Longitudinal trends in the quality, effectiveness and attributes of highly rated smartphone health apps

Mara Mercurio,[1] Mark Larsen,[2] Hannah Wisniewski,[1] Philip Henson,[1] Sarah Lagan,[1] John Torous [1]

[1]Psychiatry, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA
[2]Black Dog Institute, Randwick, New South Wales, Australia

Correspondence to
Dr John Torous, Psychiatry, Beth Israel Deaconess Medical Center, Boston, MA 02115, USA; jtorous@bidmc.harvard.edu

Received 9 December 2019
Revised 17 March 2020
Accepted 24 March 2020
Published Online First
20 April 2020

To cite: Mercurio M, Larsen M, Wisniewski H, *et al*. *Evid Based Ment Health* 2020;**23**:107–111.

ABSTRACT

Background While there are numerous mental health apps on the market today, less is known about their safety and quality. This study aims to offer a longitudinal perspective on the nature of high visibility apps for common mental health and physical health conditions.

Methods In July 2019, we selected the 10 top search-returned apps in the Apple App Store and Android Google Play Store using six keyword terms: depression, anxiety, schizophrenia, addiction, high blood pressure and diabetes. Each app was downloaded by two authors and reviewed by a clinician, and the app was coded for features, functionality, claims, app store properties, and other properties.

Results Compared with 1 year prior, there were few statistically significant changes in app privacy policies, evidence and features. However, there was a high rate of turnover with only 34 (57%) of the apps from the Apple's App Store and 28 (47%) from the Google Play Store remaining in the 2019 top 10 search compared with the 2018 search.

Discussion Although there was a high turnover of top search-returned apps between 2018 and 2019, we found that there were few significant changes in features, privacy, medical claims and other properties. This suggests that, although the highly visible and available apps are changing, there were no significant improvements in app quality or safety.

BACKGROUND

The rapid growth in availability of mental health apps presents a plethora of easily accessible tools directly to patients and clinicians. In 2018, there were reported to be over 300 000 mobile health apps, at least 10 000 of which were related to mental health.[1] Despite research that has characterised most of these apps to be of questionable quality,[2] there is a lack of data on longitudinal trends and changes in the mental health app space because most studies to date have been cross-sectional. Previous research has quantified the rate of turnover of mental health apps on app stores,[3] but changes in the apps themselves have not been examined. In the past year, health apps have come under increased scrutiny and attention from scientists and society alike, but have individual apps been updated in response to both this scientific and public pressure? As the public becomes more concerned about app privacy, new research about mental health apps is published and a greater focus on human factors and usability emerges, we assess whether such efforts are reflected in current app offerings.

In 2018, our team examined 120 popular mental and physical health-related apps and coded each across 27 features, functions and attributes in order to assess overall quality.[4] Results indicated that, even with machine learning methods, there was no clear correlation between app features and app quality, apart from time since last update. Furthermore, findings that only 70% offered a privacy policy and that 29% raised red flags for safety suggest a cautious approach when evaluating apps. The 2018 examination highlighted the well-known challenges in identifying safe and effective apps. However, in the year since we evaluated those top apps, many have been updated. In using the same metrics and code book to evaluate these apps now 12 months later, we can also assess for changes in the marketplace and identify trends in app privacy, evidence and marketing that can inform education, clinical and advocacy efforts.

Since the 2018 coding of apps, awareness about the risks of health apps and the need for further evidence has grown. In November 2018, for example, the New York Times highlighted how easily apps can capture and market people's location via smartphone GPS.[5] Furthermore, the Federal Trade Commission held national hearings and issued the largest penalties for violations around digital privacy and,[6] in April 2019, WHO released guidelines for using digital tools like apps in patient care.[7] A mental health advocacy group highlighted broad concerns on their online blog in October 2018, stating 'who owns the data collected…', 'who has access to the data…' and 'how does the tech programme actually work'.[8] While it is unreasonable to expect the entire health app landscape to dramatically improve in 1 year, as there is a lag between intent to change and actual change, the often-touted advantage of digital health tools is their ability to readily adapt and evolve to meet the needs of patients.

In this review, we aim to expand our team's 2018 review[4] and explore what features, protections, evidence, and markers of quality are present in top apps for depression, addiction, anxiety disorder, schizophrenia, hypertension, and diabetes. We hypothesise that in 2019 there will be improvements reflected in more apps having privacy policies and supporting evidence along with fewer being flagged as concerning. As with last year, we hypothesise that there will not be a simple relationship between these app features and attributes and overall app quality.

