AIMS Medical Science

AIMS Medical Science, 9(4): 512–530.
DOI: 10.3934/medsci.2022028
Received: 24 July 2022
Revised: 03 November 2022
Accepted: 21 November 2022
Published: 27 December 2022

http://www.aimspress.com/journal/medicalScience

Research article

Quality evaluation of digital voice assistants for the management of mental health conditions

Vanessa Kai Lin Chua[1], Li Lian Wong[1] and Kevin Yi-Lwern Yap[2,3,*]

[1] Department of Pharmacy, Faculty of Science, National University of Singapore, Block S4A, Level 2, 18 Science Drive 4, Singapore 117543, Singapore
[2] Department of Pharmacy, Singapore General Hospital, SingHealth Tower, 10 Hospital Boulevard, Lobby A, Level 9, Singapore 168582, Singapore
[3] Department of Public Health, School of Psychology and Public Health, La Trobe University, Melbourne (Bundoora), Victoria 3086, Australia

* **Correspondence:** Email: kevin.yap.y.l@sgh.com.sg, k.yap@latrobe.edu.au.

Abstract: Background: Digital voice assistants (DVAs) are gaining increasing popularity as a tool for accessing online mental health information. However, the quality of information provided by DVAs is not known. This study seeks to evaluate the quality of DVA responses to mental health-related queries in relation to six quality domains: comprehension ability, relevance, comprehensiveness, accuracy, understandability and reliability. **Materials and methods:** Four smartphone DVAs were evaluated: Apple Siri, Samsung Bixby, Google Assistant and Amazon Alexa. Sixty-six questions and answers on mental health conditions (depression, anxiety, obsessive-compulsive disorder (OCD) and bipolar disorder) were compiled from authoritative sources, clinical guidelines and public search trends. Three evaluators scored the DVAs from an in-house-developed evaluation rubric. Data were analyzed by using the Kruskal-Wallis and Wilcoxon rank sum tests. **Results:** Across all questions, Google Assistant scored the highest (78.9%), while Alexa scored the lowest (64.5%). Siri (83.9%), Bixby (87.7%) and Google Assistant (87.4%) scored the best for questions on depression, while Alexa (72.3%) scored the best for OCD questions. Bixby scored the lowest for questions on general mental health (0%) and OCD (0%) compared to all other DVAs. In terms of the quality domains, Google Assistant scored significantly higher for comprehension ability compared to Siri (100% versus 88.9%, $p < 0.001$) and Bixby (100% versus 94.5%, $p < 0.001$). Moreover, Google Assistant also scored significantly higher than Siri (100% versus 66.7%, $p < 0.001$) and Alexa (100% versus 75.0%, $p < 0.001$) in terms of relevance. In contrast, Alexa scored the worst in terms of accuracy (75.0%), reliability (58.3%) and

comprehensiveness (22.2%) compared to all other DVAs. **Conclusion:** Overall, Google Assistant performed the best in terms of responding to the mental health-related queries, while Alexa performed the worst. While the comprehension abilities of the DVAs were good, the DVAs had differing performances in the other quality domains. The responses by DVAs should be supplemented with other information from authoritative sources, and users should seek the help and advice of a healthcare professional when managing their mental health conditions.

Keywords: digital voice assistants; mental health information; quality evaluation; depression; anxiety; obsessive-compulsive disorder; bipolar disorder

1. Introduction

Globally, the number of persons suffering from mental health disorders is on the rise [1]. In 2015, an estimated 322 million people were living with depression worldwide [1]. With the recent COVID-19 pandemic, mental well-being was further challenged with fears of contracting an infection [2] and feelings of isolation [3]. Mental health conditions have been associated with stigma in society, causing an individual to perceive oneself as unacceptable [4,5]. The impact of stigma often results in a reduced likelihood of seeking treatment [4,6,7]. In 2018, a USA survey reported that people suffering from depression were increasingly turning to the Internet for mental health-related support [8]. Among them, 90% had researched mental health information online, while 75% had accessed others' health stories through blogs, podcasts and videos [8]. Thus, it is not uncommon that many tend to opt for online support environments, including support groups and social media channels [5,8].

In recent years, digital voice assistants (DVAs) have been increasingly adopted as digital health tools with the purpose of providing information regarding health-related queries for various health conditions, including minor ailments [9], postpartum depression [10], vaccinations [11,12], cancer screening [13] and smoking cessation advice [14]. Smartphone-based DVAs, such as Apple Siri and Google Assistant, have been particularly popular [15]. According to Google, 27% of Internet searches in 2018 came from using the voice search feature on smartphones [16], with this trend posited to grow. The artificial intelligence (AI) component in DVAs enables voice recognition and responses in natural language [10,17], thereby enabling these DVAs to participate in two-way conversations with users [18]. Given the growing popularity of using DVAs to search for online health information [8], it is crucial that DVAs are able to provide relevant, appropriate and easy-to understand responses to queries by users in relation to mental health literacy, such as symptom recognition, information sources, awareness of causes and risks and an understanding of treatment types [19,20]. While there are quality assessment tools that evaluate the quality of online health information, such as the Health-on-the-Net Code (HONcode) [21], DISCERN [22] and Quality Evaluation Scoring Tool (QUEST) [23], from our knowledge, there are no existing ones for the purpose of assessing DVAs. On the other hand, studies that have evaluated the quality of information provided by DVAs [9–14] have not focused on mental health conditions.

As we move into a post-pandemic world, it is crucial that public mental health should not be ignored [24]. There is a need to evaluate the quality of information provided by DVAs in the mental health domain. Studies have suggested that providing useful and comprehensive online information about mental health conditions in a user-friendly way can help consumers gain a better understanding

of the disease, which in turn can help prevent and/or reduce the severity of the mental health disorder [25]. Furthermore, providing high-quality information online on mental health conditions can potentially reduce the stigma and prejudice attached to these disorders [25]. With the increased popularity of consumers performing health information searches through DVAs, it is crucial that DVAs are able to provide high-quality information on mental health conditions through their responses. Our hypothesis is that DVAs are able to provide responses that are relevant, appropriate and easy-to understand in relation to mental health queries. Thus, the primary objective of this study was to evaluate the quality of DVA responses to mental health-related queries by using an in-house-developed quality assessment rubric. In this study, DVAs are defined as inanimate programs enhanced with AI that interact with human users using speech commands. These are different from other technologies such as chatbots [26] or automated telephone-response systems [27,28].

