BMJ Open

Here's something I prepared earlier: a review of the time to publication of cross-sectional reviews of smartphone health apps

Mark Larsen ,[1] Jennifer Nicholas,[2,3] Jin Han,[1] Christopher Lemon,[4,5] Kelsi Okun,[6] Michelle Torok,[1] David Wong ,[7] Iana Wong,[1] Quincy Wong,[8] Kit Huckvale[1]

To cite: Larsen M, Nicholas J, Han J, *et al*. Here's something I prepared earlier: a review of the time to publication of cross-sectional reviews of smartphone health apps. *BMJ Open* 2020;**10**:e039817. doi:10.1136/bmjopen-2020-039817

► Prepublication history and additional materials for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2020-039817).

Received 27 April 2020
Revised 09 November 2020
Accepted 17 November 2020

For numbered affiliations see end of article.

Correspondence to
Dr Mark Larsen;
mark.larsen@blackdog.org.au

ABSTRACT

Objectives Across a range of health conditions, apps are increasingly valued as tools for supporting the delivery and coordination of healthcare. Research-led cross-sectional reviews of apps are a potential resource to inform app selection in face of uncertainties around content quality, safety and privacy. However, these peer-reviewed publications only capture a snapshot of highly dynamic app stores and marketplaces. To determine the extent to which marketplace dynamics might impact the interpretation of app reviews, the current study sought to quantify the lag between the reported time of app assessment and publication of the results of these studies.

Design Searches were conducted on MEDLINE, Embase and PsycINFO to identify published cross-sectional reviews of health, fitness or wellness apps. Publication timeline metadata were extracted, allowing the primary outcome measure, the delay between app store search and manuscript publication, to be calculated. A secondary measure, the time between search and manuscript submission, was also calculated where possible.

Results After screening, 136 relevant cross-sectional app review studies were analysed. The median time to publication was 431 days (approximately 14 months, range: 42–1054 days). The median time to submission was 269 days (approximately 9 months, range: 5–874 days). Studies which downloaded apps typically took longer to publish (p=0.010), however the number of apps reviewed did not impact the time to publication (p=0.964). Studies which recommended specific apps were not published more rapidly (p=0.998).

Conclusions Most health app reviews present data that are at least a year out-of-date at the time of publication. Given the high rate of turnover of health apps in public marketplaces, it may not be appropriate, therefore, for these reviews to be presented as a resource concerning specific products for commissioners, clinicians and the public. Alternative sources of information may be better calibrated to the dynamics of the app marketplace.

Strengths and limitations of this study

- This review considers over a decade of published cross-sectional reviews of health apps.
- The age of the review findings, at the time of publication, was determined and compared with the observed rate of change of the app stores.
- The time to journal submission was also calculated, where possible, providing an indication of the quickest possible time for results to be made publicly available to inform decisions.
- Heterogeneity across reviewed clinical and technical domains may impact publication time.

INTRODUCTION

Rationale

Smartphone applications (apps) are increasingly valued as tools for supporting the delivery and coordination of healthcare. Across a range of health conditions, there is growing evidence that app-based self-care interventions can be effective at reducing symptoms,[1] supporting self-management[2 3] and promoting health behaviour change.[4 5] In 2017, half of surveyed Australian primary care doctors reported recommending apps to their patients at least once a month.[6] Across both physical[7] and mental health,[8] consumers either indicate interest in using health apps or report having already attempted to integrate apps into their health management. At a systems level, there is growing interest in the potential for digital health to enable value-based care that offers potential resource savings compared with face-to-face therapies. Examples of established initiatives include the Australian Federal Government's e-Mental Health Strategy,[9] which seeks to increase the accessibility and reach of mental health support while decreasing load on traditional services, and the state of California's Technology Suite Collaborative, which is harnessing digital technology to expand the capacity and capability of the county mental health systems, again to decrease burden on traditional care pathways.[10] Most recently, the National Health Service (NHS) in England

stated its intention for 'digitally enabled primary and outpatient care (to) go mainstream' across the entire health system as part of its long-term plan.[11]

Quality remains a key concern for healthcare providers seeking to integrate health apps into routine care. Content quality, safety and privacy deficits have been identified in a wide range of health app categories.[12–15] Despite recent progress in clarifying regulatory requirements around 'software as a medical device' in the USA and Europe, only a small fraction of available apps either fall into a category that requires formal regulatory review or have been subject to experimental evaluation. Indeed, the number of health apps evaluated through randomised studies within the research literature is dwarfed by the numbers available to consumers.[16] These apps are typically made available within the same commercial marketplaces as apps for navigation, social media and finance. This combination of prevalent quality issues and potentially large numbers of options presented without technical differentiation represents a major challenge for healthcare systems, clinicians and consumers trying to select high quality, clinically appropriate apps.

Research-led cross-sectional reviews of published apps ('app reviews') that critically appraise aspects of app quality and safety are a potential resource for healthcare practitioners, patients and the public when choosing an appropriate health app. Indeed, many app reviews either state, as aims, an intention to guide health professionals and consumers to the best apps for a given health condition or make recommendations targeting clinicians in discussion.[17] Evidence of the potential impact on clinical practice and policy of these cross-sectional studies include citations in clinical guidelines,[18] professional guidance concerning health app use[19 20] and design,[21] health system policy documents,[22 23] and expert[24] and intergovernmental[25] reports. Tools commonly used in app reviews, such as the Mobile App Rating Scale (MARS),[26] were developed with the explicit goal of providing an app evaluation resource for use by health professionals (as well as researchers.)

