Funding for this report has been generously provided by:

The RBC Foundation

Homewood Research Institute (HRI) is a registered Canadian Charity dedicated to transforming mental health and addiction treatment through research. We partner with leading scientists, universities, patients and clinicians to improve care, services and outcomes. HRI's charitable registration is # 86307 3334 RR0001.

For more information about this report, please contact:

Dr. Roy Cameron
Executive Director
Homewood Research Institute
150 Delhi Street, Riverslea Building
Guelph, ON N1E 6K9
admin@homewoodresearch.org

Suggested Citation: Quintana, Y & Torous, J. Youth Mental Health Apps in the Digital Age: A Scoping Review of Trends and Evaluations. Homewood Research Institute Report, Guelph, Ontario, Canada, November 8, 2019.

Note: The opinions and recommendations included in this report are those of the authors, and do not necessarily reflect the views of the funder.

This report can be accessed free of charge at Homewoodresearch.org

Please Consider Making A Donation.
With your support, HRI can continue to conduct research and evaluation in pursuit of our vision of a world where... No life is held back or cut short by mental illness and addiction
Preamble

Homewood Research Institute (HRI) is a nationally registered Canadian Charity dedicated to transforming mental health and addiction treatment through applied research, evaluation and innovation. We work with Canadian and international scientists, and collaborate with people who receive and provide treatment, to improve services and outcomes.

HRI seeks to address issues that can have a large impact across Canada and beyond. Use of technology to deliver treatment services is enormously important to provide broad access to timely, affordable support to those who need it, regardless of where they live.

HRI’s goal is to help create an environment where trustworthy digital technologies can be used within formal service provision systems. This report represents our first step down that path. It provides a general overview of what’s out there. It reveals the need for more rigorous evaluation of digital therapy tools and transparency in reporting evaluation results.


We are grateful to the RBC Foundation for its foresight in making digital technologies a focus of their initiative in promoting youth mental health and for the generous support for this project.

HRI is eager to partner with like-minded organizations which place value on creating an environment where high quality technology tools, backed by strong science, can be employed with confidence within service delivery systems. By aligning our efforts we will accelerate progress and enhance our collective impact in improving outcomes at individual and population levels.

Please get in touch if you see opportunities for collaboration.

Roy Cameron, PhD, FCAHS
Executive Director
Homewood Research Institute
Executive Summary

Statistics show a rapidly growing need for mental health services in Canada and worldwide. Mental health service providers have a very limited supply of professionals. Well evaluated technology can help to reach more patients, in coordination with health providers, but the technology must be rigorously evaluated. At present many apps are being marketed, but their quality is generally unknown.

In this context, there is an urgent need to take stock of what is known about existing apps. If there are well evaluated apps, this would set the stage for scaling up those apps supported by credible evidence.

A recent partnership between Homewood Research Institute (HRI) and The RBC Foundation will help to bring clarity to app users – particularly to youth who rely heavily on apps and smartphone technology.

The following report, the environmental scan, completes phase 1 of the project and informs about existing significant apps in the market place designed to address youth mental health issues. It enables us to determine what exists and what may need to be developed as we move forward.

The report indicates that at this stage, although there is some evidence related to the value of some apps, evaluations have been limited and incomplete. There is an urgent need for systematic, thorough, objective evaluation of apps to inform decisions about which apps are safe and effective.

To undertake such evaluation, there is a need to develop and use a framework that lays out how to evaluate apps. This would set the stage for sharing best practices and scaling apps that are safe and effective.

This work is broadly applicable across populations (although youth are the immediate priority), technologies (although apps are the initial focus) and countries (although Canada is the primary concern at present). HRI recruited Dr. Quintana to lead this project, given his international perspective, to ensure the framework was as well-informed and broadly applicable as possible.
1.0 Introduction

Youth mental health and mental illness is a growing global concern. The number of children and adolescents with mental illness is increasing worldwide. There are significant challenges in access to mental healthcare services due to a limited number of mental health care providers. Access issues related to clinician availability, healthcare insurance coverage, and emotional barriers in seeking help due to real and perceived stigma are just some of the hurdles faced by youth seeking care. Adapting mental health therapies to individuals is also challenging, and the evidence on treatment effectiveness and personalization remains an active area of research. One solution to these above mentioned challenges is digital health tools like mental health apps, which are easily accessible, low cost, and can support evidence-based care. Given the high prevalence of smartphone ownership among youth, mental health apps hold unique potential in this population. However, these apps are not a panacea, as careful evaluation and selection are critical to steering youth to safe and effective apps.

