BMJ Open

Conducting a systematic review and evaluation of commercially available mobile applications (apps) on a health-related topic: the TECH approach and a step-by-step methodological guide

Norina Gasteiger ,[1,2] Dawn Dowding,[1] Gill Norman ,[1] Lisa McGarrigle ,[1,3] Charlotte Eost-Telling ,[1] Debra Jones,[1] Amy Vercell,[1,4] Syed Mustafa Ali,[2] Siobhan O'Connor[1]

To cite: Gasteiger N, Dowding D, Norman G, *et al*. Conducting a systematic review and evaluation of commercially available mobile applications (apps) on a health-related topic: the TECH approach and a step-by-step methodological guide. *BMJ Open* 2023;**13**:e073283. doi:10.1136/bmjopen-2023-073283

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2023-073283).

Received 01 March 2023
Accepted 25 May 2023

Check for updates

[1]Division of Nursing, Midwifery and Social Work, The University of Manchester, Manchester, UK
[2]Division of Informatics, Imaging and Data Sciences, The University of Manchester, Manchester, UK
[3]Manchester Academic Health Science Centre, Manchester, UK
[4]The Christie NHS Foundation Trust, Manchester, UK

Correspondence to
Norina Gasteiger;
norina.gasteiger@manchester.ac.uk

ABSTRACT

Objectives To provide an overview of the methodological considerations for conducting commercial smartphone health app reviews (mHealth reviews), with the aim of systematising the process and supporting high-quality evaluations of mHealth apps.

Design Synthesis of our research team's experiences of conducting and publishing various reviews of mHealth apps available on app stores and hand-searching the top medical informatics journals (eg, The Lancet Digital Health, npj Digital Medicine, Journal of Biomedical Informatics and the Journal of the American Medical Informatics Association) over the last five years (2018–2022) to identify other app reviews to contribute to the discussion of this method and supporting framework for developing a research (review) question and determining the eligibility criteria.

Results We present seven steps to support rigour in conducting reviews of health apps available on the app market: (1) writing a research question or aims, (2) conducting scoping searches and developing the protocol, (3) determining the eligibility criteria using the TECH framework, (4) conducting the final search and screening of health apps, (5) data extraction, (6) quality, functionality and other assessments and (7) analysis and synthesis of findings. We introduce the novel TECH approach to developing review questions and the eligibility criteria, which considers the Target user, Evaluation focus, Connectedness and the Health domain. Patient and public involvement and engagement opportunities are acknowledged, including co-developing the protocol and undertaking quality or usability assessments.

Conclusion Commercial mHealth app reviews can provide important insights into the health app market, including the availability of apps and their quality and functionality. We have outlined seven key steps for conducting rigorous health app reviews in addition to the TECH acronym, which can support researchers in writing research questions and determining the eligibility criteria. Future work will include a collaborative effort to develop reporting guidelines and a quality appraisal tool to ensure transparency and quality in systematic app reviews.

STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ We leveraged the expertise of an experienced research team who have conducted various health app and systematic reviews to provide interim recommendations for good practice, based on seven key steps, to support standardised health app review methods and develop approaches to best practice in the mHealth field.
- ⇒ Relevant literature from key medical informatics journals was screened to present a robust analysis and comparison between systematic reviews and app reviews in health.
- ⇒ We propose TECH, a novel framework for constructing review questions and refining eligibility criteria in health app reviews.
- ⇒ This work focuses on mobile apps and does not include the emerging fields of virtual reality, augmented reality or mixed-reality apps.
- ⇒ The methods presented are informed by existing app reviews, which have primarily focused on client-facing apps (eg, for patients, the public or healthcare providers) rather than for the health system or data services, which are also key target users.

INTRODUCTION

With the rise in the use of smartphones and other mobile technologies, there has been an increase in the availability of health applications (mHealth apps) designed to be used by individuals for various health issues. Health apps can also support health and care professionals in their daily clinical practice by providing decision support, access to clinical guidelines and education and training.[1] In 2018, over 325 000 health apps were developed,[2] covering many health conditions and targeted behaviours. For example, mHealth apps can help to support self-management of conditions like diabetes,[3] facilitate remote monitoring of patients with chronic conditions[4] or support patients with general behaviour change such as increasing/monitoring physical activity[5] or dietary change.[6] Some health apps also support public health initiatives, such as promoting healthy lifestyles and encouraging the uptake of screening and vaccination programmes.[7 8] The World Health Organization (WHO) have released mHealth apps to educate people about road safety, sun protection measures and COVID-19.[9] However, concerns have been raised about the quality of advice and support such health apps provide.[10] Bates *et al* also highlight concerns with the accuracy of apps designed to support medical diagnosis and potential gaps in the quality of apps regarding safety and privacy.[2]

There have been attempts to provide frameworks for evaluating the quality of mHealth apps. A systematic review identified 45 frameworks for evaluating mHealth apps, which varied according to the target users (eg, developers, patients), specific conditions (eg, diabetes, mental health, cancer, pain) and various elements of evaluation that are identified in the core domains for Health Technology Assessment framework (such as safety and effectiveness).[11] Other frameworks have promoted a more holistic approach by encompassing privacy and security, the evidence base, ease of use and data integration[12] or ethical principles related to using health apps in health psychology.[13] Reviews have also identified existing methods for assessing mHealth app quality,[14] as well as guidelines for reporting evaluations of specific types of technology, such as sensors[15] and mHealth interventions, more broadly.[16] This includes the development of the Consolidated Standards of Reporting Trials of Electronic and Mobile HEalth Applications and onLine TeleHealth (CONSORT-EHEALTH) checklist, an extension of the original CONSORT checklist for reporting randomised controlled trials (RCTs).[17] This focuses on improving the reporting of evidence from research into the use and effectiveness of mHealth applications in research studies.

The existing initiatives reflect a focus on the quality of research activity, whereby mHealth interventions are evaluated in research studies, and the results of those studies are reported.[18] The process of systematically reviewing published studies then provides an overview of the evidence base for the use and effectiveness of different types of mHealth interventions for different patient populations and various purposes. However, guidance is missing on how to systematically review commercially developed mHealth apps (ie, the software products available to download from app stores), which are often not derived from research or subject to evaluation in research studies.

The process of searching, screening, extracting and analysing data, and critically appraising mHealth apps available via commercial platforms can differ from traditional approaches to reviewing published research studies of health apps. A key difference between a systematic literature review and a systematic health app review is the items evaluated. In a systematic review, reviewers attempt to identify evidence from research studies from peer-reviewed journals or grey literature to evaluate the effectiveness (RCT evidence) or other characteristics that influence the effectiveness, uptake and engagement with digital health interventions. As we have outlined, several existing systematic reviews of mHealth applications do this. In a systematic health app review, we focus on providing a transparent and replicable evaluation of the functionality, quality and purpose of mHealth apps for particular user groups or health conditions. A systematic health app review is informed by the principles and process of more traditional systematic reviews, in terms of approaches to searching, use of inclusion/exclusion criteria and explicit assessment measures of quality. However, how these are operationalised is methodologically different and is the focus of this paper. By building on our research team's experiences of conducting and publishing various reviews of commercially available mHealth apps, we provide an overview of the methodological considerations, aiming to systematise the process and support high-quality reviews of mHealth apps. In doing so, we outline the 7-step process for conducting systematic health app reviews.

METHODS

In this paper, we use examples from our previous work as case studies, supported by work from other authors to develop a new framework for conducting a review of commercially available health apps. We combine our experience (see table 1) with the results of a hand search of the top medical informatics journals (ie, The Lancet Digital Health, npj Digital Medicine, Journal of Biomedical Informatics and the Journal of the American Medical Informatics Association) over the last five years (2018–2022) to identify other reviews of commercial health apps to contribute to the discussion of this methodological approach. Based on this, we propose methods for writing the research question and aim, determining the eligibility criteria and carrying out the review and highlight and discuss the methodological issues raised at each stage.

The reviews we draw on cover a range of apps and provide examples of a number of the decisions and challenges in conducting such reviews. Two of our reviews informed wider research studies; a review of apps used to support hand hygiene to provide the focus for a subsequent research evaluation,[19] and a review of patient-facing genetics apps to inform the design and development of a genetic counselling app.[20] Our other app reviews have provided evidence alongside a more traditional systematic review of the published research literature or as an independent review to guide clinicians/patients about mHealth apps they could use to support patient care.

