 million people worldwide have experience with mHealth apps[4], and approximately one in five adults in the United States reports using one[5]. The growing interest underscores the importance of rigorous evaluation to validate the claimed benefits and to understand the real-world effectiveness and impact on health outcomes.

Agencies responsible for regulation are tasked with safeguarding the use of medical devices and establishing criteria that developers must meet. Selected examples of such agencies include the Medicines and Healthcare Products Regulatory Agency (MHRA) in the UK[6], the Therapeutic Goods Administration in Australia[7], the Federal Institute for Drugs and Medical Devices (BfArM) in Germany[8] and the U.S. Food and Drug Administration (FDA)[9]. However, due to the distinctive nature of health apps, many are considered wellness devices, and thus, in numerous countries, they remain outside the scope of healthcare regulation. This regulatory gap poses significant challenges to ensuring the quality and safety of app marketplaces, which already offer direct access to tens of thousands health of apps today[10–13]. A recent assessment of selected health apps highlighted these concerns, revealing that nearly one-fourth of apps lack a privacy policy, almost half share health data with third-party companies, and only 15% provided evidence of feasibility or efficacy for their software[14]. These

[1]Department of Psychiatry and Psychotherapy, Jena University Hospital, Jena, Germany. [2]Division of Digital Psychiatry, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA. [3]David A. Dunlap Department of Astronomy and Astrophysics, University of Toronto, Toronto, ON, Canada. [4]Center for Intervention and Research on Adaptive and Maladaptive Brain Circuits Underlying Mental Health (C-I-R-C), Jena-Magdeburg-Halle, Germany. [5]German Center for Mental Health (DZPG), Berlin, Germany. [6]Department of Psychiatry and Psychotherapy, Campus Benjamin Franklin, Charité Universitätsmedizin, Berlin, Germany. ✉email: jtorous@bidmc.harvard.edu

findings illustrate the broader issue, as the health app ecosystem continues to be described as a "Wild West" of unregulated products, characterized by a lack of quality assurance and consistent oversight[15,16].

The lack of standardized regulation has resulted in the global community developing over 130 distinct evaluation frameworks for health apps, as identified in a systematic review conducted in August 2022[17]. Defined as criteria-based tools to assess the quality of mobile health apps, these frameworks aim to enhance transparency and accountability in the evaluation process by addressing specific aspects of app quality[18]. Approaches to evaluation vary; for example, the Mobile App Rating Scale (MARS)[19], developed by the Queensland University of Technology, employs a scoring system to assess apps across five key criteria: engagement, functionality, aesthetics, information quality and subjective quality. Privacy standards are notably absent from MARS' evaluation criteria. In contrast, frameworks like Mindapps[20], which translated the American Psychiatric Association (APA)[21] app evaluation framework into a dynamic database, take a guidance-based approach. Mindapps operationalizes app evaluation through objective yes-or-no questions addressing quality standards such as privacy policies and evidence base; however, it does not employ a scoring system.

The many frameworks can serve as valuable tools for both clinicians and patients, aiding in the selection of health apps that align with individual preferences and needs. However, the proliferation of frameworks has created a paradox: while designed to bring clarity to app evaluation, their sheer abundance has made the landscape increasingly challenging to navigate. Instead of providing straightforward guidance, many frameworks introduce new complexities and challenges in selecting the most appropriate tool. Compounding this issue, there is no standardized set of evaluation domains, resulting in frameworks using different methodologies, questions, and sets of criteria to evaluate quality. The two aforementioned frameworks illustrate the discordance within the field: MARS and Mindapps rely on different criteria and adopt distinct strategies for evaluating app quality, rendering comparisons of their findings almost impossible.

In response, several reviews have been conducted to analyze the existing frameworks, making it possible to identify which frameworks may be suitable for specific needs. Reviews established domain names to summarize similar app evaluation criteria across the investigated frameworks. Moshi et al.[22] for example aligned their selection of domain names with best practice protocols for Health Technology Assessments[23], whereas Nouri et al.[24] independently developed their domain names, referred to as class names in their review. While the classification of app evaluation criteria across the 130 plus frameworks was highly valuable at the time of publication, the growing number of reviews has made it increasingly challenging to maintain an overview of the established overarching domains. Therefore, synthesizing domain names from existing reviews and establishing a standardized set of underlying questions and baseline assessments shared across all frameworks is a productive next step. This would enhance interoperability and foster greater consistency in app evaluation, which is not to suggest that there must be a single framework for app evaluation. However, identifying common aspects across frameworks can improve alignment among existing ones and inform the design of new frameworks.

A related effort was undertaken six years ago by Henson and colleagues, who convened a clinician and patient expert panel to examine 45 app evaluation frameworks[25]. The panel successfully mapped 357 unique criteria and questions assessed by these frameworks onto the five app evaluation domains of the APA[21]: Background Information, Privacy and Security, Evidence Base, Ease of Use, and Data Integration. This mapping offers a shared pool of app evaluation questions and a theoretical framework for mapping different domains, which other frameworks may consider applying or adapting. Given the significant changes in the digital health and app landscape, as well as the emergence of new frameworks since the review by Henson et al., the field of mobile health has likely evolved considerably. Additionally, when expert panels determine the naming of app evaluation domains and oversee their mapping, these critical decisions inherently carry a degree of subjectivity. While Henson's study focused on summarizing app evaluation criteria across frameworks, the increasing number of reviews on this topic now allows for a direct synthesis of domain names derived from these reviews.

The objective of this study is to identify common domains for app evaluation by synthesizing the domain names used in existing reviews on the topic. Natural Language Processing (NLP) has increasingly been used in recent years to objectively standardize language and incorporate textual data into computational analyses[26]. Its key strengths include objectivity, scalability, and the ability to process and synthesize large volumes of textual data efficiently[27]. Using NLP to analyze the domain names and their definitions employed in the reviews on app evaluation frameworks, we categorize them into standardized app evaluation clusters, assign them a standardized name, and establish a cluster definition (Fig. 1). Ultimately, we aim to compare the resulting NLP-derived clusters with the five app evaluation domains of the APA framework created by the Henson et al. review to determine whether these domains still represent a standardized and widely agreed-upon structure among experts in the field. While we recognize that five clusters may not fully encapsulate the complexity of app evaluation and that a standalone framework for patients and clinicians might need to address additional evaluation aspects, our work also aims to guide policymakers and app evaluation framework developers in integrating the standardized expert consensus on key evaluation domains into their initiatives on app evaluation.

Results

Search results

Eight review articles of mHealth app evaluation frameworks met the inclusion criteria and were included in this review. Please refer to the supplementary material to get an overview of all included review studies. Figure 2 shows the PRISMA process.

Articles published between January 1, 2008, and December 03, 2024, were included in the search. Two reviewers screened the articles.