A growing number of evaluation frameworks have been proposed to help guide the selection of apps and offer youth and clinicians help in navigating the digital mental health apps space. However, as this report will demonstrate, most of these frameworks have not been scientifically developed or validated. These frameworks may provide misleading information on the effectiveness of apps and may lead to misuse, misdiagnosis, wasted time, and even harm. The dangers of poorly formulated evaluation efforts are highlighted in a mental health app recommendation system deployed in Australia that made some youths using recommended apps feel more depressed [1]. In this review, we describe current trends in youth mental health, trends in mobile apps for mental health; discuss the strengths and limitations of available frameworks to evaluate mental health apps; and recommend future directions for more integrated approaches for design and evaluation of mental apps for youth.
2.0 Youth Mental Health Trends

The need for new solutions for youth mental health is clear [2-13]. Worldwide 10-20% of children and adolescents experience mental disorders [2]. An estimated 49.5% of adolescents in the United States have a lifetime prevalence of any mental disorder [3]. Of adolescents with any mental disorder, an estimated 22.2% had a severe impairment. DSM-IV based criteria were used to determine impairment level. [3] The data from Canada are similar [4], as an estimated 1.2 million children and youth are affected by mental illness there. By age 25, approximately 20% will have developed a mental illness. In Canada, an estimated 75% of children with mental health disorders do not access specialized treatment services [5], with average wait times of 6-12 months [9,10]. Equally concerning, in cases of severe illness, the risk of self-harm and suicide increases [11] in adolescents and young adults and is the second leading cause of death in the 15-24 age group in Canada [12]. Among underserved groups, the heightened risk is even more apparent; in some First Nations communities, the suicide rate among youth under the age of 15 is almost 50 times greater than non-indigenous. [13] Technologies like apps offer the highest potential for a solution to reach the tremendous unmet need in a scalable and affordable manner.
3.0 Mental Health Apps for Youth

There are a growing number of mobile apps for mental health for youth [14-21]. However, the majority of these apps have not been rigorously assessed or validated. The evidence for mental health apps, including those for youth, remains nascent. A review in 2017 by Lui et al. concluded that the literature remains preliminary, and there is insufficient evidence to determine which mobile apps are empirically supported [22]. A paper in 2018 by Wang et al. concluded that clinically validated evidence for most apps is unclear [23]. A review in 2019 by Larsen et al. of 73 mental health apps in the Google Play and Apple iTunes, found only one of these cited in published literature. A minority of apps (14%) described design or development involving people with lived experience, and none referenced certification or accreditation processes [24].

This lack of evidence places the burden of risk on the patient and clinician selecting the app – and thus, a need for evaluation tools. In 2012 The World Health Organization developed a framework for mental health services, but it does not include a method to evaluate the mobile apps [25]. Simple methods to evaluate apps have proved challenging. A paper in 2015 by Chan and colleagues proposed a direction for more formal evaluations and suggested that leadership was needed to develop a framework for describing apps, and guidelines are needed for both patients and mental health providers [26]. A review in 2015 by Shen et al. of apps for depression found that there is a lack of reporting of organizational affiliation and content source, which brings the credibility of the content into question [27]. A study in 2016 by Singh et al. demonstrated that the number of stars an app receives on the commercial marketplaces does not correlate with its clinical utility or validity [28]. Even certification from trusted health bodies is not a guarantee of quality. A study in 2015 by Huckvale et al. highlighted that 66% of the health apps certified as clinically safe and trustworthy by the UK NHS apps Library were sending identifying information over the internet without encryption and without disclosure that the app will do so [29].