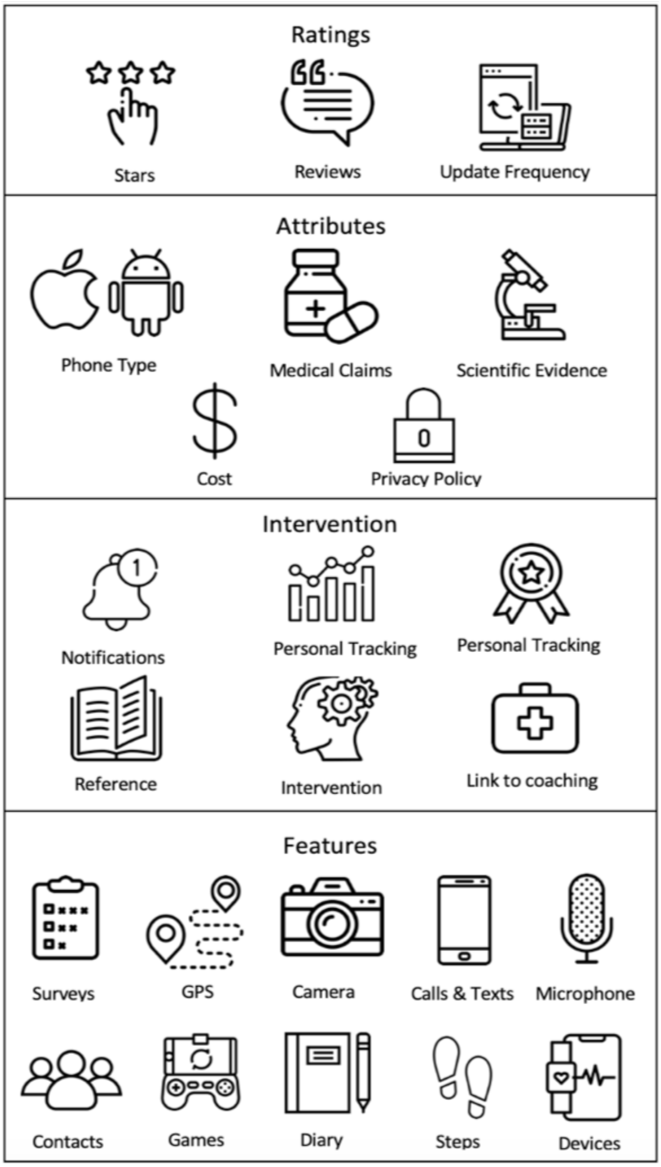

Figure 1 Diagrammatic representation of coded app features and attributes.

METHODS

We selected the top 10 search-returned apps on the US Apple iTunes App store and US Android Google Play store on 6 July 2019 for the keywords categorised across six disease states: depression, schizophrenia, anxiety, hypertension, diabetes and addiction (120 apps in total). Each app was downloaded and used by two independent coders and reviewed by Dr Torous, a board-certified psychiatrist and clinical informaticist. An evaluation of each app's features and relevant app data was entered into a code book based on the 2018 study (see online appendix I).[4] Disagreements were discussed in person until consensus was reached. Apps that were duplicated on the Apple App store and Google Play store were downloaded on both platforms and reviewed separately. As shown in figure 1, we assessed information including data on the app attributes (eg, privacy policy present, availability to delete data and any purchases), functionality for gathering data (eg, diaries, surveys, health information), returning information and engaging the user (eg, notifications, badges and social networking), measures of the apps' popularity (eg, stars and number of ratings on the app store), medical claims and scientific backing. Given the subjective nature of quality assessment in apps,[9] we assessed the presence rather than the quality of features. For example, we assessed whether a privacy policy or intervention was present but did not evaluate its quality. We assessed medical claims based on app stores' descriptions, and actual evidence for medical claims based on links to direct clinical evidence or scientific backing related to the app based on information provided and investigation by our team, which included PubMed and Google Scholar searches.

To offer an overall assessment of the app, we applied the same 3-point scale as the 2018 study as follows: 0 represented 'serious concerns regarding safety', 1 represented 'likely acceptable app' and 2 represented 'a potentially more useful app'.[4] Acknowledging this scale is itself subjective given that the utility of any app depends on the patient at hand, clinical needs and treatment goals, our analysis focused on apps that were rated with 'serious concerns regarding safety' because the clear safety concerns are far less subjective than other ratings. For example, an app that provided incorrect medical information would be scored as a 0 (a safety concern) in our ratings. Assessments of apps and scores from the 2018 study were then compared with those from the current 2019 study using t-tests.