2. Materials and methods

2.1. Definition of quality

In this study, the quality of DVAs was defined as the degree of excellence to which a DVA could fulfill the needs of mental health-related queries [29]. This definition was represented by six quality domains: comprehension ability, relevance, comprehensiveness, accuracy, understandability and reliability. The quality domains were adapted from tools evaluating the quality of online health information or sources. The relevance domain was adapted from the DISCERN [22] and CRAAP (currency, relevance, authority, accuracy and purpose) [30,31] tools. The accuracy and reliability domains were adapted from DISCERN [22], CRAAP [30,31] and HONcode [21]. In addition, the reliability domain was also adapted from the Ensuring Quality Information for Patients (EQIP) tool [32], LIDA Minervation validation instrument [33], QUEST [23] and Quality Component Scoring System [34]. The comprehensiveness domain was adapted from DISCERN and EQIP [22,32], and understandability was adapted from EQIP and LIDA [32,33].

2.2. Quality evaluation rubric

The quality domains evaluated three aspects of DVA quality: the DVAs themselves (comprehension ability), the DVAs' responses (relevance, comprehensiveness, accuracy and understandability) and the answer sources provided by the DVAs (reliability) (Figure 1). The composite score for all domains added up to a maximum of 32 points. All DVA responses were classified into four types: verbal response only, web response only, verbal and web response and no response. "Verbal response only" referred to a short verbal text that directly answered the question without providing a link. Conversely, a "web response only" referred to a link without any verbal explanation provided. A "verbal and web response" consisted of both the aforementioned parts in a single response. If the DVA did not provide any responses, it would be classified as "no response". Since understandability was evaluated for both the verbal and web responses, in cases where the DVA only provided one type of response, the composite score would be 30 points instead.

The DVA's comprehension ability was assessed based on its ability to accurately recognize and transcribe the question posed to it. Relevance of the DVA's responses was assessed based on whether the response had adequately addressed the question. For two questions, the DVAs were evaluated for

their ability to successfully refer to a contact point in cases requiring immediate intervention. Comprehensiveness was assessed based on whether the DVA's response was complete and fulfilled all of the points in the answer sheet. In addition, two quality-of-life (QoL) criteria assessed whether the DVA described impacts of treatment or treatment choices on day-to-day living or activities, and whether it supported shared decision-making regarding treatment choices. Accuracy assessed whether each point in the DVA's response correctly matched the corresponding point in the answer sheet. Understandability was assessed based on whether a layman would easily understand the DVA response according to the Simple Measure of Gobbledygook (SMOG) readability test [35,36], and whether it contained medical jargon/complex words. Lastly, the reliability of answer sources provided by the DVAs was evaluated based on six criteria: credibility of the sources and reference citations, how current/updated were the sources, presence/absence of bias and advertisements and whether there was a disclaimer stating that the information provided did not replace a healthcare professional's advice. All DVA responses were evaluated regardless of whether they were verbal or web responses.

2.3. Questions on mental health

A total of 66 questions on mental well-being and mental health conditions were compiled and categorized into five categories: general mental health, depression, anxiety, obsessive-compulsive disorder (OCD) and bipolar disorder. These conditions were chosen due to their rising prevalence in global and local data [1,37]. Besides the section on general mental health, questions in the other sections on the specific mental health conditions were classified into three subcategories: disease state, symptoms and treatment (Appendix 1).

Questions and answers were sourced primarily from the American Psychiatric Association [38], National Institute of Mental Health [39], Medline Plus [40], World Health Organization [41], USA Centers for Disease Control and Prevention [42], Mayo Clinic [43], Cleveland Clinic [44], National Alliance on Mental Illness [45], Anxiety and Depression Association of America [46] and the International Obsessive-Compulsive Disorder Foundation [47]. In addition, questions were also sourced from AnswerThePublic [48] with the following keywords: "mental health", "depression", "anxiety", "OCD" (obsessive-compulsive disorder) and "bipolar disorder". Answers were also compiled from established clinical guidelines, including the Diagnostic and Statistical Manual of Mental Disorders, 5th edition (DSM-5) [49] and the Singapore Ministry of Health Clinical Practice Guidelines [50]. The questions and answers were reviewed by three reviewers (VC, WLL, KY). Any differences in opinions were resolved through discussions until consensus was reached. Two reviewers (JC and LL) pilot-tested half of the questions to ensure that the evaluation rubric could be applied across different questions. Their feedback was used to refine the rubric for the actual evaluation.