Solutions include both improving the overall quality of the field and addressing more immediate needs to offer patients and clinicians guidance. A paper in 2018 from the developers of Health on the Net Foundation, one of the first organizations to propose a formal review of medical websites, called for greater transparency in the evaluation of health apps [30]. Many have also responded with frameworks to evaluate apps. Yet a 2019 review by Moshi et al. of 45 evaluation frameworks for mobile health application concluded that none of the frameworks could be used unaltered for health technology assessments, and only two of them evaluated the grade of evidence of the app [31]. A paper in 2019 by Henson et al. proposed a framework for categories for evaluation [32] that does not score apps but provides a useful approach to review apps systematically.
4.0 Evaluation of Mental Health Apps Youth

While there are thousands of apps, most evaluations focus on usage by early adopters, and few have rigorous scientific evaluations with large sample sizes. Below we review the evidence for some highly downloaded apps to highlight trends of what youth today are likely exposed to and using. Understanding what these apps offer, as well as what they lack, is essential for realizing the needs and challenges that successful evaluation frameworks must meet.

4.1 Popular Mental Health Apps

The Headspace App illustrates challenges related to assessing the efficacy of smartphone apps for mental health. The Headspace website makes numerous claims around efficacy, including that ten days of use will reduce stress by 14%, and three weeks will reduce aggression 57% [33]. The challenges in interpreting any efficacy claims for mental health apps were outlined in a 2017 meta-analysis of randomized controlled trials of depression apps by Torous and Firth [34]. The paper noted that the overall effect size of mental health apps appears substantial at $g=0.56$ (in range with antidepressant medications) but reduces to the more modest $g=.22$ when compared to an active control condition. That is to say that the same mental health app studied in comparison to nothing (no control) versus walking (active control) will report a different degree of impact. The studies of the Headspace app suggest that this app may not be as effective as touted. For example, one study by Noone et al. used a digital placebo version Headspace and found no benefit of using the app to improve focus in young people [35]. The digital placebo version of the app looked the same and offered narration in the same voice – but in this case, the placebo version offered non-therapeutic content. Still, students in the study could not tell the difference and found the placebo version as helpful as the actual. Another study looked at how Headspace could help people relax and reduce stress, but found no benefit when compared to playing the game Tetris [36]. The results are likely not unique to Headspace; in Fall 2019, a United Kingdom (UK) based study reported that for some people merely listening to woodland sounds might be more effective in reducing stress compared to listening to a guided meditation (which is what Headspace offers) [37]. While apps like Headspace can be effective for some young people and help them, these results force the question of how effective such apps are and the importance of asking “compared to what.”

However, even results from randomized controlled trials are not a guarantee of effectiveness of digital mental health tools like apps. A 2019 study by Huberty et al. of the mindfulness app Calm announced in the title of the paper that the app reduces stress in students [38], but closer inspection raises several concerns. In this case, the control group was instructed not to partake in any mindfulness activities; this group was compared to the group who received mindfulness for stress reduction via the
Calm app. With such an unbalanced control group, the results of the study only support that mindfulness can reduce stress compared to no mindfulness – which is a well-known fact. What is more interesting about this study is that only 56% of participating students used the app as directed, indicating that 44% were non-adherent despite volunteering to partake in the study. This lack of proper or prescribed use of the app suggests yet another challenge around mental health apps that will be discussed in other sections of this report.

Another paper on popular mental health apps also sheds light on assessing effectiveness. The classical Bradford Hill criteria for inferring causality highlight the importance of the dose-response relationship. An effective app would be expected to have more effect if used more. For example, we would expect students in the Calm study who used the app more should have a better response in stress reduction – although results were not presented by engagement, so it is not possible to answer that question. Arean et al. [39] have also studied the same app in other conditions, including depression. In one study of depression, the app was found to reduce depressive symptoms in users, which is encouraging. However, results presented in the appendix stratify decreases in depressive symptoms by engagement with the app and note that those who never engaged with the app had the same improvements as those who regularly did. This lack of a dose-effect suggests that either an unconventional mechanism of action is responsible for change or that aspects of the digital placebo effect – a term defined by Torous and Firth as response related to expectations about app use -- may be driving study results. The importance of evaluating effectiveness only after controlling for the digital placebo effect is clear.

While each of the apps mentioned above – Headspace and Calm– are popular and impressive tools that do help people feel better, it remains challenging to understand their efficacy and the best way to evaluate them. Unfortunately, most studies of smartphone apps do not present data on engagement in a consistent manner, as reported in a study by Ng. et al. [40], making it difficult to either compare studies or understand trends in engagement and adherence across these apps. Standardized reporting, as will be a recommendation in our final report, offers a simple yet effective mechanism to improve the quality and understand the effectiveness of digital mental health apps on the market today and in the future.

