A Hierarchical Framework for Evaluation and Informed Decision Making Regarding Smartphone Apps for Clinical Care

John Blake Torous, M.D., Steven Richard Chan, M.D., M.B.A., Shih Yee-Marie Tan Gipson, M.D., Jung Won Kim, M.D., Thuc-Quyen Nguyen, M.D., John Luo, M.D., Philip Wang, M.D.

With thousands of smartphone apps targeting mental health, it is difficult to ignore the rapidly expanding use of apps in the treatment of psychiatric disorders. Patients with psychiatric conditions are interested in mental health apps and have begun to use them. That does not mean that clinicians must support, endorse, or even adopt the use of apps, but they should be prepared to answer patients' questions about apps and facilitate shared decision making around app use. This column describes an evaluation framework designed by the American Psychiatric Association to guide informed decision making around the use of smartphone apps in clinical care.

Psychiatric Services 2018; 69:498–500; doi: 10.1176/appi.ps.201700423

There are at least 10,000 smartphone apps targeting mental health (1), and many patients are exploring them (2). The literature on health app ratings offers tools to help clinicians and patients pick apps (3), but so far none of these tools provides a reliable method of evaluating an app's safety and usefulness. Although simple metrics —for example, a five-star rating of a health app by a user—may appear to be a useful metric of quality, a study of 137 patient-facing apps found that star-based ratings had low correlation with the apps' clinical utility or usability (4). Clinician ratings of individual features of mental health apps also suffer from low interrater reliability, as demonstrated in a study using existing app rating metrics to evaluate popular smoking cessation and depression apps (5).

The inherently dynamic nature of apps adds to the challenge of developing reliable metrics of app quality. A study tracking the longitudinal availability of mental health apps reported that they have a half-life: after a certain amount of time, an app may no longer be available for public use (6). App creators have liberty to update apps as much or as little as they would like—some creators frequently update apps, whereas others completely abandon support and development of an app.

A further challenge in using app ratings is their use of absolute scores versus relative scores. Just as there is no single 'A+' rated therapy or medication treatment plan that is "100% effective" for all patients, apps vary in effectiveness depending on the individual user. Apps are tools that must be selected on the basis of individual needs, abilities, preferences, and many other personal patient factors.

In this column, we describe the rationale, internal testing, and release of the American Psychiatric Association's (APA) smartphone app evaluation framework. The APA framework serves as a tool to guide informed decision making and evaluation of apps, and, like any rubric, it must be reapplied for each unique patient, unique clinical context, and unique version of the app.

A Framework for Evaluation

The APA app evaluation framework offers clinicians and patients an adaptable scaffold for informed decision making. In significant ways, the framework approach adopted by the APA is unique compared with prior efforts. Instead of directly rating or scoring a particular app, the framework utilizes a simple four-stage hierarchical process, asking users to consider safety and privacy first, followed by evidence and benefit, engagement, and, finally, interoperability (Figure 1). [A color schematic of the APA app evaluation framework is available as an online supplement to this column.]

When the framework is used as intended, evaluation begins with safety and privacy, progressing to the next stage only if the particular app in question satisfies the present clinical needs surrounding that stage. For example, if an app does not satisfy the present clinical needs around privacy and safety, evaluation should stop there. The APA app evaluation framework does not offer specific criteria to judge whether an app satisfies each stage of the hierarchy. Instead, it offers a series of questions that are intended to guide a unique and personalized determination of the appropriateness of an app for each patient. Users can choose how to weigh or consider each stage, given that certain stages, such as data sharing, may not matter if the app is purely informational. The framework is an evolving tool that will be updated to reflect new knowledge about apps. The latest version is freely available through the APA Web site (https://psychiatry.org/psychiatrists/practice/mental-health-apps/app-evaluation-model).