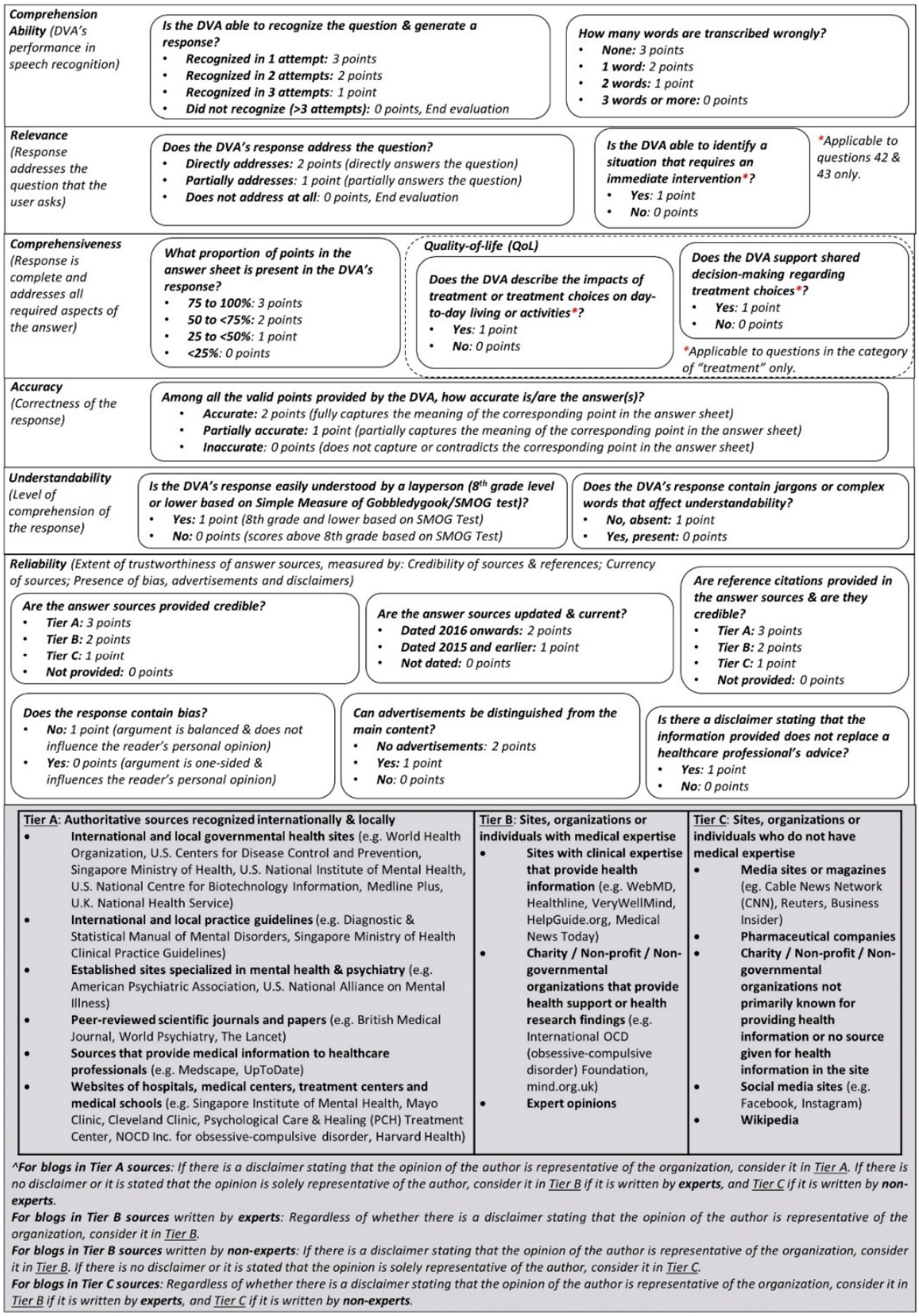

Figure 1. Quality evaluation rubric for DVAs.

2.4. Evaluation of DVAs

Four smartphone DVAs were employed for evaluation: Apple Siri, Samsung Bixby, Google Assistant and Amazon Alexa. Siri and Google Assistant were accessed by using an iPhone 6 (iOS14.7.1), while Bixby and Alexa were accessed by using a Samsung Galaxy Note 9 (OS10). All questions were posed to the DVAs in English by native English speakers—in the same order and in the exact way that the questions were phrased in Appendix 1. The evaluations and scoring were done independently on the same devices by three evaluators in a quiet room at their homes: VC (female), LSK (male) and AP (female). Each evaluator would ask all 66 questions to one DVA in one sitting. However, they would pose the questions to a different DVA in a separate sitting (i.e., four separate sessions). If the DVA was unable to capture the question and generate a response after three repeated attempts, the evaluation would end and no points would be awarded. Each evaluator completed the evaluation of all four DVAs within a week, after which, the devices were transferred to the next evaluator, who would then evaluate the DVAs on the same devices over the next consecutive week. As such, all evaluations were completed within 3 weeks. The search and internet histories for the individual DVAs were reset before and after each round of evaluation. The location function was turned on as the DVAs were evaluated for their ability to refer to a contact point. If the DVA provided more than one web link, the first web link was taken for evaluation.

2.5. Statistical analyses

Descriptive statistics (numbers and percentages) were employed to report the types of responses, proportion of successful responses and sources cited by the DVAs. The quality scores were calculated for each mental health category (general mental health, depression, anxiety, OCD, bipolar disorder) and question subcategory (disease state, symptoms, treatment), as well as for each quality domain (comprehension ability, relevance, comprehensiveness, accuracy, understandability, reliability, overall quality), by dividing the sum of points awarded for each DVA against the maximum possible number of points in each mental health category, question subcategory and quality domain (Equation 1). This calculation was also performed across all questions to generate a composite quality score. All quality scores were converted to percentages and reported as medians and interquartile ranges (IQRs). All results were taken as averages of the three evaluators.

$$\begin{pmatrix} Quality\ score\ for\ each\ mental\ health\ category, \\ question\ subcategory\ or\ quality\ domain \end{pmatrix} = \frac{\sum\ of\ points\ awarded\ for\ DVA}{\begin{pmatrix} Maximum\ number\ of\ points\ for\ the \\ category,\ \ subcategory\ or\ domain \end{pmatrix}} 100\% \quad (1)$$

All statistical analyses were performed at a significance level of 0.05 by using the Statistical Package for Social Sciences (SPSS) software (version 27). Normality tests, including Shapiro-Wilk tests ($n < 50$) and Kolmogorov-Smirnov tests ($n \geq 50$) were conducted before Kruskal-Wallis tests were applied to compare the results across all four DVAs. Post-hoc analyses using Wilcoxon rank sum tests with Bonferroni adjustments were subsequently performed for each possible pairwise comparison among the DVAs. Wilcoxon rank sum testing was also used to compare the understandability of verbal and web responses. Inter-rater reliability was calculated by using the intraclass correlation coefficient (ICC) [51] based on a mean rating of three evaluators, absolute agreement, a two-way mixed-effects model and a 95% confidence interval (95% CI).

3. Results

The majority of the responses by Siri were web responses (72.7%), while verbal responses formed the major proportion of responses by Alexa (62.1%) (Table 1). The largest proportion of responses from Google Assistant consisted of both verbal and web responses (78.8%). However, Bixby had a comparable distribution of verbal responses only (36.4%) and verbal and web responses (42.4%).

Table 1. Types of responses, proportion of successful responses and sources used for each DVA.

	Number of responses (%), N = 66 [a]			
	Apple Siri	Samsung Bixby	Google Assistant	Amazon Alexa
Types of responses by DVAs				
Verbal response only [b]	6 (9.1)	24 (36.4)	1 (1.5)	41 (62.1)
Web response only [c]	48 (72.7)	14 (21.2)	13 (19.7)	0 (0)
Verbal and web response [d]	11 (16.7)	28 (42.4)	32 (78.8)	24 (36.4)
No response	1 (1.5)	0 (0)	0 (0)	1 (1.5)
Proportion of successful responses				
Questions that were recognized [e]	63 (95.5)	46 (69.7)	66 (100.0)	60 (90.9)
Relevant responses	47 (71.2)	38 (57.6)	66 (100.0)	44 (66.7)
Proportion of sources provided in DVA responses				
Tier A	13 (19.7)	19 (28.8)	36 (54.5)	20 (30.3)
Tier B	19 (28.8)	18 (27.3)	18 (27.3)	8 (12.1)
Tier C	15 (22.7)	1 (1.5)	6 (9.1)	15 (22.7)
No sources provided, or sources that could not be evaluated	19 (28.8)	28 (42.4)	6 (9.1)	23 (34.8)

Note: [a] Results were taken from the average of three evaluators. [b] A short verbal text that directly answered the question without providing a link. [c] A link was provided in response to the question without a verbal explanation. [d] Both a verbal explanation and a link were present in the response. [e] These were questions that were captured on the smartphone screen and induced a response by the DVA. Responses such as "I'm not sure I understood that" were classified as the DVA not recognizing the question.

The proportion of responses that were successfully recognized varied across the DVAs. Responses were deemed to be recognized successfully if the questions were captured on the smartphone screen and a response was provided by the DVA. If the DVA provided a response like "I'm not sure I understood that", its response would be classified as not being recognized. Similarly, if the DVA

provided a response that was relevant to the question, it would be classified as such. For the proportion of questions that were recognized, Google Assistant performed the best (100%), followed by Siri (95.5%), Alexa (90.9%) and Bixby (69.7%). The proportion of relevant responses followed the same trend, with Google Assistant performing the best (100%) and Bixby performing the worst (57.6%).

