Journal of the American Medical Informatics Association, 28(6), 2021, 1318–1329
doi: 10.1093/jamia/ocab041
Advance Access Publication Date: 31 March 2021
Review

Review

Scoping review: Development and assessment of evaluation frameworks of mobile health apps for recommendations to consumers

Martin Hensher,[1,2] Paul Cooper,[1,2] Sithara Wanni Arachchige Dona,[1,2] Mary Rose Angeles,[1,2] Dieu Nguyen,[1,2] Natalie Heynsbergh,[1,3] Mary Lou Chatterton,[1,2], and Anna Peeters[1]

[1]Institute for Health Transformation, Faculty of Health, Deakin University, Geelong, Victoria, Australia, [2]Deakin Health Economics, School of Health and Social Development, Faculty of Health, Deakin University, Burwood, Victoria, Australia, and[3]School of Nursing and Midwifery, Faculty of Health, Deakin University, Geelong, Victoria, Australia

Corresponding Author: Martin Hensher, BA, MSc, Deakin Health Economics, BC3.110, Deakin University, 221 Burwood Highway, Burwood, VIC 3125, Australia; martin.hensher@deakin.edu.au

Received 13 December 2020; Revised 12 February 2021; Accepted 3 March 2021

ABSTRACT

Objective: The study sought to review the different assessment items that have been used within existing health app evaluation frameworks aimed at individual, clinician, or organizational users, and to analyze the scoring and evaluation methods used in these frameworks.

Materials and Methods: We searched multiple bibliographic databases and conducted backward searches of reference lists, using search terms that were synonyms of "health apps," "evaluation," and "frameworks." The review covered publications from 2011 to April 2020. Studies on health app evaluation frameworks and studies that elaborated on the scaling and scoring mechanisms applied in such frameworks were included.

Results: Ten common domains were identified across general health app evaluation frameworks. A list of 430 assessment criteria was compiled across 97 identified studies. The most frequently used scaling mechanism was a 5-point Likert scale. Most studies have adopted summary statistics to generate the total scoring of each app, and the most popular approach taken was the calculation of mean or average scores. Other frameworks did not use any scaling or scoring mechanism and adopted criteria-based, pictorial, or descriptive approaches, or "threshold" filter.

Discussion: There is wide variance in the approaches to evaluating health apps within published frameworks, and this variance leads to ongoing uncertainty in how to evaluate health apps.

Conclusions: A new evaluation framework is needed that can integrate the full range of evaluative criteria within one structure, and provide summative guidance on health app rating, to support individual app users, clinicians, and health organizations in choosing or recommending the best health app.

Key words: health apps, evaluation framework, scoring and scaling, assessment criteria

INTRODUCTION

Background and significance

Hundreds of thousands of health-related apps are now available on mobile devices, targeted toward almost every conceivable health issue. Health apps have the potential to improve health outcomes, but some authors have called into question the veracity of information provided via such apps[1] and raised the concern that they be of limited or even negative benefit.[2] Given the vast number of apps purporting to help consumers in aspects of their health, a significant challenge for consumers, clinicians, healthcare organizations, and health funders lies in choosing or recommending health apps that are most likely to be of value.[3–5] Despite their potential benefits, health apps can pose potential risks to users such as privacy and security concerns, and even more seriously the provision of incorrect information.[1,2,6]

There has so far been limited oversight by regulatory authorities with respect to health apps that are not associated with medical devices. In Australia, the Therapeutic Goods Administration only regulates those apps which meet the formal definition of a "medical device," leaving a large unregulated or partially regulated zone including a very wide range of other health apps.[7] Mobile app marketplaces (such as Google and Apple developers' guidelines) do not explicitly cover several aspects that might be considered important for health apps, such as veracity of the health information content.[7]

One response to the myriad health apps available in unregulated mobile app marketplaces has been the development of a variety of different evaluation frameworks.[6] However, to date, there is no agreed "gold standard" to evaluate the safety and usability of health apps.[8] Several systematic reviews and narrative reviews have been published in recent years on methods or standards to evaluate health apps using various domains or criteria.[9–13] However, there has not been a deep investigation of the assessment criteria (ie, questions and statements used in frameworks) for domains and scoring mechanisms used, or of the validity and reliability of the assessment methods used by these evaluation frameworks. Previous reviews have illustrated many of the questions used in app evaluation frameworks but did not provide further analysis on the advantages and disadvantages or subjectivity and objectivity of questions in a way that would be useful for developing a general evaluation framework.[10,12,13]

Objectives

The aim of this scoping review is to analyze the different assessment criteria used to evaluate each domain within existing health app evaluation frameworks and to analyze the scoring and evaluation methods used in these frameworks.

MATERIALS AND METHODS

This study's methods are based on Munn et al[14] and follow the 2015 PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines for reporting items[15] and the *JBI Manual for Evidence Synthesis for Scoping Review* (version March 2020).[16]

Search strategy and selection criteria

Medline Complete, CINAHL Complete, PubMed, Embase, Scopus, Google, and Google Scholar were searched from January 2018 to April 2020, reflecting the period after the latest systematic reviews found from a preliminary search.[9,10,12,13,17] The reference lists of the systematic and narrative reviews identified from the systematic database search were screened.[6,9–13,17,18] No limitation was applied for the publication year for this backward searching. Overall, the time frame covered by this review was from 2011 to April 2020. The search terms were synonyms of "health apps," "evaluation," and "frameworks" (Supplementary Appendix S1). Studies were included if they met the following criteria: studies related to health app evaluation frameworks or studies that have elaborated on the scaling or scoring and evaluation mechanisms applied in health app evaluation frameworks, and those studies were included only if they were related to health apps for the general population or mixed users (clinicians and the public). No restriction was applied to study design, disease area(s), or age group. We excluded studies that reported on health apps used only by clinician(s), abstracts, incomplete or ongoing studies, posters, and studies with no full text available.

Data extraction and analysis

The title and abstract of all the articles were divided into 2 groups and screened independently by 2 reviewers (S-W.A.D. and M.R.A.), using EndNote software X9.3.3 (Clarivate Analytics, Philadelphia, PA)for reference management. These were then divided equally between the same reviewers for full-text screening using the Rayyan platform for review management.[19] Excluded and included articles were checked by a third reviewer (D.N.). Any disagreement was discussed with other authors (P.C. and M.H.) to reach consensus. Data extraction was completed by 2 reviewers (S.W.A.D. and M.R.A.) independently and verified by a third reviewer (D.N.): title, authors, published year, study design, study population and sample size, country, app type, study aim, domain, results, journal or database, scaling and scoring modalities, type and numbers of evaluators, subjectivity or objectivity of the appraisal method, and assessment criteria that were included in other frameworks to evaluate the health apps.

Three reviewers (S.W.A.D., M.R.A., and D.N.) conducted a thematic analysis and synthesized the available data. The included assessment criteria that shared similar characteristics were grouped into domains. The domain names were adapted from a previously identified review.[13] Any discrepancies between the 3 reviewers were resolved through discussion with other reviewers (M.H., P.C., A.P., and M.L.C.). Scaling and scoring mechanisms used in the frameworks were also investigated and analyzed.

RESULTS

A total of 2143 studies were screened, and 34 met the inclusion criteria. During the backward reference search of reviews, 63 articles were obtained; 97 studies were therefore included in the final synthesis (Figure 1).