The eight articles included five systematic reviews and three scoping reviews. Across the eight reviews, the number of frameworks investigated per review ranged from 11 to 130. Some reviews opted to map the evaluation criteria of frameworks onto a broader range of app evaluation domains, with up to 17 domains addressed. In

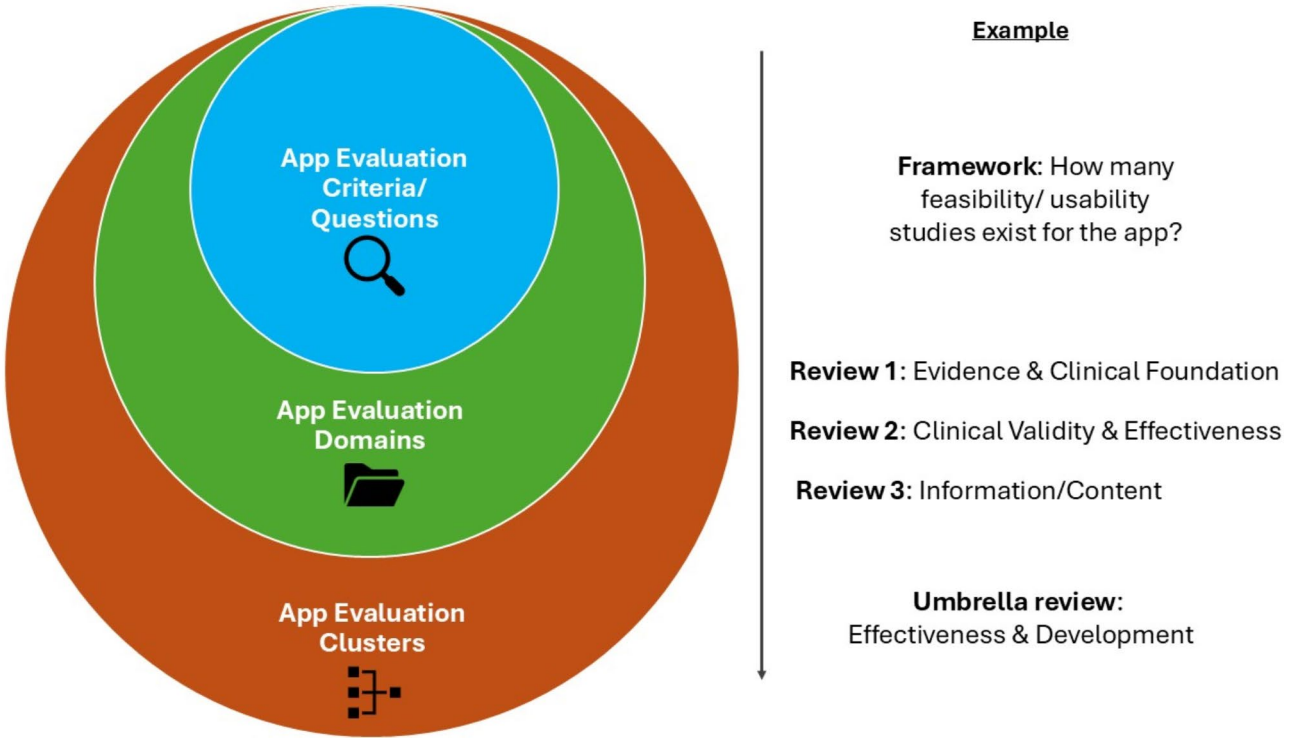

Fig. 1. In the presented figure, app evaluation criteria are hierarchically grouped into domains and broader clusters to enable a standardized evaluation. The app evaluation criteria used by individual frameworks are synthesized into app evaluation domains, as presented in various reviews. These domains are further synthesized into app evaluation clusters through our NLP-based approach.

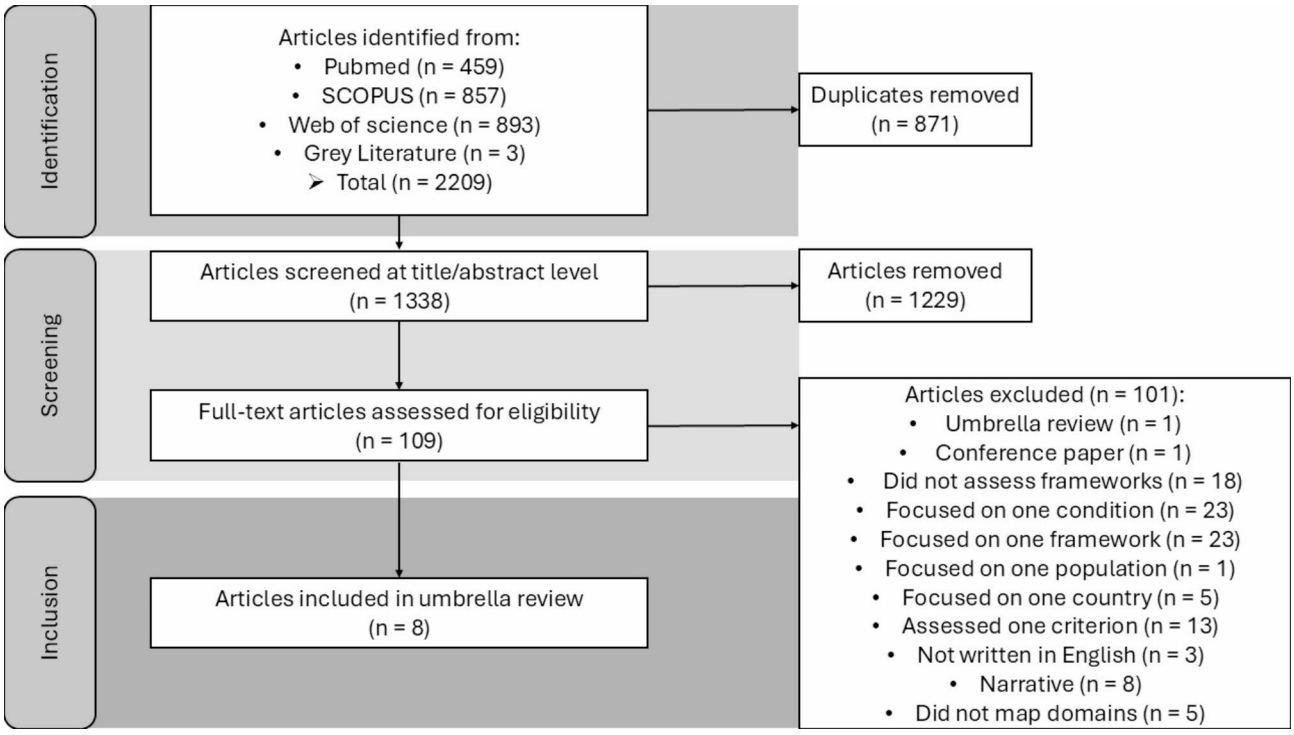

Fig. 2. The prisma diagram presents the review articles on mHealth app evaluation frameworks included in this review.

contrast, others focused more narrowly, selecting as few as six app evaluation domains. The median number of app evaluation domains in the included reviews was 11. Please refer to the supplementary materials for an overview of the findings when predefining the number of clusters as 11.