In terms of the credibility of the sources provided, Google Assistant (54.5%) and Siri (19.7%) had the highest and lowest proportions of Tier A sources, respectively. Over a quarter of the sources by Siri (28.8%), Bixby (27.3%) and Google Assistant (27.3%) were Tier B, while Siri and Alexa had the largest proportions of Tier C sources (22.7% each).

Across all 66 questions (Table 2), Google Assistant had the highest median composite quality score (78.9%) among the DVAs, while Alexa had the lowest median composite score (64.5%). Siri (83.9%), Bixby (87.7%) and Google Assistant (87.4%) scored the best for questions on depression, in contrast to Alexa (72.3%), which scored the best for OCD questions. Alexa scored significantly lower (63.0%, $p < 0.001$) than all other DVAs for questions on depression, and significantly lower (60.5%) than Bixby (75.9%, $p < 0.001$) and Google Assistant (76.4%, $p = 0.004$) for questions on anxiety. On the other hand, Bixby scored significantly lower than all other DVAs for questions on general mental health and OCD (0%, $p < 0.001$ each). Additionally, Siri scored significantly lower than Google Assistant for questions on OCD (61.7% versus 78.4%, $p = 0.002$).

Among the question subcategories, Siri (71.7%) and Google Assistant (80.5%) scored the best for questions on disease state, as compared to questions on symptoms and treatment (Table 2). On the other hand, Bixby had similar scores across all three subcategories of disease state, symptoms and treatment. In contrast, Alexa scored the highest for questions on symptoms (71.5%), but its score in the treatment subcategory (57.3%) was significantly lower than those of Bixby (78.3%, $p < 0.001$) and Google Assistant (77.3%, $p < 0.001$). Furthermore, Alexa's scores were also significantly lower than Google Assistant for questions in the subcategory of disease state (69.6% versus 80.5%, $p = 0.004$).

Table 2. Comparison of quality scores among the DVAs for all questions and across the mental health categories and question subcategories.

Classification of Questions	Median Quality Scores of DVAs [% (IQR)]				p-values*
	Apple Siri	Samsung Bixby	Google Assistant	Amazon Alexa	
Across all questions	70.4 (60.9–79.3)	72.8 (0–81.6)	78.9 (73.9–85.2)	64.5 (57.7–76.7)	<0.001
Mental Health Categories					
General mental health	77.1 (71.3–85.2)	0 (0–16.7)	80.5 (76.4–89.1)	70.7 (57.1–79.7)	<0.001
Depression	83.9 (76.0–86.9)	87.7 (83.6–89.3)	87.4 (79.4–88.7)	63.0 (61.4–72.1)	<0.001
Anxiety	71.8 (67.2–87.6)	75.9 (71.3–80.7)	76.4 (69.8–83.6)	60.5 (42.5–68.4)	0.006
Obsessive-compulsive disorder	61.7 (53.7–69.8)	0 (0–29.6)	78.4 (73.6–85.7)	72.3 (59.1–80.0)	<0.001
Bipolar disorder	66.4 (44.4–70.4)	77.5 (71.3–81.6)	75.9 (70.1–81.6)	63.0 (48.2–80.5)	0.004

Continued on next page

Classification of Questions	Median Quality Scores of DVAs [% (IQR)]				p-values*
	Apple Siri	Samsung Bixby	Google Assistant	Amazon Alexa	
Question Subcategories					
Disease state	71.7 (66.9–79.0)	71.6 (28.2–83.3)	80.5 (73.8–84.8)	69.6 (62.5–80.2)	0.031
Symptoms	66.7 (53.9–83.1)	76.7 (25.0–80.9)	77.5 (70.7–86.0)	71.5 (57.7–80.4)	0.239
Treatment	60.5 (49.4–74.2)	78.3 (63.0–85.0)	77.3 (69.6–84.3)	57.3 (30.6–62.1)	<0.001

Note: *Kruskal-Wallis test was performed among all the four DVAs with statistical significance defined as $p < 0.05$. Post-hoc analyses using the Wilcoxon rank sum test with Bonferroni adjustment were performed for each possible pairwise comparison among the DVAs, with statistical significance defined as $p < 0.00833$.

Across all quality domains, Google Assistant scored the highest while Alexa scored the lowest (Table 3). In terms of comprehension ability, Google Assistant scored significantly higher (100%, $p < 0.001$) than the other DVAs. In addition, Alexa (100%) scored significantly higher than Siri (88.9%, $p < 0.001$) and Bixby (94.5%, $p = 0.03$) in this domain. Google Assistant (100%) and Bixby (100%) also scored significantly higher than Siri (66.7%) and Alexa (75.0%) in terms of relevance. Only Google Assistant was successful in identifying situations that required immediate intervention from one evaluator (16.7%).

Alexa scored the worst among all DVAs in terms of comprehensiveness (22.2%, $p < 0.001$) and reliability (58.3%, $p < 0.001$). In addition, Alexa also performed the poorest when evaluated against the QoL criteria (10.0%), as compared to Bixby, which performed the best (76.7%). In contrast, Google Assistant scored the best (77.8%) in terms of comprehensiveness, but it had similar reliability scores as Bixby (75.0% each). In terms of accuracy, Alexa scored the lowest among the DVAs (75.0% versus 100% for other DVAs, $p = 0.003$). However, all DVAs had similar scores for understandability (50.0% each). The understandability of verbal responses was significantly lower than that of web responses (33.3% versus 50.0%, $p = 0.004$). Inter-rater reliability ranged from moderate to good for both the overall quality and the individual quality domains (Table 3).

Table 3. Comparison of quality scores among the DVAs across the quality domains.

Quality Domains	Median Quality Scores of DVAs [% (IQR)]				p-value*	Intraclass Correlation Coefficient [ICC (95% CI)] [a]
	Apple Siri	Samsung Bixby	Google Assistant	Amazon Alexa		
Comprehension ability	88.9 (70.8–100)	94.5 (0–100)	100 (100–100)	100 (88.9–100)	<0.001	0.892 (0.868–0.913)
Relevance	66.7 (50.0–100)	100 (66.7–100)	100 (83.3–100)	75.0 (33.3–100)	<0.001	0.753 (0.691–0.804)
Comprehensiveness	66.7 (44.4–83.3)	66.7 (55.6–88.9)	77.8 (55.6–88.9)	22.2 (0–66.7)	<0.001	0.747 (0.660–0.812)
Accuracy	100 (75.0–100)	100 (83.3–100)	100 (83.3–100)	75.0 (50.0–100)	0.003	0.691 (0.593–0.769)
Understandability	50.0 (25.0–75.0)	50.0 (33.3–68.8)	50.0 (33.3–66.7)	50.0 (25.0–75.0)	0.724	0.672 (0.513–0.775)
Reliability	72.9 (63.2–83.3)	75.0 (63.9–84.3)	75.0 (66.7–84.3)	58.3 (49.1–63.9)	<0.001	0.896 (0.863–0.922)
Overall quality	70.4 (60.9–79.3)	72.8 (0–81.6)	78.9 (73.9–85.2)	64.5 (57.7–76.7)	<0.001	0.848 (0.813–0.877)

Note: * Kruskal-Wallis test was performed among all four DVAs with statistical significance defined as $p < 0.05$. Post-hoc analyses using the Wilcoxon rank sum test with Bonferroni adjustment were performed for each possible pairwise comparison among the DVAs, with statistical significance defined as $p < 0.00833$. [a] ICC values and their 95% CIs were calculated using the SPSS platform based on the mean rating of three evaluators, absolute agreement and a two-way mixed-effects model. ICC values indicate moderate-to-good inter-rater reliability.