As the app review process follows much of the same steps as conducting systematic literature reviews, we also drew on some existing guidance to formulate the seven steps. This included the work by Khan *et al*[21] who name five steps for conducting systematic literature reviews: (1) framing the question, (2) identifying relevant

Table 1 Summary of our app reviews, which are used as cases to inform the methods for conducting systematic app reviews

First author(s); date published	Aim	Purpose of review	Number of apps reviewed
Paripoorani *et al*; in-progress[25]	To explore and identify menopause apps available in the UK, assess their quality, functions and content, and determine whether and to what extent they focus on menopause-related osteoporosis.	Standalone app review.	28
Vercell *et al* 2022[24]	To identify patient-facing cancer apps which can record patient-reported outcomes, and to explore their purpose, functionality, quality and ability to integrate with electronic health records.	Standalone app review.	12
Gasteiger *et al* 2022[20]	To identify patient-facing smartphone apps related to genetic or genomic conditions available in the UK and explore their purpose, functions and quality.	Inform the design of a genetics app which is being co-designed with community members.	22
Gasteiger *et al* 2021[19]	To identify smartphone apps that support hand hygiene practice and to assess their content, technical and functional features and quality. A secondary objective was to make design and research recommendations for future apps.	Background of wider project on extended reality hand hygiene training.	90
Ali *et al* 2021[26]	To explore the current state of smartphone-based pain manikins and to formulate recommendations to guide their development in the future.	To formulate recommendations to guide the development of pain manikins in the future.	28
Pearsons *et al* 2021[23]	To identify commercially available atrial fibrillation self-management apps, analyse and synthesise their characteristics, functions, privacy/security, behaviour change techniques, quality and usability.	To inform the development and testing of a new app for atrial fibrillation.	5
McGarrigle *et al* 2020[27]	To identify existing apps and websites to support independent engagement in strength and balance exercises by older people, and to evaluate evidence for effectiveness, quality and use of behaviour change techniques.	To provide evidence-based alternatives to face-to-face exercise classes.	13

publications, (3) assessing the quality of studies, (4) summarising the evidence and (5) interpreting the findings. Xiao and Watson[22] name similar steps to conducting reviews but added steps for developing and validating the review protocol, screening for inclusion, extracting data and reporting the findings.

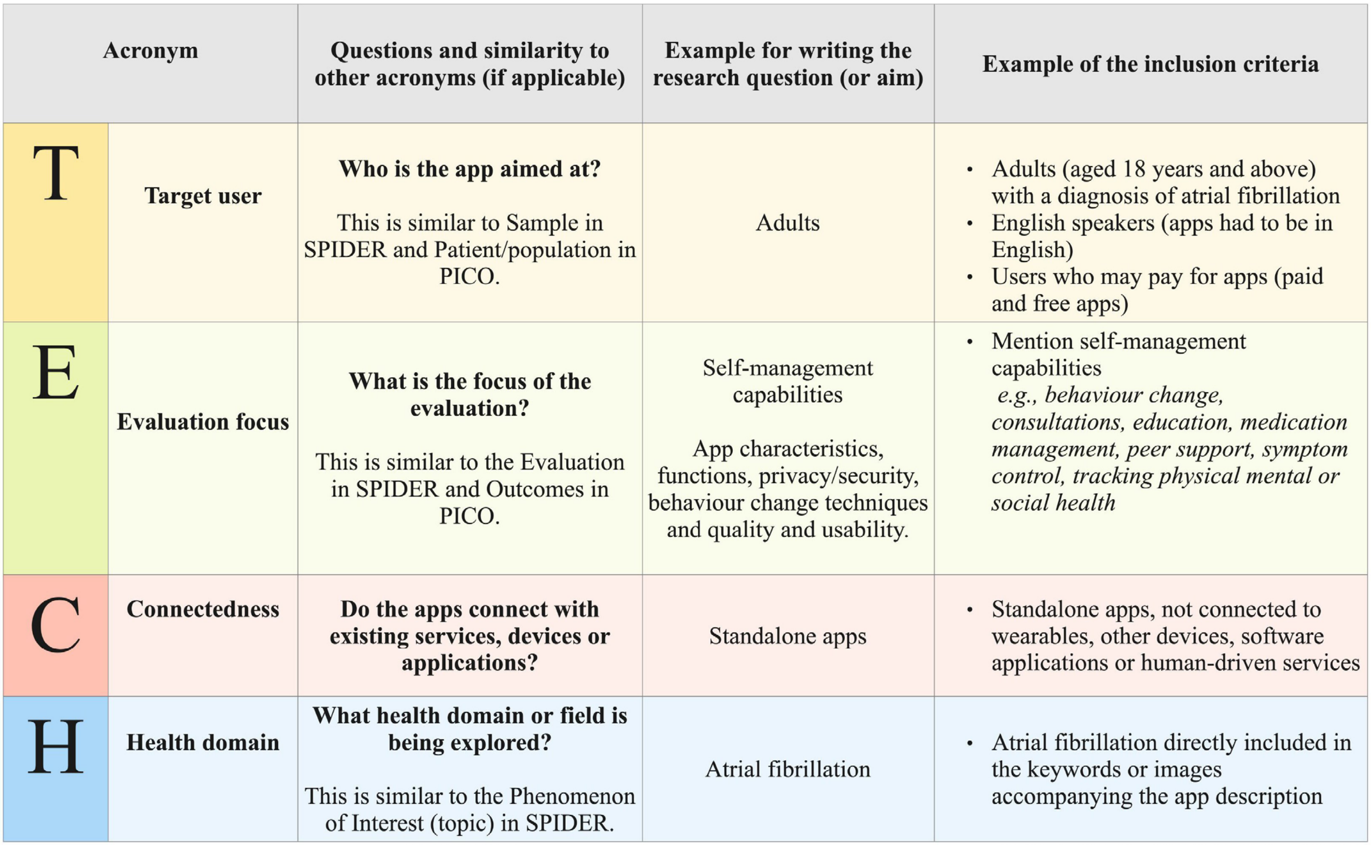
RESULTS

Through discussion within the research team and drawing on our experiences of conducting app reviews and through cross-checking with app reviews by other author teams, we have outlined seven steps to support rigour in conducting reviews of health apps available on the app market. The steps are: (1) writing a research question or aim; (2) conducting scoping searches and developing the protocol; (3) determining the eligibility criteria using the TECH framework; (4) conducting the final search and screening of health apps; (5) data extraction; (6) quality, functionality and other assessments and (7) analysis and synthesis of findings. Each step is discussed in turn.

Step 1: writing a research question (or aims)

The focus of an app review will influence the development of the research questions or aims and underpinning approach to evaluating health apps. If the purpose is to produce a standalone review to support future research and innovation in a specific health domain, understanding existing gaps can help formulate a more general research question. However, if the review is the starting point of a programme of research that aims to design, develop and evaluate a new health app with a population of patients, carers, health professionals or the public then the research questions may be more focused to examine aspects of apps such as their quality, functionality or availability. Formulating an answerable review question is essential for systematic literature reviews. While formulating a review question helps guide all stages of a systematic literature review (eg, searching, screening, extracting and synthesising), not every question format applies to systematic app reviews. For example, the PICO format is appropriate for systematic literature reviews looking at the effectiveness of interventions in a target population. However, in systematic health app reviews, reviewers can only access the results of effectiveness or evaluation studies (if conducted) if they are published. A bespoke alternative is required, just as with systematic qualitative reviews where SPIDER is used.

Therefore, we propose the acronym ‘TECH’, which represents (1) Target user, (2) Evaluation focus, (3) Connectedness and (4) Health domain, as a mechanism to develop a focused research question to guide a health app review. TECH was designed through discussion by



Acronym		Questions and similarity to other acronyms (if applicable)	Example for writing the research question (or aim)	Example of the inclusion criteria
T	Target user	**Who is the app aimed at?** This is similar to Sample in SPIDER and Patient/population in PICO.	Adults	• Adults (aged 18 years and above) with a diagnosis of atrial fibrillation • English speakers (apps had to be in English) • Users who may pay for apps (paid and free apps)
E	Evaluation focus	**What is the focus of the evaluation?** This is similar to the Evaluation in SPIDER and Outcomes in PICO.	Self-management capabilities App characteristics, functions, privacy/security, behaviour change techniques and quality and usability.	• Mention self-management capabilities *e.g., behaviour change, consultations, education, medication management, peer support, symptom control, tracking physical mental or social health*
C	Connectedness	**Do the apps connect with existing services, devices or applications?**	Standalone apps	• Standalone apps, not connected to wearables, other devices, software applications or human-driven services
H	Health domain	**What health domain or field is being explored?** This is similar to the Phenomenon of Interest (topic) in SPIDER.	Atrial fibrillation	• Atrial fibrillation directly included in the keywords or images accompanying the app description

Figure 1 TECH framework with a worked example.

the research team and by mapping key concepts against existing frameworks (eg, SPIDER and PICO). Figure 1 presents the acronym, questions which researchers may consider (and the similarity to other acronyms) and a worked example for one of our reviews which aimed to identify commercially available atrial fibrillation self-management apps, analyse and synthesise characteristics, functions, privacy/security, incorporated behaviour change techniques and quality and usability.[23] TECH is designed to capture the nuances of health domains, supporting the development of clear, answerable research questions. It is important to note that connectedness refers to connecting with other devices or applications and existing human-driven digital or other services—such as a health app for booking appointments with therapists.