There is now a substantial collection of such reviews; our searches identified at least 149 such studies published between 2008 and 2019. However, there is a critical and widely acknowledged limitation of these reviews: they are static snapshots of a volatile environment. Within app stores, app updates, additions, removals, and search result list changes are common and unpredictable, and may be further compounded by different app listings and availability in different jurisdictions. In 2016, the dynamic nature of the two leading commercial app environments was quantified.[27] Tracking the search results for depression, bipolar disorder and suicide prevention apps each day over 9 months, findings indicated that half of the Google Play search results change approximately every 4 months. Moreover, across both platforms, an app for depression became unavailable to download every 2.9 days.[27] These dynamic changes highlight the potential for information contained in cross-sectional reviews to become out-of-date, limiting its validity if used for the purposes of selecting and recommending specific health apps.

Objective

In order to explore the extent to which marketplace dynamics might impact the interpretation of research-led app reviews, the current study sought to quantify the lag between the reported time of app assessment and publication of the results of these studies. We assessed the impact on time to publication (TTP) of specific features of the review process likely to act as a proxy for researcher workload, such as whether assessment involved downloading and reviewing app content. Finally, given that some app reviews explicitly state their intention to influence professional and patient behaviour, for example by recommending specific apps for use, we tested the hypothesis that these studies would be published more rapidly. This review focuses on published reviews of health and well-being apps which could be downloaded onto a smartphone (typically, but not exclusively, native apps via the Apple App Store or Google Play Store), without limitation on app functionality. Therefore, within this study, no constraints were placed on what review authors defined as health and well-being apps, as long as the review focused on a topic related to fitness, wellness or health, and no restrictions were placed on health domain. Apps available through curated third-party lists or libraries were also considered.

METHODS

Literature search

We aimed to identify reported studies that performed cross-sectional analysis, assessments, or reviews of smartphone health and well-being apps. To identify studies, we developed a bespoke literature search strategy. Working separately, two reviewers (KH and JN) first performed exploratory searches of articles published in 2018 indexed by the MEDLINE citation database. Each reviewer used these searches to try to devise, respectively, a specificity-maximising and sensitivity-maximising search strategy (detailed in online supplemental file 1).

In order to evaluate the performance of these alternative strategies, the results of each search (specificity-maximising n=78, sensitivity-maximising n=220) were independently screened by two reviewers (ML and JN). After reconciling any differences, screening yielded a binary partition of relevant/non-relevant studies for each search strategy. Subsequent comparison of these results indicated the overall suitability of a specificity-maximising approach. Individual discrepancies in included/excluded studies were also reviewed, yielding qualitative judgements about the likely contributions of different search terms to the observed results.

We used this information to devise a unified search strategy based on the original specificity-maximising approach. This strategy was then rerun on the original

Box 1 Optimised specificity-maximising search strategy.

(apps or applications or (app adj development)).ti.
AND
(smartphone? or mobile? or cell? or cellular? or (smart adj phone?) or iphone?).ab,ti.
AND
(review or (cross adj sectional) or content or quality or survey or assessment).ti. or (mobile adj2 rating adj scale?).ab. or (google adj play).ab.
AND
(appstore? or store? or marketplace? or (market adj place?) or apple or google or android or download$ or (app$ adj rating adj scale?)).ab.

sample of 2018 citations to confirm that no relevant citations were omitted. In a final step, we broadened the search to include all years and selectively removed terms from the strategy to ascertain their impact on the final result set. Terms that did not alter the overall result count were discarded. Search results were also reviewed to ensure that studies already known to the reviewers were captured by the strategy. The final search strategy is detailed in box 1.

Searches were run on MEDLINE, Embase and PsycINFO on 30 April 2019 and included all studies published between 2008 (on the basis that this was the year in which the first commercial app store was launched) and the search date. Search results were combined and deduplicated before screening.

Eligibility criteria

Search result titles and abstracts were reviewed against a standard set of inclusion criteria. Studies were retained if they (1) focused on a topic relating to health, fitness or wellness (irrespective of whether the intended app users were consumers, carers, clinicians, researchers or some combination of these); (2) involved a cross-sectional search of an app store or library intended to generate a set of apps for subsequent examination; and (3) applied one or more methods to this set to evaluate either the metadata associated with each app (such as app store descriptions), the contents of each app or both.

Study selection

Each study was screened by two out of three reviewers (of KH, ML and JN), working independently, with any discrepancies resolved by the third reviewer. Inter-rater agreement during initial screening of studies returned by the original search strategy (n=78) was calculated using Fleiss' kappa at 0.78, indicating substantial agreement between reviewers.