Frameworks identified from the reviewed studies

Table 1 represents the distribution of evaluation framework studies, the majority of which (65%) used self-developed checklists or frameworks. Supplementary Appendix S1 represents the frameworks used in each study and their domains. Studies that utilized self-developed checklists or evaluation frameworks appeared to have based the development of these tools on a combination of literature reviews, clinical or international guidelines, and elements of existing frameworks including the Mobile App Rating Scale (MARS) (n = 1). Consistent with previous reviews,[9] MARS had been reported across included studies more frequently (n = 4) than

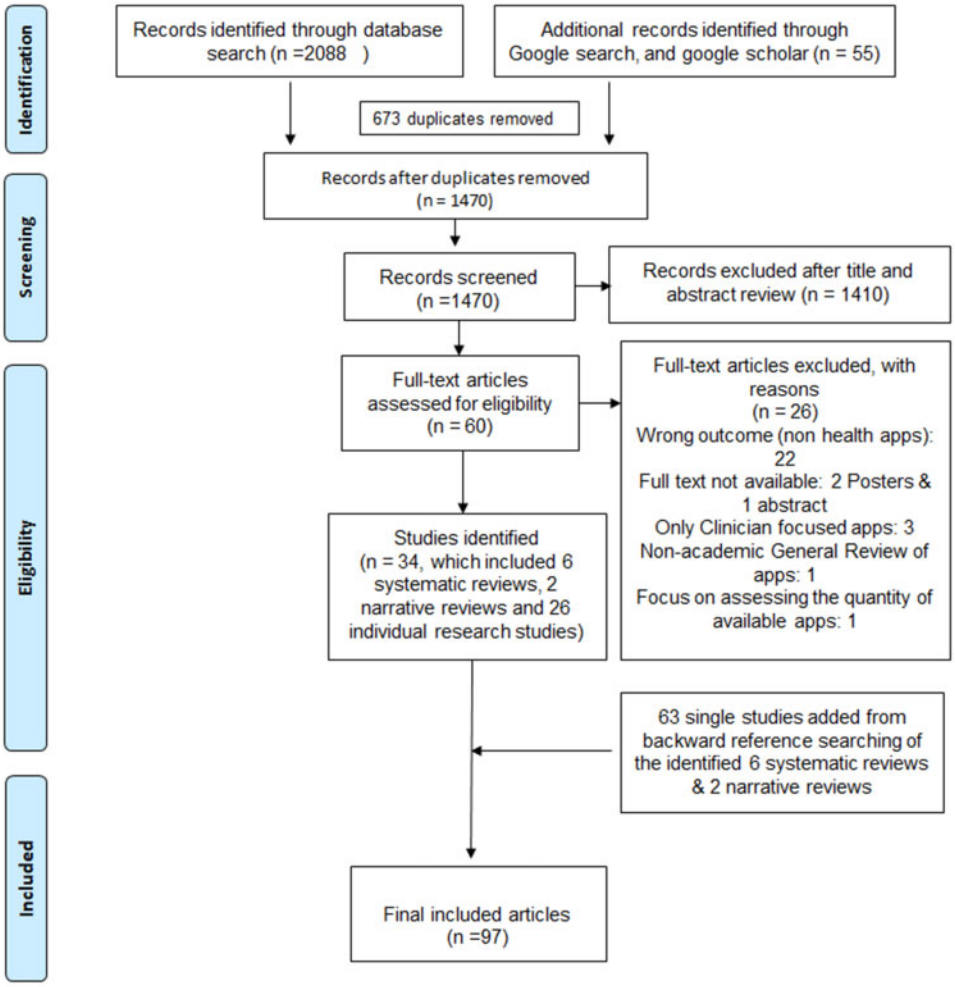

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram.

other frameworks available in the market either as a mean for evaluating health apps (n = 2) or as a guidance to develop new frameworks (n = 2).

Domains

This review identified 10 domains that were frequently used in evaluation frameworks for health apps (Table 2). Domains were identified and defined based on common themes found in the literature. Content/information validity, user experience, user engagement, interoperability, technical features and support, and privacy/security/ethics/legal were common domains assessed in most evaluation frameworks, with more than half of the articles assessing some or all of these domains. Table 3 illustrates the distribution of domains across studies by year. As shown in Table 3, content/information validity and user experience are the most frequently investigated domains across the published studies (see Supplementary Appendix S2 for details).

Assessment criteria used for analyzing health apps

The total number of assessment criteria collected from this literature was 766, with 430 unique criteria after removing duplicates. There were 269 objective questions that could be reviewed via the use of the app, published app's description, or terms and conditions and privacy documents. A total of 161 were subjective assessment criteria that allowed evaluators to review apps based on their perception or intuition. Supplementary Appendix S3 provides a "question bank" of all the identified assessment criteria.

The assessment criteria were themed into categories under each domain. Figure 2 summarizes this coding of the assessment criteria identified across studies and their relationship to the respective 10 final domains. The number of unique assessment criteria identified ranged from 4 (Interoperability domain) to 137 (user experience domain) as shown in Table 4. During the review, we were able to find some assessment criteria that can be used for domains that currently have gaps in evaluation. For example, there is limited evidence on how to evaluate the value domain in the literature. Some assessment criteria that may facilitate the evaluation of the value domain were identified, such as questions related to an app's usefulness in improving patients' quality of life,[20] improve monitoring and management of disease,[20,21] and facilitate healthcare service appointments.[22]

Reviewing the third-party sponsors of an app was deemed important in 2 of the reviewed studies, as sponsorship could provide insights related to conflict of interest that may affect the app developer's credibility.[23,24] There were also assessment criteria to assess the presence of disclaimers relating to risks or adverse events, which could be useful to evaluate an app's legality and safety.[24]

Table 1. The distribution of frameworks across studies

Framework	Name of the framework
1. Studies that used a single existing framework for app evaluation (n = 10)	MARS (n = 1)
	APA (n = 2)
	CRAAP (n = 1)
	ORCHA-24 (n = 1)
	SUMI (n = 1)
	SUS (n = 1)
	Psychological Component Checklist (n = 1)
	A synoptic framework (n = 1)
	The APPLICATION scoring system (n = 1)
2. Studies that self-developed a framework for evaluation (n = 63)	MARS (n = 1)
	uMARS (n = 1)
	The Health IT Usability, Evaluation Model (Health-ITUEM) (n = 1)
	Expert-Based Utility Evaluation (n = 1)
	The APPLICATION scoring system (n = 1)
	App Chronic Disease Checklist (n = 1)
	Nutrition App Quality Evaluation (AQEL) (n = 1)
	Enlight (tool for mobile and Web-based eHealth interventions) (n = 1)
	mHealth Emergency Strategy Index (n = 1)
	MedAd-AppQ Medication Adherence App Quality assessment tool (n = 1)
	Digital Health Scorecard (n = 1)
	Design and Evaluation of Digital Health Intervention Frameworks (n = 1)
	The mobile Health App Trustworthiness (mHAT) checklist) (n = 1)
	Ranked health (n = 1)
	PsyberGuide (n = 1)
	No particular name (n = 48)
2.1 Frameworks that influenced to develop new framework	MARS (n = 2)
	Persuasive system design principles (n = 1)
	Nielsen Usability Model (n = 1)
	Technology Acceptance Model (n = 1)
2.2 guidelines that used to develop new framework	U.S. Public Health Services Clinical Practice Guidelines (n = 1)
	UK BTS/SIGN, U.S. EPR-3, and international GINA guidelines (n = 1)
3. Studies that used a combination of self-developed and existing frameworks for evaluation (n = 6)	Brief DISCERN Instrument (n = 1)
	Silber scale (n = 2)
	Health-ITUES (n = 2)
	Tool used by Cruz—tool for measuring the compliance with Android and iOS guidelines (n = 1)
	Tool for measuring the User QoE by Martines-Perez 2013 (n = 1)
	Abbott Scale for Interactivity (n = 1)
	The Health On the Net Code Criteria (n = 1)
	The Technology Acceptance Model (n = 1)
	Usability framework of TURF (n = 1)
	Chinese Guideline for the Management of Hypertension (n = 1)
	The Anxiety and Depression Association of America (n = 1)
	PsyberGuide (n = 1)
4. Use of survey tools	Use of an existing or self- developed surveys (n = 10)
5. Not relevant	Review or opinion papers (n = 8)