Clustering mHealth app evaluation domains and deriving definitions

We transformed the domain names into vector representations and grouped them into semantic clusters. By calculating the average embeddings for each cluster, we identified two representative terms that best characterize each cluster. Table 1 provides an overview of how the original domains were mapped to their respective cluster. We further analyzed the definitions of each domain in the source material. By synthesizing them, we developed a holistic definition for each cluster, incorporating the definitions provided in each review paper (Table 1).

Temporal analysis of cluster representation

We analyzed trends in the normalized frequency of the five thematic clusters mentioned in the identified reviews. The results show that the Effectiveness & Development cluster experienced a significant upward trend over time, indicating an increasing focus of app evaluation frameworks on app development processes, effectiveness, and related outcomes (slope = 0.066, $p < 0.001$, $R^2 = 0.770$). In contrast, Technology & Functionality displayed a decline (slope = -0.079, $p < 0.001$, $R^2 = 0.844$) in normalized frequency. The Validity & Legal cluster also declined, albeit at a slower rate (slope = -0.036, $p = 0.014$, $R^2 = 0.382$). Meanwhile, Safety & Privacy showed a slight upward trend, though not statistically significant (slope = 0.017, $p = 0.310$, $R^2 = 0.790$). Lastly, Implementation & Ethics remained relatively stable, with minimal change over the years (slope = 0.003, $p = 0.719$, $R^2 = 0.010$; Fig. 3).

Cluster	Effectiveness & development	Technology & functionality	Validity & legal	Safety & privacy	Implementation & Ethics
Domains	'clinical validity & effectiveness', 'developer credibility', 'development process', 'characteristics of the development team', 'psychotherapy & usability', 'quality & subjective assessment & predisposition'	'technical robustness & requirements', 'design & usability', 'user feedback', 'basic product & company', 'design', 'usability', 'functionality', 'description & technical', 'current use of technology', 'origin & functionality', 'features & engagement', 'user experience', 'technical & support', 'cocreation of technologies & user centered design', 'features & requirements', 'functionality', 'usability & design', 'interoperability & compatibility', 'functionality & technical & technology', 'usability', 'aesthetic & virtual design', 'technical support & updates'	'credibility', 'user perceived value', 'legal', 'cost & cost-effectiveness', 'inputs & outputs', 'content & validity', 'value', 'basic', 'evaluation', 'evidence based', 'validity & added value', 'equity & equality', 'legal', 'cost-effectiveness', 'actuality', 'quality', 'therapeutic persuasion'	'clinical safety', 'data mangement & privacy', 'security', 'security & privacy', 'safety', 'privacy & security', 'privacy & security & ethical & legal', 'usability & privacy & security', 'safety', 'data privacy & data security', 'privacy & data security', 'safety of use'	'interoperability', 'alignment to users', 'content & purpose', 'information content', 'ethical', 'effectiveness', 'ethical', 'social', 'organisational', 'evidence & clinical foundation', 'interoperability & data sharing', 'clarity of purpose', 'user engagement & adherence & social support', 'interoperability', 'accessibility', 'context', 'ethical & legal & social', 'stakeholder involvement', 'scientific', 'implementation', 'adoption', 'maintenance', 'performance', 'engagement', 'technology', 'information & transparency', 'social', 'organizational', 'engagement & participation', 'accessibility', 'information', 'scientific basis', 'ethics', 'psychotherapy', 'therapeutic alliance'
Definition	**Effectiveness** of a health app is defined by its ability to *ensure* meaningful user benefits, achieved through the *accuracy* of its measurements and information, supported by *references* to clinical guidelines and evidence-based studies. **Development** encompasses a clear *aim*, defining the target population and the goals developers seek to achieve. It requires transparency about the *development* process, including the characteristics of the team involved, and their qualifications. Additionally, it involves scrutiny of *funding* sources to assess their credibility and identify potential conflicts of interest.	**Technology** refers to the *software*, maintained through regular updates and technical support, and the *development* process, which implements the app's purpose and specifies the technical requirements users must meet to use it. It also encompasses the app's role in *communicating* by integrating with electronic health records to enable interaction between users and healthcare providers. **Functionality** refers to its *design*, which prioritizes an intuitive user interface to ensure ease of use, and the provision of *information*, offering clear details about the app and its features. Additionally, they encompass *collaborative* capabilities, enabling shared decision-making by actively involving the user in the app's processes and interactions.	**Validity** refers to the app's ability to reliably fulfill the intended purpose. This involves evaluating the *credibility* of its developers, ensuring they are trusted entities with clinical expertise; assessing its *effectiveness* by determining whether its cost is justified by the quality-of-life improvements it delivers; and examining its *language* to confirm that it provides education in a clear, user-friendly, and accessible way. **Legal** refers to the adherence of a health app to *regulations*, including compliance with privacy laws, accessibility standards, and user agreements. It encompasses the *consequences* of legal implications, such as clinical accountability and copyright infringement. Additionally, it considers the *advantages* of legal compliance, such as building user trust and promoting ethical app usage.	**Safety** is defined by the app's ability to minimize potential *risks*, including the mitigation of clinical concerns and prevention of misinformation. It ensures robust *authentication* measures, such as encryption, password protection, and secure data transfer. Additionally, safety relies on thorough *testing*, including clinical trials and penetration testing, to identify and address potential vulnerabilities. **Privacy** refers to the responsible management of *data*, including its collection, storage, processing, and sharing. It involves *compliance* with legal and ethical standards to ensure user rights and transparency, supported by *security* measures that protect against unauthorized access, breaches, and misuse, fostering a safe and trustworthy environment for users.	**Implementation** is defined by the app's *clinical* integration into healthcare systems, such as connecting to electronic health records and supporting patient-relevant outcomes, while also ensuring *functionality* that seamlessly fits into the user's daily life through accessible and engaging features. Both aspects rely on *accuracy* to deliver reliable, safe, and precise performance. **Ethics** begins with prioritizing inclusivity by addressing the needs of *disadvantaged* populations, such as individuals with disabilities, vulnerable groups, or those in underserved communities, ensuring fairness and equity. It also encompasses adherence to *privacy* standards. At its core, *ethics* involves upholding principles like autonomy, beneficence, and justice, ensuring that the app's development and use align with the best interests of all users.

Table 1. The mHealth evaluation clusters and their definitions. The italicized words represent the identified definition cluster names for each cluster.

