Title: Quality assessment of digital voice assistants on information provided in eating disorders and co-existing depression

Type: Original article

Running Title: VA quality in eating disorders and co-existing depression

Authors:

Meryl KOH[1] (e0042143@u.nus.edu)

Qihuang XIE [1] (xie.qihuang123@gmail.com)

Li Lian WONG[1] (phawll@nus.edu.sg)

Kevin Yi-Lwern YAP[2] (k.yap@latrobe.edu.au, ORCID: 0000-0001-7322-4396)

Affiliations:

[1] Department of Pharmacy, Faculty of Science, National University of Singapore, Block S4A, Level 2, 18 Science Drive 4, Singapore 117543

[2] Department of Public Health, School of Psychology and Public Health, La Trobe University, Melbourne (Bundoora), Victoria 3086, Australia

Corresponding Author:

Kevin Yi-Lwern Yap, PhD, SRPharmS, CBDSA

Senior Lecturer in Public Health (Digital Health)

Department of Public Health, School of Psychology and Public Health

La Trobe University, Melbourne (Bundoora), VIC 3086, Australia

Email: kevinyap.ehealth@gmail.com; k.yap@latrobe.edu.au

Tel: +61 (0)3 9479 6068

Conflicts of Interest: No competing financial interests exist.

Financial Disclosure: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.
Conflicts of Interest: No competing financial interests exist.

Contributorship: KY and LLW conceived and designed the study. MK conducted the study. MK and QX analyzed the results. MK, QX, LLW and KY wrote the manuscript. QX and KY revised the manuscript. All authors agreed to the publication of the manuscript.

Keywords: Digital voice assistants; Eating disorders and co-existing depression; Quality assessment; Reliability; Accuracy; Comprehensiveness

Abstract

Background

Patients with eating disorders and co-existing depression often rely on the Internet, and digital voice assistants (VAs) as methods of searching for health-related information regarding their conditions. However, the information quality provided by VAs is questionable. We evaluated the quality of information on eating disorders and co-existing depression from 4 commonly-used VAs (Google Assistant, Siri, Cortana and Bixby) and
Google Search.

Methods

Forty-four questions on eating disorders and co-existing depression were evaluated. Their responses were evaluated by two raters for accuracy (score: 2), source expertise (score: 1), underlying references cited (score: 2) and comprehensiveness (score: 2) using a scoring matrix (score: 8). Descriptive statistics and odds ratios were used for analysis. Cohen Kappa was used to measure inter-rater agreement.

Results

Cortana (mean=5.23±2.01) and Siri (mean=4.42±2.50) scored the highest and lowest for overall quality respectively. Cortana (41/44, 93.2%) and Bixby (32/44, 72.7%) provided the most and least number of relevant sources (41/44, 93.2% versus 32/44, 72.7%, $p<0.0001$), and the highest and lowest mean accuracy scores (1.82±0.54 versus 1.43±0.89, $p=0.0016$) respectively. Bixby was the most reliable in terms of source expertise (mean=0.43±0.50) and underlying references cited (mean=0.93±0.50). Google Search scored the highest in terms of comprehensiveness, while Siri performed the worst for comprehensiveness, source expertise and underlying references cited.

Conclusion

Most of the sources provided by the VAs were accurate and comprehensive, but not as reliable. Patients should be cautious when using VAs to search for information on eating disorders and co-existing depression.

Introduction

Eating disorders are characterised by severe disturbances to an individual's eating behaviour.[1] Patients with eating disorders usually present with a preoccupation with body weight and restriction in food intake.[2] The physiological changes secondary to dieting and undernourishment are hypothesised to contribute to the development of depressive symptoms.[3, 4] As a result, patients with eating disorders are often diagnosed with co-existing depression.[5] Individuals suffering from eating disorders and depression tend to conceal their symptoms and are often unwilling to seek help due to barriers like stigma and shame.[6] Consequently, the Internet which provides anonymity and privacy, is commonly used by patients with mental health illnesses to seek information regarding their conditions.

In addition to obtaining information through tradition web searches, digital voice assistants (VAs) like Siri, Google Assistant, and Cortana have become widely used as a means to perform voice-activated Internet searches for information.[7] As of 2019, 3.25 billion of the world's population used VAs.[8] A survey by Microsoft showed that 72% of respondents in the United States (US) reported using VAs.[9] Voice searches present added benefits over manual text searches such as increased convenience and speed.[7] According to Google, 20% of queries on the Google mobile app were shown to be voice searches.[10] Given that patients with eating disorders are likely to search for information online, the increase in popularity and accessibility of VAs mean that VAs may potentially be used as information searching tools by these patients. However, a recent study has shown that the accuracy of information provided by VAs has declined over the past few years.[11] The difficulty to distinguish accurate and reliable health information provided by VAs can potentially pose serious health risks to these patients.[12] In a study which investigated how medical and medication-related information provided by the VAs affected users' health behaviors, it was reported that 29.2% of the reported health behaviors could potentially cause harm, out of which 16.1% could have resulted in death.[13] Patients with eating disorders and co-existing depression who follow the inaccurate and unreliable information provided by the VAs may potentially delay seeking medical attention, undertake risky health actions and ultimately lead to poor prognosis.[14]