In order to understand the rationale for the selection and ordering of the hierarchical stages, it is useful to briefly explore the current state of mental health apps. At first glance, it may be difficult to imagine how apps cause harm; yet there is a growing literature on potential dangers of app use. Because many apps exist outside the scope of federal privacy laws (for example, HIPAA), given that they are marketed directly to consumers, apps can be used to collect the personal mental health data of app users, and these data can often be sold, traded, marketed, and indefinitely stored by app companies. Evidence also suggests that the majority of health apps currently lack even basic privacy policies, meaning that simply checking for the existence of a privacy policy will help identify many questionable apps (7). Beyond privacy concerns, apps have been known to offer dangerous and harmful advice (8).

When evaluating efficacy, it is important to realize that although many apps appear useful, the actual evidence for clinical efficacy is nascent. This does not mean apps cannot be helpful, but it highlights the importance of considering whether the current evidence for the app in question is sufficient or relevant for a particular patient. Together, stages 1 and 2 of the framework (risk/privacy and safety, and evidence and benefit) constitute basic medical decision making centered on nonmaleficence.

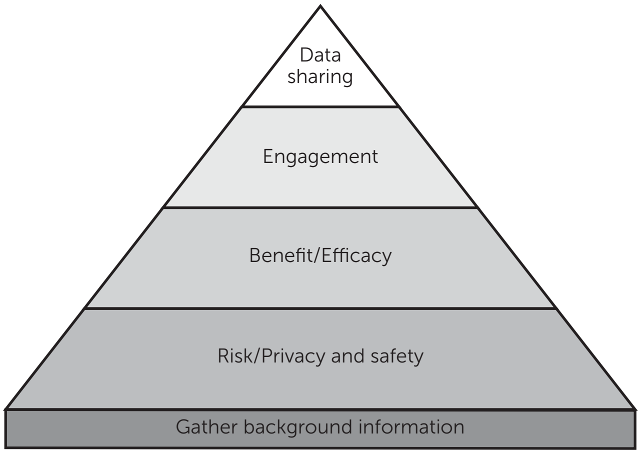
The engagement stage represents the growing awareness that many patients do not stick with apps or may find them difficult to use (9). This likely reflects the lack of patient involvement in the development of mental health apps.

Data sharing, the final stage, reflects the need to ensure that app data are available to the treatment team. Poor interoperability can fragment care by limiting appropriate data sharing and access to information that is necessary to guide care and make treatment decisions.

Preliminary Internal Testing

Like any framework and tool for informed decision making, the APA app evaluation framework will evolve based on user feedback and evaluation. In order to gain an early understanding of the reliability of the framework and generate foundational data, we conducted internal interrater reliability testing with five psychiatrists (JBT, SRC, SYG, JWK, and TN). Each psychiatrist was presented with three mood tracking apps (MoodTrack, MoodTools, and T2 Mood Tracker) that closely duplicate apps used in a recent study of app usability among patients with depression (9). The psychiatrists were asked to rate the app for use in two clinical situations, using the app evaluation framework to rate the app at all four stages.

FIGURE 1. Steps in the smartphone app evaluation framework[a]



[a] A color schematic of the framework is available as an online supplement. The online version lists specific questions that should be asked at each step of the framework. The questions are also available on the Web site of the American Psychiatric Association (https://psychiatry.org/psychiatrists/practice/mental-health-apps/app-evaluation-model).

The first clinical case involved a patient “who is tech savvy, in his twenties, suffering from moderate depression, without suicidal ideation, and interested in using an app to monitor mood while on a selective serotonin reuptake inhibitor.” The second clinical case involved a patient “who is less tech savvy but owns a smartphone her daughter gave her, in her late sixties, and suffering from moderate depression. She has two apps on her phone that she rarely uses but would like to monitor her mood while on a selective serotonin reuptake inhibitor.”

The psychiatrists were not provided any further information about the cases. Each reviewer downloaded the apps in October 2016 and were instructed to use each one for at least 15 minutes before reviewing it as well as to search for any research studies on the apps. We analyzed concordance among all five raters in ratings of each stage of the framework. We used a Kendall’s coefficient of concordance of greater than .667 to indicate agreement among the raters.