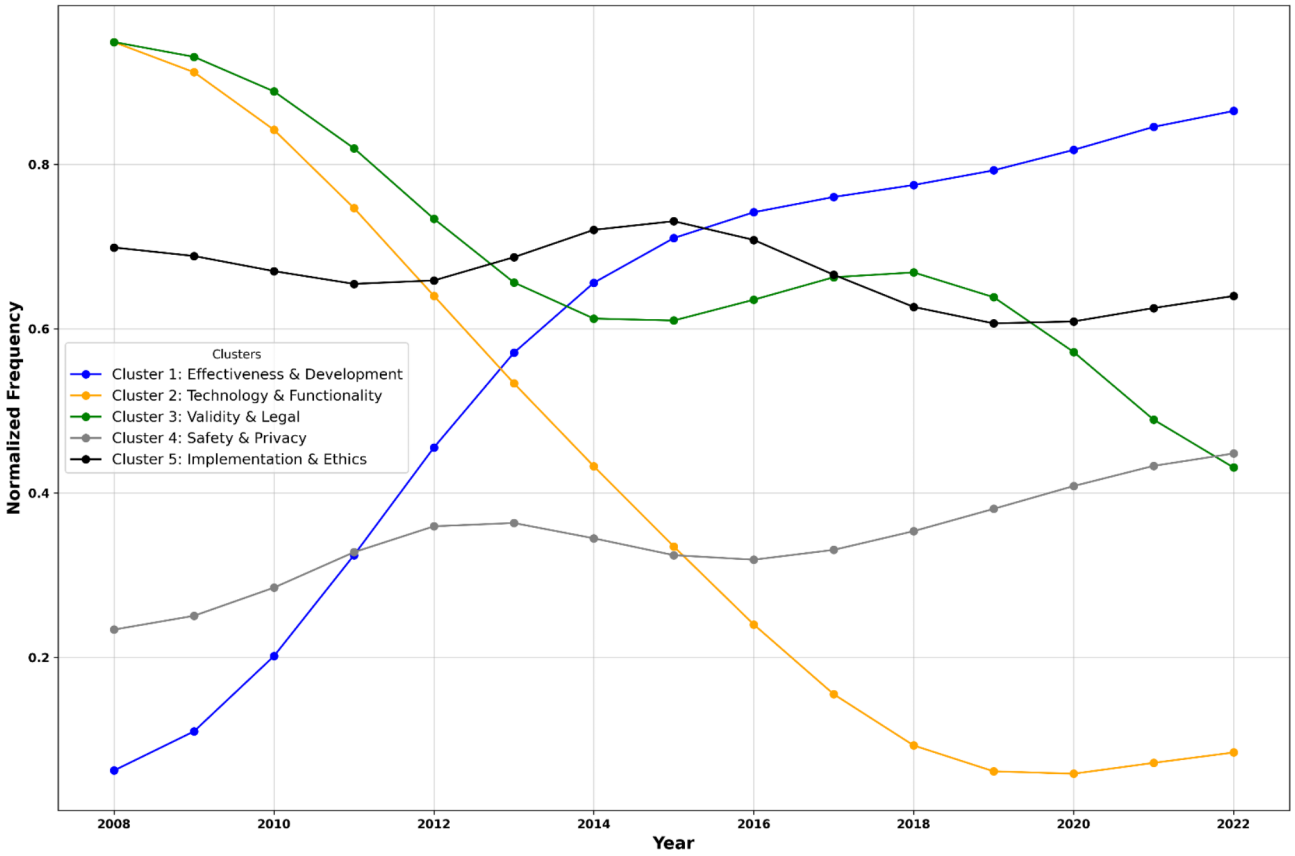

Fig. 3. The figure visualizes temporal trends in standardized mHealth app evaluation cluster occurrences (2008–2022). The Effectiveness & Development cluster experienced an upward trend over time. Technology & Functionality and Validity & Legal declined in popularity.The mean number of reviews per year included was 4.33 (SD = 1.70), with a minimum of 2 and a maximum of 7 reviews per year.

Mapping clusters to APA framework domains

The identified clusters are comparable to the app evaluation criteria established by the APA. The *Validity and Legal* cluster captures a construct similar to the ground level of the APA framework (Level 0; *Background Information*), as they both ensure the credibility of app developers and app accessibility. The *Safety & Privacy* cluster closely aligns with the *Privacy and Security* criterion of the APA framework (Level 1; *Risk*), but extends beyond addressing the risks associated with the sharing and security of sensible data. It also encompasses the dangers posed by misinformation provided by apps. The *Effectiveness & Development* cluster examines whether an app provides evidence-based studies to support its methodology and ensures its approaches accurately achieve user benefits. This aligns with the APA framework's Evidence Base criterion (Level 2; *Evidence*). Level 3 (*Ease of Use*) of the APA framework is indirectly reflected by the clusters. Instead of evaluating ease of use, the *Functionality & Technology* cluster focuses on app features that may enhance subjective criteria like usability. Lastly, Level 4 of the APA framework (*Interoperability*) is closely related to the *Implementation & Ethics* cluster (Fig. 4).

Discussion

By applying NLP to app evaluation domains to develop standardized terms and definitions, we identified five main overarching clusters (Effectiveness & Development, Technology & Functionality, Validity & Legal, Safety & Privacy, Implementation & Ethics) that can provide patients, clinicians, and regulators with critical and unique aspects to consider when evaluating apps across the spectrum, from safety to ethics. From over 130 frameworks identified in eight review articles, we synthesized a set of common metrics, questions, and domains that can inform the current strengths and weaknesses of current frameworks as well as guide the development of new ones. While previous reviews have summarized app evaluation criteria into overarching domains, this study represents the first NLP-based synthesis of reviews. In line with our assumptions, NLP proved to be a valuable tool for standardizing language, minimizing subjectivity in the naming of evaluation metrics and their definitions, and synthesizing findings from various review articles on app evaluation domains. Effectiveness and Privacy & Safety continue to serve as fundamental pillars of app evaluation.

The identified clusters also introduce new considerations that could guide health app regulators and framework developers in refining their evaluation criteria: a foundational level of app evaluation (referred to as the Ground Level in the APA framework and the *Validity & Legal* cluster) should emphasize inclusivity for

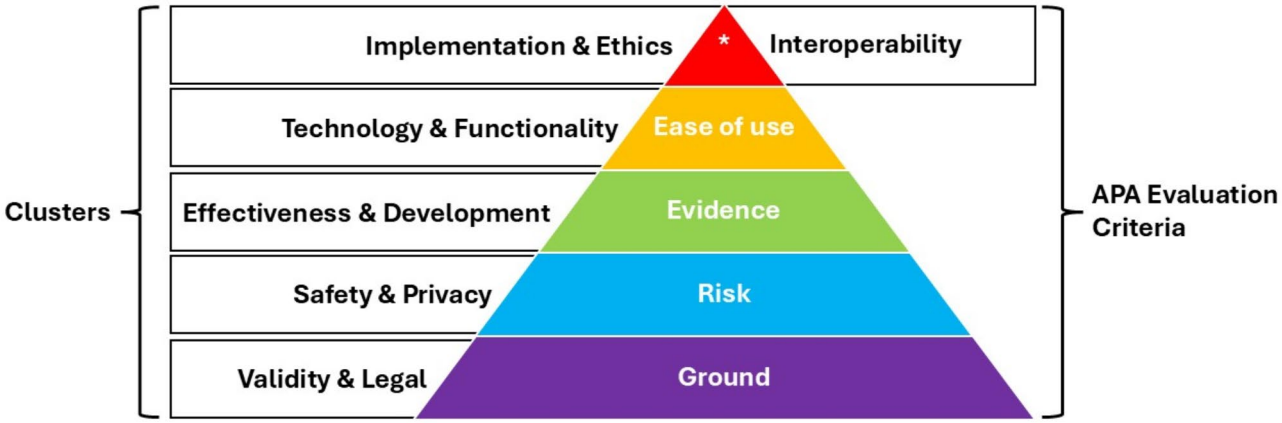

Fig. 4. The figure compares the identified clusters with the app evaluation framework of the American Psychiatric Association (APA).