To date, studies that have been conducted on the information provided by VAs have concentrated on other health-related topics such as vaccines, smoking cessation advice, and sexual health advice.[15-17] However, studies on their abilities to provide quality information regarding eating disorders and co-existing depression are limited. In addition, studies that had evaluated the quality of VAs on mental health had concentrated on sucide-related queries. For example, in a US study that assessed the quality of VA responses on suicide-related queries, it was reported that VAs had failed to recognise and respond adequately to those queries and also failed to redirect the user to a suicide prevention helpline.[18] Therefore, our study aimed to evaluate the performance of four VAs (Apple Siri, Google Assistant, Microsoft Cortana and Samsung Bixby) in relation to the accuracy, reliability and comprehensiveness of the

information provided about eating disorders and co-existing depression. A secondary objective was to compare the quality of their responses with Google Search.

Methods

Definition of quality

The overall quality of the VAs was defined in terms of the accuracy, reliability and comprehensiveness of the information sources provided. Accuracy was defined as the degree of concordance of the VA responses with a predefined answer key compiled from reputable health organizations, such as the United States National Institute of Mental Health (NIMH), American Psychiatric Association (APA), United Kingdom National Health Service, medical journals from PubMed, and other biomedical databases such as MedlinePlus and UpToDate (**Appendix 1**). Reliability was determined by the source expertise and the underlying references cited in the sources (e.g. expert opinions or evidence-based guidelines). Comprehensiveness was determined by the percentage coverage of the number of points provided by the VAs to the total number of points provided by the answer key.

Selection of questions

A total of 44 questions (**Appendix 1**) were developed on the topics of eating disorders and co-existing depression. These questions were adapted from patient education resources from the NIMH, expert question and answer pages from the APA and a patient education brochure on depression from the Singapore Ministry of Health. Additionally, some questions were generated using AnswerThePublic.com, which was a website that provided common search phrases from the suggested autocomplete searches of both Google and Bing. Hence, AnswerThePublic.com offered a greater database compared to other tools like Google Trends, which only provided popular search queries from Google search.

Data collection

Three devices were used to access the VAs. An iPhone 6S (iOS 13.3.1) was used to access Siri and a Windows 10 laptop for Microsoft Cortana, while Bixby and Google Assistant were accessed using a Samsung Galaxy S8 smartphone (Android version 9). Google Search was done on a Windows 10 laptop using a private browsing window. The location services of the devices were disabled to prevent the results from being influenced by location. Furthermore, the search histories of the voice assistants were cleared before data collection. The data collection was conducted by one female (MK, rater 1) and one male rater (SK, rater 2) on six separate days between 2nd and 7th March 2020 in Singapore. Independently, the two raters asked the 44 questions in English to each VA and the same questions were manually entered into Google Search. If the VA provided a source that was relevant to the question when asked

on the first attempt, that source would be used for evaluation. If not, a second attempt would be made. If the VA failed to provide a relevant source again, evaluation for that question would be voided. The first non-advertisement source that each rater obtained was used for evaluation and each rated the quality of the source independently.

Assessing the quality of the information provided by VAs and Google Search

A scoring matrix was adapted from the quality rubrics developed by Alagha and Helbeing on assessing the accuracy and reliability of VAs.[15] Additionally, our matrix evaluated the comprehensiveness of the source provided by the VAs (**Figure 1**). The sources provided by the VAs and Google Search would first be evaluated for their relevance. If the VAs and Google Search provided a source that was irrelevant to the question asked, the evaluation process for that question would be terminated, resulting in a score of 0. However, if the source provided was relevant to the question, the rater will continue to evaluate for its accuracy, reliability and comprehensiveness.

Accuracy of the sources were classified into three categories: correct (score: 2), partially correct (score: 1) and incorrect (score: 0). In terms of reliability, a source was considered an expert source (score: 1) if it was from government, university, hospital, non-profit health organizations, or medical journals, whereas, sources from crowd-sourced sites, commercial sites, or non-health sites were considered non-expert sources (score: 0). Additionally, the underlying references cited in the source were identified as using evidence-based guidelines (score: 2) or expert opinions (score: 1). If no underlying reference was stated, no point was awarded (**Figure 1**). Therefore, the reliability score ranged from 0 to 3. Lastly, a source was considered comprehensive if it managed to obtain 68-100% of the answer key (score: 2), less comprehensive if it provided 34-67% of the answer key (score: 1), and not comprehensive if it provided less than 33% of the answer key (score: 0). Overall, the possible scores for a given source ranged from 0 to 8.