Data extraction

Following screening, the full text of each included study was obtained for data extraction. Table 1 details the data elements that were extracted, if available. Coding aimed to quantitate the time taken to publish each study and identify proxy measures of the effort required for its completion (eg, number of apps included in the study, and whether apps were downloaded as part of the review process.) A final parameter concerning whether study authors identified specific apps in their results or discussion (eg, to recommend for, or caution against, use)

Table 1 Data extracted from each included study

Category	Item	Description
Publication timeline	Earliest search date	The earliest date authors report searching the app stores.
	Latest search date	If app store searches were conducted over a period of time, the latest date authors report searching the app stores.
	Updated search date	If subsequent app store searches were performed, for example to update the initial search results, the latest date authors report conducting the updated search.
	Submission date	Date of manuscript submission to journal.
	Publication date	Earliest identified date at which the accepted, peer-reviewed manuscript is made available to the public—which may be an online-first/electronic preprint. Preprints prior to manuscript acceptance were not considered.
	Dates imputed	A Boolean variable coded as: FALSE if both the search date and publication date were specified precisely, or TRUE if either date was imputed. Imputation was based on the midpoint of the specified date range, for example if a search month is specified rather than a search date, then the 15th day of the month was the imputed search date.
Review parameters	Number of apps reviewed	Number of apps reported for analysis, after any screening or filtering for relevance.
	Apps downloaded	Ordinal variable coded as 'no apps downloaded' (eg, analysis was based on only app store metadata), 'some apps downloaded' (eg, a targeted or random sample), or 'all apps downloaded'.
	Apps recommended	A Boolean variable coded as: TRUE if individual apps were named and described in a manner which suggests or recommends their use, or FALSE otherwise.

was collected to investigate whether inclusion of such recommendations influenced publication speed. The data extraction schema was developed through an initial pilot phase in which n=60 studies were reviewed by three reviewers (KH, ML and JN) to identify relevant data items and confirm the feasibility of extraction.

Extraction was completed in a two-phase process. In the first phase, we attempted to automatically extract study metadata, app store search and publication history dates. We used a heuristic text matching strategy to locate and excerpt relevant text from study full text, published study metadata and citation database records. Any matched text was used to pre-populate a standardised data extraction form for subsequent review.

In the second phase, each study was reviewed manually to verify automatically identified data and populate coding items not suitable for automation. Only studies which reported at least one search date (ie, earliest search date) and a publication date were retained for analysis. Each study was reviewed by a pair of reviewers (from JH, KH, CL, ML, JN, KO, MT, DW, IW, QW), each working independently. Any discrepancies were resolved by a third reviewer (from KH, ML and JN) not involved in the original review. Because data extraction included items with non-categorical assignments, we assessed inter-rater agreement using raw agreement (the proportion of scored data items where each reviewer pair assigned the same value). Overall agreement was 0.79 (n=1273/1618 data items) versus 0.83 (n=538/646) for those items extracted in the pilot phase. Considering extraction of study publication dates alone, being the data items intended to inform the primary analysis of study, agreement was 0.84 (n=478/570) versus 0.90 (n=207/230) extracted in the pilot phase.

Data analysis

The primary reported outcome is TTP, calculated as the difference in days between the earliest search date and the date of publication. This window was justified on the basis that it reflects a conservative upper bound on the 'staleness' of information contained in any review at the earliest time it becomes accessible to a public audience. A secondary measure, the time to submission (TTS), was calculated as the difference between the earliest search date and the submission date. Descriptive statistics are reported for both TTP and TTS.

Data concerning review parameters (number of apps reviewed, whether apps were downloaded and whether apps were recommended) are presented descriptively. Due to non-normalcy, non-parametric tests were used to assess the relationship between these parameters and the primary outcome, TTP. The correlation between TTP and the number of apps reviewed was measured using Spearman's correlation coefficient (ρ). The impact of downloading some or all apps on TTP was assessed using a Kruskal-Wallis test, with follow-up tests to identify differences between specific groups. Whether studies which recommended specific apps were published more rapidly was assessed using a Wilcoxon rank sum test.

Two sensitivity analyses were specified a priori. We anticipated that studies would exist where date information was reported only partially, for example, reporting only the month and year in which app searches were performed. In these cases (n=93), we imputed the date as the 15th of the stated month or, if the authors reported a range of actual dates, selected a single date lying in the middle of this range. The first sensitivity analysis assessed the consequences of partial date reporting by comparing TTP for those studies where date imputation was and was not required.

A second sensitivity analysis aimed to explore the consequences of assuming that no important app changes occurred between the earliest and most recent app search/update dates. To do this, TTP was recalculated using the *last available* search date for each study (ie, the latest reported value of 'earliest search date', 'latest search date' and 'updated search date') and compared with the original TTP measure. For both sensitivity analyses, TTP values were compared using Wilcoxon rank sum tests.

A *post hoc* exploration of the relationship between the two effort-related review parameters (number of apps reviewed and whether apps were downloaded) was conducted. The number of apps reviewed was compared across the subgroups based on whether none, some, or all of the apps were downloaded in the review process and tested with a Kruskal-Wallis test. An additional metric, the review time per app, was defined as the TTP divided by the number of apps reviewed. This metric was again compared across downloaded subgroups and compared using a Kruskal-Wallis test.

Finally, a second *post hoc* analysis was undertaken to examine whether the TTP has changed over time. This may reflect, for example, that the methodology for app reviews has developed and normalised in recent years. To investigate this effect, a linear regression of the TTP against the earliest search date was performed.

All analyses were conducted using MATLAB V.8.6.

Patient and public involvement

As this was an analysis of previously published literature, patients or the public were not involved in the design, or conduct, or reporting, or dissemination plans of our research.