Step 2: conducting scoping searches and developing the protocol

A preliminary (scoping) search of the health app market via the Apple, Google and Microsoft app stores is an essential first step to help determine whether the number of commercial health apps available is feasible to review. It is worth noting that the language used in descriptions of commercial mHealth apps can vary widely and differ from the scientific language used in published research studies. Hence, a broad search using a range of terminology should be employed initially to avoid missing relevant health apps. We recommend that researchers use basic keywords focused on the health domain/topic (see figure 1) as the search function within app stores is limited. For example, for our hand hygiene app review we only used two keywords: hand hygiene and hand washing.[19] In our cancer app review,[24] we used more keywords, but all were related to the health domain and only one focused on the target user (patients): cancer, cancer patient, cancer treatment, cancer management and cancer side effects.

If too few health apps are returned, this might allow for broadening the scope of the topic or adding more keywords, while too many apps will likely require the scope and language used to be narrowed. This means that the research question and the eligibility criteria may need to be refined iteratively, with multiple scoping searches performed until a reasonable number of apps are identified. The number of potential apps that may be included in the review can be counted by reading the app's name and description and judging its relevance to the topic.

To give an indication of how many apps is reasonable to review, we previously identified 236,[25] 405,[24] 555,[23] 668,[19] 754[20] and 3938[26] health apps from initial searches, before screening or deduplication took place. One of our reviews identified 7561 apps before screening[27] due to the topic (exercise), for which many apps exist. Following the initial screening of app titles and app store descriptions, this number was significantly reduced, and only 13 were included in the review. However, each research team should decide what number is appropriate by considering resources (eg, time, budget, and the number of reviewers available) and the topic of interest.

Scoping searches can also help to identify important considerations for refining the inclusion and exclusion criteria. For example, in our review of hand hygiene apps,[19] we initially intended to target healthcare providers as the population of interest. However, the scoping search highlighted that few apps specified their intended users. We, therefore, removed healthcare providers as the intended audience from the inclusion criteria and replaced this with adults more generally.

It is important to note that the scoping search should be conducted when the team is almost ready to begin the final searches, as health apps can disappear and emerge quickly from app stores. This means that the numbers determined from the scoping searches will likely differ from the number of apps identified in the final search. Longer periods between the scoping and final search will result in more substantial differences in the number of apps available.

In addition to scoping searches in the app stores, we recommend conducting initial searches in databases (eg, MEDLINE and SCOPUS) and protocol registration databases to identify whether similar app reviews have been published or are underway. We also strongly recommend that researchers prospectively develop a protocol to guide their methods. Unlike systematic reviews which should usually be registered on PROSPERO (https://www.crd.york.ac.uk/prospero/) or OSF (https://osf.io/), there have not been any formal requirements to publish protocols of systematic health app reviews. However, we recommend that future protocols for commercial health app checks be published (in advance of the searches) on OSF to reduce the likelihood that the review is unnecessarily duplicated and ensure greater research transparency. This is becoming accepted practice for other reviews (eg, scoping reviews) for which PROSPERO registration is currently not possible.[28]

Step 3: determining the eligibility criteria using the TECH framework

Inclusion and exclusion criteria should be carefully defined using the information obtained in the scoping searches. Frameworks such as PICO or SPIDER may help develop eligibility criteria for a review so that characteristics such as the population/sample, type of intervention or phenomenon of interest and outcomes related to mHealth apps are considered.[29] However, we propose using a more focused framework for health app reviews to support the nuances of undertaking this type of systematic review, as the aim is not to examine the effectiveness of apps or to synthesise the findings of qualitative studies on health apps. As described above, the 'TECH' acronym considers the Target user, Evaluation focus, Connectedness and Health domain (see figure 1 for the worked example). Well-thought-out eligibility criteria using the TECH framework may support the development of an appropriate search strategy considering four aspects of commercial health apps, which can lead to a systematic and unbiased selection of appropriate apps. Additionally, reviewers should consider searching the literature to identify published and relevant app reviews to refine the eligibility criteria further.

A health app may be characterised by its intended use and target audience. Therefore, we suggest separating the eligibility criteria into two components: app characteristics (evaluation focus, connectedness, health domain) and target audience characteristics (eg, age, gender, race/ethnicity, geographic location). App characteristics are likely to include the type of health intervention or health prevention method, such as self-management, that can be captured under the 'Evaluation focus', along with the target disease, problem or focus of the health app, which falls under the 'Health domain'. It may be helpful to state whether the health condition is specific or if it covers a broad category of health-related issues. Some health apps can link to other software applications, connect to hardware devices (eg, wearable technologies) or rely on additional external devices (eg, virtual reality headsets or smartwatches) to function correctly, which should be captured under the 'Connectedness' criteria. The target audience characteristics will likely include the population type, including patients, healthcare professionals, healthcare students, carers, the public, or particular organisations. Audience characteristics may also include age ranges or if the app is aimed at children or adults; whether the app is aimed at individuals or groups; whether it is assumed that users will pay to access the app (or content within it) and whether the app is available in certain languages or locations.

Step 4: conducting the final search and screening of health apps

Once the scoping searches are complete and the search terms have been collated, the final search can be run on the main app stores (ie, Apple, Google Play and Microsoft) to identify potentially relevant health apps. Several third-party app stores or repositories are also available for Android-based apps (eg, Amazon Appstore, F-Droid and Samsung Galaxy Apps), some of which are open source making the download process easier. The volume of app stores that can be searched will depend on the time and resources available to the review team. Other approaches include using a proprietary software database which enables searching for mobile health apps across iOS and Google Play app stores[30] or publicly available online rating frameworks for health apps that use expert reviewers[31] such as the Organisation for the Review and Care of Health Apps (ORCHA, https://orchahealth.com/), the PsyberGuide by One Mind (https://onemindpsyberguide.org/) and MindTools. This approach could be combined with independent searching and evaluating of commercial mHealth apps to ensure an exhaustive assessment is conducted.

In contrast to established bibliographical research databases such as MEDLINE or PubMed, which enable complex searches (eg, use of Boolean operators and filtering options), the search function within app stores is limited. Basic filters may exclude apps that cost or only include child-friendly apps. Some stores (eg, the Google Play store) also enable for users to identify family-friendly apps and distinguish the type of app being searched (ie, phone, tablet, TV, Chromebook, watch or car). The Apple app store also has basic filters for the price (any or free), category (including health and fitness) and sorting (relevance, popularity, ratings or release date). Other factors app stores use may also affect search results (eg, app rating or use of adverts). To overcome this, searching across multiple app stores is advisable.

The search on an app store results in a list of available apps and their descriptions. Unlike research literature databases, there are no options to export the results in a useful format (eg, RIS) for uploading to specialised screening and data management software (eg, Covidence[32] or Rayyan[33]). Hence, the research team must review each app on the results page of each app store to determine if it meets the inclusion criteria for the review. Like systematic literature reviews, using two or more researchers to reach a consensus on eligibility from the review enhances rigour and study quality. Screening should ideally be conducted on the same day to avoid differences in search results on the different app stores, which can vary day-to-day and across countries. A workaround can be to log the app name, version number and link to the webpage on the app store hosting each app to capture the search results and ensure these are used consistently by the review team.

The second screening stage involves downloading all apps deemed to have met the inclusion criteria, requiring at least one Android and one iPhone smartphone between the review team. Strategies for this can include having separate researchers download apps using different devices or sharing one device between reviewers. Some health apps also require user accounts to be set up and verified before allowing access to their full functionality, which may be required to assess eligibility. In one of our reviews, we approached app developers for full app access, finding that they were more than happy to allow this.[27] Additional researchers can be consulted to resolve differences during the second screening phase. Finally, modifying a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram[34] can provide a transparent overview of the search and screening process. This also requires clearly stating the number of duplicates across the searches in addition to how many apps were excluded at each screening stage, with the reasons outlined at the second stage.