Data analysis

All results were analyzed using SPSS version 25. Descriptive statistics were used to describe the overall qualities of the VAs and Google Search, the quality of information sources in terms accuracy, reliability and comprehensiveness, and the difference in sources obtained by the two raters. The differences in overall quality scores and the scores for accuracy, reliability and comprehensiveness were compared using Kruskal Wallis H-test. For statistically significant results, Mann-Whitney U-test with Bonferroni adjustments was used as a post-hoc test. Inter-rater reliability was measured using Cohen's kappa. Additionally, Mann-Whitney U-test was used to compare the differences in scores between male and female raters for each VA and Google Search. Odds ratios were used to compare the frequency of obtaining the highest-scoring sources for VAs and Google Search.

Results

Overall, Cortana scored the highest for the number of relevant sources (mean=0.93±0.25) while Bixby scored the lowest (mean=0.73±0.45), p<0.0001. For Siri, Google Assistant, and Google search, more than 80% of the sources provided were relevant to the question (**Table 1**). In general, the overall quality scores and the scores for each quality parameter for the VAs provided by the 2 raters were similar **(Table 1)**. In terms of the five quality parameters of the VAs, males seemed to provide slightly higher scores for Siri (3/5 parameters) and Bixby (4/5 parameters), while females seemed to score Cortana (3/5 parameters) and Google Assistant (5/5 parameters) better. In general, males seemed to provide higher scores for the VAs' abilities to provide comprehensive and expert sources, while females scored higher for the relevance and accuracy of the sources, and the underlying references cited. However, the differences in scores were not statistically significant. When the scores given by both raters for all questions were compared, the Cohen's kappa was 0.651 (p<0.0001), which showed substantial agreement between the two raters. The kappa statistic might have been lowered by variation in answers offered by the tools. When only questions that resulted in the same sources provided to both raters were included, the kappa value increased to 0.827 (p<0.0001).

Overall performance of the tools

Cortana (mean=5.23±2.01) and Google search (mean=5.01±2.51) obtained the highest overall quality scores. Siri, Bixby and Google Assistant acquired lower overall quality scores (**Table 1**). However, the differences in the overall quality scores were not statistically significant (p=0.24). When stratified according to gender, the female rater gave higher overall quality scores for Cortana (mean=5.27±1.87 versus 5.18±2.16, p=0.94), Google Assistant (mean=5.02±2.49 versus 4.70±2.68, p=0.64) and Google Search (mean=5.05±2.51 versus 4.98±2.54, p=0.92). On the other hand, the male rater provided higher overall quality scores for Siri (mean=4.52±2.57 versus 4.32±2.46, p=0.61) and Bixby (mean=4.93±2.93 versus 4.70±3.35, p=0.80). However, the differences were not statistically significant **(Table 1)**. Despite obtaining the lowest overall quality score, Bixby provided the most number of highest-scoring sources (24/44, 54.5%). On the other hand, Cortana, which had a higher overall quality score than Bixby, provided fewer highest-scoring sources (18/44, 40.9%). In contrast, Siri performed the worst in terms of the overall quality score (mean=4.42±2.50) and the frequency of providing sources with the highest score (17/44, 38.6%). When comparing the ability of VAs and Google Search to provide sources with the highest score, no statistically significant differences were observed. (**Table 2**).

Reliability of sources

The most cited source for Siri was Wikipedia.org (6/38, 15.8%) and this VA provided the least number of expert sources (9/38, 23.7%). At the other end of the spectrum, Bixby cited Mayoclinic.org most of the time (7/32, 21.9%) and more than half of its answers were from expert sources (19/32, 59.4%). On the other hand, Cortana cited WebMD.com most of the time (14.6%, 6/41) and Aware.org.sg was cited by Google Assistant (8/37, 21.6%) and Google Search (9/38, 23.7%) most of the time (**Table 1**).

Overall, VAs and Google Search scored poorly for their source expertise and underlying references cited in the source (**Table 1**). Among the VAs and Google Search, Bixby scored the highest for source expertise (mean=0.43±0.50), which was significantly higher than Siri (mean=0.21±0.41), which scored the lowest (p=0.0012). Furthermore, Bixby scored the highest for the underlying reference cited (mean=0.93±0.83) while Siri scored the lowest (mean=0.74±0.78).

Accuracy and Comprehensiveness of sources

Google Search and the VAs managed to obtain a mean score of above 1 for accuracy and comprehensiveness of the sources (**Table 1**). Among Google Search and the VAs, Cortana (mean=1.82±0.54) scored significantly higher in accuracy compared to Bixby (mean=1.43±0.89, p=0.0016). For comprehensiveness, Cortana (mean=1.35±0.80) scored slightly lower than Google Search (mean=1.36±0.82), while Siri scored the lowest (mean=1.02±0.88).

Differences in search results

Of the 176 questions (44 questions for each VA), the two raters obtained different sources for 53 questions (30.1%). Siri (17/53, 32.1%), Cortana (17/53, 32.1%) and Bixby (15/53, 28.3%) collectively accounted for most of the differences (45/53, 84.9%). Three different reasons were identified to describe the differences that were observed. Majority of the differences (41/53, 77.4%) were due to three of the VAs' failure (Siri, Cortana and Bixby) to recognize the questions posed by the raters accurately. (**Figure 2**). Occasionally, it was observed that the VAs provided different sources, even though they managed to recognize the questions posed by the raters accurately (10/53, 18.9%). For the remaining question (1/53, 1.9%), a system interruption occurred whereby the VA (i.e. Cortana) started to process the words being said before the sentence was completed by the two raters. Among the VAs, Google Assistant was the most consistent in the sources provided to both raters – it had the least number of questions in which different sources were provided to both raters (8/53, 15.1%).

Compared to the VAs, Google Search was significantly more consistent in the sources that it provided to both raters (38 questions, 86.4%, OR=2.7, p=0.027).

Discussion

In this study, the performance of four VAs (Apple Siri, Google Assistant, Samsung Bixby and Microsoft Cortana) in providing quality information sources regarding eating disorders and co-existing depression was studied and compared to Google Search. Cortana achieved the highest overall quality score and scored the highest in terms of accuracy. Despite having the lowest overall quality score, Bixby still managed to provide the most number of highestscoring sources and was also able to provide the most reliable sources. Google Search scored the highest for comprehensiveness, but among the VAs, the sources provided by Cortana were the most comprehensive. In contrast, Siri performed the worst in overall quality scores, comprehensiveness, and reliability.

Interestingly, unlike previous studies which found that Google Search provided better quality health information than the VAs,[16, 17] our study showed that some VAs were able to perform better than Google Search in some aspects, such as the overall quality score, accuracy, and reliability. The difference between our study and those studies was that they had only compared two VAs (Siri and Google Assistant) with Google Search, while ours compared among four VAs. Furthermore, the health domains that were evaluated in all of the studies were different (i.e. smoking cessation and sexual health versus mental health disorders and co-existing depression), therefore the VAs might have provided different varieties and qualities of sources for the different health domains based on their artificial intelligence algorithm. Evaluating a greater variety of VAs on a wider variation of health conditions/domains in future studies might provide further insight to their consistency and quality of the information sources provided.

Although Cortana and Siri provided lesser expert sources and scored lower than Bixby in terms of their underlying references cited, the sources provided by Cortana and Siri still scored higher than Bixby for accuracy. Even though Cortana and Siri provided sources from WebMD.com and Wikipedia.org most of the time, the information provided on these sites had adopted a major proportion of their references from reputable sources, as suggested by a study that investigated the quality of information provided by Wikipedia, which found that 56% of the references cited were from reputable sources.[19] In general, the VAs and Google

Search performed poorly in terms of reliability of the sources provided. However, the accuracy and the comprehensiveness of the information on eating disorders and co-existing depression were not compromised.

Regarding the consistency of responses provided, the two raters obtained different information sources for one-third of the questions. Failure to accurately recognize the questions posed by the raters accounted for most of the differences. For example, Bixby recognised "Will I die from Bulimia nervosa" incorrectly as "Will I die from ballima Navassa" which led to a response of "I didn't understand that". These natural language processing errors made up the majority of the obstacles encountered when using voice user interfaces.[20] This failure to accurately recognize the questions and provide a relevant source might have led to Bixby getting a lower overall quality score despite it being able to provide the most number of highest scoring answers.

The recognition errors could have also been influenced by the accents of the raters.[21] Cortana and Bixby were not specifically designed to recognize Singaporean accents,[22, 23] which could have contributed to their recognition errors. On the other hand, Google Assistant provides support for Singaporean English, which may explain its lower frequency of recognition errors. However, Siri also supports Singapore English,[24] but its frequency of recognition errors leading to inconsistent results was still high. Other factors, such as the tone or pitch of the voice, could also have influenced the accuracy of voice recognition.

Besides recognition errors, system interruptions were another type of voice input error that caused the failed searches.[25] System interruptions might have been caused by a short pause when the query was being vocalised, and this could be affected by how the rater had spoken.[25] Our study observed only one system interruption whereby Cortana recognized "What is orthorexia nervosa" as "What is orthorexia". In this case, the definition of orthorexia was provided which was still relevant, but the answer had a lower score compared to when the question was correctly recognized.

Surprisingly, the VAs provided different responses even when there were no voice input errors. As the voice searches for the two raters were not conducted on the same day, we postulate that the variations in the sources obtained could have been due to updates in the search algorithms over time. Furthermore, as the voice search histories were not cleared after each question, the sources provided by subsequent voice searches could have been influenced by the results from previous voice searches.