4.2 Studies with Larger Sample Sizes

Studies with larger sample sizes also offer a valuable perspective on mental health apps for youth and suggest how evaluation frameworks must approach apps backed by more clinical data. One investigation by Kauer et al. [41] studied 163 youth (mean age 20) over 2 to 4 weeks using MobileType (an app tracking mental health relevant thoughts and behaviors) or using a control app that tracks irrelevant behaviors not
related to mental health. Participants reviewed information gathered by the MobileType app with their general practitioner and were given guides for managing mental health with outcomes measured on the depression anxiety stress scale. The study reported that those monitoring their symptoms with the MobileType app ended the study with lower scores on the depression and anxiety stress scale, which is interesting - but the true impact of the paper lies with the reported mechanism of action. Study results suggest that self-monitoring via the app increased emotional self-awareness, which in turn decreased depression. Thus, this study suggests that apps offering ecological momentary assessment alone may also be offering therapeutics effects - suggesting 1) that even simple apps have the potential to be effective and 2) that this effect must be taken into consideration when evaluating the effectiveness of apps that do not even offer explicit interventions.

Another study by Proudfoot et al. [42] utilized the depression and anxiety stress scale to evaluate the MyCompass app – which enabled self-monitoring of problematic moods, thoughts, and behaviors, and receiving feedback advice and mental health management tips by text messages (n=126) vs. attention-matched (n=195) and waitlist control (n=198) over seven weeks. Roughly 20% of the sample were students. With a total sample size of 519, this study represents one of the largest smartphone based digital health studies to date. While the MyCompass app group did have statistically significant improvements in both depression and anxiety scores at the end of the active period of the study (7 weeks), during the 12-week follow-up period, these results became non-significant compared to the attentional control group. These results underscore that while it is possible to assess the immediate impact of app use, less is known about longer-term impact and sustainability of improvements gained over time. The majority of app studies today do not offer any data on long term outcomes, meaning that little is known about the actual efficacy of these digital tools.

Another large study sought to understand if apps targeting certain behaviors based on behavior change theories led to changes in those targets. Building off Kauer et al. [41] above, which investigated emotional self-awareness (ESA) as a mechanism of action in apps for youth, Bakker et al. [43] studied three popular mood apps MoodPrism, MoodMission, MoodKit. Each of these apps is available on commercial app stores and popular with many current users. The study investigated three common therapeutic elements across these apps, which are reviewed here:

1. Emotional self-awareness = ESA, an individual’s ability to comprehend their own emotions, leading to positive mental health outcomes. The focus of Kauer et al. [41] above.
2. Coping self-efficacy = CSE, a measure of an individual's confidence in their coping ability

3. Mental Health Literacy (Psychoeducation), an individual's understanding of mental disorders, which can lead to mental health improvements via recognition, prevention, or management of dysfunction

In this 30 day study, 78 were assigned to a waitlist and 78 to MoodPrism, which offered ESA + PE, 78 to MoodMission, which offered CSE + PE, and 78 to MoodKit which offered CSE + ESA + PE. While all groups, even the control, improved in depression and anxiety outcomes after 30 days, CSE and not ESA or PE was reported to drive improvements. While the absence of ESA effects in MoodKit and MoodMission groups may be due to participants not utilizing the mood-tracking features within each app, the difference from Kauer et al. [41] discussed above which found ESA to be the putative mechanism is notable. Having two of the largest app studies present contradicting results of the mechanism of action of mental health apps suggests the need for further underlying research.

From these three large studies, it is possible to better appreciate the challenges in assessing the efficacy of mental health apps as well as evaluating them. Kauer et al. [41] and Bakker et al. [43] report on important mechanisms of action but ultimately highlight that the actual mechanisms of action remain unknown with contradictory results. Assessing efficacy thus becomes more challenging when we do not yet know exactly how to capture the fidelity of the intervention or the process underlying it. A paper by Proudfoot et al. [42] highlights the importance of assessing not just immediate results on mental health outcomes related to app use, but also longer-term follow up to understand the durability or decay of improvements. Results of that study suggest efficacy must be measured at multiple time points and that results may not be enduring. While it is possible to find smaller pilot studies that offer many other interesting interpretations of the mechanism of action and duration of action, the three larger sample size studies presented here warrant attention and suggest caution when formulating app evaluation frameworks.