RESULTS

Search and selection

Searches of the published literature were performed on 30 April 2019 and yielded 439 study reports. After deduplication, screening and full-text review (summarised in figure 1), 136 reports were included in the final analysis (see online supplemental file 2).

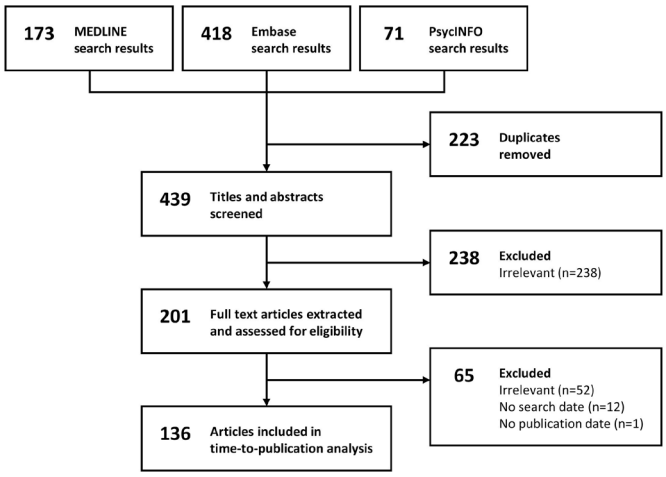

Figure 1 Study selection flowchart.

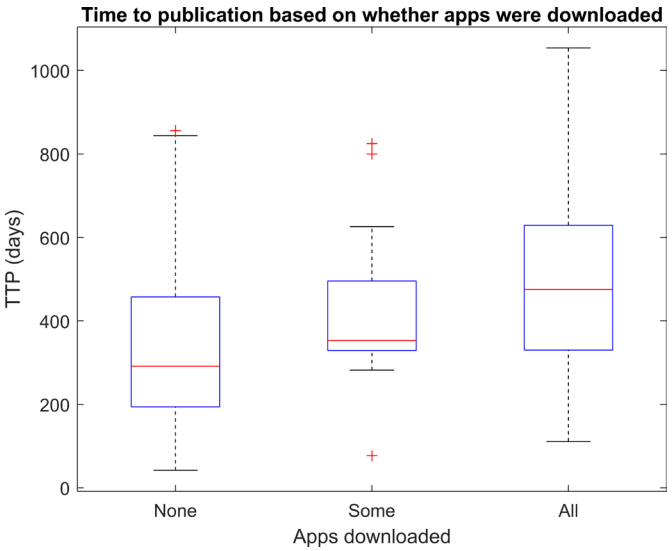

Figure 3 Distribution in the time to publication (TTP) based on whether apps were downloaded as part of the review process.

TTP and TTS

The median TTP was 431 days (approximately 14 months, range: 42–1054 days) from the earliest search date. A total of 100 papers reported a submission date, however in eight cases the submission date was prior to the search date. Logically the search date should precede the submission, therefore these eight papers were excluded from the analysis as the accuracy of the reported dates is uncertain. From the 92 remaining studies, the median TTS was 269 days (approximately 9 months, range: 5–874 days). The distribution of TTP and TTS is shown in figure 2.

Review parameters

The median number of apps reviewed in the 136 included studies was 52, although there was large variation between studies (range: 4–1806 apps). A near-zero correlation was found between the number of apps reviewed and TTP (ρ=−0.004, p=0.964).

Authors typically downloaded the apps for review, rather than relying on app store descriptions: 72.1% (n=98/136) papers indicated all apps were downloaded for review versus 17.6% (n=24/136) where no apps were downloaded. A targeted or random sample of apps was downloaded in 9.6% (n=13/136) of studies, and in one study it was not possible to determine whether or not apps were downloaded. There was a significant difference in TTP between the subgroups (p=0.010). Figure 3 shows the distribution of TTP for each subgroup. Follow-up tests identified a longer publication time when all apps compared with no apps were downloaded (median TTP: 476 days vs 292 days).

Specific apps were named and recommended for use in 15.4% (n=21/136) of the reviewed studies. Studies which included recommendations for specific named apps were published marginally more quickly than other studies, but this difference was not significant (median: 425 vs 440 days, p=0.998, see figure 4)

Sensitivity analyses

Two-thirds of the studies (92/136, 67.6%) did not specify an exact app store search date, and two (2/136, 1.5%) were not associated with a precise publication date. In combination, dates were imputed in 93 of the studies (68.4%). The difference in TTP was not significantly different between papers with precise or imputed dates (457 vs 430 days, respectively, p=0.648).

Approximately a quarter of the app reviews (37/136, 27.2%) reported a latest search date, and 5.1% (7/136) reported an updated search date. Using the latest of the three reported dates, the median TTP reduced to 387 days, which did not reach significance for difference from the primary outcome (p=0.063).