4. Discussion

In relation to our hypothesis, this study has shown that DVAs are able to provide relevant and appropriate responses to mental health-related queries. However, the understandability of their responses was relatively low. Furthermore, not all DVAs fared the same in terms of the different quality domains, and they also varied across the various mental health conditions. Overall, Google Assistant performed the best among all DVAs, suggesting that it was able to comprehend the queries and provide responses that were relevant and accurate across the various mental health categories. In comparison, Bixby fared the worst in terms of responding to questions on general mental health and OCD. On the other hand, Alexa's responses were the least comprehensive and reliable across all questions, as well as in the categories of depression, anxiety and bipolar disorder.

All DVAs performed well in terms of comprehension ability. This result was similar to a study by Yang and colleagues, who investigated the abilities of Siri, Google Assistant, Alexa and Cortana in terms of responding to questions on postpartum depression [10]. In their study, all DVAs performed well in terms of recognizing the postpartum depression questions, with scores ranging from 79% (Alexa) to 100% (Siri and Google Assistant). However, in our study, Siri and Bixby performed poorer than Google Assistant and Alexa. For Bixby, a quarter of the questions posed (27.3%, n = 18/66) were

scored as 0%. In particular, Bixby often transcribed "OCD" as "o CD" (two separate words), resulting in a large proportion of questions failing to be recognized. In addition, while Bixby could accurately transcribe questions on general mental health, it could not generate responses for many of these questions (80%, n = 8/10) and frequently answered with "I'm not sure I understood that". We postulate that our observations could be due to Bixby's primary design intent, which was to assist users in operating the phone via voice commands, rather than provide accurate responses to questions, as in the case of other DVAs [52]. On the other hand, while Siri could successfully capture all questions, it was penalized for transcribing errors. Siri tended to cut off the user before the entire question was posed, resulting in incomplete prompts being captured on the screen. Examples included "Can depression…" and "What is the difference between…", when the entire questions that were meant to be asked were "Can depression be genetic?" and "What is the difference between normal behavior and OCD?". respectively.

In regard to relevance, Siri and Alexa performed more poorly than Google Assistant and Bixby due to the irrelevant responses provided. For example, Siri responded with answers about medications when the question posed was "How are anxiety disorders diagnosed?" Similarly, Alexa responded with the effects of bipolar disorder to the question of "Who does bipolar disorder affect?" When the DVAs were evaluated for their ability to refer cases that required immediate intervention, only Google Assistant managed to respond appropriately to one evaluator. Interestingly, our observations differed from a study by Kocaballi and colleagues [53], who reported that Siri scored the highest for safety-critical prompts when compared to Google Assistant, Bixby and Alexa. In another study by Miner et al. [17], even though Google Now and Samsung S Voice (predecessor of Bixby) [54] managed to recognize queries on suicide as a cause for concern, Google Now did not recognize the cause for concern for queries on depression, while the responses from S Voice varied, with the cause of concern being recognized only in some instances. Nonetheless, the authors of both studies agreed that there was an inconsistency in the responses of the DVAs and that their abilities to recognize causes for concern should improve. It is unclear whether the inability of DVAs to respond to queries appropriately is due to system failure, a failure of the natural language understanding, a misrecognized prompt, the DVA being unable to find a response or the DVA deliberately not responding to particular types of queries [53]. However, we agree with Kocaballi and colleagues and advocate that the DVAs' capabilities should be made more transparent to users so that it can improve user experience and reduce confusion.

For comprehensiveness, Alexa performed the worst among the DVAs. It also scored significantly lower than Bixby and Google Assistant in terms of accuracy. In contrast, Alexa performed well in terms of comprehension ability, suggesting that, even though it could comprehend the questions being posed, it did not provide comprehensive and accurate responses. Our findings were consistent with a study by Alagha and Helbing, who evaluated the quality of responses to questions on vaccines by Google Assistant, Siri and Alexa [11]. In their study, the authors indicated that Alexa lacked in its ability to process health queries and generate responses from high-quality sources. Furthermore, in our study, Alexa performed significantly poorer than the other DVAs in terms of reliability. One reason was its tendency to only provide verbal responses, such as "Here's something I found on Mayo Clinic", while the other DVAs provided specific links to webpages. In addition, Alexa provided invalid links to "reference.com", which could not be accessed on several occasions. Our observations were also in line with the DVA vaccine information study by Alagha and Helbing [11], who reported that Google Assistant and Siri were more capable of directing the user to authoritative sources than Alexa, which

did not provide answers from the same sources as the other DVAs. Hence, our recommendation is to supplement Alexa's responses to mental health queries with those of another DVA or other external resources so that any lack of or discrepancies in health-related information provided can be identified by the user.

There was a significant difference between the understandability of verbal responses versus web responses. Verbal responses were less easily understood, as according to the SMOG readability test, and contained more jargon than web responses. However, both types of responses also scored poorly, indicating that the responses of the DVAs to mental health queries are less likely to be understood by a layperson. Our results concurred with a study assessing the readability of online health information, which showed that, among 12 health conditions, the information on dementia and anxiety were the hardest to read [55]. As the understandability of health-related information is important to raise one's awareness and knowledge of mental health issues and self-care, we advocate that the information provided by DVAs should be complemented with other information online and shared between the patient and caregiver (or someone whom the patient trusts) in a close and private setting that is comfortable for the patient.

Across the mental health conditions, Siri, Bixby and Google Assistant scored the highest for questions on depression. Our results were similar to the study by Miner et al., which investigated the responses of Siri, Google Now, S Voice and Cortana to questions on depression [17]. In their study, the DVAs were generally able to recognize prompts, but they were not able to refer the user to a depression helpline. On the contrary, a study by Kocaballi et al. showed that DVAs had the lowest ratio of appropriate responses to mental health prompts, including those of depression [53]. Even though there have been studies investigating the quality of conversational agents on mental health conditions [56,57], these studies focused on other types of conversational agents, such as chatbots and mobile apps, instead of DVAs. To the best of our knowledge, there is a paucity of studies that explore the quality of DVAs in relation to mental health conditions, especially OCD and bipolar disorder. While Google Assistant seems to be one of the top two DVAs that can potentially be recommended for queries on OCD and bipolar disorder (Figure 2), its ability to answer questions on these two conditions may not be as well established as that for general mental health and depression queries. Interestingly, Siri did not perform as well on either of these mental health conditions. As such, we recommend Apple users who seek information about OCD and/or bipolar disorder from Siri to supplement their responses with other online resources from Google Assistant or Google searches. In any case, our study presents new insight into the quality of DVAs across the span of these four mental health conditions—depression, anxiety, OCD and bipolar disorder.