Step 5: data extraction

Like systematic literature reviews, the data extraction process in app reviews requires identifying relevant information from the eligible apps. Data are extracted into a pre-defined data extraction (coding) sheet by using the app. The length of use to extract the information depends on the types of apps, number of data extraction items and the focus of the review. For example, some apps will take longer to review as they may require more comprehensive information to be extracted, users to register personal profiles or send push notifications at specific times of the day (eg, behaviour change apps).

A range of data extraction items may be used. Henson *et al*[12] developed a five-level framework for evaluating health apps, concluding that background information, privacy and security, evidence, ease of use and data integration are vital components to consider. Across our previous work, we have categorised the items as descriptive information, technical information and content (see table 2). We have included additional items regarding gamification principles and tactics used in an app review by Rajani *et al*[35] and levels of personalisation, security and privacy, which Parmar *et al*[30] proposed. Other approaches used in one of our reviews[23] include extracting information on the Online Trust Alliance Best Practices Privacy Recommendations.[36] This considers four elements: (1) basic notice/disclosure, (2) key compliance policies, (3) protected privacy and protected sharing criteria and (4) miscellaneous privacy elements. Lastly, our osteoporosis app review considered ratings by the ORCHA (https://orchahealth.com/). ORCHA objectively reviews health apps, giving scores for three domains (data privacy, professional assurance and usability/accessibility) and an overall score (%). Scores below 65% imply the presence of some issues, and below 45% indicate considerable issues.

We encourage researchers to develop their own data extraction items relevant to their topics of interest. For example, in our hand hygiene app review and like other app reviews,[37 38] we developed criteria to assess the comprehensiveness of the content by identifying themes across reputable sources and guidelines on hand hygiene. Other examples include extracting information about health app security and privacy, including HIPPA or COPPA compliance, whether a medical disclaimer was provided, encrypted data disclaimer, and user verification strategies during login.[30]

We also note that sometimes the information sought is not readily available or transparently reported within apps. In this case, researchers should note where information is missing, using acronyms like N/R (not reported) or N/A (not available). This can also be an interesting finding and an opportunity for apps to be improved. For example, excluding information about data sharing may be concerning for health apps that collect and record personal medical information.

Readability metrics can also help determine how appropriate the language used in each app is. Researchers can determine readability by copying a paragraph into a Microsoft Word document and using two Flesch-Kincaid metrics that are built into the word processing software.[39 40] The Flesch-Kincaid Reading Ease score ranges from 0 to 100, with higher scores indicating that the material is easier to read.[39] The Flesch-Kincaid Grade Level

Table 2 Example of data extraction items for a commercial health app review

Items	Description
Descriptive information	
App name	Name of the mobile app
Version number	Version of the app reviewed
Developer	Name of the developer
Market/s available	Name of the markets where the app is available
Cost	Select which apply: free to download, cost to download (in GBP); in-app purchases available
Affiliated with a professional medical/health association, charity or government body	No; Yes (name of organisation)
Average user rating from app markets	Not rated; average number of public ratings (maximum 5 points)
Number of user ratings	Total number of user ratings on app markets
Technical information	
Privacy strategy	Select which apply: privacy policy, login, password, two-factor authentication, compliance with data protection acts (eg, HIPAA or GDPR), no privacy strategy
Security	Select which apply: email verification, text verification, social media verification, no strategy
Third-party authorisations (eg, data sharing)	Yes; No
Works offline	Yes; No
Works in the background	Yes; No
Asks to enable push notifications	Yes; No
Content	
Purpose	► Diagnose, record data/track, educate/inform, instruct, remind, analyse (ie, DNA sample/test data)* ► Remind, track, record data, educate/instruct, inform, convince, provide feedback† ► Educational; risk assessor; tracking‡ ► Diagnose, record data/track, educate/inform, instruct, remind, analyse§ ► Educate, instruct¶
Description	Summary of the app's content
Behaviour change techniques used	► Select which apply: self-monitoring, feedback, goal-setting and action-planning, social support, reward/threat, prompt practice* ► Note applicable Behaviour Change Techniques (BCTs) using Michie's 93 item BCT taxonomy as a guide[70]¶
Best practice guidelines mentioned† Scientific studies mentioned¶	No; Yes (name, for example, WHO five moments for hand hygiene) No; Yes (which, what claims are made)
Comprehensiveness of the content†	Select which criteria were met (from pre-determined criteria)
Innovative/personalisation features used†¶	► Select which apply: personal profile, tracking/reminders, virtual reality, augmented reality,** app-based community, chatbots or gamification (avatars/characters, competition, levels, scores, marketplace/coins, rules, collaboration/teamwork). ► Summary of the features.
Gamification techniques	► Principles used (select which apply): goal-setting, capacity to overcome challenges (eg, growth or learning), feedback on performance, reinforcement, compare progress (eg, monitoring), social connectivity, fun and playfulness (eg, alternative reality), no gamification principles used ► Tactics used (select which apply): providing clear goals, offering a challenge, levels or incremental challenges, allocating points, showing progress, providing feedback, rewards, badges for achievements, showing game leaders (eg, leaderboard), story/theme

Continued

Table 2 Continued

Items	Description
Levels of personalisation	► Select which apply: implicit personalisation (information needed for personalisation was obtained by the system), explicit personalisation (information required active use/engagement), no personalisation ► Type of personalisation (select which apply): individuated (targets an individual user), categorical (targets groups) ► Aspects of personalisation (select which apply): content, user interface (the manner in which information is presented), delivery channel (the media through which information is delivered), functionality
Data can be integrated into electronic health record§	No; Yes
Readability	Flesch Reading Ease: scored 0–100; Flesch-Kincaid grade level—corresponds with USA education grades level
ORCHA ratings‡	Reviewed by ORCHA: yes; no. If rated, scores for three domains (%)—data privacy, professional assurance and usability/accessibility; overall score: %

*Item used for the genetic app review.
†Item used for hand hygiene app review.
‡Item used for the osteoporosis app review.
§Item used for the oncology app review.
¶Item used for strength and balance exercises for older adults app review.
**Some apps require additional devices such as headsets and may not be appropriate to be included in a general app review.
ORCHA, Organisation for the Review of Care and Health Apps.

gives a score that refers to the equivalent grade level of education in the USA.[40] For example, a score of 12 indicates that a twelfth grader (aged 17 or 18) in the USA should be able to understand the content.

Step 6: quality, functionality, and other assessments

Quality

Evaluating the quality of apps requires a different approach to using critical appraisal tools or risk of bias measures commonly used in systematic literature reviews. Quality can be assessed using the Mobile App Rating Scale (MARS), which comprises 19 items across four objective scales (engagement, functionality, aesthetics and information quality) and an additional 4-item subjective quality scale.[41] Each item is rated on a 5-point Likert scale: (1) inadequate, (2) poor, (3) acceptable, (4) good and (5) excellent. MARS has been translated into several languages, including French, Spanish, German and Italian[42–45] and is suitable for assessing mobile apps for health conditions due to its reliability, validity and objectivity.[46] In all but two of the app reviews we have conducted to date, we have excluded the subjective quality scale to ensure that assessments are as objective as possible. Nevertheless, Stoyanov *et al*[41] reported that the objective measures correlated well with the subjective measures. A step-by-step training video on the use of the MARS is available on YouTube.[47]

The MARS question 'Has the app been trialled/tested?' can be answered by searching for literature on evaluation (eg, usability, satisfaction or effectiveness) and using more traditional methods, such as risk of bias, to evaluate the quality of the evidence.[48] However, some review teams may wish to take this a step further if the review aims to recommend evidence-based apps to their target population. For example, in our review on strength and balance exercises for older adults,[27] we also visited the app/developer websites and contacted the developers directly for information on any evaluations that had taken place concerning the effectiveness of the apps in preventing falls. Due to the absence of evaluations, we compared the interventions promoted by the app with those used in known 'gold standard' strength and balance programmes to determine if they had an evidence base.

Researchers may also build on the MARS items to assess the quality of each app in more detail. In one of our reviews, we added predetermined criteria to further evaluate the current state of development of apps for pain assessment and to provide future directions to developers.[26] For example, for the item about customisation, we looked at whether the app provides a setting tool allowing users to change the interface to suit them best. For the interactivity item, we also extracted data on the manikin regarding dimension (2-dimensional or 3-dimensional), orientation (left/right) and gender (male, female or neutral). It is important to note that directly amending the MARS may impact the validity of the tool. However, additional relevant items can further explore the dimensions.