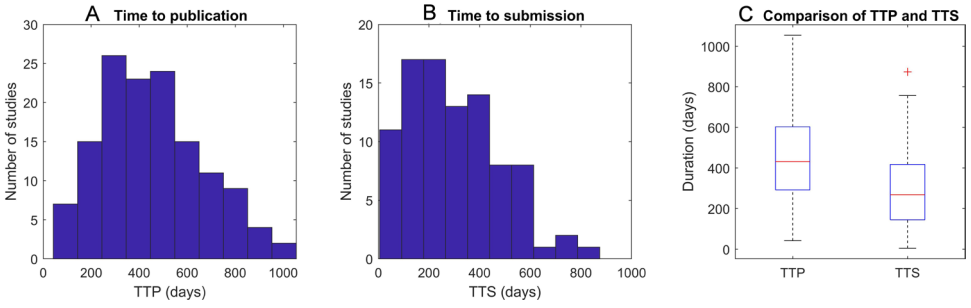

Figure 2 Frequency distributions for (A) the time to publication (TTP), and (B) time to submission (TTS), with (C) a side-by-side comparison.

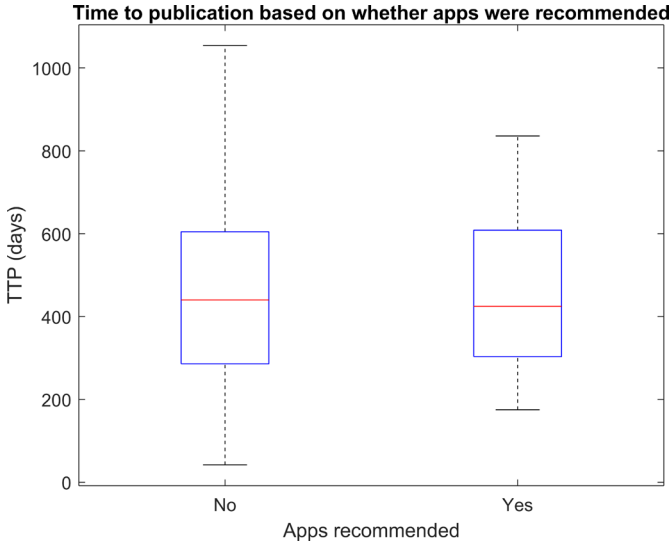

Figure 4 Distribution in the time to publication (TTP) based on whether specific apps were named and recommended.

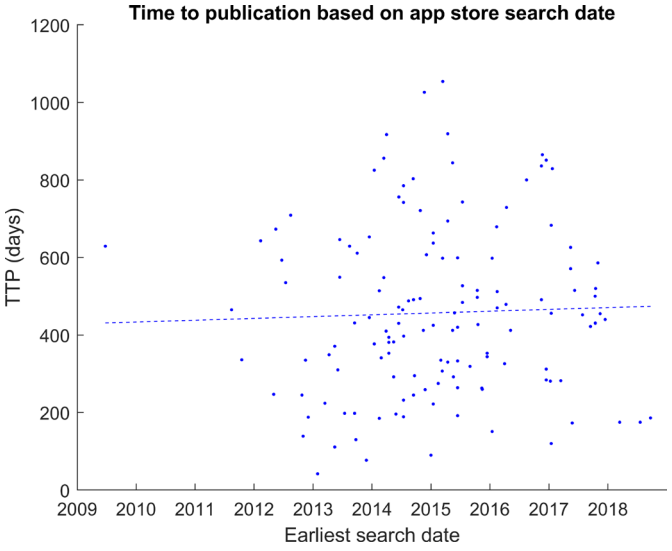

Figure 6 Variation in the time to publication (TTP) based on the app store search date.

Post hoc analyses

As there appeared to be a relationship between whether apps were downloaded and the TTP, but no such relationship for the number of apps reviewed in the studies, we conducted a *post hoc* analysis to examine whether there was a relationship between these two review features. Figure 5 shows the variation in the number of apps reviewed, based on whether apps were downloaded as part of the review process, and a significant difference was found (p<0.001). Follow-up tests identified significantly more apps were reviewed when no or some apps were downloaded, compared with all apps being downloaded (median number of apps: 130, 175 and 43, respectively).

When the TTP was normalised by the number of apps included in the review, the review time per app was 8.2 days. There were significant differences between studies based on whether apps were downloaded (p<0.001)—with those studies where all apps were downloaded taking significantly longer (median 12.1 days/app) compared with the no downloaded (2.9 days/app) and some downloaded (4.2 days/app) apps.

The second *post hoc* analysis examined the change in TTP over time, as shown in figure 6. The TTP has increased by 4.6 days each year, however this is not significant (p=0.69).

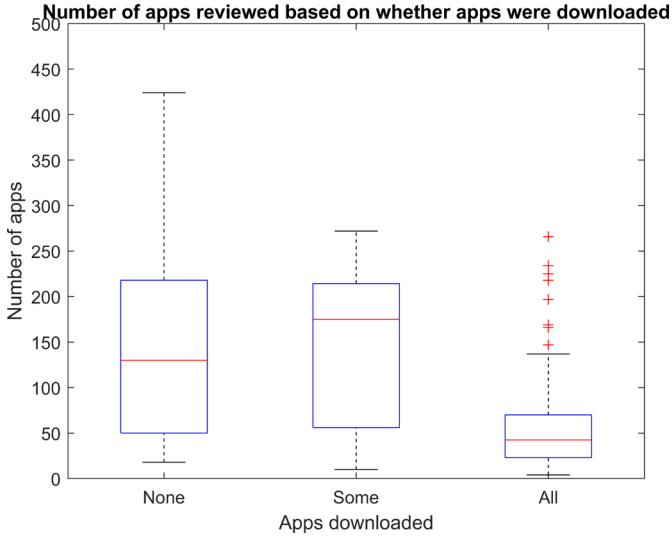

Figure 5 Variation in the number of apps reviewed, depending on whether the apps were downloaded as part of the review process. The y-axis has been truncated, and not all outlier values are shown.