The methods of assessing the relationship between app features and reviewer quality flags were similar to those used in the 2018 paper.[4] Specifically, we used variable selection using the Lasso method to obtain models that relied on fewer app metrics. In applying regression to all metrics within each disease state, we applied a penalty using the number of ratings as weights with a ceiling of 1000 for apps with >1000 ratings. Tuning parameters were chosen by fivefold cross-validation and we repeated the process 100 times to account for our relatively small sample size.

RESULTS

We coded a total of 120 apps, with 20 for each condition (10 iOS and 10 Android). On both the Apple App Store and the Google Play store, three apps appeared in both the depression and anxiety searches (on Apple App Store: Moodpath: Depression & Anxiety, AntiStress Anxiety Relief Game and Pacifica for Stress & Anxiety; on the Google Play store: Moodpath: Depression & Anxiety, Youper—Emotional Health and Wysa: stress, depression & anxiety therapy chatbot). Compared with 1 year prior, the top 10 apps across each of the 6 conditions were largely different. Only 34 (57%) of the apps from the Apple App Store and 28 (47%) of the apps from the Google Play Store in our 2018 search were still in the top 10 search-returned apps in the 2019 search.

Compared with the apps identified in 2018, more apps made medical claims in every disease state except addiction, which decreased from one to none. However, the absolute number of apps now making medical claims (50%) was not statistically significantly higher than the 2018 claims data (30.8%). Likewise, the number of apps offering privacy policies changed across all conditions: in 2018, 70% of the apps coded contained privacy policies compared with 87.5% in 2019, a change that was not statistically significant. While we did not evaluate the privacy policies themselves, we did assess for ability to delete data, which actually decreased for depression, anxiety and diabetes apps. Other results are shown in table 1.

We found overall few changes in app features for information/data collection and interventions provided in 2019 compared with 2018. Pop-up messages offering information or returning summarised/analysed data (such as average steps taken per day) remain the most common intervention mode. Looking at

Table 1 App attributes for 2019 (2018).

	Anxiety n=20 2019 (2018)	Schizophrenia n=20 2019 (2018)	Depression n=20 2019 (2018)	Diabetes n=20 2019 (2018)	Addiction n=20 2019 (2018)	Hypertension n=20 2019 (2018)	Average n=120 2019 (2018)
User star ratings	4.63 (4.29)*	2.83 (3.56)	4.46 (4.41)	4.48 (4.13)	4.70 (4.22)	3.92 (3.49)	4.17 (4.02)
Presence of a privacy policy	95% (85%)	70% (50%)	100% (85%)	100% (85%)	90% (70%)	70% (45%)	87.5% (70%)†
Ability to delete data	70% (70%)	20% (20%)	55% (70%)	45% (60%)	35% (45%)	20% (25%)	40.8% (48.3%)*
Costs associated with the app	95% (70%)	25% (45%)	25% (45%)	70% (55%)	70% (80%)	50% (60%)	55.8% (59.2%)
Days since last update	20 (55)	514 (392)	155 (138)	86 (35)	173 (157)	321 (652)*	211.5 (238.2)
Medical claims by app	60% (15%)†	40% (30%)	70% (45%)	60% (45%)	0% (5%)	70% (45%)	50% (30.8%)*
Specific evidence to support medical claims	25% (5%)	10% (10%)	10% (0%)	20% (5%)	0% (0%)	0% (0%)	10.8% (3.3%)

*p<0.05

†p<0.01

in, data that is often collect by or into the app; out, outputs and results shared outward by the app.

disease-specific apps, there were few significant changes in the 12-month period covered by this study. More apps included privacy policies, and fewer apps offered the ability to delete data, both changes which were insignificant. Apps associated with schizophrenia had the least number of features, the lowest star ratings in the stores and the highest number of days since last update.

Using Lasso regression, we replicated the prior study's finding that apps which had not been updated in over 180 days were associated with our rating for serious concerns regarding safety (two-sided t-test, $p<0.01$). Results did not change when we weighted apps to account for number of reviews. The schizophrenia apps which had not been updated for a mean of 514 days had a mean flag value of 0.55 flag, and the hypertension apps that were not updated for a mean of 321 days a mean flag value 0.75. We did not find any other clear association between individual app metrics recorded and quality, a finding also in line with the results from the 2018 study.