These studies also highlight another challenge of app evaluation – the lack of longer-term follow-up data. A review of the most extensive studies in terms of the number of patients enrolled showed that most studies do not have interventions beyond four weeks and the follow up period is rarely more than eight weeks after the intervention (Table 1). This is not a long enough period to see if behavioral changes are sustained. Most large studies have been done in adults. Few studies have more than 50 patients. The apps also do not report on the cognitive or behavioral models used in the design of the app. Thus it is not clear if the app is using validated cognitive models or evidence-based therapeutic protocols.
<table>
<thead>
<tr>
<th>App(s) [Reference]</th>
<th>Year</th>
<th>Condition</th>
<th>Country</th>
<th>Target</th>
<th>Size of Control Group</th>
<th>Size of Intervention Group</th>
<th>Outcome Measures</th>
<th>Study Period</th>
<th>Follow-up Period</th>
<th>Content Validation</th>
<th>Funder Disclosed</th>
<th>Cognitive Behavior Model Described</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project EVO [39]</td>
<td>2016</td>
<td>Depression</td>
<td>USA</td>
<td>Adults with mild-moderate depression</td>
<td>206</td>
<td>211 / 209</td>
<td>PHQ-9</td>
<td>12 weeks</td>
<td>None</td>
<td>PHQ-9</td>
<td>NIMH</td>
<td>None</td>
</tr>
<tr>
<td>MyCompass [42]</td>
<td>2013</td>
<td>Depression</td>
<td>Australia</td>
<td>Adults with mild-moderate depression</td>
<td>198</td>
<td>126 / 195</td>
<td>DASS</td>
<td>7 weeks</td>
<td>12 weeks</td>
<td>DASS</td>
<td>Australian Health and Ageing</td>
<td>None</td>
</tr>
<tr>
<td>MoodHacker [48]</td>
<td>2016</td>
<td>Depression</td>
<td>USA</td>
<td>Adults with mild-moderate depression</td>
<td>150</td>
<td>150</td>
<td>PHQ-9</td>
<td>6 weeks</td>
<td>10 weeks</td>
<td>PHQ-9</td>
<td>NIH / NIMH</td>
<td>None</td>
</tr>
<tr>
<td>CBM Active [49]</td>
<td>2014</td>
<td>Social Anxiety</td>
<td>USA</td>
<td>Adults</td>
<td>141</td>
<td>158</td>
<td>DASS</td>
<td>4 weeks</td>
<td>8 weeks</td>
<td>DASS</td>
<td>NR</td>
<td>None</td>
</tr>
<tr>
<td>SuperBetter [50]</td>
<td>2015</td>
<td>Depression</td>
<td>USA</td>
<td>Adults with significant depression</td>
<td>93</td>
<td>93 / 97</td>
<td>CES-D</td>
<td>4 weeks</td>
<td>6 weeks</td>
<td>CES-D</td>
<td>NR</td>
<td>None</td>
</tr>
<tr>
<td>SuperBetter Plus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>None</td>
</tr>
<tr>
<td>Mobiletype [51]</td>
<td>2013</td>
<td>Australia</td>
<td>Primary Care patients aged 14-24</td>
<td></td>
<td>46</td>
<td>68</td>
<td>DASS ESA</td>
<td>2-4 weeks</td>
<td>6 weeks</td>
<td>DASS ESA</td>
<td>Telstra Foundation</td>
<td>None</td>
</tr>
<tr>
<td>The Toolbox [52]</td>
<td>2017</td>
<td>Depression</td>
<td>Australia</td>
<td>Young adults in Australia</td>
<td>195</td>
<td>192</td>
<td>Well-being</td>
<td>up to 6 months</td>
<td>None</td>
<td>Mental Health Continuum Short Form (MHC-SF)</td>
<td>Young and Well Cooperative Research Centre</td>
<td>None</td>
</tr>
</tbody>
</table>

Table 1 - Evaluations of Mental Health apps with larger sample sizes
5.0 Current Evaluation Frameworks

There are a growing number of groups proposing systems for evaluations of apps. However, most of these groups do not identify the method for their evaluations. Most of these app review sites do not report who is doing the review, what data were used to evaluate the app, how the data was collected, and if the data are representative of the target audience. Further, most of those reviews are outdated by the time they are published. These problems are discussed in a 2019 review paper [44] which noted that the average time to review apps was between 109 and 714 days. This is especially concerning since many apps are often updated every 30-90 days.