APA: American Psychiatric Association; BTS: British Thoracic Society; CRAAP: Currency, Relevance, Authority, Accuracy, and Purpose; EPR-3: Expert Panel Report 3; GINA: Global Initiative for Asthma; Health-ITUES: Health Information Technology Usability Evaluation Scale; MARS: Mobile App Rating Scale; ORCHA-24: Organisation for the Review of Care and Health Applications–24-Question Assessment; SIGN: Scottish Intercollegiate Guidelines Network; SUMI: Standardized Software Usability Measurement Inventory; SUS: System Usability Scale; TURF: Task, User, Representation and Function;

In the reviewed frameworks, some assessment criteria were designed to be answered by expert reviewers (n = 15). For instance, some assessment criteria were technical, which were most suitable for evaluators with academic, information technology, or clinical backgrounds. Other assessment criteria were too general, vague, or nonspecific to be useful.[23,25] Some focused on health apps for specific conditions or issues, such as mental health, pregnancy, diabetes, asthma, and chronic disease, while other assessment criteria were for more general health or wellness apps. However, assessment criteria with no focus on specific health conditions were found to be useful for general health app evaluation frameworks. For instance, De Sousa Gomes et al[26] did not use disease-specific questions in their framework evaluating mobile apps for health promotion of pregnant women with preeclampsia.

Table 2. Commonly identified domains from health app evaluation frameworks

No	Domain	Coverage/definition
01	Clarity of purpose of the app	A clear statement of the intended purpose of the app as well as the specificity of the users or the disease.
02	Developer credibility	Transparency of the app development and testing process, and accountability and credibility of the app developer, funders, affiliations, and sponsors.
03	Content/information validity	Readability, credibility, characteristics, quality, and accuracy of the information in the health app. The ability to tailor the app content per user preference and using simple language.
04	User experience	The overall experience of using an app in terms of its user friendliness, design features, functionalities, and ability to consider user preference through personalization function.
05	User engagement/adherence and social support	The extent of how apps maintain user retention using functionalities such as gamification, forums, and the use of behavior techniques as well as the extent of social support.
06	Interoperability	Data sharing and data transfer capabilities of the health apps.
07	Value	Perceived benefits and advantages associated with the use of health app.
08	Technical features and support	Health apps that are free from defects, errors, bugs, and quantity and timely updates. Technical support and service quality provided within the app.
09	Privacy/security/ethical/legal	Privacy and security domains pertain to data protection, cybersecurity, and encryption mechanisms for the storage and data transmission. Legalities of the health app that look at whether the health apps adhere to guidelines and have disclaimers concerning on clinical accountability.
10	Accessibility	This pertains to the ability of health apps to capture a wider audience and bridge the gap in access to health apps and healthcare services for vulnerable populations/people with disabilities.

Scaling and rating mechanisms

Frameworks have used different methods for scoring and rating assessment criteria (Figure 3; Supplementary Appendix S4). The most frequently reported scaling method was a point system (n = 34). Twenty-two studies used a 5-point Likert scale for each assessment criteria.[20,26–46] The other scales used were 3-point (n = 6),[47–52] 4-point (n = 2),[22,53] 7-point (n = 3),[54–56] or 10-point (n = 1)[21] scales or dichotomous questions (n = 13) that were answerable by a "yes or no" option or "presence or absence" option.[23,25,57–67] Nineteen studies used a mixed approach, which included a combination of point scales (2-, 3-, 5-, and 7-point scales), dichotomous type, and open-ended questions.[68–86]

Eight studies did not use numerical values in their evaluation; rather, they were filter based (n = 2)[3,23] or checked against set criteria or availability of the items (n = 1),[24] descriptive analysis (n = 2),[87,88] scorecard based with no explanation on scoring (n = 1),[89] qualitative methods such as review of user comments (n = 1),[90] or pictorial schemes (n = 1).[91] Other studies did not elaborate on their scaling method (n = 23).[2,4–6,9–13,17,92–104]

For the scoring modalities, the most popular approach taken was the calculation of mean or average scores (n = 22 studies, 23% of the total number of studies) (Figure 4; Supplementary Appendix S4).[20,21,27,30,32,33,35,37–39,41,43,45,46,50,54,56,70,71,81,98] Thirteen (13%) studies presented their scores as a sum or total, and 11 (11%) studies used a mixed of mean, median, interquartile range, percentage, or total scoring.[25,26,36,53,57,60,61,64,65,67,72–74] Nine (9%) studies employed different approaches such as adjustment of scores, percentage scoring, interquartile, frequency count, and summation of ordinal answers.[28,29,48,62,68,69,75,83,84] Six (6%) studies did not employ any scoring mechanism.[3,23,24,89–91] Thirty-six (37%) studies did not report the scoring mechanism or its reporting was not applicable (reviews or opinion articles).[2,4–6,9–13,17,22,23,34,40,42,44,52,55,58,66,78,79,86,88,92–97,99–104]

Most of the frameworks (n = 49) calculated the total score using equal weighting across domains, while 6 studies calculated the app's scores using different weightings of domains.[33,51,72–74,98] The weighted scores were mainly based on the primary goal of the evaluation framework. For instance, higher weights were allocated to the

Table 3. Distribution of domains discussed across studies by year

Domain	Number of studies reported on each domain by year									
	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Clarity of purpose of the app	0	1	0	2	1	2	2	5	4	2
Developer credibility	0	1	0	3	1	6	1	3	6	3
Content/information validity	1	2	2	5	8	10	4	13	9	4
User experience	1	2	4	8	10	13	9	16	12	5
User engagement/adherence and social support	1	0	1	1	3	8	3	6	8	1
Interoperability	0	2	1	0	1	1	1	6	3	2
Value	0	1	1	3	3	5	5	6	5	2
Technical features and support	0	1	1	1	1	2	1	6	7	2
Privacy/security/ethical/legal	0	1	2	1	2	3	2	11	7	3
Total identified studies	1	3	4	10	11	16	10	20	16	6

