Medical Science

AIMS Medical Science, 10(1): 80–106.
DOI: 10.3934/medsci.2023008
Received: 04 September 2022
Revised: 02 March 2023
Accepted: 14 March 2023
Published: 30 March 2023

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

Research article

Quality evaluation of digital voice assistants for diabetes management

Joy Qi En Chia[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 increasingly used to search for health information. However, the quality of information provided by DVAs is not consistent across health conditions. From our knowledge, there have been no studies that evaluated the quality of DVAs in response to diabetes-related queries. The objective of this study was to evaluate the quality of DVAs in relation to queries on diabetes management. **Materials and methods:** Seventy-four questions were posed to smartphone (Apple Siri, Google Assistant, Samsung Bixby) and non-smartphone DVAs (Amazon Alexa, Sulli the Diabetes Guru, Google Nest Mini, Microsoft Cortana), and their responses were compared to that of Internet Google Search. Questions were categorized under diagnosis, screening, management, treatment and complications of diabetes, and the impacts of COVID-19 on diabetes. The DVAs were evaluated on their technical ability, user-friendliness, reliability, comprehensiveness and accuracy of their responses. Data was analyzed using the Kruskal-Wallis and Wilcoxon rank-sum tests. Intraclass correlation coefficient was used to report inter-rater reliability. **Results:** Google Assistant (n = 69/74, 93.2%), Siri and Nest Mini (n = 64/74, 86.5% each) had the highest proportions of successful and relevant responses, in contrast to Cortana (n = 23/74, 31.1%) and Sulli (n = 10/74, 13.5%), which had the lowest successful and relevant responses. Median total scores of the smartphone DVAs (Bixby 75.3%, Google Assistant 73.3%, Siri 72.0%) were comparable to that of Google Search (70.0%, $p = 0.034$), while median total scores of non-smartphone DVAs (Nest Mini 56.9%, Alexa 52.9%, Cortana 52.5% and Sulli the Diabetes Guru 48.6%) were significantly lower ($p < 0.001$). Non-smartphone DVAs had much lower median comprehensiveness (16.7% versus 100.0%, $p < 0.001$) and reliability scores (30.8% versus 61.5%, $p < 0.001$) compared to Google Search.

Conclusions: Google Assistant, Siri and Bixby were the best-performing DVAs for answering diabetes-related queries. However, the lack of successful and relevant responses by Bixby may frustrate users, especially if they have COVID-19 related queries. All DVAs scored highly for user-friendliness, but can be improved in terms of accuracy, comprehensiveness and reliability. DVA designers are encouraged to consider features related to accuracy, comprehensiveness, reliability and user-friendliness when developing their products, so as to enhance the quality of DVAs for medical purposes, such as diabetes management.

Keywords: diabetes mellitus; digital voice assistants; quality evaluation; accuracy; comprehensiveness; reliability; technical ability; user-friendliness

1. Introduction

Digital voice assistants (DVAs) are gaining popularity after Siri, the first smartphone virtual assistant, was launched in 2011 [1]. Common smartphone DVAs include Google Assistant, Apple Siri and Samsung Bixby, while non-smartphone DVAs include smart speakers, such as Amazon Alexa and Google Nest Mini, and the laptop-based Microsoft Cortana [1]. DVAs are increasingly becoming part of peoples' everyday lives and they can assist in various tasks, such as receiving reminders, making phone calls, sending messages, finding locations (e.g. businesses, restaurants) and answering questions, among others [2]. In 2020, ~24% of Americans (~60 million people) owned a smart speaker [2]. It is forecasted that 8.4 billion people worldwide will use DVAs by 2024 [3]. A previous survey in over 5,000 consumers in the US, UK, France and Germany reported that 52% of respondents preferred using DVAs over websites or apps because it was more convenient and 48% valued its hands-free function, which allowed them flexibility to multi-task [4].

DVAs have been increasingly used for healthcare purposes. As of 2019, one-in-13 people in the US used DVAs for health-related matters, such as asking about illness symptoms, medication-related information, as well as seeking the location of healthcare providers and care facilities, among others [5]. However, the quality of health-related information provided by DVAs were inconsistent [6–8]. Studies showed that DVAs were limited in their ability to advise on lifestyle and safety-critical prompts [9], provided inconsistent and incomplete responses to mental health, physical health and interpersonal violence-related queries, and were unable to recognize emergency situations that required referral [10]. In fact, there was also a possibility that the responses provided by DVAs could result in harm or death [11].