Functionality

We recommend using the IMS Institute for Healthcare Informatics functionality score to assess the functionality of apps.[49] This records the availability of 11 different functions within an app, rated 1 if they are present and 0 if otherwise (see table 3). It complements the MARS functionality score, which measures the quality

Table 3 IMS Institute for Healthcare Informatics functionality score items and descriptions

Items	Description
1. Inform	Provides information in a variety of formats (eg, text, photo, video.
2. Instruct	Provides instructions to the user.
3. Record	Captures user-entered data.
3.1 Collect data	Able to enter and store health data on individual phone.
3.2 Share data	Able to transmit health data.
3.3 Evaluate data	Able to evaluate the entered data by patient and provider, provider and administrator, or patient and caregiver.
3.4 Intervene	Able to send alerts based on the data collected or propose behavioural intervention or changes.
4. Display	Graphically display user-entered data/output user-entered data.
5. Guide	Provide guidance based on user-entered information, and may further offer help (eg, diagnosis, recommend consultation with a doctor or a course of treatment).
6. Remind or alert	Provide reminders to the user.
7. Communicate	Provide communication between users, consumers or others and/or provide links to social networks.

of performance, ease of use, navigation and design of an app using rating scales. The IMS functionality score is calculated from seven main criteria (inform, instruct, record, display, guide, remind or alert and communicate) and four subcategories under the 'record' item (collect, share, evaluate, intervene). An overall functionality score, between 0 and 11 for the full scale, is calculated by summing the scores across the individual items. The IMS scale may be tailored to ensure relevance for a specific review; for instance, in our review of hand hygiene apps, we omitted the 'evaluate data' criteria because it was irrelevant to the topic.[19]

Written results from the IMS scale are generally supported with visual representations, such as radar graphs/charts, which map variables onto axes protruding from a central point (see figure 2). Each axis can represent a different item of the IMS, with values plotted onto each axis. These types of data visualisations could be helpful for clinicians, patients, developers and other stakeholders so they can quickly see which health apps are

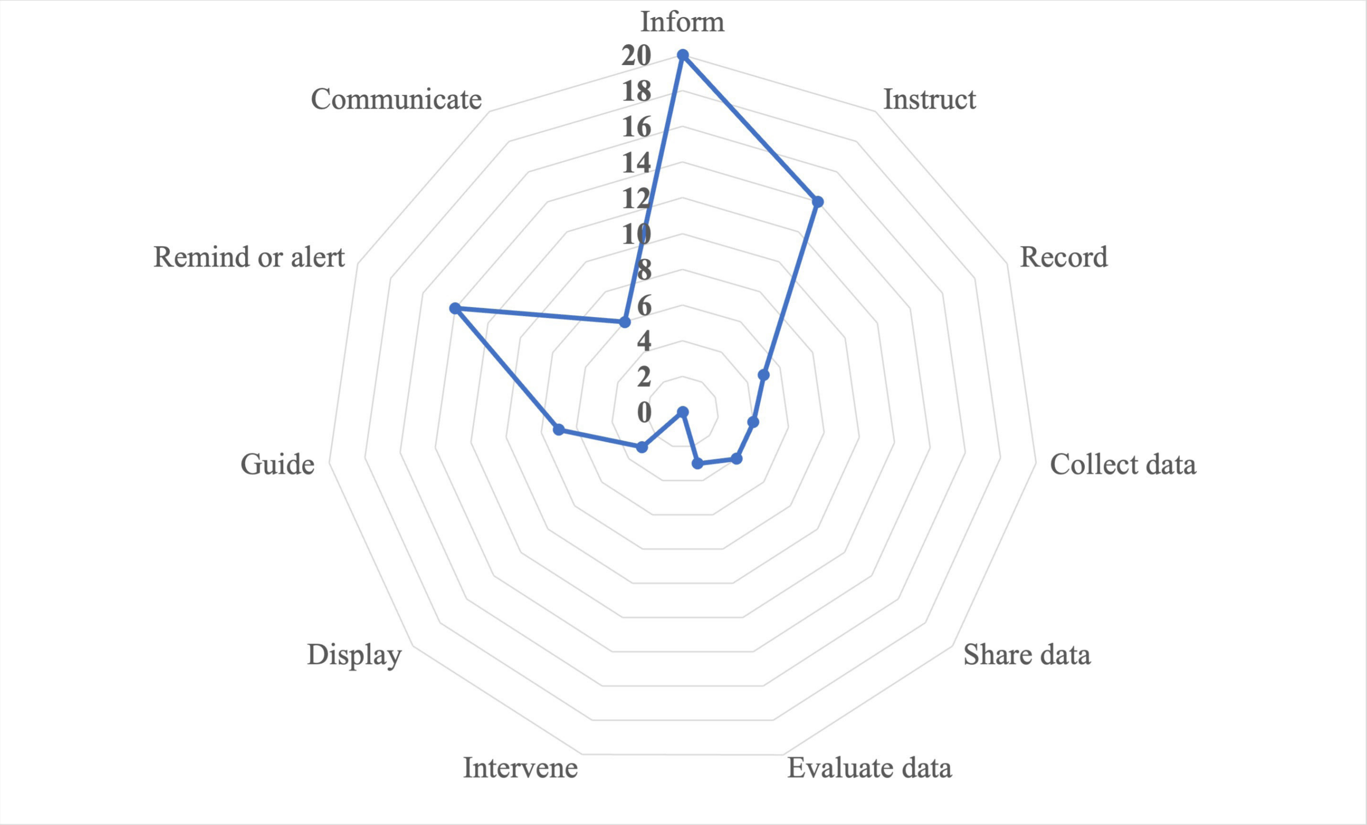


Figure 2 Simulated radar graph mapping the 11 IMS criteria.

ranked high or low across a range of evaluation metrics to inform decision-making about which, if any, to use.

Other assessments

Other ways to evaluate health apps include examining the user reviews of each app on the app stores. Plante *et al*[50] took this approach when reviewing blood pressure measuring smartphone apps by downloading the ratings and reviews from the iTunes store and developing a series of narrative themes linked to the high and low-rated apps. Some themes associated with high user ratings included accuracy, login functionality, convenience and successful measurement. In contrast, lower-rated apps were associated with inaccuracy, inability to produce a successful reading and refunds requested by users. A qualitative approach to understanding the experiences of a range of users can help inform the final evaluation of a range of apps.

Other rating tools include THESIS, developed by evaluating over 200 mHealth apps with a panel of experts.[51] THESIS encompasses six domains: (1) transparency, (2) health content, (3) technical content, (4) security/privacy, (5) usability and (6) subjective rating, which considers other factors like software stability, interoperability, bandwidth and application size. Additionally, health apps have been evaluated for their security features (eg, app signing security, encryption schemes, malware presence, permissions and secure communication adoption) along with users' subjective perceptions of app security.[52] However, this approach requires technical expertise to undertake static and dynamic analysis techniques that may be outside some research teams' scope.

Other reviews have also evaluated criteria such as ethical values and medical claims. This includes beneficence, non-maleficence, autonomy, justice and legal obligation in COVID-19 mobile phone apps following the Systems Wide Analysis of mobile health-related Technologies provided in the NHS Digital Assessment Questionnaire[53] and the medical claims of mental health apps such as scientific language, technical expertise and lived experience perspectives.[10]

Step 7: analysis and synthesis of findings

Data synthesis may be performed by generating descriptive statistics (sums, averages, standard deviations and percentages) on relevant items or combining these with forms of qualitative synthesis. Previously, we identified the highest-scoring apps regarding functionality and quality, presenting these with a written description of their main features. Inter-rater reliability can be calculated for the binary IMS Institute for Healthcare Informatics functionality scores using Cohen's Kappa statistic[54] and an intraclass correlation coefficient (ICC) can be used to calculate inter-rater reliability for the ordinal MARS scores.[55] The ICC is the most commonly used statistic for assessing inter-rater reliability for ordinal variables. We have typically used an absolute agreement 2-way mixed-effects, average-measures model,[56] which assumes that the raters are fixed and that systematic differences between raters are relevant.