disadvantaged populations. This includes using language that delivers education in a clear, user-friendly, and accessible manner while considering the potential advantages of app use for a broad spectrum of users—a factor increasingly emphasized in recent literature[18]. The temporal trend of the cluster aligns with recent findings in the literature, showing that factors related to accessibility and inclusivity have diminished in importance in recent years[18].

The scope of *Safety & Privacy* has expanded beyond the risks of data sharing and privacy violations to include the dangers of misinformation presented by health apps. The introduction of chatbots in health apps has especially been associated with an increased risk of offering harmful advice and encouraging destructive behaviors[28]. Given the added complexity introduced to the field of *Safety & Privacy*, it is concerning that the cluster has not demonstrated a significant increase in focus over time, as reflected in the temporal change figure. We emphasize the importance of prioritizing this domain to inform and guide new framework developers in addressing emerging challenges effectively.

The assessment of evidence supporting an app is a cornerstone of many evaluation frameworks, as shown in Fig. 4 by the increasing attention *Effectiveness & Development* have received in recent years. However, the standards required to fulfill this criterion remain undefined. Is published scientific literature demonstrating app effectiveness sufficient, as is already achieved by only a small minority of health apps[29]? Or should a comparative study proving a positive impact on care, as the DiGA framework suggests[30], be the benchmark? While addressing this question is beyond the scope of this review, our definition of effectiveness provides a framework for a more nuanced understanding of what it entails, going beyond merely the presence of evidence. It incorporates the ability to ensure meaningful outcomes for users, the accuracy of information, and alignment with clinical guidelines, recognizing that its application will inevitably vary across regional contexts and specific clinical needs. *Ensuring* user benefits emphasizes the app's capacity to deliver tangible and meaningful improvements to the user's health or well-being. The *accuracy* of measurements and information encompasses the precision of health tracking features, the validity of diagnostic tools, and the consistency of the information presented. Without such accuracy, the app risks delivering misleading or harmful advice, undermining its overall effectiveness. Lastly, *references* to clinical guidelines and evidence-based studies is fundamental to ensuring that an app's content and functionalities are rooted in scientifically validated principles. However, these references must also be adaptable to region-specific clinical guidelines, reflecting the diversity of healthcare systems and practices worldwide.

One of the most significant differences between the APA framework and the identified clusters lies in their approach to evaluating user engagement. While the APA framework emphasizes ease of use, the clusters we created prioritize *Technology & Functionality*. Assuming that *Technology & Functionality* include factors that could enhance user engagement with an app, as suggested by recent literature[31,32], this shift in focus provides valuable insights into the elements that contribute to sustained user interaction and satisfaction. Technology encompasses regularly updated and supported software, ensuring reliability and user trust. The *development* process is critical in implementing the app's purpose and specifying the technical requirements users must meet, directly influencing accessibility and usability. Furthermore, technology facilitates *communication* by integrating with electronic health records, enabling seamless interaction between users and healthcare providers. These aspects collectively highlight that robust technological infrastructure is essential for creating an app users can depend on and engage with effectively. Functionality, on the other hand, focuses on the user experience. Its *design* prioritizes an intuitive interface that ensures ease of use, minimizing barriers to engagement and fostering user confidence. Providing clear, accessible *information* about the app and its features further enhances understanding and trust. *Collaboration* enables shared decision-making by actively involving users in the app's processes and interactions. By combining the three pillars of technology with the three pillars of functionality, we learn that reliable technology establishes the foundation for trust and usability. In contrast, thoughtful functionality ensures a positive and empowering user experience. Together, they create a comprehensive framework for predicting and fostering sustained engagement with health apps. Given the substantial retention

challenges faced by health apps, providing deeper insights into the factors that could drive user engagement is crucial in shaping the future of the digital health landscape[31,32]. However, to gain a holistic understanding of what drives these factors, it is essential to have access to real-life, real-time user uptake data from app developers, as they hold the key to understanding how apps are being adopted and utilized in practice. We advocate for stronger collaboration between academic researchers, clinicians, and app developers.

Lastly, *Implementation* emphasizes the *clinical* interconnectedness between integrating the app into users' daily lives and providing *accurate* information derived from app usage to healthcare providers. This aligns closely with what Henson et al.[25] have described as interoperability. Defined as "the ability of two or more systems or components to exchange information and to use the information that has been exchanged"[33], interoperability has become highly desirable in healthcare[34]. However, medical data often originates as a collection of fragmented, disconnected small data points, making the goal of achieving interconnectedness a significant challenge within medicine[35,36]. Moreover, an increasingly interconnected healthcare system, while enhancing productivity and communication between all stakeholders, raises privacy concerns that appear to be a significant source of apprehension among patients. Our cluster of *Implementation & Ethics* underscores the importance and potential of fostering an increasingly interconnected exchange between technology and healthcare providers, aligning with findings from previous literature[37]. However, at the same time, it emphasizes the need to uphold *ethical* principles, such as rigorous privacy and security testing for data exchanges between the two systems, and to consider the needs of *disadvantaged* populations. Users should not experience any disadvantages from sharing information with an app and having that information shared with healthcare providers. Equally important is the requirement for full transparency regarding such exchanges. Transparency is crucial for establishing user trust and securing sustained consent for longitudinal data sharing. Furthermore, we aim to address regulators and app evaluation framework developers, urging them to integrate new insights of app evaluation into their assessment efforts. The five clusters, which effectively reflect expert opinions, can serve as a foundation in app evaluation. While additional aspects may warrant consideration, they represent the essential minimum criteria that need to be considered and addressed.

A practical use case for our presented clusters could be as follows: A clinician searching for a suitable health app for their patient may feel overwhelmed by the vast number of available options and question which apps are beneficial, and which might even pose risks. Using our cluster-based approach, they could systematically evaluate an app as follows: First, they could assess whether the developers behind the app represent a trustworthy and credible entity and whether the app adheres to quality and legal standards. This could serve as an initial benchmark: Does the app appear reliable, user-friendly, and trustworthy? Next, they could evaluate whether the app has a clear privacy policy and assess the potential risk of sensitive health information being compromised. Regarding effectiveness, the clinician would need to decide whether the app's claims are convincing or if its benefits should be substantiated by rigorous evidence, such as a randomized controlled trial. Finally, the clinician might consider how the app's functionalities could support professional therapy, such as whether it integrates with electronic health records or facilitates collaboration by providing updates on the patient's progress.