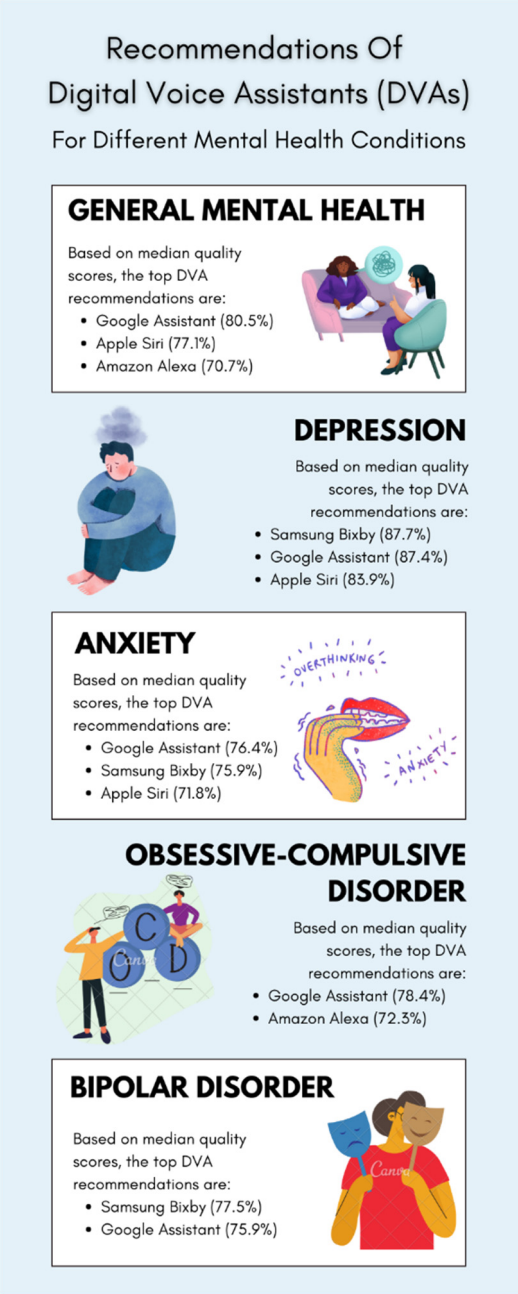

Figure 2. DVA recommendation list for the different mental health conditions.

5. Limitations

The main limitation of this study is that we were only able to evaluate a subset of four DVAs and four mental health conditions. Therefore, our results might not be representative of the DVAs' performances for other mental health conditions, nor of the quality of other DVAs (e.g., Google Home Mini and Microsoft Cortana). Furthermore, as the location function of the DVAs were switched on during our evaluations, the search results might have been adapted to the local context, and minor variations could exist depending on the country and location of the user. Studies have shown that the responses of DVAs provided to the same questions can differ [17,58]. Although the qualitative

responses of the DVAs were not compared in this study, we tried to minimize this variability by having each evaluator use the same devices for their evaluations. In order to account for the variations in evaluation scores of the same DVA response by the different evaluators, we calculated the ICC values for each quality domain (Table 3) to determine the inter-rater reliability; our results indicated moderate-to-good reliability. Similarly, inter-rater reliability for the overall quality scores of the DVAs was good. Nonetheless, we acknowledge that this bias may exist in the DVA responses, and our study results should be interpreted with this limitation in mind. In addition, our evaluation protocol might not be reflective of real-life usage of DVAs by the layperson. In our study, when the question posed to the DVAs was not recognized on the first attempt, there would be two more attempts made before the evaluation ended. However, in real-life, users might forgo repeatedly asking the same question multiple times if they encountered an unsuccessful response on their first try. Next, due to time limitations, only the first web link provided by the DVAs was evaluated in this study, but, in reality, users might access other links as well if more than one link was provided by the DVAs. Lastly, our results only provide the quality of the DVAs in a snapshot of time. With advancements in voice recognition technologies, natural language processing and other AI-based algorithms, we expect that the quality of the DVAs will also improve over time. As such, we advise caution when extrapolating the results of this study to other DVAs, other countries/states, other mental health conditions or over time.

6. Conclusions

Overall, Google Assistant performed the best in terms of responding to mental health-related queries, while Alexa performed the worst. In terms of specific mental health conditions, Bixby performed the worst for questions on general mental health and OCD. While the comprehension abilities of the DVAs were generally good, our study showed that the DVAs had differing performances in the domains of relevance, comprehensiveness, accuracy and reliability. Moreover, the responses of the DVAs generally lacked in understandability. Based on our quality evaluations, we have provided a DVA recommendation list that users can potentially consider for the different mental health conditions (Figure 2). While Google Assistant generally works well across all of the included mental health conditions, Siri and Bixby can also be used for depression and anxiety. On the other hand, Alexa and Bixby may potentially be used for OCD and bipolar disorder, respectively. However, when depending on the DVA responses to their mental health-related queries, we caution the general public to supplement the information provided by the DVAs with other online information from authoritative healthcare organizations, and to always seek the help and advice of a healthcare professional when managing their mental health condition(s). In light of many organizations adapting to the post-pandemic world, future research should focus on other types of mental health conditions (e.g., stress) in patients, caregivers and healthcare professionals resulting from specific circumstances, such as workplace disruptions, loss of healthcare services and the accumulation of new job roles as healthcare undergoes a major digital transformation worldwide. In addition, further research can also be done to evaluate other types of DVAs' performance for mental health conditions that are relevant to the researchers' communities.

Acknowledgments

The authors would like to thank Mr. Christopher Chua and Mr. Luke Si Sheng Lim for loaning their Samsung Galaxy Note 9 and iPhone 6, respectively, for the evaluation of the digital voice assistants; Ms. Joy Qi En Chia and Mr, Luke Si Sheng Lim for pilot testing the evaluation rubric; and Mr. Sheng Kiat Lee and Ms. Alyssa Pua for their time and effort in conducting the actual evaluation with all of the digital voice assistants.