Patient and public involvement and engagement

None of the commercial health app reviews generated by our research team actively included patients or the public due to pragmatic reasons, including time, resources and funding constraints. Additionally, we did not identify any health app review recently published in the top medical informatics journals that took this approach. However, patient and public involvement and engagement (PPIE) is viewed favourably by many funders, researchers and policymakers as it can add value to health research and facilitates its dissemination and impact. As outlined above, all stages of a commercial app review could benefit from the perspectives of patients, carers and members of the public towards the health apps being evaluated. As with traditional systematic reviews, they could assist with reviewing the protocol, searching, screening, extracting and analysing data[57] and co-production by undertaking quality, usability or other assessments and participating in various dissemination activities. PPIE could provide another valuable dimension to the process and enrich the results of a commercial health app review.

Key differences between systematic literature reviews and systematic health app reviews

Here, we summarise the main differences between a traditional approach to systematic reviewing literature versus undertaking a commercial health app review, as outlined in this methodological discussion (see table 4).

DISCUSSION

This methods paper outlines the 7-step process for conducting systematic reviews of commercial health apps. Through comparison with systematic literature reviews, we explore the complexities of each stage of an app review and provide suggestions on how to formulate a research question, develop and run scoping searches, register the protocol, determine the eligibility criteria, conduct the final search and screening, extract data, perform quality assessments and synthesise the findings. We also propose that the novel TECH framework is adopted to allow a standardised specification to be developed and applied in health app reviews, similar to using PICO, SPICE or SPIDER in traditional systematic reviews.[29] Additionally, we highlight the potential for PPIE activities within health app reviews.

Although health app reviews share core features with systematic reviews, three key differences warrant discussion. First, commercial health apps, and reviews of these, are more transitory with rapid changes in the mHealth landscape that private industry providers dominate. For example, the geographical and price specificity of app sources are not replicable in the same way that evidence sources for systematic reviews are, and apps can appear, change and disappear quickly, impacting the replicability

Table 4 Summary of the review stages, contrasting systematic literature review methods with systematic health app review methods

Review stage	Systematic literature review (of effectiveness) with a focus on quantitative reviews	Systematic commercial health app review
Scoping work	► Scoping searches of the literature are usually necessary to inform protocol development and ensure the review addresses appropriate questions. ► Sometimes used to ensure manageable size.	► Scoping searches of some app stores are essential to determine whether the number of apps available are feasible to review. ► The research question and the eligibility criteria may be refined iteratively, with multiple scoping searches performed until a reasonable number of apps are identified.
Protocol development	► Journals increasingly require pre-registration of protocols. ► There are dedicated registries for many review types (eg, PROSPERO). ► Alternatives such as OSF are sometimes used.	► There is no formal requirement for protocols to be registered. ► Registration on OSF is appropriate and we recommend this.
Stakeholder engagement	► Varies from none through protocol review to co-development or co-production. ► This is recommended but not required unless a specific design is used, or it is a Cochrane review where consumer peer review is required.	► Not formally required and most have had no stakeholder engagement. ► All stages could benefit from stakeholder engagement for example, co-production by searching, screening, extracting and analysing data.
Inclusion criteria	► Most reviews include primary research studies. ► PICO is usually used to define key eligibility criteria.	► The novel TECH framework can be used to help determine the eligibility criteria. TECH considers the Target user, Evaluation focus, Connectedness and Health domain.
Search	► Searching multiple databases of published literature plus (often) trial registries and/or grey literature. ► Search strategy, dates and number of records reported for each database. ► De-duplication of search results may include using a reference manager. ► Citation searching is also often used. ► There is extensive literature on multiple aspects of searching. ► Information specialists should be involved.	► Searching the app market via multiple app stores using basic keywords. ► Additional sources may include a proprietary software database or publicly available online rating frameworks for health apps that use expert reviewers (eg, ORCHA). ► The search information (eg, app market, date of search and number of apps identified) is recorded on Excel. ► De-duplication of search results also often takes place on Excel. ► Information specialists are not generally involved, given that the search process is simple.
Screening	► Screening of search results exported from databases. Uses tools including Rayyan, Covidence, Endnote, EppiReviewer. ► Two-stage process conducted in duplicate at each stage; disagreements resolved through consensus/consulting third reviewer. Full text excludes listed with reasons or available on request. ► A PRISMA flowchart is used to visually report the literature search and screening process.	► Screening of search results manually extracted into an Excel sheet. ► Two-stage process in which stage 1 includes screening the apps title and description on the app store. Stage 2 includes downloading the app and assessing eligibility. ► Two reviewers are generally involved and a third may help to reach consensus on any disagreements. Studies excluded at stage 2 are listed with reasons for exclusion. ► The PRISMA flowchart is often amended and used to report the app search and screening process.
Data extraction	► Data are extracted into a pre-specified and piloted form. Tools include Excel, Covidence, Revman, Eppi-Reviewer. ► Usually, data are extracted by one reviewer and checked by a second, sometimes duplicate extraction is used for some or all data.	► Data are manually extracted into a pre-specified form on Excel. ► Data may be extracted by one reviewer and checked by a second, or the task may be shared between reviewers.

Continued

Table 4 Continued

Review stage	Systematic literature review (of effectiveness) with a focus on quantitative reviews	Systematic commercial health app review
Data management	► Data may be transformed in various ways and processes may be implemented for the handling of missing data, such as assumption or imputation. ► There is extensive guidance on methodological approaches to challenges in data management.	► Not generally relevant for commercial health app reviews. ► Researchers may contact developers for more information about any evaluations that have taken place.
Quality appraisal	► A wide range of tools for assessment of risk of bias depending on study design and purpose; usually carried out in duplicate with disagreements resolved through consensus/ consulting third reviewer. ► Recorded in Excel, Revman, EppiReviewer. ► Risk of bias plots can be generated on RevMan or RobVis. ► Increasingly, reviews will also use Grading of Recommendations, Assessment, Development and Evaluation (GRADE) to rate the certainty of the evidence (from high to very low); GRADE assessment may use GradePRO.	► Quality is generally assessed using the MARS and recorded in Excel. ► Good practice requires each app to be reviewed independently by two raters. ► Inter-rater reliability is analysed and presented in the review.
Synthesis	► Meta-analysis may be conducted; Cochrane Handbook is a usual source of methods guidance; extensive literature exists on various aspects of this. ► Many reviews use narrative synthesis, which Cochrane offers guidance on, and there is recent guidance on SWiM (Synthesis without Meta-analysis).	► Data synthesis is generally performed descriptively by generating statistics (sums, averages, standard deviation and percentages) on relevant items. ► The highest-scoring apps (regarding functionality and quality) are identified. ► Descriptive summaries may be written for text-based items (eg, description of the main features). ► Inter-rater reliability can be calculated for the IMS Institute for Healthcare Informatics functionality scores using Cohen's Kappa statistic and an intraclass correlation coefficient for MARS scores.
Data presentation	► Meta-analyses (and sometimes non-pooled data) are presented using forest plots, often with risk of bias plots displayed alongside. ► Risk of bias results displayed using bespoke figures (see above); GRADE results typically displayed using Summary of Findings Tables.	► Data tend to be presented as descriptive summaries and Tables. Bespoke figures can also be created. ► Data pertaining to the IMS Institute for Healthcare Informatics functionality score is often presented as a radar graph/chart. ► Inter-rater reliability statistics are presented for both the MARS quality appraisals and the IMS Institute for Healthcare Informatics functionality scores.
Updating and currency	► Reviews should generally have a search date within the last 12 months at submission for publication. Searches can be updated by re-running searches with the relevant date limit: new records will be identified but old one will not be lost (the process is additive) (an exception may be the grey literature).	► Apps emerge, are updated, and disappear very quickly, so app reviews should be conducted and published as promptly as possible. ► New or updated searches in app stores will likely yield very different results, so updating a review is difficult.
Reporting	► PRISMA checklist	► No formal reporting guidelines exist for health app reviews.
Guidance	► The Cochrane Handbook and other guidance exists for specific reviews.	► A YouTube video shows how to use the MARS to assess quality.

MARS, Mobile App Rating Scale; ORCHA, Organisation for the Review of Care and Health Apps; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

of the search results. Authors must be aware of the need to report granular details of their searches and to publish their review within a reasonable time from the search date, such as by using preprint servers while peer-review in a scientific journal takes place. Furthermore, the resources available to the review team are more critical than in systematic reviewing, so it is essential to transparently document scoping searches and iterative adjustments to review scope and inclusion criteria.[58]

Second, the critical appraisal process for health app reviews requires a radically different approach to quality assessments for literature reviews, with multiple approaches and tools being used to explore app functionality and quality. This proliferation of assessment approaches means there is no 'gold standard' equivalent to the Cochrane Risk of Bias tool for RCTs in systematic reviews.[48 59] There is also no equivalent to the wider consideration of evidence certainty which is provided by Grading of Recommendations, Assessment, Development and Evaluation (GRADE) in systematic reviews.[60] Researchers must select an approach that best fits the review aims. Additionally, they may need to consider national standards within some countries, such as those from the UK National Institute for Health Research on digital health technologies evaluation,[61] which may require specific approaches for the review context.