This study has several limitations that should be acknowledged. First, while our search covered reviews published up to the end of 2024, the included reviews only examined frameworks developed up to 2022. For example, the most recent review included in our study was published in May 2024[38]. However, it examined frameworks developed between 2016 and 2021. As a result, developments from the past three years may not be fully reflected in our study. Additionally, while we conducted a systematic search, it is possible that we did not identify all relevant reviews. Our exclusion of conference papers may have contributed to publication bias, as important findings presented in such settings were not included in this analysis.

Second, we recognize the potential for confirmatory bias towards the APA framework, given that our clustering approach predefined the number of clusters to align with its structure. To present an alternative result, we included a supplementary analysis in which the number of clusters was predefined based on the median number of app evaluation domains reported in the included reviews. While the NLP-based clustering method offers an objective means to group domain names, it has challenges. The accuracy of clustering depends on model parameters, the preprocessing steps applied, and the quality of the data input, which we did not have an impact on. Furthermore, simplifying domain terms during preprocessing, though necessary, may have led to a loss of nuance in some cases.

Our temporal analysis also has limitations. We examined the development of domain mentions over time but not the total frequency of mentions across all years. As seen in the number of domains mapped onto each cluster, those emphasizing usefulness and engagement—particularly *Functionality & Technology* and *Implementation & Ethics*—received the most attention overall. This finding aligns with observations from recent studies[25,39,40].

Despite the abundance of health app evaluation frameworks developed over the past decade, we are still far from a standardized and reliable system that helps to identify safe and effective health apps. The diversity and inconsistency among existing frameworks, often varying in terminology, assessment criteria, and methodologies, create confusion and hinder comparability. Our NLP approach can guide various stakeholders, clinicians, and users in identifying individually suitable apps, as well as policymakers and app evaluation framework developers in gaining orientation on the key aspects of app evaluation that experts widely agree upon.

Methods

This review was conducted according to the protocol registered on PROSPERO (CRD42024614460), following established guidelines for systematic reviews.

For the systematic search of literature, PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines were followed. The search was conducted in PubMed, Scopus, and Web of Science databases, focusing on identifying systematic and scoping reviews that evaluate mHealth app evaluation frameworks and mapped app evaluation criteria to app domains. Articles from the grey literature, including

references from retrieved articles and previously known studies familiar to the research team, were also included. Articles published between January 1, 2008, and December 03, 2024, were included in the search (Table 2). A detailed description of the search strategy is included in the supplementary materials.

Studies were selected for inclusion in a two-step process. First, titles and abstracts were screened by two reviewers to identify articles meeting the inclusion criteria. Eligible articles were then subjected to a full-text review. Discrepancies in the final list of reviews included were resolved through discussion. A third reviewer was available to resolve any disagreements in the inclusion of studies that could not be settled by discussion.

We employed a multi-step natural language processing (NLP) approach to analyze the reviews' original domains of mHealth app evaluation and cluster them into standardized categories (clusters). NLP techniques allow for the objective analysis of large volumes of text data, reducing the influence of individual biases that may affect subjective assessments. Moreover, by clustering similar domains based on semantic similarities, NLP can help establish more consistent groupings that might not be immediately apparent through manual evaluation.

To enhance the precision of the NLP model, domain names containing compound terms (e.g., *ethical aspects*, *legal information*, *social considerations*, *technical features*) were simplified by removing the secondary descriptor (e.g., *aspects*, *information*, *considerations*, *features*). This preprocessing step known as text normalization is a common practice in NLP[41] and ensured that the model focused on the core concept, improving its ability to capture semantic relationships. We used the SentenceTransformers model (https://huggingface.co/sentence-transformers; 'all-MiniLM-L6-v2'[42–44]) to encode the names into high-dimensional vector representations. These embeddings captured the semantic relationships between domain terms, enabling the grouping of similar concepts. Clustering was performed using the K-Means algorithm with the number of clusters predefined as five (similar to the APA's app evaluation framework). Five clusters allow for a comprehensive categorization without overcomplicating the structure, ensuring that each cluster captures a distinct area of evaluation. The algorithm assigned each domain to a cluster, grouping semantically similar terms together. Using the K-means algorithm with cosine similarity[45], we calculated the average embedding of each domain assigned to a cluster to identify the two most representative terms for each cluster. (Fig. 5).

As a supplementary analysis, we applied the same clustering method but predefined the number of clusters based on the median number of app evaluation domains reported in the included reviews.

Most reviews use definitions or examples to clarify what each app evaluation domain represents in mHealth apps. For example, Hensher and colleagues[11] define *Interoperability* as "Data sharing and data transfer capabilities of the health apps". We examined the definitions of domains assigned to each cluster and performed tokenization[46] to identify keywords within these definitions. Using the same NLP model as in 2.2, we encoded these keywords into vector representations and identified the six most representative terms for each cluster by computing the average embedding of all words within the cluster and then selecting the closest word from the vocabulary based on cosine similarity. The closest words served as the basis for developing standardized definitions that comprehensively reflect the original domain content. To ensure accuracy and inclusivity, the synthesized cluster definitions were manually cross-checked against the original domain definitions provided in the reviews[47–49].

To explore how app evaluation domains have been represented in the literature over time, we analyzed the temporal distribution of domains associated with the five clusters. For each cluster, we aggregated the yearly frequency of domains mentioned in the literature reviews from 2008 to 2024, corresponding to the periods during which the reviews conducted their respective searches. To account for variability in dataset density, we normalized the frequency of each cluster by the total number of mentions across all clusters in each respective year. Additionally, we divided the raw frequency of terms identified in each cluster by the total number of terms associated with that cluster. After normalizing the raw frequencies, we rescaled the values for each cluster to a standardized range between 0 and 1. This rescaling facilitated a direct comparison of trends across clusters independent of their absolute frequencies. Linear regression models were then applied to identify longitudinal trends. The independent variable was the year, while the dependent variable was the normalized frequency of mentions for each cluster.

Ultimately, we sought to manually map the five identified clusters and their respective definitions onto the five evaluation domains of the APA framework given this approach was utilized in the predicate review by Henson et al.

Picos element	Description
Population	This review does not involve specific participants or populations. Instead, it synthesizes existing reviews of mHealth app evaluation frameworks. These reviews focus on frameworks designed to assess app quality across healthcare domains, aiming to provide guidance for clinicians, healthcare providers, and end-users.
Intervention	This review focuses on existing reviews of evaluation frameworks for mobile health (mHealth) applications.
Comparator	Not applicable.
Outcome	Primary outcomes include identifying key quality clusters, analyzing the consistency and variability of app evaluation domains with reviews and documenting changes in cluster focus over time.
Study Design	**Inclusion Criteria**: Systematic and scoping reviews that analyze evaluation frameworks for mHealth applications. Reviews must be written in English and published between January 1, 2008, and December 03, 2024. Studies must have mapped the evaluation questions or criteria of mHealth app evaluation frameworks to distinct domains or quality dimensions. **Exclusion Criteria**: Reviews focusing on single populations such as people with a specific condition (e.g. depression or heart failure), single app types, specific criteria, or single frameworks, as well as reviews limited to one country or region. Conference papers were excluded from the analysis.