<table>
<thead>
<tr>
<th></th>
<th>Mindtools.io</th>
<th>Psyberguide</th>
<th>ORCHA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraction (%) of top 25 apps reviewed</td>
<td>14/25 (56)</td>
<td>19/25 (76)</td>
<td>22/25 (88)</td>
</tr>
<tr>
<td>Website total global visits over 3-month period(a)</td>
<td>22,780</td>
<td>68,041</td>
<td>40,527</td>
</tr>
<tr>
<td>Average age of review (days)(b)</td>
<td>Mean 714</td>
<td>598</td>
<td>109</td>
</tr>
<tr>
<td></td>
<td>Median 776</td>
<td>424</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>Standard deviation 178</td>
<td>305</td>
<td>170</td>
</tr>
</tbody>
</table>

Figure 1 –Review times for mobile apps. Source: NPJ Digit Med. 2019 Jun 17;2:54. [44]

In another study done in Canada [42], a group of 25 participants reviewed mobile apps. This group included app developers and users, health care providers, mental health advocates, people with lived experience of a mental health problem or mental illness, policymakers, and researchers. In response to the contextual questions posed in Round 1, almost half of the participants (12/25, 48%) rated the overall quality of the health apps as poor or fair, whereas 24% (6/25) gave a good or excellent rating. The remainder of the participants (7/25, 28%) indicated that they were uncertain of the overall quality of the health apps. This shows that unless there are clear guidelines and criteria for evaluation, ratings will reflect a wide range of contradictory opinions.

The Mental Health Commission of Canada developed another evaluation toolkit for mobile mental health apps [46]. However, this toolkit does not provide a clear way to evaluate the effectiveness of apps. Most criteria are not specific enough to allow reliable evaluation (Figure 2). The report notes that “few quality control mechanisms exist to ensure e-mental health tools are user-friendly, accurate, or efficacious.” Below is a summary of the evaluation criteria. Note that usability is described as an app being “user-friendly and engaging enough to make people want to keep using it.” However, a gambling app can be made to be friendly, engaging, and want to keep people using it, but it would only reinforce addictive behavior.
### Evaluative Criteria

<table>
<thead>
<tr>
<th>Effectiveness</th>
<th>Clinical Claims</th>
<th>Usability</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is the app’s intended purpose? Can it actually do what it says it will? Is there proof?</td>
<td>If the app makes certain clinical claims (e.g., reducing stress or anxiety), does it give proof of its efficacy?</td>
<td>Is the app user-friendly and engaging enough to make people want to keep using it?</td>
</tr>
<tr>
<td>User desirability</td>
<td>Security and privacy</td>
<td></td>
</tr>
<tr>
<td>Will the people the app is designed for actually want (or be able) to use it?</td>
<td>Does the app clearly state how it will collect, store, use and protect personal health information? Is this information easy to find or hidden deep within the app? Does the app meet all applicable federal and provincial/territorial legislative standards and requirements regarding personal health information?</td>
<td></td>
</tr>
</tbody>
</table>

### Informative Criteria

<table>
<thead>
<tr>
<th>Functionality</th>
<th>Interoperability</th>
<th>Supported platforms</th>
</tr>
</thead>
<tbody>
<tr>
<td>What functions does the app offer (e.g., journaling, mood tracking, guided exercises)?</td>
<td>Does the app use open standards allowing it to exchange data with other health apps or tools (if applicable)?</td>
<td>Is the app exclusive to one platform, which may create accessibility barriers? Or is it available to many users across Android, iOS and other devices?</td>
</tr>
<tr>
<td>Target users</td>
<td>Price</td>
<td>Transparency</td>
</tr>
<tr>
<td>Who is the intended audience for the app? Is it clear who should or should not be using it?</td>
<td>Is the app upfront about its cost or are there hidden/extra fees? Will the price create accessibility barriers for the intended users?</td>
<td>Does the app clearly state the individuals or organizations involved in its development? Does it clearly state who provided the funding for its development?</td>
</tr>
</tbody>
</table>

### Inclusion

Were the target users involved in the development and testing of the app to ensure it responds to their needs and expectations? How diverse was the user input? Were people from a variety of populations with unique mental health challenges involved (e.g., immigrant, refugee, ethnocultural and racialized communities, First Nations, Inuit, Métis, LGBTQ2+, people who are homeless, seniors, youth)?