The Kendall’s coefficient of concordance was .93 ($p \leq .01$) for the risk/privacy and safety stage, .95 ($p < .01$) for evidence and benefit, .67 ($p \leq .01$) for engagement, and .77 ($p < .01$). for interoperability.

Conclusions

An evaluation framework for informed decision making is a useful solution to the current challenges involved in ratings of apps. In presenting initial and internal reliability metrics of the APA app evaluation framework, we underscore the potential of this simple four-stage hierarchical process model—as well as opportunities to improve it. Although this column focuses on a depression example, we note that this framework is intended for use with patients and apps focused on other conditions, such as schizophrenia (2). For patients with lower literacy, impaired cognition, and apathy the same evaluation process and stages are equally important and relevant. In an effort to better understand how clinicians use this model and to gain data for its further improvement, we have recently begun allowing users to share their app evaluations on the APA Web site.

App evaluation is a complex process involving the input of numerous stakeholder groups (10). Although these initial efforts were developed and tested with psychiatrists, efforts are under way to incorporate diverse stakeholder input into this framework, including the voices of patients and family members. Like apps themselves, app ratings are a dynamic and evolving process. We hope the APA efforts presented here will stimulate discussion and encourage informed decision making around using apps in clinical care.

AUTHOR AND ARTICLE INFORMATION

Dr. Torous and Dr. Nguyen are with the Department of Psychiatry, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston. Dr. Chan is with the Department of Medicine, University of California, San Francisco (UCSF). Dr. Gipson and Dr. Kim are with the Department of Psychiatry, Children's Hospital Boston, Boston. Dr. Luo is with the Department of Psychiatry, University of California, Riverside. Dr. Wang is with the Division of Research, American Psychiatric Association, Washington, D.C. Dror Ben-Zeev, Ph.D., is editor of this column. Send correspondence to Dr. Torous (e-mail: jtorous@bidmc.harvard.edu).

Dr. Chan writes for The Doctor Weighs In, for which he is admitted free to conferences but does not receive financial reimbursement, equity, or compensation. He also advises but does not receive financial remuneration from Whitekoat, Valera Health, Doximity, and the UCSF Catalyst Program. In 2017, he was financially compensated by Guidewell for a talk on mental health innovation, and he taught CME courses for Psych Congress, North American Center for Continuing Medical Education, L.L.C., for which he was financially compensated. The other authors report no financial relationships with commercial interests.

Received September 26, 2017; revisions received November 1 and November 7, 2017; accepted November 17, 2017; published online February 15, 2018.

REFERENCES

1. Torous J, Roberts LW: Needed innovation in digital health and smartphone applications for mental health: transparency and trust. JAMA Psychiatry 74:437–438, 2017
2. Sandoval LR, Torous J, Keshavan MS: Smartphones for smarter care? Self-management in schizophrenia. American Journal of Psychiatry 174:725–728, 2017
3. Torous J, Powell A, Knabble M: Quality assessment of self-directed software and mobile applications for the treatment of mental illness. Psychiatric Annals 46:579–583, 2016
4. Singh K, Drouin K, Newmark LP, et al: Many mobile health apps target high-need, high-cost populations, but gaps remain. Health Affairs 35:2310–2318, 2016
5. 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 and uHealth 4:e15, 2016
6. Larsen ME, Nicholas J, Christensen H: Quantifying app store dynamics: longitudinal tracking of mental health apps. JMIR mHealth and uHealth 4:e96, 2016
7. Rosenfeld L, Torous J, Vahia IV: Data security and privacy in apps for dementia: an analysis of existing privacy policies. American Journal of Geriatric Psychiatry 25:873–877, 2017
8. Nicholas J, Larsen ME, Proudfoot J, et al: Mobile apps for bipolar disorder: a systematic review of features and content quality. Journal of Medical Internet Research 17:e198, 2015
9. Sarkar U, Gourley GI, Lyles CR, et al: Usability of commercially available mobile applications for diverse patients. Journal of General Internal Medicine 31:1417–1426, 2016
10. Ben-Zeev D: Creating new knowledge and inventing the future of services. Psychiatric Services 68:107–108, 2017