Table 2. PICOS methodology for systematic search.

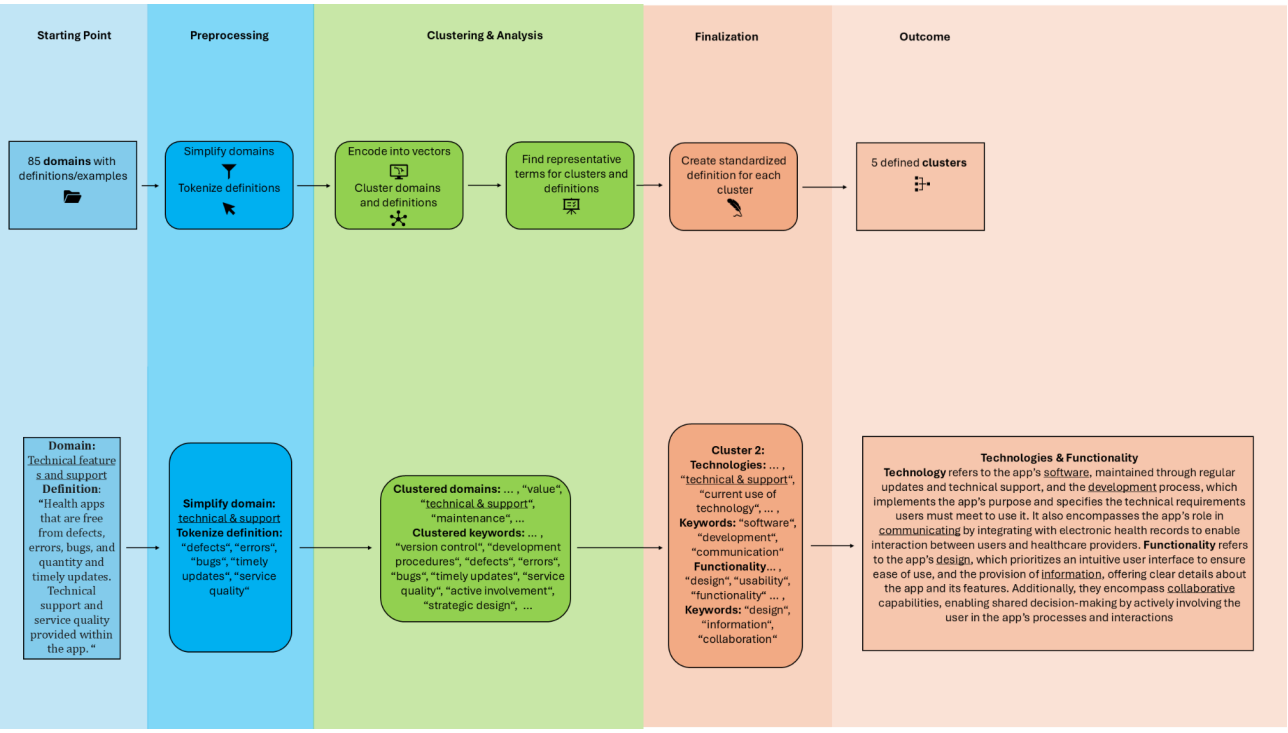

Fig. 5. The figure visualizes the process of clustering and defining the mHealth app evaluation domains used in the included reviews. The top rows shows the theoretical process, the row below presents an example.

Data availability

Data supporting the findings of this study, including details of the included reviews and their characteristics, are available in the supplementary materials.

Received: 16 February 2025; Accepted: 27 March 2025
Published online: 06 April 2025

References

1. Google How google play works (2024). https://play.google/howplayworks/ (2024).
2. Apple, A. App Store. https://www.apple.com/app-store/ (2024).
3. Bates, D. W., Landman, A. & Levine, D. M. Health Apps and Health Policy: What Is Needed? *JAMA* **320**, 1975 (2018).
4. Business of Apps. Health App Revenue and Usage Statistics. (2024). https://www.businessofapps.com/data/health-app-market/?utm_source=chatgpt.com
5. Serpwatch App Usage Statistics 2024: Downloads, Revenue, Popularity. https://serpwatch.io/blog/app-usage-statistics/?utm_source=chatgpt.com (2024).
6. Morrison, I. et al. Is the medicines and healthcare products regulator agency (Mhra) guidance on sodium valproate acceptable to women of childbearing age?? *J. Royal Coll. Physicians Edinb.* **50**, 114–117 (2020).
7. Ghosh, D., Skinner, M. & Ferguson, L. R. The role of the therapeutic goods administration and the medicine and medical devices safety authority in evaluating complementary and alternative medicines in Australia and new Zealand. *Toxicology* **221**, 88–94 (2006).
8. Federal institute for Drugs and Medical Devices. Medical Devices Assessing Risks. Protecting patients.
9. Cruz Rivera, S. et al. Patient-reported outcomes in the regulatory approval of medical devices. *Nat. Med.* **27**, 2067–2068 (2021).
10. Torous, J., Stern, A. D. & Bourgeois, F. T. Regulatory considerations to keep Pace with innovation in digital health products. *Npj Digit. Med.* **5**, 121 (2022).
11. Hensher, M. et al. Scoping review: development and assessment of evaluation frameworks of mobile health apps for recommendations to consumers. *J. Am. Med. Inform. Assoc.* **28**, 1318–1329 (2021).
12. Department of Health Therapeutic Goods Administration. Understanding regulation of software-based medical devices. (2022).
13. U.S & Food & Drug Administration. Policy for Device Software Functions and Mobile Medical Applications. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/policy-device-software-functions-and-mobile-medical-applications (2022).
14. Camacho, E., Cohen, A. & Torous, J. Assessment of mental health services available through smartphone apps. *JAMA Netw. Open.* **5**, e2248784 (2022).
15. Nebeker, C., Torous, J. & Bartlett Ellis, R. J. Building the case for actionable ethics in digital health research supported by artificial intelligence. *BMC Med.* **17**, 137 (2019).
16. Gagnon, L. Time to rein in the wild West of medical apps. *CMAJ* **186**, E247–E247 (2014).
17. Frey, A. L. et al. Domain coverage and criteria overlap across digital health technology quality assessments: a systematic review. *Preprint Https://doi*. https://doi.org/10.31219/osf.io/qg9vd (2024).
18. Ramos, G., Ponting, C., Labao, J. P. & Sobowale, K. Considerations of diversity, equity, and inclusion in mental health apps: A scoping review of evaluation frameworks. *Behav. Res. Ther.* **147**, 103990 (2021).