---

**Figure 2 - Canadian Mental Health Commission Evaluation Criteria for Mobile Apps**

Source: CMHC. Available at URL: https://www.mentalhealthcommission.ca/English/media/4137

The Anxiety and Depression Association of America has webpages with app reviews [47]. Their criteria have several categories but not a clear definition of how to evaluate the app in those categories (Figure 3). For example, for “effectiveness” the website does not specify who is qualified to be a rater and how the rating scale differs from 1 to 5. The Effectiveness category says that it covers Education, Self-Monitoring, and Treatment, but does not describe how to do those evaluations.
An initiative of the American Psychiatric Association, led by John Torous, MD, from our team, has created a framework with more specific category definitions, based on a position paper published in 2019 [32]. This approach has a hierarchical rating system and rubric to make APA members aware of very important information that should be considered when picking an app that is not the same as the information used to judge a medication or therapy. The foundation of the evaluation model rests in the maxim ‘do no harm’ as well as a risk-benefit analysis. The APA does not explicitly rate apps but provides a useful way to review apps. The four areas comprising the model (beyond gathering basic background information) are Safety/Privacy, Evidence (i.e., effectiveness), Ease of Use, and Interoperability. This a better approach, but further ongoing work is needed to develop the scientific method for evaluating each criterion, including therapeutic effectiveness, that can yield reliable results.
In summary, most apps are not evaluated for therapeutic intent, and most reviews are not based on scientific evaluation for clinical effectiveness. Private vendor evaluations do not reveal what data were used for evaluation. Behavioral outcomes need to be independently analyzed with a large enough number of participants and an appropriate follow-up evaluation period.
6.0 Towards a Scientific Evaluation Framework

Given the lack of a detailed scientific methodology for evaluating mobile mental health apps, the group is currently developing a more detailed, rigorous framework that can be used consistently and reliably. Below are the major categories of this framework.

- Effectiveness
- Design Model with Youth Input
- Usability
- Privacy
- Security
- Implementation
- Cost
- Disclosures of Source of Funding, Conflicts of Interest
- Potential Harms of the mobile app

For the effectiveness category, below are key components for investigation and evaluation.

- Does the app have clear defined effectiveness metrics?
- Does the app have therapeutic intent?
- Has it been tested with enough patients to measure clinically significant outcomes?
- Were the evaluation methods appropriate?
- Were there validated instruments and surveys used to collect the data?
- Was the follow-up period observation sufficient to observe a change?
- Were the results published in a peer-reviewed journal with a reputable editorial board?

For the cognitive model category, below are key aspects,

- Did app disclose the cognitive models used to promote behavior change?
- Are these cognitive models previously known and validated?
- Are these models appropriate for this particular patient population and condition?
- Is there evidence the models were actually implemented in the app, not just mentioned?

For usability, we will identify scientifically proven methods to measure usability with the intended audience.

- Does the app use known scientific methods to evaluate usability?
- Were there an appropriate number of test subjects with the intended condition that were used in the evaluation?
- Was observational usability data collected and analyzed?
- Were the results published in a reputable peer-reviewed journal?

For privacy, there needs to be clarity and verification of the privacy features.

- Are users able to understand the privacy policy?
- Do users have the right to prevent further dissemination of their data beyond their phone?
• Can users request their data not be used by third parties?
• Can users request their data be permanently deleted?
• Is there independent monitoring that the company/app is following for its privacy practices?
7.0 Conclusions

This review shows there is a clear and growing need to address mental health and mental illness. Current services are not sufficient to meet current or future demand due to limitations of the number of available qualified staff and the geographical distribution of service providers. Technologies, like apps, are thus of great interest as they may be able to increase access to care given their high rates of prevalence among youth.