Another notable finding was that the two raters obtained different sources from Google Search despite using an Incognito window. Google Search might have modified the search results according to the rater's activity in the Incognito browsing session during data collection,[26] since the raters did not reopen a new Incognito browser after every search. Furthermore, the constant updating of its search algorithms by Google would imply that a specific query might not return the same search results on different occasions.[27] Nonetheless,

Google Search provided more consistent answers compared to the VAs and the differences in answers did not greatly affect its quality of answers.

Limitations and Future Work

The VAs used in this study were limited to only Google Assistant, Siri, Cortana and Bixby. Therefore, the results presented in this study cannot be extrapolated to predict the performance of other VAs. Future studies should include other VAs such as Google Home, Amazon Alexa and Amazon Echo. Furthermore, this study only involved two raters and hence might not be representative or externally valid to the general population. A larger study involving more participants can be conducted in the future. In addition, our study only evaluated the first non-advertisement weblink provided. As people tend to look through multiple sources when finding health information, future studies can include a more thorough evaluation of the health information provided by evaluating the first page of search results. Lastly, the search histories of the voice assistants were not deleted after every question. We also did not reopen a new Incognito window after every Google search, which could have influenced the results. Therefore, future studies can ensure these additional steps are performed.

Conclusion

Overall, Cortana provided the best quality information sources on eating disorders and coexisting depression, followed by Google Search. Siri performed the poorest for overall quality. VAs and Google Search were generally able to provide accurate and comprehensive information sources. Cortana scored the highest for accuracy, while Google Search scored the highest for comprehensiveness. However, the reliability of the sources provided by the VAs and Google Search should be improved. Patients should be cautious when using VAs to search for information on eating disorders and co-existing depression and verify the information that they obtained from VAs with credible health organizations. Furthermore, healthcare professionals should educate these patients on their medical conditions, so that they can use the information that they obtain from VAs and the Google Search to supplement their knowledge. As VA technology becomes more advanced, developers should enhance their search algorithms to improve the voice recognition abilities of VAs, so that more reliable sources for mental health disorders, such as eating disorders and co-existing depression, can be provided to patients and healthcare professionals.

Acknowledgement: We would like to acknowledge Mr Shaun Koh for helping out in this study as the second rater of the voice assistants.

References

1. National Health Service: Eating disorders; 2018 [Internet]. Available from: https://www.nhs.uk/conditions/eating-disorders/ [cited 2020, Feb 7].
2. American Psychiatric Association. Diagnostic and statistical manual of mental disorders (DSM-5). Fifth ed. Arlington, VA: American Psychiatric Association; 2013.
3. Mattar L, Huas C, Duclos J, Apfel A, Godart N. Relationship between malnutrition and depression or anxiety in anorexia nervosa: A critical review of the literature. J Affect Disord 2011;132(3):311-8.
4. Garner D. The effects of starvation on behavior: Implications for dieting and eating disorders. Healthy Weight Journal 1998;12:68-72.
5. Blinder BJ, Cumella EJ, Sanathara VA. Psychiatric comorbidities of female inpatients with eating disorders. Psychosom Med 2006;68(3):454-62. [PMID: 16738079]
6. Becker AE, Hadley Arrindell A, Perloe A, Fay K, Striegel-Moore RH. A qualitative study of perceived social barriers to care for eating disorders: Perspectives from ethnically diverse health care consumers. Int J Eat Disord 2010;43(7):633-47. [PMID: 19806607]
7. Enge E: Mobile voice usage trends in 2019; 2019 [Internet]. Available from: https://www.perficientdigital.com/insights/our-research/voice-usage-trends [cited 2020, Mar 12].
8. Statista: Number of digital voice assistants in use worldwide from 2019 to 2023; 2020 [Internet]. Available from: https://www.statista.com/statistics/973815/worldwide-digitalvoice-assistant-in-use/ [cited 2020, Mar 22].
9. Olson C, Kemery K: Voice report: From answers to action: Customer adoption of voice technology and digital assistants; 2019 [Internet]. Available from: https://advertiseonbingblob.azureedge.net/blob/bingads/media/insight/whitepapers/2019/04%20apr/voicereport/bingads_2019_voicereport.pdf [cited 2020, Feb 27].
10. Lawson M: 4 Things you need to know about the future of marketing; 2017 [Internet]. Available from: https://www.thinkwithgoogle.com/intl/en-apac/trends-and-insights/4-thingsfuture-marketing/ [cited 2020, Mar 12].
11. Enge E: Rating the smarts of the digital personal assistants in 2019; 2019 [Internet]. Available from: https://www.perficientdigital.com/insights/our-research/digital-personalassistants-study [cited 2020, Mar 25].
12. Morahan-Martin JM. How internet users find, evaluate, and use online health information: A cross-cultural review. Cyberpsychol Behav 2004;7(5):497-510. [PMID: 15667044]
13. Bickmore TW, Trinh H, Olafsson S, O'Leary TK, Asadi R, Rickles NM, et al. Patient and consumer safety risks when using conversational assistants for medical information: An observational study of Siri, Alexa, and Google Assistant. J Med Internet Res