19. Stoyanov, S. R. et al. Mobile app rating scale: A new tool for assessing the quality of health mobile apps. *JMIR mHealth uHealth*. **3**, e27 (2015).
20. Lagan, S. et al. Actionable health app evaluation: translating expert frameworks into objective metrics. *NPJ Digit. Med.* **3**, 100 (2020).
21. App Advisor An American Psychiatric Association Initiative. The App Evaluation Model. https://www.psychiatry.org/psychiatrists/practice/mental-health-apps/the-app-evaluation-model
22. Moshi, M. R., Tooher, R. & Merlin, T. Suitability of current evaluation frameworks for use in the health technology assessment of mobile medical applications: a systematic review. *Int. J. Technol. Assess. Health Care*. **34**, 464–475 (2018).
23. Busse, R. et al. Best practice in undertaking and reporting health technology assessments: Working group 4 report. *Int. J. Technol. Assess. Health Care*. **18**, 361–422 (2002).
24. Nouri, R., Niakan Kalhori, R., Ghazisaeedi, S., Marchand, M., Yasini, M. & G. & Criteria for assessing the quality of mHealth apps: a systematic review. *J. Am. Med. Inform. Assoc.* **25**, 1089–1098 (2018).
25. Henson, P., David, G., Albright, K. & Torous, J. Deriving a practical framework for the evaluation of health apps. *Lancet Digit. Health*. **1**, e52–e54 (2019).
26. Kumar, S. Text normalization. in Python for Accounting and Finance 133–145 (Springer Nature Switzerland, Cham, doi:https://doi.org/10.1007/978-3-031-54680-8_9. (2024).
27. Jurafsky, D. & Martin, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models* (2025).
28. Laestadius, L., Bishop, A., Gonzalez, M., Illenčík, D. & Campos-Castillo, C. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New. Media Soc.* **26**, 5923–5941 (2024).
29. Terhorst, Y. et al. Validation of the mobile application rating scale (MARS). *PLoS ONE*. **15**, e0241480 (2020).
30. Schlieter, H. et al. Digitale gesundheitsanwendungen (DiGA) Im Spannungsfeld von fortschritt und kritik: Diskussionsbeitrag der fachgruppe „digital health der gesellschaft für informatik e. *V Bundesgesundheitsbl*. **67**, 107–114 (2024).
31. Kopka, M., Camacho, E., Kwon, S. & Torous, J. Exploring how informed mental health app selection May impact user engagement and satisfaction. *PLOS Digit. Health*. **2**, e0000219 (2023).
32. Torous, J., Nicholas, J., Larsen, M. E., Firth, J. & Christensen, H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. *Evid. Based Mental Health*. **21**, 116–119 (2018).
33. IEEE Standard Computer Dictionary. A compilation of IEEE standard computer glossaries. https://doi.org/10.1109/IEEESTD.1991.106963
34. Kiourtis, A., Mavrogiorgou, A. & Kyriazis, D. Aggregating heterogeneous health data through an ontological common health language. in *10th International Conference on Developments in eSystems Engineering (DeSE)* 175–181 (IEEE, 2017). https://doi.org/10.1109/DeSE.2017.9.
35. Lehne, M., Sass, J., Essenwanger, A., Schepers, J. & Thun, S. Why digital medicine depends on interoperability. *Npj Digit. Med.* **2**, 79 (2019).
36. Schulz, S., Stegwee, R. & Chronaki, C. Standards in healthcare data. in Fundamentals of Clinical Data Science (eds Kubben, P., Dumontier, M. & Dekker, A.) 19–36 (Springer International Publishing, Cham, doi:https://doi.org/10.1007/978-3-319-99713-1_3. (2019).
37. Kyriazis, D. et al. The crowdhealth project and the Hollistic health records: collective wisdom driving public health policies. *Acta Inf. Med.* **27**, 369–373 (2019).
38. Giebel, G. D. et al. Quality assessment of mHealth apps: a scoping review. *Front. Health Serv.* **4**, 1372871 (2024).
39. Lagan, S., Sandler, L. & Torous, J. Evaluating evaluation frameworks: a scoping review of frameworks for assessing health apps. *BMJ Open*. **11**, e047001 (2021).
40. Carlo, A. D., Ghomi, H., Renn, R., Areán, P. A. & B. N. & By the numbers: ratings and utilization of behavioral health mobile applications. *Npj Digit. Med.* **2**, 54 (2019).
41. Institute of Data. Cleaning the Corpus: Text Pre-Processing in NLP. https://www.institutedata.com/blog/cleaning-the-corpus-text-pre-processing-in-nlp/ (2024).
42. Galli, C., Donos, N. & Calciolari, E. Performance of 4 Pre-Trained sentence transformer models in the semantic query of a systematic review dataset on Peri-Implantitis. *Information* **15**, 68 (2024).
43. Kurek, J. et al. Zero-Shot recommendation AI models for efficient Job–Candidate matching in recruitment process. *Appl. Sci.* **14**, 2601 (2024).
44. Venkatesh Sharma, K., Ayiluri, P. R., Betala, R., Jagdish Kumar, P. & Shirisha Reddy, K. Enhancing query relevance: leveraging SBERT and cosine similarity for optimal information retrieval. *Int. J. Speech Technol.* **27**, 753–763 (2024).
45. Khan, M. K., Sarker, S., Ahmed, S. M. & Khan, M. H. A. K-Cosine-Means Clustering Algorithm. in *International Conference on Electronics, Communications and Information Technology (ICECIT)* 1–4 (IEEE, Khulna, Bangladesh, 2021). https://doi.org/10.1109/ICECIT54077.2021.9641480.
46. Menzli, A. Tokenization in NLP: Types, Challenges, Examples, Tools. https://neptune.ai/blog/tokenization-in-nlp?utm_source=chatgpt.com (2023).
47. Ribaut et al. Developing a comprehensive list of criteria to evaluate the characteristics and quality of eHealth smartphone apps: systematic review. *JMIR Mhealth Uhealth*. **12**, e48625. https://doi.org/10.2196/48625 (2023).
48. Giebel et al. Quality assessment of digital health applications: protocol for a scoping review. *JMIR Res. Protoc.* **11** (7), e36974. https://doi.org/10.2196/36974 (2022).
49. Delgado-Morales, C. & Duarte-Hueros, A. Una Revisión sistemática de instrumentos que evalúan la calidad de aplicaciones móviles de salud. *Pixel-Bit, Revista de Medios y Educación* **67**, 35–58. https://doi.org/10.12795/pixelbit.97867 (2023).

Acknowledgements

This work was supported by a gift from the Argosy Foundation.

Author contributions

JH and JTo conducted the study. JH, JTa and JTo conducted the primary analysis. JH and JTo wrote the main manuscript. BD, JTa, and NO helped write, edit, and revise the paper. All authors approved the final version.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-96369-w.

Correspondence and requests for materials should be addressed to J.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.