DISCUSSION

Our review found a high degree of turnover for top search-returned apps across diabetes, hypertension, depression, anxiety, addiction and schizophrenia but overall little evidence for change in their privacy, safety, features and functions. The rate of stability of apps from the top 10 (57% and 47% of iOS and Android apps, respectively, remaining after 12 months) appears higher than previously reported (95.8% and 82.4% of the top 25 iOS and Android apps, respectively, remaining after 9 months in 2015).[3] This may indicate that turnover is more frequent than previously reported among the top search-returned apps, or that the app store marketplace is now even more dynamic and volatile than it was 4 years ago.

While public debate on digital privacy and new research on evidence-based interventions evolves, it appears that those apps most accessible to consumers are not evolving as quickly. Our finding that there was not a clear association between any measure except for time since last update >180 days with app quality suggests that there is no simple formula to assess the clinical safety and potential of these digital tools.

There were few improvements in the 2019 app space compared with our 2018 results. Compared with 2018, more apps in 2019 did offer privacy policies, although we did not evaluate the content of these privacy policies and the absolute changes were small and not statistically significant. The fact that any of the top search-returned health apps still fail to include a privacy policy is a cause for concern. Second, there were no significant changes in the ability to delete data, which is one of the more objective and easily assessed metrics related to digital privacy. Third, the number of apps claiming scientific evidence did not significantly change apart from anxiety apps, the percentage of which claiming scientific evidence increased from 15% to 60%. However, no significant changes in actual evidence supporting these apps in any disease state are noteworthy. This finding is in line with recent research suggesting that, while over 50% of apps may make medical claims, <2% of such claims were validated with a clinical study.[10] Lastly, despite the potential of digital phenotyping and apps using sensors to help understand context and provide environmentally triggered responses, surveys remain the most frequent data input mode across all app categories. Overall, the lack of significant changes in these top search-returned apps between 2018 and 2019 reflects room for further improvement and an important area for focus in 2020.

Our results also highlight the divide between the potential of apps and their current offerings available to the public. As shown in table 2, there were few significant changes between 2018 and 2019 in the types of data these top apps collected or the means they used to return data or offer interventions to users. The majority of top apps still capture data via surveys or diaries and return that data via popup messages in a summarised format or with a disease-related fact. This model of use fails to take advantage of novel means to capture clinical state like digital phenotyping or smartphone features,[11] which would enable greater understanding of the context and environments surrounding the person to deliver individualised care. For some diseases like schizophrenia, top apps continue to offer mainly reference information, and much of that is out of date and of concerning quality.

Apps for physical health conditions (diabetes and hypertension) showed little observable difference in quality measures from mental health apps. Overall, more apps for diabetes and hypertension included medical claims than in mental health apps (65% vs 42.5%); however, this difference is largely due to a lack of medical claims in the addiction apps. Physical health apps also more commonly included step counters as well as integrations with other health data and smart devices. Given the importance of physical activity for both cardiovascular and mental health, this could reflect an opportunity for improved tracking of physical activity in the context of mental health apps.

Just as in the 2018 study, we did not identify a strong association between coded app attributes and our quality flag. Acknowledging that our quality flag metric is itself subjective, we did observe decreases in anxiety and addiction apps that were

Table 2 App attributes/features from 2019 (2018).

	Anxiety n=20 2019 (2018)	Schizophrenia n=20 2019 (2018)	Depression n=20 2019 (2018)	Diabetes n=20 2019 (2018)	Addiction n=20 2019 (2018)	Hypertension n=20 2019 (2018)	Average n=120 2019 (2018)
Surveys (in)	70% (60%)	35% (35%)	75% (70%)	65% (40%)	45% (30%)	40% (40%)	55% (45.8%)
GPS (in)	20% (30%)	5% (20%)	40% (30%)	35% (70%)*	15% (30%)	0% (15%)	19.2% (32.5%)
Call/Text logs (in)	5% (10%)	0% (0%)	0% (10%)	0% (10%)	0% (15%)	0% (0%)	0.8% (7.5)%*
Camera (in)	20% (25%)	15% (0%)	20% (15%)	40% (45%)	20% (10%)	5% (0%)	20% (15.8%)
Microphone (in)	15% (25%)	0% (5%)	15% (20%)	5% (5%)	5% (5%)	0% (5%)	6.7% (10.8%)*
Device integration (eg, smartwatch) (in)	10% (30%)	15% (10%)	5% (25%)	55% (55%)	0% (0%)	15% (25%)	16.7% (24.2%)
Diary (in)	55% (40%)	15% (20%)	55% (50%)	55% (25%)	40% (30%)	20% (25%)	40% (31.7%)
Contact list (in)	15% (25%)	15% (5%)	5% (25%)	15% (45%)*	5% (15%)	5% (5%)	10% (20%)
Steps/Other Apple HealthKit or Google Fit Data (in)	20% (25%)	10% (10%)	15% (25%)	65% (60%)	0% (0%)	55% (35%)	27.5% (25.8%)
Games (in)	15% (10%)	10% (10%)	5% (10%)	0% (0%)	20% (15%)	5% (10%)	9.2% (9.2%)
Pop-up messages (out)	85% (75%)	25% (20%)	85% (80%)	70% (60%)	75% (70%)	30% (30%)	61.7% (55.8%)*
Reference information (out)	70% (65%)	85% (90%)	55% (80%)	70% (70%)	30% (70%)*	30% (70%)*	56.7% (74.2%)
Social network connections (out)	20% (30%)	20% (20%)	10% (25%)	30% (25%)	40% (40%)	0% (0%)	20% (23.3%)
Analysing data to return insights (out)	80% (80%)	15% (20%)	80% (80%)	65% (75%)	75% (75%)	55% (50%)	61.7% (63.3%)
Linking to formal care or coaching	35% (30%)	5% (20%)	30% (45%)	5% (30%)	30% (25%)	0% (5%)	17.5% (25.8%)
In app rewards or badges	30% (40%)	0% (5%)	10% (20%)	20% (0%)*	55% (70%)	0% (0%)	19.2% (22.5%)
In app interventions (eg, CBT)	80% (65%)	5% (10%)	60% (55%)	30% (60%)	25% (30%)	5% (5%)	34.2% (37.5%)
Mean flag rating	1.05 (1.2)*	0.55 (0.3)*	0.95 (0.9)	1.4 (1.4)	0.75 (0.9)*	0.75 (0.55)	0.91 (0.88)

CBT, cognitive behavioural therapy.

significant (table 2) and an increase in the mean flag rating of schizophrenia apps between 2018 and 2019. In attempting to build models to predict app quality from features and attributes in tables 1 and 2, we only sought to predict low-quality flag metrics (<1), as it is easier to determine what is a poor-quality app than a high-quality app given the numerous use cases and personal preferences related to app engagement. Our result that apps not updated in over 180 days are likely of poor quality is helpful in screening out concerning apps, but does not offer a simple formula or feature set with which to identify high-quality apps. These results call into question app rating and curation efforts which attempt to rank apps in part based on various app features and attributes—which may not be able to keep current with frequent app turnover,[12] or account for the myriad of ways people use apps. Prior research on existing app evaluation systems has also questioned the validity of metrics used to calculate recommendations scores.[13] Without well-validated metrics to guide app evaluation, we suggest that a more holistic and informed approach to picking the right app for the right patient may make sense today.

Our results are well aligned with other recent studies examining mobile app features. A recent study evaluating data security and privacy policies of mobile apps for depression found that most policies lacked information about the ability to edit and delete personal information[14] and although there was an absence of information in privacy policies, other studies similarly showed there was an increase in the number of privacy policies for the apps evaluated a year later.[15] Another recent study determined that 64% of mental health apps evaluated made claims of effectiveness,[16] which corresponds to the growing number of apps that made medical claims in our study.

LIMITATIONS

Like all studies, there are limitations which must be considered. First, we coded for the presence of app features and attributes, but not for the quality of their implementation. While this helps avoid bias and identifies apps with clear concerns, such as not having a privacy policy, it does not help answer the equally important question of what that privacy policy actually offers. Second, we selected only the top 10 apps from each app store and thus our results may not generalise to other apps outside the top 10. Still, given that consumers are most likely to see these highly visible apps in search results, we believe our results remain broadly relevant. Third, without a gold standard for app quality, our finding that apps not updated in over 180 days are associated with more concerning apps is a useful finding, but its contrary does not mean an app updated more frequently will be of higher quality. Fourth, we do not know how the app store ranks apps and why they feature some highly in search results and others not. Thus, our search from the year prior may not be identical to this year's. Also, we only looked at free apps, although in the future will consider paid versions as well. Although free apps are the majority, they are only a portion of the apps available and may offer less efficacy than purchased apps—although evidence for this is lacking. Lastly, since the ease of use varies between people, we attempted to measure features of apps that were most invariant and more objective by using a set of metrics that could be generalisable. However, even these metrics will vary between actual app users based on their background and clinical needs.