The third significant difference between systematic literature and health app reviews is the extent to which guidance, guidelines and infrastructure support them. The methodological and reporting guidance for health app reviews is in its infancy. In contrast, an extensive body of literature outlining methods for different types of systematic and now scoping and rapid reviews exists.[62–64] There are also clear reporting guidelines for systematic reviews, which have been expanded to scoping reviews but are lacking for health app reviews.[34 65] While we are undertaking research to develop methods and reporting guidance for app reviews, systematic review guidance should be referred to and adapted as necessary as an interim measure. Ultimately, there may also be a need for a tool which will allow critical appraisal of these types of mHealth reviews to parallel those that exist for systematic reviews such as AMSTAR-2[66] or ROBIS.[67]

We propose that this outline will guide the conduct of good quality app reviews that can inform healthcare practitioners, patients, carers, health service managers, educators, and policymakers. We also recommend the prospective registration of an app review protocol on OSF as a suitable alternative to PROSPERO where systematic review protocols are held,[68] and the use of preprint servers to make app reviews openly available online, allowing for rapid dissemination of findings ahead of journal publication.

Implications

This outline will help ensure that others can easily replicate the methods and that future app reviews are conducted in a standardised and rigorous manner. However, while outlining the methods, we noted gaps in conducting and reporting commercial health app reviews that need addressing. Hence, we are developing reporting guidelines for systematic health app reviews and plan to subsequently develop a quality appraisal tool. Similar to the 27-item PRISMA guideline for systematic reviews[34] and the 22-item PRISMA-Scr guideline for scoping reviews,[65] our guideline will consist of a structured list of items that should be included when reporting commercial health app reviews.

For those conducting commercial health app reviews, there is an opportunity for the inclusion of stakeholders to strengthen the quality and impact of their findings. This is particularly beneficial if the intended target audience experiences barriers when using health apps, as clear recommendations from an app review can help to improve the design and function of future versions of an app. Researchers should also be aware of the context in which the review is being conducted. Namely, companies owning the apps may use the review for business development and promotion opportunities or contest the quality scores. However, this highlights an opportunity for further stakeholder engagement: researchers could collaborate or consult with developers to ensure that the product aligns with the research assessment process of an app's quality. This has the potential to influence and promote accessibility and quality as aspects of development that might not be considered otherwise. While industry developers focus on creating a commercially viable product, understanding this review process will potentially enhance and refine their development process to create a superior app than initially proposed. Ultimately, it is important to be aware of any conflicts of interest between researchers who are conducting reviews in systematic and robust ways, and industry who may wish to promote their work and financially benefit from the review findings. As with systematic reviews, collaborations which have the potential to generate such conflicts of interest should be fully and transparently reported in reviews, and review methods which minimise their potential impact should be implemented.

Strengths and limitations

This methodological discussion has numerous strengths, such as an experienced research team who have conducted various health app reviews and various types of traditional literature and systematic reviews. This was supplemented by identifying and including relevant app reviews from the top medical informatics journals and a robust analysis and comparison between traditional systematic reviews and commercial health app reviews. However, a systematic search of health app reviews was not undertaken (this will form part of our work in developing reporting guidance), nor did we focus on the emerging field of extended reality (ie, virtual, augmented and mixed reality) and their corresponding apps, many of which are health-related and available in other app stores (eg, Steam and Oculus/Meta). Additionally, our app reviews focused on apps for clients (eg, patients or the public) and healthcare providers rather than for the health system or data services, which are also target users of digital interventions.[69] It is likely that the recommendations in this discussion about evaluating commercially available mHealth apps will also apply to other health apps, including extended reality, and could support researchers working in these fields.

CONCLUSION

Reviews of commercial apps can provide insights into the availability of apps for a specific health topic, including their quality and functionality. We have proposed a 7-step method in an effort to standardise the process of conducting mHealth reviews. At each step, we have discussed the methods in contrast to systematic literature reviews, given that the process should similarly be systematic and robust. We have also introduced the novel TECH acronym, which will assist researchers with writing research questions for app reviews and determining the eligibility criteria. Through ongoing collaboration, we will continue to advocate for transparency and quality in app reviews by working on reporting guidelines and a quality appraisal tool.

Twitter Siobhan O'Connor @shivoconnor

Contributors NG and DD conceptualised the paper. NG, DD, GN, LM, CE-T, DJ, AV, SMA and SO made intellectual contributions and contributed to writing the paper. All authors approved the final manuscript.

Funding NG, DD, GN, CE-T and SMA are funded by the National Institute for Health and Care Research Applied Research Collaboration Greater Manchester (NIHR ARC GM). LM is supported by a National Institute for Health and Care Research Senior Investigator Award to Professor Chris Todd (Reference NIHR200299). The views expressed in this publication are those of the authors and not necessarily those of the National Institute for Health and Care Research or the Department of Health and Social Care. Grant numbers: N/A.

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs

Norina Gasteiger http://orcid.org/0000-0001-7801-7417
Gill Norman http://orcid.org/0000-0002-3972-5733
Lisa McGarrigle http://orcid.org/0000-0002-0533-3029
Charlotte Eost-Telling http://orcid.org/0000-0002-9568-3195

REFERENCES

1 Mayer MA, Rodríguez Blanco O, Torrejon A. Use of health Apps by nurses for professional purposes: web-based survey study. *JMIR Mhealth Uhealth* 2019;7:e15195.
2 Bates DW, Landman A, Levine DM. Health Apps and health policy: what is needed *JAMA* 2018;320:1975–6.
3 Dsouza SM, Shetty S, Venne J, *et al*. Effectiveness of self-management applications in improving clinical health outcomes and adherence among diabetic individuals in low and middle-income countries: a systematic review. BMJ Open 2022;12:e060108.
4 Doumen M, De Cock D, Van Lierde C, *et al*. Engagement and attrition with eHealth tools for remote monitoring in chronic arthritis: a systematic review and meta-analysis. *RMD Open* 2022;8:e002625.
5 De Santis KK, Jahnel T, Matthias K, *et al*. Evaluation of digital interventions for physical activity promotion: Scoping review. JMIR Public Health Surveill 2022;8:e37820.
6 Chew HSJ, Koh WL, Ng JSHY, *et al*. Sustainability of weight loss through Smartphone Apps: systematic review and meta-analysis on Anthropometric, metabolic, and dietary outcomes. *J Med Internet Res* 2022;24:e40141.
7 Dasgupta N, Lazard A, Brownstein JS. Covid-19 vaccine Apps should deliver more to patients. *Lancet Digit Health* 2021;3:e278–9.
8 Lee M, Lee H, Kim Y, *et al*. Mobile App-based health promotion programs: a systematic review of the literature. Int J Environ Res Public Health 2018;15:2838.
9 World Health Organisation. Applications, 2022. Available: https://www.who.int/news-room/apps
10 Larsen ME, Huckvale K, Nicholas J, *et al*. Using science to sell Apps: evaluation of mental health App store quality claims. *NPJ Digit Med* 2019;2:18.
11 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;34:464–75.
12 Henson P, David G, Albright K, *et al*. Deriving a practical framework for the evaluation of health Apps. *Lancet Digit Health* 2019;1:e52–4.
13 Lagan S, Aquino P, Emerson MR, *et al*. Actionable health App evaluation: translating expert Frameworks into objective Metrics. *NPJ Digit Med* 2020;3:100.
14 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;25:1089–98.
15 Manta C, Mahadevan N, Bakker J, *et al*. EVIDENCE publication checklist for studies evaluating connected sensor Technologies: explanation and elaboration. *Digit Biomark* 2021;5:127–47.
16 Agarwal S, LeFevre AE, Lee J, *et al*. Guidelines for reporting of health interventions using mobile phones: mobile health (mHealth) evidence reporting and assessment (mERA) checklist. BMJ 2016;352:i1174.
17 Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. *J Med Internet Res* 2011;13:e126.
18 Weisel KK, Fuhrmann LM, Berking M, *et al*. Standalone Smartphone Apps for mental health—a systematic review and meta-analysis. *NPJ Digit Med* 2019;2:118.
19 Gasteiger N, Dowding D, Ali SM, *et al*. Sticky apps, not sticky hands: a systematic review and content synthesis of hand hygiene mobile apps. J Am Med Inform Assoc 2021;28:2027–38.
20 Gasteiger N, Vercell A, Davies A, *et al*. Patient-facing genetic and genomic mobile apps in the UK: a systematic review of content, functionality, and quality. J Community Genet 2022;13:171–82.
21 Khan KS, Kunz R, Kleijnen J, *et al*. Five steps to conducting a systematic review. J R Soc Med 2003;96:118–21.
22 Xiao Y, Watson M. Guidance on conducting a systematic literature review. *J Plan Educat Res* 2019;39:93–112.
23 Pearsons A, Hanson CL, Gallagher R, *et al*. Atrial fibrillation self-management: a mobile telephone APP scoping review and content analysis. Eur J Cardiovasc Nurs 2021;20:305–14.
24 Vercell A, Gasteiger N, Yorke J, *et al*. Patient-facing cancer mobile apps that enable patient reported outcome data to be collected: a systematic review of content, functionality, quality, and ability to integrate with electronic health records. *Int J Med Inform* 2023;170:104931.
25 Paripoorani D*et al*. A systematic review of menopause apps with an emphasis on osteoporosis in progress.
26 Ali SM, Lau WJ, McBeth J, *et al*. Digital manikins to Self-Report pain on a Smartphone: a systematic review of mobile Apps. *Eur J Pain* 2021;25:327–38.
27 McGarrigle L, Boulton E, Todd C. Map the apps: a rapid review of digital approaches to support the engagement of older adults in strength and balance exercises. BMC Geriatr 2020;20:483.
28 Booth A, Clarke M, Dooley G, *et al*. The nuts and bolts of Prospero: an international prospective register of systematic reviews. Syst Rev 2012;1:2.
29 Cooke A, Smith D, Booth A. Beyond PICO: the SPIDER tool for qualitative evidence synthesis. *Qual Health Res* 2012;22:1435–43.
30 Parmar P, Ryu J, Pandya S, *et al*. Health-focused conversational agents in person-centered care: a review of apps. NPJ Digit Med 2022;5:21.
31 Carlo AD, Hosseini Ghomi R, Renn BN, *et al*. By the numbers: ratings and utilization of behavioral health mobile applications. npj Digital Medicine 2019;2:54.