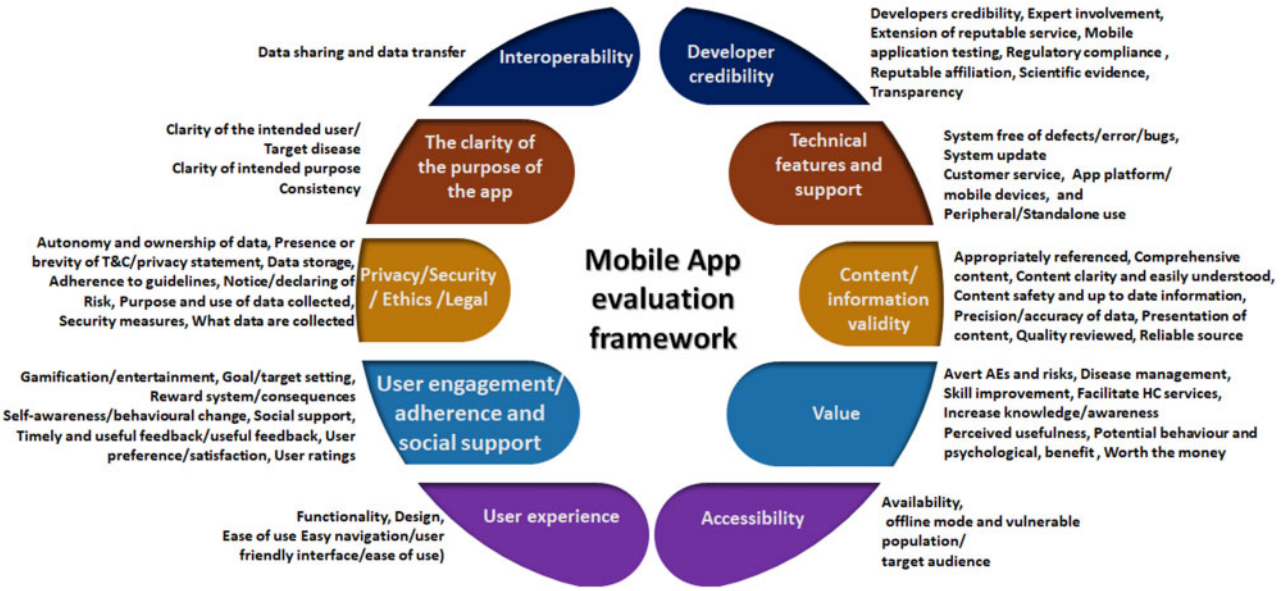

Figure 2. Categories of app assessment criteria respective to identified domains. AE: adverse events; HC: health care.

content (20%), transparency (20%), and evidence (60%) in Butcher et al's[72] framework as their objective was to evaluate the quality, validity, and reliability of the resources used in apps.

Six steered away from using numerical values and did not use any scoring or ranking system.[3,23,52,58,68,103] Non-numeric approaches included a filter approach, narrative review, categorical assessment, or a requirement or criteria-based approach.[3,23,58,103] A pyramid approach was one method used in filtering apps, and none of the studies that employed this approach incorporated a scoring scale.[3,104] For example, the American Psychiatric Association (APA) App Evaluation Framework adopted a pyramid approach, which filtered apps based on 5 levels from background information (level 1) at the bottom to data integration or data sharing (level 5) at the top of the pyramid.

In terms of resourcing the process of assessment, evaluators were either the authors, end users, experts in information technology, or health professionals. Most of the evaluation studies (n = 29) were assessed by end users, while some studies utilized either experts in the field (n = 9), other professionals (n = 3) or authors (n = 24) as evaluators. Five studies used various mixes of these evaluator types. Twenty-seven studies did not elaborate on the type of evaluators. Sixty studies used a minimum of 2 reviewers, mostly with a third reviewer to resolve discrepancies as a strategy to ensure accurate responses. One study developed a "user manual."[62] Interrater reliability testing (ie, the degree of agreement between raters) to address consistency between 2 assessors was undertaken in 28 of 97 studies, and 9 analyzed internal consistency (ie, extent to which the items of a framework measures the same construct) to address the reliability of the framework or scale.[12,25,29–31,33,35–37,42,44,45,48,49,53,54,59,61–63,66,69,71,73,75,79,81,85,90,92,103] The content validity index defined as "to identify the extent to which a scale has an appropriate sample of items to represent the construct of interest" was used in 2 studies.[32,53]

The process and timing of evaluating apps varied across studies. Three studies explicitly timed their use of health apps for the purpose of evaluation, while the rest did not provide further details. Wisniewski et al,[52] Torous et al,[4] and Mani et al[37] allowed use of the app for 10, 15, or 30 minutes, respectively, to obtain information about the app prior to evaluation.

We also identified a number of strategies to ensure accurate responses to assessment criteria. These included involving 2 or more assessors for the evaluation, or considering the following strategies:

Table 4. The number of unique assessment criteria per domain

Domain	Number of unique questions per domain	Number of objective questions	Number of subjective questions
Clarity of purpose of the app	13	10	3
Developer credibility	24	23	1
Content/information validity	77	52	25
User experience	137	75	62
User engagement/adherence and social support	51	24	27
Interoperability	4	3	1
Value	48	15	33
Technical features and support	14	13	1
Privacy/security/ethical/legal	51	43	8
Accessibility	11	11	0
Total	430	269	161

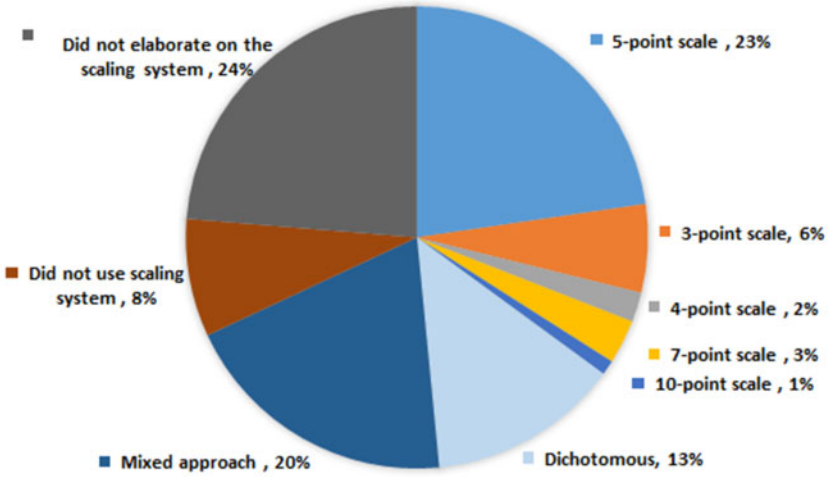

Figure 3. Frequency distribution of evaluative scaling methods (N = 97).

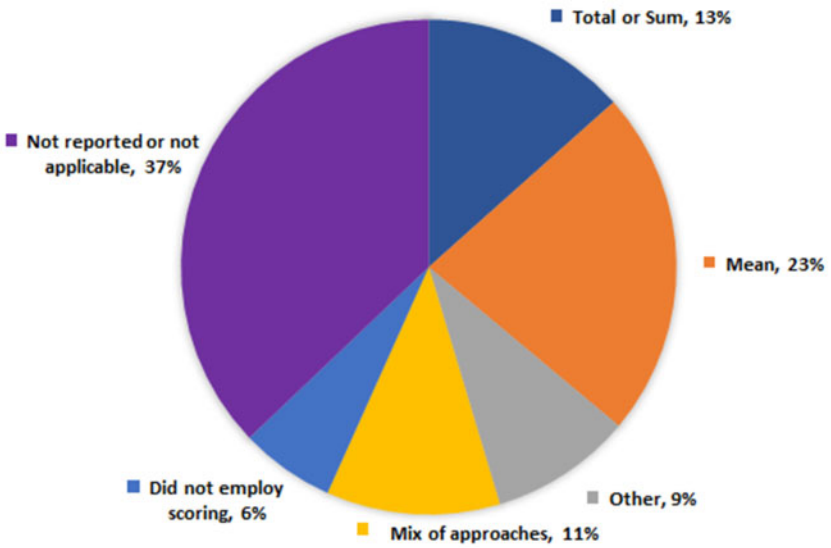

Figure 4. Frequency distribution of scoring mechanisms (N = 97).

reviewing the terms and conditions, privacy statement, and app description; undertaking a literature search for further investigation of content validity; using the readability statistics within Microsoft Word (Microsoft, Redmond, WA) to review the readability of the content (ie, Flesch-Kincaid Grade Level)[83]; reviewing the app metrics; using benchmark criteria to properly scale or score the

domains; and downloading and installation of apps to further investigate the key health app domains.