2018;20(9):e11510. [PMID: 30181110]

14. Rowe E. Early detection of eating disorders in general practice. Aust Fam Physician 2017;46(11):833-8. [PMID: 29101919]

15. Alagha EC, Helbing RR. 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 2019;26(1):e100075. [PMID: 31767629]

16. Boyd M, Wilson N. Just ask Siri? A pilot study comparing smartphone digital assistants and laptop Google searches for smoking cessation advice. PLoS One 2018;13(3):e0194811. [PMID: 29590168]

17. Wilson N, MacDonald EJ, Mansoor OD, Morgan J. In bed with Siri and Google Assistant: A comparison of sexual health advice. BMJ 2017;359:j5635. [PMID: 29237603]

18. Miner AS, Milstein A, Schueller S, Hegde R, Mangurian C, Linos E. Smartphonebased conversational agents and responses to questions about mental health, interpersonal violence, and physical health. JAMA Intern Med 2016;176(5):619-25. [PMID: 26974260]

19. Haigh CA. Wikipedia as an evidence source for nursing and healthcare students. Nurse Educ Today 2011;31(2):135-9. [PMID: 20646799]

20. Myers C, Furqan A, Nebolsky J, Caro K, Zhu J. Patterns for how users overcome obstacles in voice user interfaces. Conference on Human Factors in Computing; April 2018; Montreal, Canada 2018. p. 1-7.

21. Palanica A, Thommandram A, Lee A, Li M, Fossat Y. Do you understand the words that are comin outta my mouth? Voice assistant comprehension of medication names. NPJ Digit Med 2019;2:55. [PMID: 31304401]

22. Microsoft: Cortana's regions and languages; 2019 [Internet]. Available from: https://support.microsoft.com/en-us/help/4026948/cortanas-regions-and-languages [cited 2020, Mar 17].

23. Samsung: Bixby [Internet]. Available from: https://www.samsung.com/sg/apps/bixby/ [cited 2020, Mar 17].

24. Apple: iOS and iPadOS feature availability [Internet]. Available from:

https://www.apple.com/sg/ios/feature-availability/ [cited 2020, Mar 17].

25. Jiang J, Jeng W, He D. How do users respond to voice input errors?: Lexical and phonetic query reformulation in voice search. 36th International ACM SIGIR Conference on Research & Development in Information Retrieval; Dublin, Ireland: ACM; 2013. p. 143-52. 26. Google Chrome Help: How private browsing works in Chrome [Internet]. Available from: https://support.google.com/chrome/answer/7440301 [cited 2020, Mar 17].

27. Google Search: Rigorous testing [Internet]. Available from:

https://www.google.com/search/howsearchworks/mission/users/ [cited 2020, Mar 18].