DISCUSSION

This study aimed to quantify the extent to which data presented in cross-sectional app reviews are up-to-date by examining the delay between the selection of apps for review and the time of publication. By the time that most app reviews become available for use, a considerable period of time has elapsed (median: 431 days, 14.2 months). This measure still exceeds 1 year (387 days, 12.7 months) when a more lenient measure of the recency of findings at the time of publication is used. This delay is not wholly attributable to factors outside the control of researchers, such as the peer review process: when the estimated TTS was calculated to provide a crude metric of the time to conduct the study, excluding journal peer-review and editing processes, there was still substantial time between app search and manuscript submission (269 days, 8.8 months). The time taken to publish app reviews was influenced by the nature of the analysis, with findings indicating that reviews that downloaded apps for analysis took significantly longer to publish than those that did not. Surprisingly, the number of apps reviewed did not influence publication time, however a *post hoc* analysis indicated papers that downloaded apps reviewed significantly fewer apps than those that did not.

Given previous research indicating a high rate of turnover in the app marketplace, with 50% of mental health search results changing within approximately 3 months,[27]

the observed delay in publication raises questions about the validity of study findings at the time they become available to the research, clinical and broader community, particularly where such reviews focus on recommendations concerning specific products. Reviews may recommend the use of apps which are no longer supported by the developers, have been withdrawn from the app stores or, conversely, have been updated substantially since the review. Contrary to our hypothesis, we did not find that studies making specific recommendations had a shorter TTP than other reviews. Recommendations for specific apps in published reviews cannot, therefore, automatically be considered reliable. The delay in publication may also mean that more recent, potentially high-quality, apps are not made known to the research or clinical communities.

Our finding of substantial delays between initial assessment and publication provides a counterpoint to recent commentaries from academic clinicians discussing how to introduce apps into clinical practice that have emphasised a role for this kind of research to guide clinicians by identifying unsafe and poor quality apps[28] while simultaneously identifying deficiencies in alternatives to framework-based, structured app reviews such as certification programs[28] and user reviews.[29] Our data show that cross-sectional app reviews are also subject to important limitations concerning how up-to-date the information they contain is (and perhaps can be, given the academic publication process.) Because academic reviews are not designed to be continuously updated, healthcare professionals cannot assume that conclusions concerning the quality and safety of specific apps are still valid.

There are a number of potential strategies that could mitigate this issue. The first is for review authors to refresh their results prior to publication. Prepublication update is a standard practice in systematic reviews. For example, the Cochrane Collaboration will not publish reviews unless the most recent search date is less than 12 months (and ideally less than 6 months) old.[30] As part of efforts to improve the quality of cross-sectional app reviews,[31] editors and peer reviewers should consider at least asking for justification where there is a long period between search and submission. The practicality of this solution must nevertheless consider review-specific factors that may affect the TTP, the feasibility of update, and whether the review *intends* to guide clinical and public uses.

The second potential strategy is for authors to adjust the numbers of apps incorporated in their reviews. The relationship between downloading of apps and the number of apps reviewed may indicate that study authors hold some shared perceptions about what represents a 'publishable unit' of work. This may be achieved by either downloading a smaller number of apps, or by reviewing the app store descriptions for a larger number of apps. Both appear to result in publication in approximately 1 year (±3 months). While both have merit, it seems likely that studies that scrutinise app content directly are likely to yield richer insights than those relying on summary information presented in app stores for the purposes of marketing. However, with the longer TTP associated with downloading, it may be appropriate to focus on a smaller sample of the most popular, most used or top-ranked apps, which can be published more quickly. Authors should also consider how cross-platform apps should be handled. Apps which are available for both Android and iOS may share common features and functionality, however some aspects may be unique to one platform. There may therefore be a trade-off between comprehensively reviewing all versions and streamlining the review of a single version.

A third possible strategy is to remove app assessment and review from the academic sphere, to organisations whose resources are not subject to the constraints of the academic publication process and are, at least in principle, resourced to be able to respond to app dynamics such as update and withdrawal. Indeed, continuous app scanning and review approaches have now been adopted by a number of health organisations, including the NHS Apps Library[32] and the American Psychiatric Association.[33] However, despite the intention to provide continuous review, it is unclear how often reviews and recommendations that appear within these portals are actually updated. Further, even within app portals, the large number of available apps often necessitates that thoroughness be balanced with expediency,[34] potentially still limiting broad utility of such resources.

In parallel with the development of the academic literature regarding the quality of health apps, different jurisdictions have developed regulatory frameworks to govern the distribution of apps, particularly those which may be considered to be medical devices. While these frameworks differ across jurisdictions, harmonisation of quality criteria may help further refine and improve the wide range of quality assessment methodologies employed across the literature.[35]

Importantly, we do not suggest that our findings imply that cross-sectional assessments of health apps have no utility. App reviews may not be the optimal source of timely information about the function and quality of specific apps, but research-based methods are well suited to identifying and providing unbiased estimates of the nature and extent of thematic issues affecting specific populations of apps. Such insights can and have guided systemic responses to app quality problems. Research-led studies have arguably been important both in identifying *new* issues affecting apps, particularly where identification of issues involves complex exploratory and technical analysis,[36] and in devising systematic strategies for their identification, such as MARS. Unless specifically resourced to do so, it seems unlikely that continuous scanning programmes and app portals will be able to fulfil this discovery function.