DISCUSSION

This study analyzes and reports for the first time 430 unique assessment criteria used in existing health app evaluation frameworks. We identified 10 unique domains that represent the breadth and specificity of the various existing frameworks. While many studies used similar overall domains, there was little uniformity in the precise components of each domain. Our review also identified the assessment criteria required within each domain, along with a variety of scoring modalities. Finally, our review identified a number of key principles and processes for a health app evaluation framework to ensure usability, reliability and internal consistency.

Our analysis suggests that there is considerable flexibility within frameworks and that organization of domains is not a standardized process. For example, some have incorporated behavior change techniques under the design and functionalities domain, but some have included it under engagement. However, based on the most common themes that emerged, our review identified 10 domains that can be used for a future framework.

One of the key gaps identified in frameworks across articles was the lack of or difficulty in assessing the value domain (often referred to as perceived value in the evaluation frameworks). This is due to the current landscape of health apps (fast and evolving market and subjectivity of value), and because studies to demonstrate apps' efficacy and value for money are often not undertaken.[36] In addition, our findings indicated that no existing framework has considered assessing all the domains we identified in one structure. For example, privacy and security domain was not included in MARS. Self-developed frameworks or checklists by reviewed studies also did not cover all the domains. This suggests the potential to improve on existing frameworks by developing a comprehensive approach that includes all the domains identified in this review.

In the studies we reviewed, most of the assessment criteria used in the evaluation frameworks were objective in nature. More effective assessment criteria displayed clarity and comprehensiveness of structure, to enhance readability and understandability for the app assessor. Our review highlights the advantage of using properly constructed assessment criteria in app evaluation frameworks to facilitate app assessors' ability to understand them clearly, concisely, and more easily, enabling a more easily replicable evaluation of apps. The unique assessment criteria we identified can be useful in developing app evaluation frameworks as well as developing guidelines for framework users. We found that assessment criteria used across frameworks were not always understandable by nonexperts and some were disease specific.

For the scoring modality, the point-scale method was the most popular approach, using numerical values to facilitate health apps evaluation. The most common scale used was 5-point scale, and some authors suggested that simpler scales yield greater validity while bigger scales pose response bias resulting in lower data quality.[52,105,106] Summary statistics such as mean or total of these scales were generally used for scoring, and weighting based on the importance of domains was also adopted by some studies.[35,59,73–75] However, there are arguments over scoring systems due to interrater reliability issues,[104] and debates over whether different rating scales are more likely to increase response bias.[105] For example, Torous et al[6] discussed several existing frameworks (MARS, APA, Enlight, PsyberGuide, Anxiety and Depression Associaton of America) and pointed to the limitation of these frameworks as having lower interrater reliability in practice. Therefore, it is not surprising that some authors tried to evaluate apps without adopting a scoring approach. However, the outcome of such evaluations often seemed to be subjective, which can reduce the credibility and validity of the evaluation process such as in the pyramid approach. Our findings are consistent with previous studies, which highlighted the importance of a point-scale approach.[6]

A contemporary example of using point-scale and mean scoring for domains is MARS,[45] which is a frequently adopted general health app evaluation framework among existing established frameworks.[9] Other examples of point scales are the "Health Protected Information" checklist, Design and Evaluation of Digital Health Intervention Frameworks,[102] Ranked Health,[34] and PsyberGuide[78]; however, they lacked a clear explanation of scoring, which was a key limitation of several studies reviewed. Outside the research literature, commercial evaluation frameworks may be more likely to have rather opaque methods and scoring systems—yet, lack of transparency is an obvious criticism that needs to be answered.

There was little consistency in terms of composition of evaluators in applying these frameworks across studies. While end users were the evaluators in most studies, some of the evaluations were conducted by researchers related to the study, and assessment criteria in the frameworks were designed to be answered only by expert reviewers or researchers, which makes it difficult for the public to evaluate apps. Most end user frameworks were self-developed by researchers to answer their study objectives. One established framework, MARS, was later modified by its developers as uMARS, with the aim of reducing technical content to facilitate ease of use[44] Only 3 frameworks utilized a mix of end users and experts or researchers.[76,77,84] The other methods used were interviews or focus groups or surveys to receive user feedback. Developing a framework that incorporates all the domains identified as important by this review, and which is suitable for any evaluator to use, is a challenge that will need to be overcome in the future. In addition, interrater reliability and internal consistency were measured in 37 studies to ensure agreement among the app raters. Verification may depend on the evaluator: certain assessment criteria used in frameworks were verifiable only by experts or researchers but not by the end user. Future frameworks can be validated and improved by undertaking pilot testing with randomly selected apps and statistical analyzes such as interrater reliability or internal consistency.[44]

Other limitations of the currently available frameworks also show the need for improvement and the potential for development of a new framework. The MARS and uMARS frameworks did not evaluate the privacy and security domains, even though privacy and security are integral domains in health app evaluation, as protecting users' information is required by law.[3,10] It was evident that self-developed frameworks were not always subjected to a validation process. Our findings are consistent with previous studies that highlighted the limitations of various evaluation frameworks.[11] These included the uncertainty of the validity and the reliability of the self-developed checklists, the subjective nature of the assessment of the raters, disparities in the results due to the setting that was considered during the evaluation of the apps (ie, clinical setting), and the applicability of behavioral change theories employed in the framework.[57] Mathews et al[89] also recognized that their Digital Scorecard framework may not be useful for a specific context (payers' perspective) because it was mainly for supporting development of digital health products that bring maximum benefits to users, but such product developments could be costly and may not

be practical. The pyramid filtering style adopted from the APA that was highlighted in some studies[3,104] had disadvantages, such as its dependence on the original choice and ordering of priority domains within the pyramid, and the evaluator's subjective assessment, although this approach provides a streamlined process via a filtering method and its visual illustration that facilitates ease in evaluation. Another limitation encountered in the existing frameworks was the lack of clear descriptions of the methodology underpinning framework scaling and scoring modalities. Therefore, a thorough description is needed for a future framework.

A limitation of our review's methodology was that we did not consider commercially available app evaluation guidelines or frameworks that were not indexed in our search resources (databases and reference lists). Another limitation was that we restricted our search to studies published in English: therefore, non-English evidence was not reviewed.

CONCLUSION

Our scoping review is part of a larger research project developing a general health app evaluation framework for Australian individual- or mixed-user (individual and clinician) applications and for healthcare organizations, which will be validated through interrater reliability or internal consistency testing and published upon its completion. Our review suggests that a new evaluative framework is needed that can integrate the full range of evaluative criteria within one structure, and provide summative guidance on health app assessment, to support choosing or recommending the best health app for individual app users and health organizations.

Findings of this scoping review have important implications that lead us to make the following recommendations.

1. An ideal health app evaluation framework should integrate the 10 identified domains within one structure, to support individual users or organizations in choosing the best health apps for disease management and promoting healthy lifestyles. This would overcome the limitations of earlier frameworks and would cover the evaluation of health apps for quality, safety, and patient's utility.
2. Evaluation criteria to assess an app should be clear, concise, specific, and objective. Our study has collated a library of specific assessment criteria from the studies we reviewed. Our reference "question bank" should be used in drafting assessment criteria for domains as a guide.
3. The selection of assessment criteria from the "question bank" for app evaluation frameworks should be carefully conducted based on factors including but not limited to the structure, depth, and expected outcome from the assessment criteria, and its subjectivity or objectivity, because individual perceptions on the quality of the assessment criteria can vary from one end user to another.
4. A comprehensive objective framework requires future testing on various platforms across many health conditions to determine a low-burden approach to completing health app assessments cheaply and efficiently.