32 Veritas Health Innovation. Covidence systematic review software. 2023. Available: www.covidence.org
33 Ouzzani M, Hammady H, Fedorowicz Z, *et al*. Rayyan-a web and mobile APP for systematic reviews. Syst Rev 2016;5:210.
34 Page MJ, McKenzie JE, Bossuyt PM, *et al*. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71.
35 Rajani NB, Weth D, Mastellos N, *et al*. Use of gamification strategies and tactics in mobile applications for smoking cessation: a review of the UK mobile APP market. BMJ Open 2019;9:e027883.
36 Internet Society. Best practices: privacy. 2019. Available: https://www.internetsociety.org/resources/ota/2019/best-practices-privacy
37 Lalloo C, Shah U, Birnie KA, *et al*. Commercially available smartphone Apps to support postoperative pain self-management: Scoping review. JMIR Mhealth Uhealth 2017;5:e162.
38 Grainger R, Townsley H, White B, *et al*. Apps for people with rheumatoid arthritis to monitor their disease activity: a review of Apps for best practice and quality. JMIR Mhealth Uhealth 2017;5:e7.
39 Flesch R. How to write plain English: a book for lawyers and consumers. New York: Harper & Row, 1979.
40 Kincaid J*et al*. Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted personnel. research branch report 8-75, 1975. Naval Technical Training, U. S. Naval Air Station: Millington, TN.
41 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:e72.
42 Domnich A, Arata L, Amicizia D, *et al*. Development and validation of the Italian version of the mobile application rating scale and its generalisability to apps targeting primary prevention. BMC Med Inform Decis Mak 2016;16:83.
43 Martin Payo R, Fernandez Álvarez MM, Blanco Díaz M, *et al*. Spanish adaptation and validation of the mobile application rating scale questionnaire. Int J Med Inform 2019;129:95–9.
44 Messner E-M, Terhorst Y, Barke A, *et al*. The German version of the mobile APP rating scale (MARS-G): development and validation study. JMIR Mhealth Uhealth 2020;8:e14479.
45 Saliasi I, Martinon P, Darlington E, *et al*. Promoting health via mHealth applications using a French version of the mobile APP rating scale: adaptation and validation study. JMIR Mhealth Uhealth 2021;9:e30480.
46 Terhorst Y, Philippi P, Sander LB, *et al*. Validation of the mobile application rating scale (MARS). *PLoS ONE*One 2020;15:e0241480.
47 Stoyanov S. MARS training Video. 2016. Available: https://www.youtube.com/watch?v=25vBwJQIOcE
48 Higgins JPT, Altman DG, Gøtzsche PC, *et al*. The Cochrane collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011;343:d5928.
49 Aitken M, Gauntlett C. IMS Institute for Healthcare informatics: patient Apps for improved healthcare. From Novelty to Mainstream 2013 http://ignacioriesgo.es/wp-content/uploads/2014/03/iihi_patient_apps_report_editora_39_2_1.pdf
50 Plante TB, O'Kelly AC, Macfarlane ZT, *et al*. Trends in user ratings and reviews of a popular yet inaccurate blood pressure-measuring smartphone app. J Am Med Inform Assoc 2018;25:1074–9.
51 Levine DM, Co Z, Newmark LP, *et al*. Design and testing of a mobile health application rating tool. NPJ Digit Med 2020;3:74.
52 Tangari G, Ikram M, Sentana IWB, *et al*. Analyzing security issues of android mobile health and medical applications. J Am Med Inform Assoc 2021;28:2074–84.
53 Chidambaram S, Erridge S, Kinross J, *et al*. Observational study of UK mobile health apps for COVID-19. Lancet Digit Health 2020;2:e388–90.
54 McHugh ML. Interrater reliability: the kappa statistic. Biochemia Medica 2012;22:276–82.
55 Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. Tutor Quant Methods Psychol 2012;8:23–34.
56 Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420–8.
57 Serrano-Aguilar P, Trujillo-Martín MM, Ramos-Goñi JM, *et al*. Patient involvement in health research: a contribution to a systematic review on the effectiveness of treatments for degenerative ataxias. Soc Sci Med 2009;69:920–5.
58 Nussbaumer-Streit B, Ellen M, Klerings I, *et al*. Resource use during systematic review production varies widely: a scoping review. *Journal of Clinical Epidemiology* 2021;139:287–96.
59 Sterne JAC, Savović J, Page MJ, *et al*. Rob 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* 2019;2:l4898.
60 Guyatt G, Oxman AD, Akl EA, *et al*. GRADE guidelines: 1. introduction—GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol* 2011;64:383–94.
61 National Institute for Health and Care Excellence. Evidence standards framework (ESF) for Digital health Technologies. 2022. Available: https://www.nice.org.uk/about/what-we-do/our-programmes/evidence-standards-framework-for-digital-health-technologies#how-use
62 Higgins J*et al*. *Cochrane Handbook for systematic reviews of interventions version 6.3*. 2022. Available: www.training.cochrane.org/handbook
63 Garritty C, Gartlehner G, Nussbaumer-Streit B, *et al*. Cochrane rapid reviews methods group offers evidence-informed guidance to conduct rapid reviews. J Clin Epidemiol 2021;130:13–22.
64 Peters MDJ, Marnie C, Tricco AC, *et al*. Updated methodological guidance for the conduct of scoping reviews. *JBI Evid* Synth 2020;18:2119–26.
65 Tricco AC, Lillie E, Zarin W, *et al*. PRISMA extension for Scoping reviews (PRISMA-SCR): checklist and explanation. *Ann Intern Med* 2018;169:467–73.
66 Shea BJ, Reeves BC, Wells G, *et al*. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of Healthcare interventions, or both. *BMJ* 2017;358:j4008.
67 Whiting P, Savović J, Higgins JPT, *et al*. ROBIS: A new tool to assess risk of bias in systematic reviews was developed. *J Clin Epidemiol* 2016;69:225–34.
68 Booth A, Clarke M, Dooley G, *et al*. PROSPERO at one year: an evaluation of its utility. *Syst Rev* 2013;2:4.
69 World Health Organization. Classification of digital health interventions. Geneva, 2018.
70 Michie S, Richardson M, Johnston M, *et al*. The behavior change technique taxonomy (V1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. ann. behav. med. 2013;46:81–95.