CONCLUSION

The digital health app space and scope continues to rapidly evolve,[17–19] with many new apps appearing as others disappear. As a group, however, the quality, features and attributes of the top apps we examined does not appear to be changing as quickly. Ensuring that the current international efforts around digital health privacy and recent research findings are rapidly disseminated into available apps represents a challenge that the digital health field must now embrace if it is to fulfil its potential of offering safe and effective tools.

Twitter John Torous @JohnTorousMD

Contributors All authors contributed equally.

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 None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available. The goal of this project is not to call out any single app but rather note larger trends. It is possible to recreate these data for current apps with materials offered in the paper.

ORCID iD

John Torous http://orcid.org/0000-0002-5362-7937

REFERENCES

1 Torous J, Roberts LW. Needed innovation in digital health and smartphone applications for mental health: transparency and trust. *JAMA Psychiatry* 2017;74:437–8.

2 Firth J, Torous J, Nicholas J, *et al*. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. *World Psychiatry* 2017;16:287–98.

3 Larsen ME, Nicholas J, Christensen H. Quantifying APP store dynamics: longitudinal tracking of mental health Apps. *JMIR Mhealth Uhealth* 2016;4:e96.

4 Wisniewski H, Liu G, Henson P, *et al*. Understanding the quality, effectiveness and attributes of top-rated smartphone health apps. *Evid Based Ment Health* 2019;22:4–9.

5 Valentino-DeVries J, Singer N, Keller MH, *et al*. Your apps know where you were last night, and they're not keeping it secret, 2018. Available: https://www.nytimes.com/interactive/2018/12/10/business/location-data-privacy-apps.html?smid=pl-share [Accessed Mar 2020].

6 FTC. FTC imposes $5 billion penalty and sweeping new privacy restrictions on facebook, 2019. Available: https://www.ftc.gov/news-events/press-releases/2019/07/ftc-imposes-5-billion-penalty-sweeping-new-privacy-restrictions

7 WHO. *Who guideline: recommendations on digital interventions for health system strengthening*. Geneva: World Health Organization, 2019. https://apps.who.int/iris/bitstream/handle/10665/311941/9789241550505-eng.pdf?ua=1

8 Digital phenotyping in the public mental health system: what clients and advocates need to know, 2018. Available: https://272d6681-17ea42d09bbcbc096b89055a.filesusr.com/ugd/c82a51_7c79bfa991224f0c8250541395a24ab6.pdf

9 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.

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 Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. *NPJ Digit Med* 2019;2:88.

12 Carlo AD, Hosseini Ghomi R, Renn BN, *et al*. By the numbers: ratings and utilization of behavioral health mobile applications. *NPJ Digit Med* 2019;2:54.

13 Powell AC, Torous J, Chan S, *et al*. Interrater reliability of mHealth APP rating measures: analysis of top depression and smoking cessation Apps. *JMIR Mhealth Uhealth* 2016;4:e15.

14 Parker L, Halter V, Karliychuk T, *et al*. How private is your mental health APP data? an empirical study of mental health APP privacy policies and practices. *Int J Law Psychiatry* 2019;64:198–204.

15 O'Loughlin K, Neary M, Adkins EC, *et al*. Reviewing the data security and privacy policies of mobile apps for depression. *Internet Interv* 2019;15:110–5.

16 Buechi R, Faes L, Bachmann LM, *et al*. Evidence assessing the diagnostic performance of medical smartphone apps: a systematic review and exploratory meta-analysis. *BMJ Open* 2017;7:e018280.

17 Faurholt-Jepsen M, Frost M, Christensen EM, *et al*. Validity and characteristics of patient-evaluated adherence to medication via smartphones in patients with bipolar disorder: exploratory reanalyses on pooled data from the MONARCA I and II trials. *Evid Based Ment Health* 2020;23:2–7.

18 Punukollu M, Marques M. Use of mobile apps and technologies in child and adolescent mental health: a systematic review. *Evid Based Ment Health* 2019;22:161–6.

19 Naslund JA, Aschbrenner KA. Digital technology for health promotion: opportunities to address excess mortality in persons living with severe mental disorders. *Evid Based Ment Health* 2019;22:17–22.