Conflict of interest

No competing financial interests exist. All authors have no conflict of interest with any of the evaluated apps or the companies in this study. This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

References

1. World Health Organization (2017) Depression and other common mental disorders: Global health estimates, Geneva, Switzerland: WHO Document Production Services, 24 pp. Available from: https://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf. Accessed 10 Jun 2022.
2. World Health Organization. Mental health and COVID-19. Available from: https://www.who.int/teams/mental-health-and-substance-use/mental-health-and-covid-19. Accessed 10 Jun 2022.
3. Centers for Disease Control and Prevention. Mental health: Coping with stress. U.S. Department of Health & Human Services, 2022. Available from: https://www.cdc.gov/coronavirus/2019-ncov/daily-life-coping/managing-stress-anxiety.html. Accessed 10 Jun 2022.
4. Borenstein J (2020) Stigma, prejudice and discrimination against people with mental illness. American Psychiatric Association. Available from: https://www.psychiatry.org/patients-families/stigma-and-discrimination. Accessed 10 Jun 2022.
5. DeAndrea DC (2015) Testing the proclaimed affordances of online support groups in a nationally representative sample of adults seeking mental health assistance. *J Health Commun* 20: 147–156. https://doi.org/10.1080/10810730.2014.914606
6. Gulliver A, Griffiths KM, Christensen H (2010) Perceived barriers and facilitators to mental health help-seeking in young people: A systematic review. *BMC Psychiatry* 10: 113. https://doi.org/10.1186/1471-244X-10-113
7. Rickwood DJ, Deane FP, Wilson CJ (2007) When and how do young people seek professional help for mental health problems? *Med J Aust* 187: S35–S39. https://doi.org/10.5694/j.1326-5377.2007.tb01334.x
8. Well Being Trust. Digital health practices, social media use, and mental well-being among teens and young adults in the US. Well Being Trust, 2018. Available from: https://wellbeingtrust.org/bewell/digital-health-practices-social-media-use-and-mental-well-being-among-teens-and-young-adults-in-the-u-s/. Accessed 10 Jun 2022.

9. Jee C (2019) Amazon Alexa will now be giving out health advice to UK citizens. MIT Technology Review. Available from: https://www.technologyreview.com/2019/07/10/134244/amazon-alexa-will-now-be-giving-out-health-advice-to-uk-citizens/. Accessed 10 Jun 2022.
10. Yang S, Lee J, Sezgin E, et al. (2021) Clinical advice by voice assistants on postpartum depression: Cross-sectional investigation using Apple Siri, Amazon Alexa, Google Assistant, and Microsoft Cortana. *JMIR Mhealth Uhealth* 9: e24045. https://doi.org/10.2196/24045
11. Alagha EC, Helbing RR (2019) Evaluating the quality of voice assistants' responses to consumer health questions about vaccines: An exploratory comparison of Alexa, Google Assistant and Siri. *BMJ Health Care Inform* 26: e100075. https://doi.org/10.1136/bmjhci-2019-100075
12. Figueiredo CMS, de Melo T, Goes R (2022) Evaluating voice assistants' responses to COVID-19 vaccination in Portuguese: Quality assessment. *JMIR Hum Factors* 9: e34674. https://doi.org/10.2196/34674
13. Hong G, Folcarelli A, Less J, et al. (2021) Voice assistants and cancer screening: A comparison of Alexa, Siri, Google Assistant, and Cortana. *Ann Fam Med* 19: 447–449. https://doi.org/10.1370/afm.2713
14. Boyd M, Wilson N (2018) Just ask Siri? A pilot study comparing smartphone digital assistants and laptop Google searches for smoking cessation advice. *PLoS One* 13: e0194811. https://doi.org/10.1371/journal.pone.0194811
15. Kinsella B (2019) Voice assistant demographic data—Young consumers more likely to own smart speakers while over 60 bias toward Alexa and Siri. Voicebot.ai. Available from: https://voicebot.ai/2019/06/21/voice-assistant-demographic-data-young-consumers-more-likely-to-own-smart-speakers-while-over-60-bias-toward-alexa-and-siri/. Accessed 11 Jun 2022.
16. Think with Google. Marketing strategies. Global Web Index, Voice Search Insight Report, 2018. Available from: https://www.thinkwithgoogle.com/marketing-strategies/search/voice-search-mobile-use-statistics/. Accessed 11 Jun 2022.
17. Miner AS, Milstein A, Schueller S, et al. (2016) Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA Intern Med* 176: 619–625. https://doi.org/10.1001/jamainternmed.2016.0400
18. Sezgin E, Militello LK, Huang Y, et al. (2020) A scoping review of patient-facing, behavioral health interventions with voice assistant technology targeting self-management and healthy lifestyle behaviors. *Transl Behav Med* 10: 606–628. https://doi.org/10.1093/tbm/ibz141
19. Jorm AF, Korten AE, Jacomb PA, et al. (1997) "Mental health literacy": A survey of the public's ability to recognise mental disorders and their beliefs about the effectiveness of treatment. *Med J Aust* 166: 182–186. https://doi.org/10.5694/j.1326-5377.1997.tb140071.x
20. Zachrisson HD, Rodje K, Mykletun A (2006) Utilization of health services in relation to mental health problems in adolescents: A population based survey. *BMC Public Health* 6: 34. https://doi.org/10.1186/1471-2458-6-34
21. Health On The Net (HON) Foundation. HONcode health sites certification guidelines. Health On The Net, 2020. Available from: https://web.archive.org/web/20220119005932/https://www.hon.ch/imgs/2020/EN-Guidelines-Sites_compressed.pdf. Accessed 11 Jun 2022.