To conclude, challenges exist in the investigation of health apps due to the absence of a comprehensive and "gold standard" evaluation framework. To date, there is no universal and rigorous framework to investigate health apps that encompasses all the domains that our scoping review identified as important for testing. Our review has demonstrated considerable diversity of approaches and rigor with respect to the systematic use of assessment criteria, scoring, and rating methodologies in the field of health app evaluation.

FUNDING

This work was supported by a Medibank Better Health Foundation grant ("Developing and piloting a framework to evaluate Health APPs to enable the promotion of a curated set of evidence-based Health APPs to consumers in the Australian setting").

AUTHOR CONTRIBUTIONS

MH, PC, SWAD, MRA, and DN made substantial contributions to the conception, design, acquisition, drafting, and critical revisions of the literature review. AP, MLC, and NH made substantial contributions to the analysis and interpretation of data. All authors critically reviewed the manuscript. All authors provided final approval and agree to be accountable for all aspects of this work.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

DATA AVAILABILITY STATEMENT

The data underlying this article are available in the article and in its online supplementary material.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Schulke DF. The regulatory arms race: mobile-health applications and agency posturing. *Boston Univ Law Rev* 2013; 93 (5): 1699–752.
2. Zhang Y, Li X, Luo S, *et al.* Exploration of users' perspectives and needs and design of a type 1 diabetes management mobile app: mixed-methods study. *JMIR Mhealth Uhealth* 2018; 6 (9): e11400–e00.
3. Henson P, David G, Albright K, *et al.* Deriving a practical framework for the evaluation of health apps. *Lancet Digit Health* 2019; 1 (2): e52–4.
4. Torous J, Andersson G, Bertagnoli A, *et al.* Towards a consensus around standards for smartphone apps and digital mental health. *World Psychiatry* 2019; 18 (1): 97–8.
5. Wyatt J. How can clinicians, specialty societies and others evaluate and improve the quality of apps for patient use? *BMC Med* 2018; 16 (1): 225.
6. Torous J, Firth J, Huckvale K, *et al.* The emerging imperative for a consensus approach toward the rating and clinical recommendation of mental health apps. *J Nerv Ment Dis* 2018; 206 (8): 662–6.
7. Department of Health Therapeutic Goods Administration. Regulation of software as a medical device. Australian Government. 2020. https://www.tga.gov.au/regulation-software-medical-device. Accessed May 1, 2020.
8. Schoenfeld AJ, Sehgal NJ, Auerbach A. The challenges of mobile health regulation. *JAMA Intern Med* 2016; 176 (5): 704–5.
9. Azad-Khaneghah P, Neubauer N, Miguel CA, *et al.* Mobile health app usability and quality rating scales: a systematic review. *Disabil Rehabil Assistive Technol* 2020 Jan 8 [E-pub ahead of print].
10. Llorens-Vernet P, Miró J. Standards for mobile health-related apps: systematic review and development of a guide. *JMIR Mhealth Uhealth* 2020; 8 (3): e13057.

11. McKay FH, Cheng C, Wright A, *et al.* Evaluating mobile phone applications for health behaviour change: a systematic review. *J Telemed Telecare* 2018; 24 (1): 22–30.
12. Moshi MR, 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* 2018; 34 (5): 464–75.
13. Nouri R, R Niakan Kalhori S, Ghazisaeedi M, *et al.* Criteria for assessing the quality of mHealth apps: a systematic review. *J Am Med Inform Assoc* 2018; 25 (8): 1089–98.
14. Munn Z, Peters MDJ, Stern C, *et al.* Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Med Res Methodol* 2018; 18 (1): 143.
15. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). PRISMA Statement 2015. http://www.prisma-statement.org/PRISMAStatement/CitingAndUsingPRISMA. Accessed May 15, 2020.
16. Peters MDJ, Godfrey C, McInerney P, Munn ZTricco, ACKhalil, H. Chapter 11: scoping reviews (2020 version). In: Aromataris E, Munn Z, eds. *JBI Manual for Evidence Synthesis*. 2020. https://wiki.jbi.global/display/MANUAL/Chapter+11%3A+Scoping+reviews. Accessed May 15, 2020.
17. Jeminiwa RN, Hohmann NS, Fox BI. Developing a theoretical framework for evaluating the quality of mhealth apps for adolescent users: a systematic review. *J Pediatr Pharmacol Ther* 2019; 24 (4): 254–69.
18. Scott K, Deborah R, Rajindra A. A review and comparative analysis of security risks and safety measures of mobile health apps. *Aust J Inform Syst* 2015; 19: 1210.
19. Ouzzani M, Hammady H, Fedorowicz Z, *et al.* Rayyan-a web and mobile app for systematic reviews. *Syst Rev* 2016; 5 (1): 210.
20. Martínez-Pérez B, De la Torre-Díez I, López-Coronado M. Experiences and results of applying tools for assessing the quality of a mHealth App named heartkeeper. *J Med Syst* 2015; 39 (11): 142.
21. Shah N, Jonassaint J, De Castro L. Patients welcome the sickle cell disease mobile application to record Symptoms via Technology (SMART). *Hemoglobin* 2014; 38 (2): 99–103.
22. Lim D, Norman R, Robinson S. Consumer preference to utilise a mobile health app: a stated preference experiment. *PLoS One* 2020; 15 (2): e0229546–e46.
23. Scott IA, Scuffham P, Gupta D, *et al.* Going digital: a narrative overview of the effects, quality and utility of mobile apps in chronic disease self-management. *Aust Health Rev* 2020; 44(1): 62–82.
24. Huckvale K, Car M, Morrison C, *et al.* Apps for asthma self-management: a systematic assessment of content and tools. *BMC Med* 2012; 10 (1): 144.
25. Leigh S, Ouyang J, Mimnagh C. Effective? Engaging? Secure? Applying the orcha-24 framework to evaluate apps for chronic insomnia disorder. *Evid Based Mental Health* 2017; 20 (4): e20.
26. De Sousa Gomes ML, Rios Rodrigues I, dos Santos Moura N, *et al.* Evaluation of mobile Apps for health promotion of pregnant women with preeclampsia. *Acta Paulista Enfermagem* 2019; 32 (3): 275–81.
27. Al Ayubi SU, Parmanto B, Branch R, *et al.* A persuasive and social mHealth application for physical activity: a usability and feasibility study. *JMIR Mhealth Uhealth* 2014; 2 (2): e25.
28. Alnasser A, Kyle J, Alkhalifah A, *et al.* Relationship between evidence requirements, user expectations, and actual experiences: usability evaluation of the Twazon Arabic weight loss app. *JMIR Hum Factors* 2018; 5 (2): e16.
29. Carpenter DM, Geryk LL, Sage A, *et al.* Exploring the theoretical pathways through which asthma app features can promote adolescent self-management. *Transl Behav Med* 2016; 6 (4): 509–18.
30. Cho MJ, Sim JL, Hwang SY. Development of smartphone educational application for patients with coronary artery disease. *Healthc Inform Res* 2014; 20 (2): 117–24.
31. Demidowich AP, Lu K, Tamler R, *et al.* An evaluation of diabetes self-management applications for Android smartphones. *J Telemed Telecare* 2012; 18 (4): 235–8.
32. Gaziel-Yablowitz M, Schwartz D. A review and assessment framework for mobile-based emergency intervention apps. *ACM Comput Surv* 2018; 51 (1): 1–32.
33. Guo Y, Bian J, Leavitt T, *et al.* Assessing the quality of mobile exercise apps based on the American College of Sports Medicine Guidelines: a reliable and valid scoring instrument. *J Med Internet Res* 2017; 19 (3): e67.
34. Hacking Medicine Institute. RANKED-Curated apps providers can recommend, and patients can use. 2020. http://www.rankedhealth.com/approach/. Accessed May 1, 2020.
35. Kuhn E, Greene C, Hoffman J, *et al.* Preliminary evaluation of PTSD coach, a smartphone app for post-traumatic stress symptoms. *Mil Med* 2014; 179 (1): 12–8.
36. Lin Y-H, Guo J-L, Hsu H-P, *et al.* Does "hospital loyalty" matter? Factors related to the intention of using a mobile app. *Patient Prefer Adherence* 2019; 13: 1283–94.
37. Mani M, Kavanagh DJ, Hides L, *et al.* Review and evaluation of mindfulness-based iPhone apps. *JMIR Mhealth Uhealth* 2015; 3 (3): e82.
38. Martínez-Pérez B, De la Torre-Díez I, Candelas-Plasencia S, *et al.* Development and evaluation of tools for measuring the quality of experience (QoE) in mHealth applications. *J Med Syst* 2013; 37 (5): 9976.
39. Mattson DC. Usability evaluation of the digital anger thermometer app. *Health Inform J* 2017; 23 (3): 234–45.
40. Meedya S, McGregor D, Halcomb E, *et al.* Developing and testing a mobile application for breastfeeding support: the Milky Way application. *Women Birth* 2021; 34 (2): e196–203.
41. Rizvi SL, Hughes CD, Thomas MC. The DBT Coach mobile application as an adjunct to treatment for suicidal and self-injuring individuals with borderline personality disorder: a preliminary evaluation and challenges to client utilization. *Psychol Serv* 2016; 13 (4): 380–8.
42. Sage A, Roberts C, Geryk L, *et al.* A self-regulation theory–based asthma management mobile app for adolescents: a usability assessment. *JMIR Hum Factors* 2017; 4(1): e5.
43. Spook JE, Paulussen T, Kok G, *et al.* Monitoring dietary intake and physical activity electronically: feasibility, usability, and ecological validity of a mobile-based ecological momentary assessment tool. *J Med Internet Res* 2013; 15 (9): e214.
44. Stoyanov SR, Hides L, Kavanagh DJ, *et al.* Development and validation of the User Version of the Mobile Application Rating Scale (uMARS). *JMIR Mhealth Uhealth* 2016; 4 (2): e72.
45. Stoyanov SR, Hides L, Kavanagh DJ, *et al.* Mobile app rating scale: a new tool for assessing the quality of health mobile apps. *JMIR Mhealth Uhealth* 2015; 3 (1): e27.
46. Wang C-J, Chaovalit P, Pongnumkul S. A breastfeed-promoting mobile app intervention: usability and usefulness study. *JMIR Mhealth Uhealth* 2018; 6 (1): e27.
47. Abroms LC, Padmanabhan N, Thaweethai L, *et al.* iPhone apps for smoking cessation: a content analysis. *Am J Prev Med* 2011; 40 (3): 279–85.
48. Anderson K, Burford O, Emmerton L. App chronic disease checklist: protocol to evaluate mobile apps for chronic disease self-management. *JMIR Res Protoc* 2016; 5 (4): e204.
49. Cruz Zapata B, Hernández Niñirola A, Idri A, *et al.* Mobile PHRs compliance with android and iOS usability guidelines. *J Med Syst* 2014; 38 (8): 81.
50. O'Malley G, Dowdall G, Burls A, *et al.* Exploring the usability of a mobile app for adolescent obesity management. *JMIR Mhealth Uhealth* 2014; 2 (2): e29.
51. Robustillo Cortés MA, Cantudo Cuenca MR, Morillo Verdugo R, *et al.* High quantity but limited quality in healthcare applications intended for HIV-infected patients. *Telemed J E Health* 2014; 20 (8): 729–35.
52. Wisniewski H, Liu G, Henson P, *et al.* Understanding the quality, effectiveness and attributes of top-rated smartphone health apps. *Evid Based Ment Health* 2019; 22 (1): 4–9.
53. Jin M, Kim J. Development and evaluation of an evaluation tool for healthcare smartphone applications. *Telemed J E Health* 2015; 21 (10): 831–7.