Studies in recent years have focused on the use of DVAs for the management of chronic diseases, since DVAs can be easily accessed through personal devices and are available 24 hours a day [12]. According to the World Health Organization, diabetes is a chronic condition that has affected 422 million people worldwide [13]. In fact, the World Health Organization's Global Diabetes Compact launched at the Global Diabetes Summit 2021 aims to focus efforts on reducing the risk and burden of diabetes, as well as increase access to affordable and quality treatment internationally [14]. In Singapore, over 400,000 citizens are living with diabetes and this number is projected to increase to nearly one-fifth of its population (over one million people) by 2050 [15]. As a result, the Singapore government declared a War on Diabetes in 2016 to rally a nationwide effort to reduce the diabetes burden in its population [15]. The COVID-19 pandemic has seen an increase in the adoption of DVAs [2].

In addition to the common smartphone, smart speaker and laptop-based DVAs that provide general and health information, diabetes-specific DVAs have also been developed by third-party organizations, which can be accessed by Google Assistant and/or Alexa [16,17]. Users have the option to harness the skills of Alexa and Google Assistant to track their blood glucose levels [18–20], ask for diabetes management tips [21,22], or answer their diabetes-related questions [21,23–25]. In particular, Sulli the Diabetes Guru (which shall henceforth be referred to as Sulli), is a diabetes-specific DVA that is available for patients with Type 2 diabetes through both Google Assistant and Alexa [26]. The advancement in DVAs can potentially aid users manage their pre-diabetes and diabetes conditions through lifestyle changes and therefore, ease the healthcare burden in hospitals [27,28].

To our knowledge, there have been no studies that evaluated the use of DVAs for diabetes management. The closest research that we could find were evaluation studies on a prototyped Alexa skill and mobile applications with DVA integration for diabetes self-management [29–31]. Hence, the objective of this study is to evaluate the quality of commonly available DVAs and a diabetes-specific DVA (i.e. Sulli) in terms of their technical ability, user-friendliness, accuracy, comprehensiveness and reliability of their responses to diabetes-related queries. The purpose of evaluating the reliability and accuracy of the DVAs is to ensure that users are not misinformed by the DVA responses and can trust the information provided, while evaluation of their comprehensiveness and user-friendliness can ensure that the information is sufficiently adequate and tailored to the level of understanding of the general public. Internet Google Search (which shall be referred to as Google Search) was evaluated as a comparison due to its widespread use [32].

2. Materials and methods

2.1. Definition of quality

Following the quality definition for online drug databases described by Yap and colleagues [33], the quality of DVAs in this study was defined as the level of excellence which characterizes the DVAs in terms of its ability to satisfy the diabetes-related queries of the users. There are many quality tools, such as the Health on the Net Foundation Code of Conduct (HONcode) [34], DISCERN [35,36], Quality Evaluation Scoring Tool (QUEST) [37] and the Patient Education Materials Assessment Tool (PEMAT) [38], which have been developed to evaluate the quality of health information on the internet and audio/visual materials. However, these tools do not assess the technical abilities of DVAs and may also contain criteria that are not applicable to the DVAs. Hence, the evaluation rubric used in this study was adapted from Goh et al. [8], who had evaluated COVID-19 information provided by DVAs (Figure 1). The rubric was refined based on the criteria in QUEST [37] and PEMAT tools [39], the MedlinePlus tutorial guide from the US National Library of Medicine [40] and a study which evaluated COVID-19 vaccine information on video-sharing platforms [41]. The rubric was refined for suitability for evaluation of diabetes-related information.