22. Charnock D (1998) *The DISCERN Handbook: Quality criteria for consumer health information on treatment choices*, Abingdon, Oxon: Radcliffe Medical Press, 55 pp. Available from: https://web.archive.org/web/20220621053038/http://www.discern.org.uk/discern.pdf. Accessed 11 Jun 2022.
23. Robillard JM, Jun JH, Lai JA, et al. (2018) The QUEST for quality online health information: Validation of a short quantitative tool. *BMC Med Inform Decis Mak* 18: 87. https://doi.org/10.1186/s12911-018-0668-9
24. Ren FF, Guo RJ (2020) Public mental health in post-COVID-19 era. *Psychiatr Danub* 32: 251–255. https://doi.org/10.24869/psyd.2020.251
25. Nemoto K, Tachikawa H, Sodeyama N, et al. (2007) Quality of Internet information referring to mental health and mental disorders in Japan. *Psychiatry Clin Neurosci* 61: 243–248. https://doi.org/10.1111/j.1440-1819.2007.01650.x
26. Inkster B, Sarda S, Subramanian V (2018) An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. *JMIR Mhealth Uhealth* 6: e12106. https://doi.org/10.2196/12106
27. Kaminer Y, Litt MD, Burke RH, et al. (2006) An interactive voice response (IVR) system for adolescents with alcohol use disorders: A pilot study. *Am J Addict* 15: 122–125. https://doi.org/10.1080/10550490601006121
28. Rose GL, Skelly JM, Badger GJ, et al. (2012) Interactive voice response for relapse prevention following cognitive-behavioral therapy for alcohol use disorders: A pilot study. *Psychol Serv* 9: 174–184. https://doi.org/10.1037/a0027606
29. Yap KY, Raaj S, Chan A (2010) OncoRx-IQ: A tool for quality assessment of online anticancer drug interactions. *Int J Qual Health Care* 22: 93–106. https://doi.org/10.1093/intqhc/mzq004
30. Farkas C, Solodiuk L, Taddio A, et al. (2015) Publicly available online educational videos regarding pediatric needle pain: A scoping review. *Clin J Pain* 31: 591–598. https://doi.org/10.1097/AJP.0000000000000197
31. Garcia M, Daugherty C, Khallouq BB, et al. (2018) Critical assessment of pediatric neurosurgery patient/parent educational information obtained via the Internet. *J Neurosurg Pediatr* 21: 535–541. https://doi.org/10.3171/2017.10.PEDS17177
32. Moult B, Franck LS, Brady H (2004) Ensuring quality information for patients: Development and preliminary validation of a new instrument to improve the quality of written health care information. *Health Expect* 7: 165–175. https://doi.org/10.1111/j.1369-7625.2004.00273.x
33. Minervation. The LIDA Instrument: Minervation validation instrument for health care web sites—Full version (1.2) containing instructions, 2007. Available from: http://www.minervation.com/wp-content/uploads/2011/04/Minervation-LIDA-instrument-v1-2.pdf. Accessed 11 Jun 2022.
34. Martins EN, Morse LS (2005) Evaluation of internet websites about retinopathy of prematurity patient education. *Br J Ophthalmol* 89: 565–568. https://doi.org/10.1136/bjo.2004.055111
35. WebFX. Readability test: Quick and easy way to test the readability of your work. Available from: https://www.webfx.com/tools/read-able/. Accessed 11 Jun 2022.
36. Brown DM (2021) Simple Measure of Gobbledygook (SMOG) formula for calculating readability. Network of the National Library of Medicine/NNLM Region 4. Available from: https://news.nnlm.gov/region_4/simple-measure-of-gobbledygook-smog-formula-for-calculating-readability/. Accessed 11 Jun 2022.

37. Institute of Mental Health. Media release: Latest nationwide study shows 1 in 7 people in Singapore has experienced a mental disorder in their lifetime. Institute of Mental Health, 2018. Available from: https://www.imh.com.sg/Newsroom/News-Releases/Documents/SMHS%202016_Media%20Release_FINAL_web%20upload.pdf. Accessed 10 Jun 2022.
38. American Psychiatric Association. Patients and families. Available from: https://www.psychiatry.org/patients-families. Accessed 11 Jun 2022.
39. National Institute of Mental Health. Health topics. Available from: https://www.nimh.nih.gov/health/topics. Accessed 11 Jun 2022.
40. Medline Plus. Mental health and behaviour. National Library of Medicine. Available from: https://medlineplus.gov/mentalhealthandbehavior.html. Accessed 11 Jun 2022.
41. World Health Organization. Fact sheets. Available from: https://www.who.int/news-room/fact-sheets. Accessed 11 Jun 2022.
42. Centers for Disease Control and Prevention. About mental health. U.S. Department of Health & Human Services. Available from: https://www.cdc.gov/mentalhealth/learn/index.htm. Accessed 11 Jun 2022.
43. Mayo Clinic. Mental illness. Available from: https://www.mayoclinic.org/diseases-conditions/mental-illness/symptoms-causes/syc-20374968. Accessed 11 Jun 2022.
44. Cleveland Clinic. Mental health, 2022. Available from: https://health.clevelandclinic.org/topics/health-a-z/mental-health/. Accessed 11 Jun 2022.
45. National Alliance on Mental Illness. Frequently asked questions. Available from: https://www.nami.org/FAQ. Accessed 11 Jun 2022.
46. Anxiety and Depression Association of America. Understand anxiety and depression: Take the first step—Understand the facts. Available from: https://adaa.org/. Accessed 10 Jun 2022.
47. International OCD Foundation. What is OCD? Available from: https://iocdf.org/about-ocd/. Accessed 10 Jun 2022.
48. NP Digital. Discover what people are asking about…. AnswerThePublic. Available from: https://answerthepublic.com/. Accessed 11 Jun 2022.
49. Substance Abuse and Mental Health Services Administration (2016) DSM-5 Child Mental Disorder Classification, In: *DSM-5 Changes: Implications for Child Serious Emotional Disturbance [Internet]*, Rockville, MD. Available from: https://www.ncbi.nlm.nih.gov/books/NBK519712/. Accessed 11 Jun 2022.
50. Ministry of Health. MOH Clinical Practice Guidelines—Depression, 2012. Available from: https://www.moh.gov.sg/docs/librariesprovider4/guidelines/depression-cpg_r14_final.pdf. Accessed 11 Jun 2022.
51. Koo TK, Li MY (2016) A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 15: 155–163. https://doi.org/10.1016/j.jcm.2016.02.012
52. Seifert D (2017) Samsung's new virtual assistant will make using your phone easier. The Verge. Available from: https://www.theverge.com/2017/3/20/14973742/samsung-bixby-virtual-assistant-announced-galaxy-s8. Accessed 11 Jun 2022.
53. Kocaballi AB, Quiroz JC, Rezazadegan D, et al. (2020) Responses of conversational agents to health and lifestyle prompts: Investigation of appropriateness and presentation structures. *J Med Internet Res* 22: e15823. https://doi.org/10.2196/15823

54. Farooqui A (2020) Samsung's old S Voice assistant is being discontinued on June 1, 2020. Sammobile. Available from: https://www.sammobile.com/news/samsungs-s-voice-assistant-being-discontinued-june-1-2020/. Accessed 3 Nov 2022.
55. Cheng C, Dunn M (2015) Health literacy and the Internet: A study on the readability of Australian online health information. *Aust N Z J Public Health* 39: 309–314. https://doi.org/10.1111/1753-6405.12341
56. Vaidyam AN, Linggonegoro D, Torous J (2021) Changes to the psychiatric chatbot landscape: A systematic review of conversational agents in serious mental illness: Changements du paysage psychiatrique des chatbots: une revue systematique des agents conversationnels dans la maladie mentale serieuse. *Can J Psychiatry* 66: 339–348. https://doi.org/10.1177/0706743720966429
57. Martinez-Miranda J, Martinez A, Ramos R, et al. (2019) Assessment of users' acceptability of a mobile-based embodied conversational agent for the prevention and detection of suicidal behaviour. *J Med Syst* 43: 246. https://doi.org/10.1007/s10916-019-1387-1
58. Sander L, Kuhn C, Bengel J, et al. (2019) Responses of German-speaking voice assistants to questions about health issues. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* 62: 970–980. https://doi.org/10.1007/s00103-019-02979-x