Figures

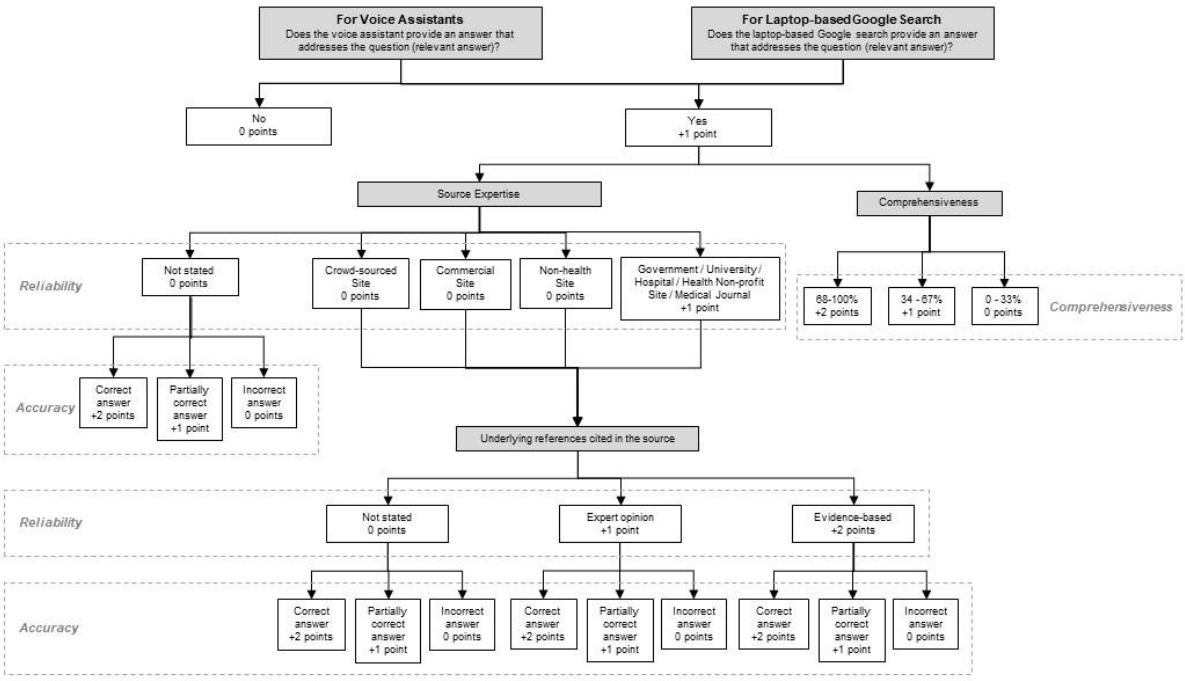

Figure 1. Rubric for assessing the quality of the responses from voice assistants and Google search.

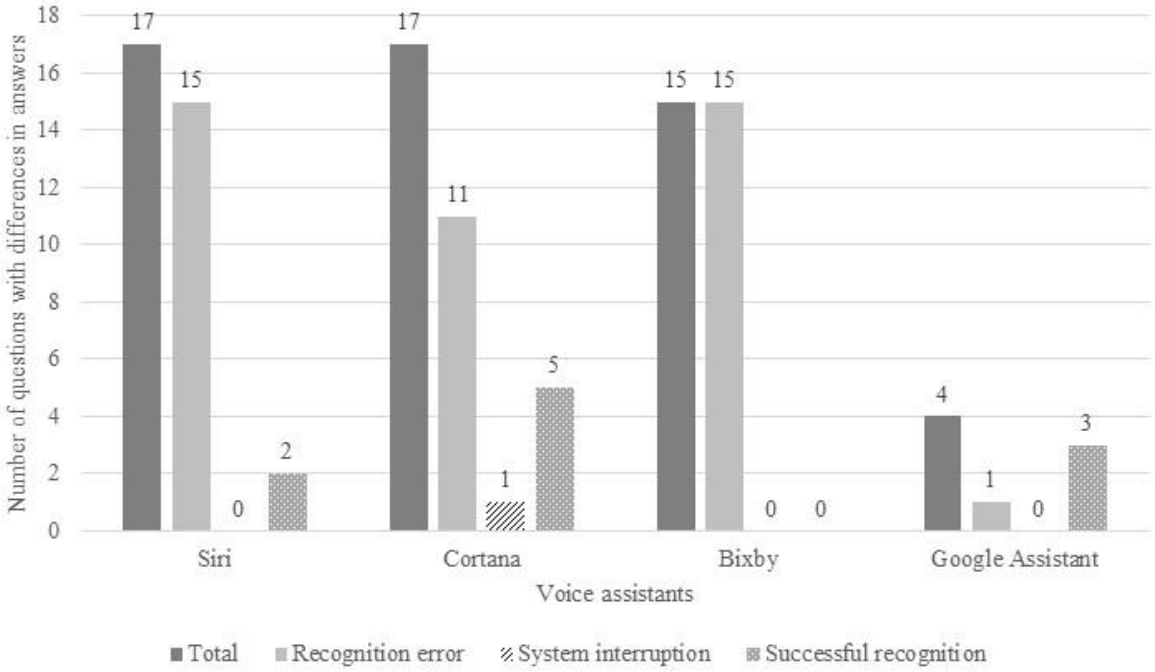

Figure 2. Total number of questions with different responses for each voice assistant and the number of questions with different responses categorised into recognition errors, system interruptions and successful recognition.

Table 1. Summary of performance of the VAs and Google Search.