Limitations

While this study examined app reviews across all health domains, to characterise the publication delay generally for the mobile health field of research, it did not examine

differences across specific subdomains. It is possible that different outcomes would be observed for different health conditions, for example due to resourcing availability/constraints across clinical domains, or due to differences in self-management approaches for different conditions. These differences were not considered in the current study due to substantial observed heterogeneity in the scope of reviews: some considered only technical domains of app quality (for example, data privacy), some considered broad categories (for example, mental health or physical health) and some considered specific conditions. Furthermore, the databases selected for the literature searches may not have provided complete coverage across all reviews focused on health and well-being domains (for example, those reported in allied health publications) or technical domains (for example, those focusing on data privacy and security). However, the databases selected are likely to capture the papers most likely to have an impact on clinical practice.

The TTS was calculated in addition to the primary outcome, TTP, to provide an approximate measure of the time to conduct the review process. However, this is only a crude estimate as it cannot account for manuscripts submitted to multiple journals prior to acceptance.

Some journals offer a fee-for-service option to expedite the peer review process, which would be expected to result in a quicker TTP. It is possible that authors using this facility may also conduct the reviews in a shorter period of time, resulting in quicker TTS. We observed no markers of whether articles had been expedited, so it was not possible to assess the impact of this publication model.

Finally, given the research question addressed by this study, we acknowledge the time taken to conduct and publish this review. However, the body of peer-reviewed, published academic literature is more stable and develops at a slower pace than the highly dynamic app stores. Furthermore, our findings show that the TTP has been stable for the past decade, therefore it is unlikely that the findings reported here have lost relevancy since the literature search was conducted.

CONCLUSIONS

The majority of health app reviews present data that are at least a year out-of-date at the time of publication. Given the high rate of observed turnover of health apps in public marketplaces, it may not be appropriate, therefore, for these reviews to be presented as a resource concerning specific products for commissioners, clinicians and the public. Authors of such reviews should, where possible, take steps to minimise the delay to publication, update their results prior to publication and consider whether making specific product recommendations is appropriate. App reviews may nevertheless fulfil important functions to identify novel and thematic issues and guide policy and systemic responses to health app quality and safety. App users should consider alternative sources of information about apps that are better calibrated to the dynamics of the app marketplace, such as continuous scanning services offered by dedicated health app portals.

Author affiliations
[1]Black Dog Institute, UNSW Sydney, Randwick, New South Wales, Australia
[2]Orygen, Parkville, Victoria, Australia
[3]Centre for Youth Mental Health, University of Melbourne, Parkville, Victoria, Australia
[4]St Vincent's Hospital, Sydney, New South Wales, Australia
[5]Faculty of Medicine, UNSW Sydney, Sydney, New South Wales, Australia
[6]Stanford University School of Humanities and Science, Stanford, California, USA
[7]Centre for Health Informatics, The University of Manchester, Manchester, UK
[8]School of Psychology, Western Sydney University, Sydney, New South Wales, Australia

Twitter Jennifer Nicholas @jenmnicholas1 and David Wong @drdavecwong

Contributors ML and KH conceived the study and are guarantors. JN and KH devised the search strategies and performed the searches. ML, JN and KH performed initial screening and ML collated the results. KH developed and executed the automatic data extraction process. KH prepared the manual data extraction tools, allocated reviewer tasks and collated results. ML, JN, JH, CL, KO, MT, DW, IW, QW and KH performed data extraction. ML performed principal data analysis. ML, JN and KH wrote the first draft of the manuscript. ML, JN, JH, CL, KO, MT, DW, IW, QW and KH provided comments and/or contributed revisions to the draft manuscript. ML, JN, JH, CL, KO, MT, DW, IW, QW and KH reviewed and approved the final manuscript version.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests KH, ML and JN are authors of studies that were included in this review. They declare no other competing financial or non-financial interests. All other authors declare no competing financial or non-financial interests.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

ORCID iDs
Mark Larsen http://orcid.org/0000-0002-0272-2053
David Wong http://orcid.org/0000-0001-8117-9193

REFERENCES

1 Kitsiou S, Paré G, Jaana M, *et al*. Effectiveness of mHealth interventions for patients with diabetes: an overview of systematic reviews. *PLoS One* 2017;12:e0173160.
2 Coorey GM, Neubeck L, Mulley J, *et al*. Effectiveness, acceptability and usefulness of mobile applications for cardiovascular disease self-management: systematic review with meta-synthesis of quantitative and qualitative data. *Eur J Prev Cardiol* 2018;25:505–21.
3 Whitehead L, Seaton P. The effectiveness of self-management mobile phone and tablet apps in long-term condition management: a systematic review. *J Med Internet Res* 2016;18:e97.