However, while there are a large number of mental health apps, few have scientific evaluations or strong evidence to support their claims. The scientific evaluations that exist are limited in value due to small numbers of patients, short intervention periods, and limited follow up periods needed to assess sustained changes. Thus patients and clinicians seeking apps are presented with a glut of options and yet little objective data.

The evaluation frameworks that exist to guide patients and clinicians lack detail and methodologies needed to provide reliable results. Most of the frameworks have not been scientifically tested. With the need clearly defined, we will now develop a more rigorous evaluation framework and apply it to the largest available studies that have their data publicly accessible in peer-reviewed publications or verified data sets.

Once this framework is completed, we propose that a full systematic review be done of all studies with 50 or more patients to create the most comprehensive review of the published literature. This will provide the best available evidence on the effectiveness of apps for youth mental health and their current limitations. This review could be updated annually to provide continuing unbiased guidance to healthcare providers and funders on the state-of-the-art mental health apps.

We also propose a co-design project with youth, holding meetings involving a broad range of healthcare providers and our core team to design an app that takes into account the lessons learned. An implementation plan will be developed to pilot the app and evaluate it using our evaluation framework. The intent is to design the mental health app so it can be scaled provincially or nationally with tight integration with providers. Using a rigorous methodology will ensure the app is active, scalable, and can serve as a reference model for the nation and the rest of the world.
Acknowledgements

Yuri Quintana and John Torous prepared the final report, with contributions from the following people in alphabetical order:

- Alaa Ali E. Abd-Alrazaq, PhD, Postdoc, Health Informatics, Hamad Bin Khalifa University, Qatar
- Nashva Mohammed Ali, Senior Researcher, Hamad Bin Khalifa University, Qatar
- Darren Fahy, Research Coordinator, Division of Clinical Informatics, Beth Israel Deaconess Medical Center
- Mowafa Househ, PhD, Associate Professor, Hamad Bin Khalifa University, Qatar
- Alejandro R. Jadad, MD, Professor, Dalla Lana School of Public Health, University of Toronto, Canada
- Yuri Quintana, PhD, Chief, Division of Clinical Informatics, Beth Israel Deaconess Medical Center, Assistant Professor of Medicine, Harvard Medical School, Harvard Medical School, Boston, MA, USA.
- Charles Safran, MD, Professor of Medicine, Harvard Medical School, Boston, MA, USA.
- John Torous, MD, Chief, Division of Digital Psychiatry, Instructor in Psychiatry, Harvard Medical School, Harvard Medical School, Boston, MA, USA.

During the development of the project we had consultations with the following people:

- Elizabeth Borycki, PhD, RN, Professor Health Information Science, Director - Social Dimensions of Health Program, University of Victoria
- Roy Cameron, PhD, Executive Director, Homewood Research Institute, Guelph, Canada
- Cyndy Forsyth, Chief Development Officer, Homewood Research Institute, Guelph, Canada
- Joanna Henderson, PhD, CPsych, Director, Margaret and Wallace McCain Centre for Child, Youth & Family Mental Health; and Senior Scientist in the Child, Youth and Emerging Adult Program, Centre for Addiction and Mental Health (CAMH), Toronto, Ontario, Canada
- Andre Kushniruk, PhD, Professor and Director of the School of Health Information Science
- Vimla L. Patel, PhD, Senior Research Scientist and Director, Center for Cognitive Studies in Medicine and Public Health, The New York Academy of Medicine, New York, NY, Adjunct Professor Department of Biomedical Informatics, Columbia University, New York, NY
- Brian Rush, PhD - Senior Scientist, Homewood Research Institute, Scientist Emeritus, Institute for Mental Health Policy Research, Centre for Addiction and Mental Health (CAMH), Toronto, Ontario, Canada
- Stanley Kutcher, MD, a Canadian psychiatrist, formerly faculty member at Dalhousie University, appointed to the Senate of Canada on December 12, 2018.
- David Wiljer PhD, Executive Director, Education Technology Innovation at University Health Network, Toronto, Associate Professor, Psychotherapies, Humanities and Educational Scholarship, University of Toronto

Please note that the analyses and conclusions in the present document do not necessarily reflect those of the individuals or organizations mentioned above.
References


199-210. DOI: 10.1037/pro0000122 URL:


