PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	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
AUTHORS	Gasteiger, Norina; Dowding, Dawn; Norman, Gill; McGarrigle, Lisa; Eost-Telling, Charlotte; Jones, Debra; Vercell, Amy; Ali, Syed; O'Connor, Siobhan

VERSION 1 – REVIEW

REVIEWER	Rajani, Nikita Imperial College London School of Public Health, Department of Primary care and Public Health
REVIEW RETURNED	15-Mar-2023

GENERAL COMMENTS	The paper is well-written and does provide valuable content that could be of interest to academics/researchers but also industry professionals. Whilst no comprehensive framework for the process of conducting mobile app reviews has been published, it is unclear how exactly the steps described were formulated. Although it says that the authors developed the steps based on their own reviews and handsearched reviews of other apps, it is not very clear how exactly they created this 7 step approach. Did they use some sort of systematic methodology to come up with these steps? Were there any differences between the methodologies that were adopted by various reviews? Were the 7 steps they describe followed by all the reviews that they have personally conducted (I think 7 reviews are specifically mentioned in table 1). The methods section is quite unclear and unstructured with little information on how the 7-steps have actually been derived and also how the TECH framework was developed. It would be nice to have a separate subtitle before step 1 is introduced under methods as the 7 steps are actually the “results” of the paper rather than the methods. Is the methods section about the methodology that the authors adopted to derive the framework/steps for conducting a mobile app review or the actual method itself for conducting a review? Some clarity would be useful for readers. Step 1 – it is unclear how TECH framework was developed and what is meant by target user, evaluation focus, connectedness, and health domain. These four components are listed and the purpose of TECH is articulated. However, what does target user mean in this context? And how does it help formulate the research question? Perhaps an example of this worked out would be valuable for readers.

	Step 2 – information about what a scoping search is and why it is important is clear. However, some information about how to conduct a scoping search would be useful and valuable for readers. How is a scoping search done? Generally, you need search terms – how many search terms should there be and how are these determined/what are they dependent on? Step 3 – is the TECH framework used to develop research question and eligibility criteria then? Here more information about TECH is provided (maybe this can be mentioned before as a response to comment for step 1). Step 4 – search terms are mentioned here as if they were a result of step 3 but there is no mention of search terms in step 3. If search terms are determined in step 3, how is this actually done? Step 5 – useful to see different items that can be “extracted” from apps when being tested. Some information on how long extraction takes or how long apps are tested/reviewed for by researchers would also be interesting and useful. Should the apps be reviewed for 10 minutes? 30 minutes? What is this time period dependent on? Often the number of apps tested can be chosen based on the number of criteria and extent of testing that is required? E.g. a general overview of the market vs. in-depth content extraction. Discussion – interesting point about working together with developers and their possible use of mobile app review results for business develop and commercial purposes. Would also be interesting to mention conflicts of interest in terms of research goals and app developer interests.

REVIEWER	Nittas, Vasileios University of Zurich, Epidemiology, Biostatistics and Prevention Institute
REVIEW RETURNED	09-Apr-2023

GENERAL COMMENTS	This is an important attempt to provide some form of guidance for the screening of commercial health apps. My concerns are outlined below 1. Abstract. Please make clear(er) you address the review within app stores. 2. Step 2. Scoping reviews can get time consuming. Here I would like to see some recommendations of where to put the boundaries and how deep/nuanced that step should be. Also, you mention " Following the initial screening of app titles and app store descriptions, this number was significantly reduced, and only 13 were included in the review." -> this sentence got me confused as it related to the actual searches correct? I do assume that the scoping stage does not include any detailed screening? please clarify 4. Step 4: I would like to see more explicit examples of filters in in the different app stores...there limitations and how searches can be created to be sensitive enough without getting out of hand 5. Step 4: is there any software / AI tools out there that could support with the steps of exracting / exporting? If yes, would be nice to report them here…

	6. Step 4: Do you recommend recording excluded apps (in both stages) with reasons? Please clarify 7. Step 5: often information is not readily available / transparently reported in apps. Any recommendations on how to mitigate that and get the required information? 8. Table 2 I would place in the appendix. 9. Table 4: A bit too lengthy. To improve readability I would shorten and use bullet points.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Ms. Nikita Rajani, Imperial College London School of Public Health

Comments to the Author:

The paper is well-written and does provide valuable content that could be of interest to academics/researchers but also industry professionals. Whilst no comprehensive framework for the process of conducting mobile app reviews has been published, it is unclear how exactly the steps described were formulated.

Although it says that the authors developed the steps based on their own reviews and handsearched reviews of other apps, it is not very clear how exactly they created this 7 step approach. Did they use some sort of systematic methodology to come up with these steps? Were there any differences between the methodologies that were adopted by various reviews? Were the 7 steps they describe followed by all the reviews that they have personally conducted (I think 7 reviews are specifically mentioned in table 1). The methods section is quite unclear and unstructured with little information on how the 7-steps have actually been derived and also how the TECH framework was developed.

> In the Methods we have mentioned that the case studies were used to develop the new framework, outline the methods and for the discussion of the methodological issues:
>
> 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.
>
> 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.
>
> We have also explained that we drew on some existing guidance to formulate the 7 steps:

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 7 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 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.

We have added some information on how TECH was developed in Step 2, when we first mention it. We have amended Figure 1 to highlight some of the similarities of the acronym to other existing frameworks (SPIDER and PICO). We now state:

TECH was designed through discussion by the research team and by mapping key concepts against existing frameworks (e.g., 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[26].