4 Han M, Lee E. Effectiveness of mobile health application use to improve health behavior changes: a systematic review of randomized controlled trials. *Healthc Inform Res* 2018;24:207–26.
5 Zhao J, Freeman B, Li M. Can mobile phone apps influence people's health behavior change? an evidence review. *J Med Internet Res* 2016;18:e287.
6 Byambasuren O, Beller E, Glasziou P. Current knowledge and adoption of mobile health apps among Australian general practitioners: survey study. *JMIR Mhealth Uhealth* 2019;7:e13199.
7 Boyle L, Grainger R, Hall RM, *et al*. Use of and beliefs about mobile phone apps for diabetes self-management: surveys of people in a hospital diabetes clinic and diabetes health professionals in New Zealand. *JMIR Mhealth Uhealth* 2017;5:e85.
8 Rubanovich CK, Mohr DC, Schueller SM. Health APP use among individuals with symptoms of depression and anxiety: a survey study with thematic coding. *JMIR Ment Health* 2017;4:e22.
9 Australian Government Department of Health and Ageing. *E-Mental health strategy for Australia*. Canberra, Australia: Commonwealth of Australia, 2012.
10 California Mental Health Services Authority. *'The technology suite' Driving access to behavioral health care thru innovation*, 2018.
11 NHS England. *The NHS long term plan*. London: HM Government, 2019.
12 Huckvale K, Adomaviciute S, Prieto JT, *et al*. Smartphone apps for calculating insulin dose: a systematic assessment. *BMC Med* 2015;13:106.
13 Huckvale K, Car M, Morrison C, *et al*. Apps for asthma self-management: a systematic assessment of content and tools. *BMC Med* 2012;10:144.
14 Larsen ME, Nicholas J, Christensen H. A systematic assessment of smartphone tools for suicide prevention. *PLoS One* 2016;11:e0152285.
15 Nicholas J, Larsen ME, Proudfoot J, *et al*. Mobile Apps for bipolar disorder: a systematic review of features and content quality. *J Med Internet Res* 2015;17:e198.
16 Donker T, Petrie K, Proudfoot J, *et al*. Smartphones for smarter delivery of mental health programs: a systematic review. *J Med Internet Res* 2013;15:e247.
17 Thornton L, Quinn C, Birrell L, *et al*. Free smoking cessation mobile apps available in Australia: a quality review and content analysis. *Aust N Z J Public Health* 2017;41:625–30.
18 Goodwin GM, Haddad PM, Ferrier IN, *et al*. Evidence-based guidelines for treating bipolar disorder: revised third edition recommendations from the British association for psychopharmacology. *J Psychopharmacol* 2016;30:495–553.
19 Wyatt JC, Thimbleby H, Rastall P, *et al*. What makes a good clinical APP? introducing the RCP health informatics unit checklist. *Clin Med* 2015;15:519–21.
20 Torous JB, Chan SR, Gipson SY-MT, *et al*. A hierarchical framework for evaluation and informed decision making regarding smartphone apps for clinical care. *Psychiatr Serv* 2018;69:498–500.
21 Stephens H, Uccellini M, McKay F. *Guidelines for creating healthy living apps*. Melbourne, Australia: Dialog Consultking, 2015.
22 Centers for Disease Control and Prevention. *Best practices user guide: health communications in tobacco prevention and control*. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 2018.
23 Rose KJ, Petrut C, L'Heveder R, *et al*. IDF Europe's position on mobile applications in diabetes. *Diabetes Res Clin Pract* 2019;149:39–46.
24 Ronchi E, Woskie LR, Milstein JA. *Mobile technology-based services for global health and wellness: opportunities and challenges*. Boston, USA: Organisation for Economic Cooperation and Development (OECD), 2017.
25 World Bank Group. *Empowering households and individuals to co-produce positive health outcomes for dignified, person-centered care amidst demographic change*. Washington, DC: World Bank, 2018.
26 Stoyanov SR, Hides L, Kavanagh DJ, *et al*. Mobile APP rating scale: a new tool for assessing the quality of health mobile apps. *JMIR Mhealth Uhealth* 2015;3:e27-e:e27.
27 Larsen ME, Nicholas J, Christensen H. Quantifying APP store dynamics: longitudinal tracking of mental health Apps. *JMIR Mhealth Uhealth* 2016;4:e96.
28 Gordon WJ, Landman A, Zhang H, *et al*. Beyond validation: getting health apps into clinical practice. *NPJ Digit Med* 2020;3:14.
29 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.
30 Tariq S, Akhtar N, Afzal H, *et al*. A novel co-training-based approach for the classification of mental illnesses using social media posts. *IEEE Access* 2019;7:166165–72.
31 BinDhim NF, Hawkey A, Trevena L. A systematic review of quality assessment methods for smartphone health apps. *Telemed J E Health* 2015;21:97–104.
32 NHS. NHS Apps library. Available: https://www.nhs.uk/apps-library/ [Accessed 25 Feb 2020].
33 American Psychiatric Association. App evaluation model. Available: https://www.psychiatry.org/psychiatrists/practice/mental-health-apps/app-evaluation-model [Accessed 25 Feb 2020].
34 Meek T. *Tom's digital disruptors: testing times for apps*. Digital Health, 2015.
35 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.
36 Huckvale K, Torous J, Larsen ME. Assessment of the data sharing and privacy practices of smartphone Apps for depression and smoking cessation. *JAMA Netw Open* 2019;2:e192542.