54. Barrio P, Ortega L, López H, *et al.* Self-management and shared decision-making in alcohol dependence via a mobile app: a pilot study. *Int Behav Med* 2017; 24 (5): 722–7.
55. English LL, Dunsmuir D, Kumbakumba E, *et al.* The PAediatric Risk Assessment (PARA) mobile app to reduce postdischarge child mortality: design, usability, and feasibility for health care workers in Uganda. *JMIR Mhealth Uhealth* 2016; 4 (1): e16.
56. Price M, Sawyer T, Harris M, *et al.* Usability evaluation of a mobile monitoring system to assess symptoms after a traumatic injury: a mixed-methods study. *JMIR Ment Health* 2016; 3 (1): e3.
57. Bachiri M, Idri A, Fernández-Alemán JL, *et al.* Evaluating the privacy policies of mobile personal health records for pregnancy monitoring. *J Med Syst* 2018; 42 (8): 144.
58. Chernetsky Tejedor S, Sharma J, Lavallee DC, *et al.* Identification of important features in mobile health applications for surgical site infection surveillance. *Surg Infect (Larchmt)* 2019; 20 (7): 530–4.
59. Hoppe CD, Cade JE, Carter M. An evaluation of diabetes targeted apps for Android smartphone in relation to behaviour change techniques. *J Hum Nutr Diet* 2017; 30 (3): 326–38.
60. Iribarren SJ, Schnall R, Stone PW, *et al.* Smartphone applications to support tuberculosis prevention and treatment: review and evaluation. *JMIR Mhealth Uhealth* 2016; 4 (2): e25.
61. Jeon E, Park H-A, Min YH, *et al.* Analysis of the information quality of Korean obesity-management smartphone applications. *Healthc Inform Res* 2014; 20 (1): 23–9.
62. McMillan B, Hickey E, Patel MG, *et al.* Quality assessment of a sample of mobile app-based health behavior change interventions using a tool based on the National Institute of Health and Care Excellence behavior change guidance. *Patient Educ Couns* 2016; 99 (3): 429–35.
63. Middelweerd A, Mollee JS, van der Wal CN, *et al.* Apps to promote physical activity among adults: a review and content analysis. *Int J Behav Nutr Phys Act* 2014; 11 (1): 97.
64. Portelli P, Eldred C. A quality review of smartphone applications for the management of pain. *Br J Pain* 2016; 10 (3): 135–40.
65. van Haasteren A, Gille F, Fadda M, *et al.* Development of the mHealth App Trustworthiness checklist. *Digit Health* 2019; 5:2055207619886463.
66. Vollmer Dahlke D, Fair K, Hong YA, *et al.* Apps seeking theories: results of a study on the use of health behavior change theories in cancer survivorship mobile apps. *JMIR Mhealth Uhealth* 2015; 3 (1): e31.
67. Xiao Q, Wang Y, Sun L, *et al.* Current status and quality assessment of cardiovascular diseases related smartphone apps in China. *Stud Health Technol Inform* 2016; 225: 1030–1.
68. Ainsworth MC, Pekmezi D, Bowles H, *et al.* Acceptability of a mobile phone app for measuring time use in breast cancer survivors (Life in a Day): mixed-methods study. *JMIR Cancer* 2018; 4 (1): e9.
69. Ali EE, Teo AKS, Goh SXL, *et al.* MedAd-AppQ: A quality assessment tool for medication adherence apps on iOS and android platforms. *Res Soc Admin Pharm* 2018; 14 (12): 1125–33.
70. Arnhold M, Quade M, Kirch W. Mobile applications for diabetics: a systematic review and expert-based usability evaluation considering the special requirements of diabetes patients age 50 years or older. *J Med Internet Res* 2014; 16 (4): e104.
71. Baumel A, Faber K, Mathur N, *et al.* Enlight: a comprehensive quality and therapeutic potential evaluation tool for mobile and web-based ehealth interventions. *J Med Internet Res* 2017; 19 (3): e82.
72. Butcher R, MacKinnon M, Gadd K, *et al.* development and examination of a rubric for evaluating point-of-care medical applications for mobile devices. *Med Ref Serv Q* 2015; 34 (1): 75–87.
73. Chen J, Cade JE, Allman-Farinelli M. The most popular smartphone apps for weight loss: a quality assessment. *JMIR Mhealth Uhealth* 2015; 3 (4): e104.
74. Chyjek K, Farag S, Chen KT. Rating pregnancy wheel applications using the APPLICATIONS Scoring System. *Obstet Gynecol* 2015; 125 (6): 1478–83.
75. DiFilippo KN, Huang W, Chapman-Novakofski KM. A new tool for nutrition App Quality Evaluation (AQEL): development, validation, and reliability testing. *JMIR Mhealth Uhealth* 2017; 5 (10): e163.
76. Fiks AG, Fleisher L, Berrigan L, *et al.* Usability, acceptability, and impact of a pediatric Teledermatology Mobile Health Application. *Telemed e-Health* 2018; 24 (3): 236–45.
77. Loy JS, Ali EE, Yap KY-L. Quality assessment of medical apps that target medication-related problems. *J Manag Care Spec Pharm* 2016; 22 (10): 1124–40.
78. One Mind PsyberGuide. About One Mind PsyberGuide. 2020. https://onemindpsyberguide.org/about-psyberguide/. Accessed May 2, 2020.
79. Powell AC, Torous J, Chan S, *et al.* Interrater reliability of mHealth app rating measures: analysis of top depression and smoking cessation apps. *JMIR Mhealth Uhealth* 2016; 4 (1): e15.
80. Reynoldson C, Stones C, Allsop M, *et al.* Assessing the quality and usability of smartphone apps for pain self-management. *Pain Med* 2014; 15 (6): 898–909.
81. Schnall R, Cho H, Liu J. Health Information Technology Usability Evaluation Scale (Health-ITUES) for usability assessment of mobile health technology: validation study. *JMIR Mhealth Uhealth* 2018; 6 (1): e4.
82. Shaia KL, Farag S, Chyjek K, *et al.* An evaluation of mobile applications for reproductive endocrinology and infertility providers. *Telemed J E Health* 2017; 23 (3): 254–58.
83. Taki S, Campbell KJ, Russell CG, *et al.* Infant feeding websites and apps: a systematic assessment of quality and content. *Interact J Med Res* 2015; 4 (3): e18.
84. Tay I, Garland S, Gorelik A, *et al.* Development and testing of a mobile phone app for self-monitoring of calcium intake in young women. *JMIR Mhealth Uhealth* 2017; 5 (3): e27.
85. Van Singer M, Chatton A, Khazaal Y. Quality of smartphone apps related to panic disorder. *Front Psychiatry* 2015; 6: 96.
86. Yasini M, Beranger J, Desmarais P, *et al.* mHealth quality: a process to seal the qualified mobile health apps. *Stud Health Technol Inform* 2016; 228: 205–9.
87. Kharrazi H, Chisholm R, VanNasdale D, *et al.* Mobile personal health records: an evaluation of features and functionality. *Int J Med Inform* 2012; 81 (9): 579–93.
88. McNiel P, McArthur EC. Evaluating health mobile apps: information literacy in undergraduate and graduate nursing courses. *J Nurs Educ* 2016; 55 (8): 480.
89. Mathews SC, McShea MJ, Hanley CL, *et al.* Digital health: a path to validation. *NPJ Digit Med* 2019; 2 (1): 38.
90. Brown W, Yen P-Y, Rojas M, *et al.* Assessment of the Health IT Usability Evaluation Model (Health-ITUEM) for evaluating mobile health (mHealth) technology. *J Biomed Inform* 2013; 46 (6): 1080–7.
91. Basilico A, Marceglia S, Bonacina S, *et al.* Advising patients on selecting trustful apps for diabetes self-care. *Comput Biol Med* 2016; 71: 86–96.
92. Aji M, Gordon C, Peters D, *et al.* Exploring user needs and preferences for mobile apps for sleep disturbance: mixed methods study. *JMIR Ment Health* 2019; 6 (5): e13895.
93. Bauer AM, Iles-Shih M, Ghomi RH, *et al.* Acceptability of mHealth augmentation of Collaborative Care: a mixed methods pilot study. *Gen Hosp Psychiatry* 2018; 51: 22–29.
94. Chaudhry BM, Faust L, Chawla NV, eds. From design to development to evaluation of a pregnancy app for low-income women in a community-based setting. In: *MobileHCI '19: Proceedings of the 21st International Conference on Human–Computer Interaction with Mobile Designs and Services*; 2019: 7.
95. De Korte EM, Wiezer N, Janssen JH, *et al.* Evaluating an mHealth app for health and well-being at work: mixed-method qualitative study. *JMIR Mhealth Uhealth* 2018; 6 (3): e72.
96. Deshpande AK, Shimunova T. A comprehensive evaluation of tinnitus apps. *Am J Audiol* 2019; 28 (3): 605–16.