It would be nice to have a separate subtitle before step 1 is introduced under methods as the 7 steps are actually the “results” of the paper rather than the methods. Is the methods section about the methodology that the authors adopted to derive the framework/steps for conducting a mobile app review or the actual method itself for conducting a review? Some clarity would be useful for readers.

In the Methods we have mentioned that the case studies were used to develop the new framework, outline the methods and for the discussion of the methodological issues:

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.

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.

We have also added a Results title and introduced the 7 steps:

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 – it is unclear how TECH framework was developed and what is meant by target user, evaluation focus, connectedness, and health domain. These four components are listed and the purpose of TECH is articulated. However, what does target user mean in this context? And how does it help formulate the research question? Perhaps an example of this worked out would be valuable for readers.

We have added some information on how TECH was developed, in addition to some information about the components and a worked example in Figure 1.

TECH was designed through discussion by the research team and by mapping key concepts against existing frameworks (e.g., 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[26].

Figure 1. TECH framework with a worked example.

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

Step 2 – information about what a scoping search is and why it is important is clear. However, some information about how to conduct a scoping search would be useful and valuable for readers. How is a scoping search done? Generally, you need search terms – how many search terms should there be and how are these determined/what are they dependent on?

We have clarified that researchers should use basic keywords for the searches and that these should be refined iteratively (by changing the scope of the topic or keywords) based on the number of apps that are returned. Some examples are provided. We also added that researchers can determine the number of potential apps by considering the relevance to the topic by reading the app's name and description and then counting those which are relevant. This section now reads:

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 focussed 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 focussed 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.

Step 3 – is the TECH framework used to develop research question and eligibility criteria then? Here more information about TECH is provided (maybe this can be mentioned before as a response to comment for step 1).

Yes- We have added some information on how TECH was developed, in addition to some information about the components and worked example for the research question/aim and inclusion criteria in Figure 1 (see above).

TECH was designed through discussion by the research team and by mapping key concepts against existing frameworks (e.g., 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[26].

Step 4 – search terms are mentioned here as if they were a result of step 3 but there is no mention of search terms in step 3. If search terms are determined in step 3, how is this actually done?

As described above, we have added more information to Step 2 on the keywords used for the searches:

We recommend that researchers use basic keywords focussed 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 focussed on the target user (patients): cancer, cancer patient, cancer treatment, cancer management and cancer side effects.

Step 5 – useful to see different items that can be “extracted” from apps when being tested. Some information on how long extraction takes or how long apps are tested/reviewed for by researchers would also be interesting and useful. Should the apps be reviewed for 10 minutes? 30 minutes? What is this time period dependent on? Often the number of apps tested can be chosen based on the number of criteria and extent of testing that is required? E.g. a general overview of the market vs. in-depth content extraction.

It is difficult to give an exact timeframe, due to many factors that influence how long an app needs to be used for, in order to extract the relevant information. We have added a paragraph to Step 5 that acknowledges this:

Data is extracted into a predefined 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 to send push notifications at specific times of the day (e.g., for behaviour change apps).

Discussion – interesting point about working together with developers and their possible use of mobile app review results for business develop and commercial purposes. Would also be interesting to mention conflicts of interest in terms of research goals and app developer interests.

We have stated that an awareness of the conflicts of interest is important:

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. Whilst 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.

Reviewer: 2

Mr. Vasileios Nittas, University of Zurich

Comments to the Author:

This is an important attempt to provide some form of guidance for the screening of commercial health apps. My concerns are outlined below:

1. Abstract. Please make clear(er) you address the review within app stores.

We have made this clearer in two instances (the Design section and Results):

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 (e.g., 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.

2. Step 2. Scoping reviews can get time consuming. Here I would like to see some recommendations of where to put the boundaries and how deep/nuanced that step should be. Also, you mention " Following the initial screening of app titles and app store descriptions, this number was significantly reduced, and only 13 were included in the review." -> this sentence got me confused as it related to the actual searches correct? I do assume that the scoping stage does not include any detailed screening? please clarify

To clarify, this section is about scoping searches, not scoping reviews. The scoping searches are to help determine whether the number of commercial health apps available is feasible to review. We have provided the example so that others have an idea of what might be feasible for their project and to highlight that although a large number of apps might initially be identified, only a few may actually be included in the review. We have clarified this and also

stated that researchers can count the number of potential apps by considering the app's name and description:

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[23], 405[24], 555[26], 668[19], 754[20] and 3938[25] health apps from initial searches, before screening or deduplication took place.

4. Step 4: I would like to see more explicit examples of filters in in the different app stores...there limitations and how searches can be created to be sensitive enough without getting out of hand

We have added some more information to Step 4 about the filters available in different app stores:

Basic filters may exclude apps that cost or only include child-friendly apps. Some stores (e.g., the Google Play store) also enable for users to identify family-friendly apps and distinguish the type of app being searched (i.e., 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).

5. Step 4: is there any software / AI tools out there that could support with the steps of exracting / exporting? If yes, would be nice to report them here…

No- we are currently unaware of any software that could help with extracting the information or exporting the results.

6. Step 4: Do you recommend recording excluded apps (in both stages) with reasons? Please clarify

We have clarified that the screening process should follow the PRISMA guidance:

Finally, modifying a 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.

7. Step 5: often information is not readily available / transparently reported in apps. Any recommendations on how to mitigate that and get the required information?

We have acknowledged that information is not always readily available within apps. In the section on Step 5 we now state:

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.

8. Table 2 I would place in the appendix.

We would like to keep this in the main text, as readers may be unlikely to look in the appendix. We believe the examples of data extraction items will be useful for researchers conducting systematic app reviews.

9. Table 4: A bit too lengthy. To improve readability I would shorten and use bullet points.

We have used bullet points and shortened the content.