Parameters	Search Tools					
	Siri	Cortana	Bixby	Google Assistant	Google search	pvalue
	Mean Score (SD), n=44					
Overall quality	4.42 (2.50)	5.23 (2.01)	4.82 (3.13)	4.86 (2.58)	5.01 (2.51)	0.24
Female	4.32 (2.46)	5.27 (1.87)	4.70 (3.35)	5.02 (2.49)	5.05 (2.51)	
Male	4.52 (2.57)	5.18 (2.16)	4.93 (2.93)	4.70 (2.68)	4.98 (2.54)	
Quality parameters:						
Relevance of the sources provided	0.85 (0.36)	0.93 (0.25) *	0.73 (0.45) *	0.84 (0.37)	0.86 (0.35)	0.006 ~~59~~
Female	0.86 (0.35)	0.95 (0.21)	0.68 (0.47)	0.86 (0.35)	0.86 (0.35)	
Male	0.84 (0.37)	0.91 (0.29)	0.77 (0.42)	0.82 (0.39)	0.86 (0.35)	
Accuracy of the sources	1.60 (0.77)	1.82 (0.54) *	1.43 (0.89) *	1.66 (0.74)	1.72 (0.69)	0.017
Female	1.61 (0.75)	1.86 (0.46)	1.36 (0.94)	1.70 (0.70)	1.73 (0.69)	
Male	1.59 (0.79)	1.77 (0.61)	1.50 (0.85)	1.61 (0.78)	1.70 (0.70)	
Comprehensiveness of the sources	1.02 (0.88)	1.35 (0.80)	1.26 (0.92)	1.31 (0.84)	1.36 (0.82)	0.059
Female	0.95 (0.89)	1.30 (0.85)	1.25 (0.94)	1.34 (0.81)	1.36 (0.81)	
Male	1.09 (0.88)	1.41 (0.76)	1.27 (0.90)	1.27 (0.87)	1.36 (0.84)	
Source Expertise	0.21 (0.41) *	0.30 (0.46)	0.43 (0.50) *	0.28 (0.45)	0.32 (0.47)	0.026
Female	0.18 (0.39)	0.27 (0.45)	0.41 (0.50)	0.30 (0.46)	0.34 (0.48)	
Male	0.23 (0.42)	0.32 (0.47)	0.45 (0.50)	0.27 (0.45)	0.30 (0.46)	
Underlying references cited in the sources	0.74 (0.78)	0.83 (0.73)	0.93 (0.83)	0.77 (0.80)	0.75 (0.79)	0.47
Female	0.70 (0.80)	0.89 (0.72)	1.00 (0.86)	0.82 (0.82)	0.75 (0.78)	
Male	0.77 (0.77)	0.77 (0.74)	0.86 (0.80)	0.73 (0.79)	0.75 (0.81)	
	Number of responses (%)					

Frequency of providing an answer with the highest score (%)	17/44[1] (38.6%)	18/44 (40.9%)	24/44 (54.5%)	20/44[1] (45.5%)	22/44[1] (50.0%)	NA
Female	17/44 (38.6%)	18/44 (40.9%)	24/44 (54.5%)	20/44 (45.5%)	21/44 (47.7%)	
Male	16/44 (36.4%)	18/44 (40.9%)	24/44 (54.5%)	19/44 (43.2%)	22/44 (50.0%)	
Frequency of providing a	38/44[1] (86.4%)	41/44 (93.2%)	32/44 (72.7%)	37/44 (84.1%)	38/44 (86.4%)	NA
relevant source (%)						
Female	38/44 (86.4%)	42/44 (95.5%)	30/44 (68.2%)	38/44 (86.4%)	38/44 (86.4%)	
Male	37/44 (84.1%)	40/44 (90.9%)	34/44 (77.3%)	36/44 (81.8%)	38/44 (86.4%)	
Most cited source (%)	Wikipedia.org	WebMD.com	Mayoclinic.org	Aware.org.sg	Aware.org.sg	NA
Frequency of the most cited source[2] (%)	6/38 (15.8%)	6/41 (14.6%)	7/32[1] (21.9%)	8/37 (21.6%)	9/38 (23.7%)	NA
Female	6/38 (15.8%)	5/42 (11.9%)	7/30 (23.3%)	8/38 (21.1%)	9/38 (23.7%)	
Male	6/38 (15.8%)	7/40 (17.5%)	6/34 (17.6%)	8/36 (22.2%)	9/38 (23.7%)	
Frequency of obtaining an answer from an expert source[2] (%)	9/38 (23.7%)	13/41 (31.7%)	19/32 (59.4%)	13/37[1] (35.1%)	14/38 (36.8%)	NA
Female	8/38 (21.1%)	12/42 (28.6%)	18/30 (60.0%)	13/38 (34.2%)	15/38 (39.5%)	
Male	10/37 (27.0%)	14/40 (35.0%)	20/34 (58.8%)	12/36 (33.3%)	13/38 (34.2%)	

[1] Mean of two raters rounded up to the whole number

[2] The total number of relevant sources provided by each search tool was used in the calculating the percentage *

Mann-Whitney U-test: $p<0.05$

Table 2. Odds ratios of the VAs and Google Search for providing sources with the highest scores for overall quality.

Search tools	Search tools				
	Siri	**Cortana**	**Bixby**	**Google Assistant**	**Google search**
	Odds ratios of providing sources with the highest scores for overall quality when comparing VA in column against VA in row [1] (OR, p-value)				
Siri	NA	1.1, p=0.83	1.9, p=0.14	1.3, p=0.52	1.6, p=0.28
Cortana	0.9, p=0.83	NA	1.7, p=0.20	1.2, p=0.67	1.4, p=0.39
Bixby	0.5, p=0.14	0.6, p=0.20	NA	0.7, p=0.39	0.8, p=0.67
Google Assistant	0.8, p=0.52	0.8, p=0.67	1.4, p=0.39	NA	1.2, p=0.67
Google Search	0.6, p=0.28	0.7, p=0.39	1.2, p=0.67	0.8, p=0.67	NA

[1] As an example, OR=1.1 implies that Cortana has 1.1 times higher odds of obtaining a source that scored the highest for overall quality compared to Siri.