97. Edney S, Ryan JC, Olds T, *et al*. User engagement and attrition in an app-based physical activity intervention: secondary analysis of a randomized controlled trial. *J Med Internet Res* 2019; 21 (11): e14645.
98. IMS Institute for Healthcare Informatics. Patient apps for improved healthcare: from novelty to mainstream. 2013. http://moodle.univ-lille2.fr/pluginfile.php/215345/mod_resource/content/0/IIHI_Patient_Apps_Report.pdf. Accessed May 2, 2020.
99. Liang J, He X, Jia Y, *et al*. Chinese Mobile Health APPs for hypertension management: a systematic evaluation of usefulness. *J Healthc Eng* 2018; 2018: 7328274.
100. Liew MS, Zhang J, See J, *et al*. Usability challenges for health and wellness mobile apps: mixed-methods study among mHealth experts and consumers. *JMIR Mhealth Uhealth* 2019; 7 (1): e12160.
101. Meedya S, Sheikh MK, Win Kt, *et al*. Evaluation of breastfeeding mobile health applications based on the persuasive system design model. In: Karapanos E, Kyza E, Oinas-Kukkonen H, Karppinen P, Win KT, eds. *PERSUASIVE 2019: Persuasive Technology: Development of Persuasive and Behavior Change Support Systems*. Berlin, Germany: Springer Verlag; 2019: 189–201.
102. Tobias K, Lena O, Samira H, *et al*. A design and evaluation framework for digital health interventions. *it - Inform Technol* 2019; 61 (5–6): 253–63.
103. Torous J, Nicholas J, Larsen ME, *et al*. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. *Evid Based Ment Health* 2018; 21 (3): 116–9.
104. Torous JB, Chan SR, Yee M, Tan Gipson S, *et al*. A hierarchical framework for evaluation and informed decision making regarding smartphone apps for clinical care. *Psychiatr Serv* 2018; 69 (5): 498–500.
105. Dolnicar S, Grun B, Leisch F, *et al*. Three good reasons not to use five and seven point Likert items. In: *CAUTHE 2011: 21st CAUTHE National Conference*; 2011: 1–4.
106. Revilla MA, Saris WE, Krosnick JA. Choosing the number of categories in agree–disagree scales. *Soc Methods Res* 2014; 43 (1): 73–97.