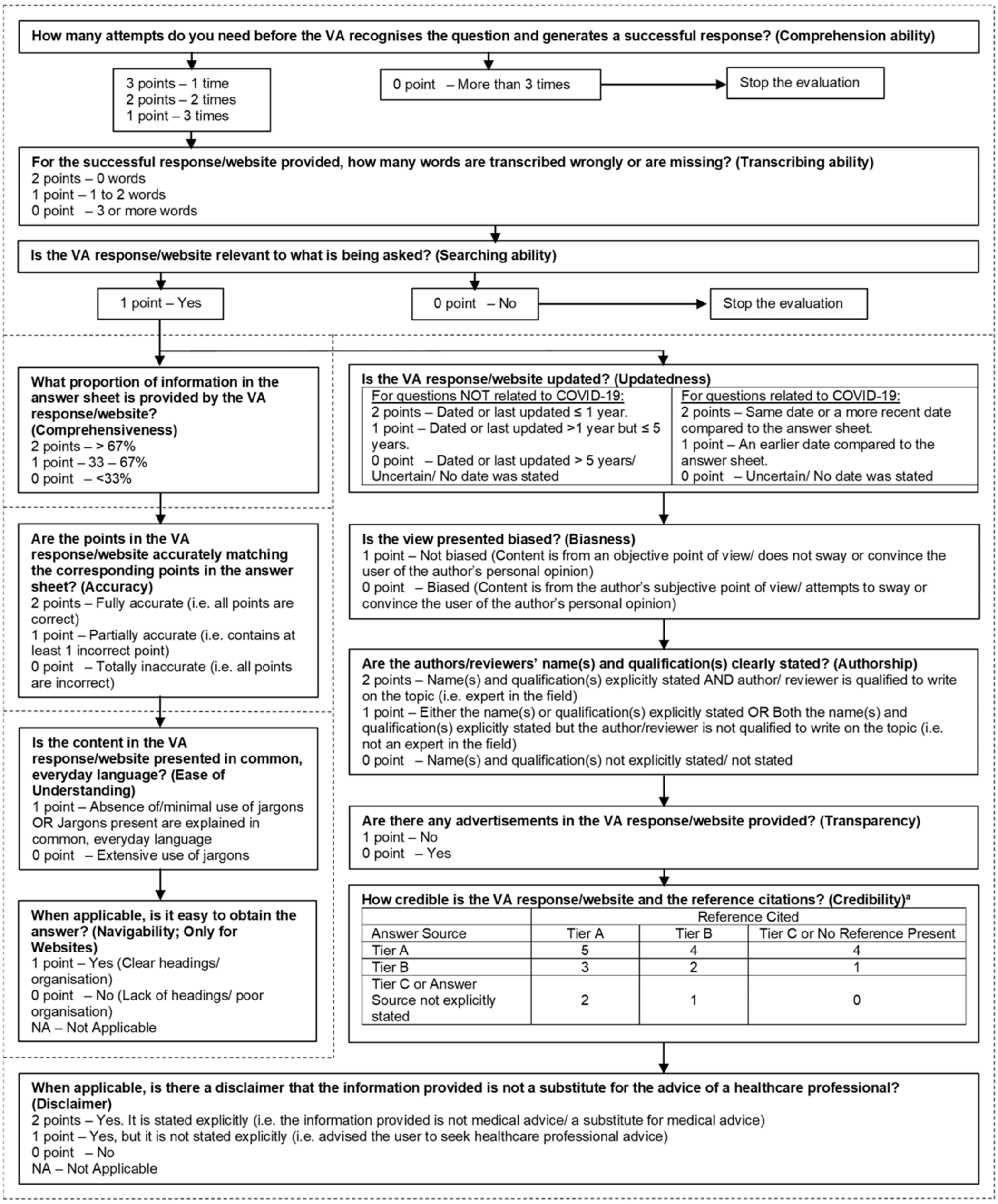
2.2. Development of the quality assessment rubric

The rubric comprised of 5 quality parameters: technical ability, user-friendliness, reliability, comprehensiveness and accuracy (Figure 1). For technical ability, the DVAs were scored based on the number of attempts needed before they recognized the question and generated a successful response (comprehension ability). The responses were considered unsuccessful if the DVAs replied with “Sorry, I didn’t understand that”. The DVAs were also assessed on the number of words transcribed wrongly or missing (transcribing ability) and whether the responses were relevant or not (searching ability). The evaluation would end if the DVAs did not provide successful responses after 3 attempts or if the responses were irrelevant.

For user-friendliness, the responses of the DVAs were assessed based on whether they were presented in common, everyday language (ease of understanding) and if a web search was provided, whether it was easy to obtain the answers (navigability).

For reliability, the DVAs were scored based on how updated the responses were (updatedness), whether the responses were biased or unbiased (biasness), whether name(s) and qualification(s) of the author/reviewer(s) were provided and if they were qualified to write on the topic (authorship). Author/reviewers were considered qualified if they were medical experts. The DVAs were also assessed on whether there were advertisements present (transparency), and if the answer source and reference were trustworthy (credibility). Credibility was assessed according to a 3-tier system. Tier A included sites of recognized health authorities (e.g., World Health Organization (WHO) and American Diabetes Association (ADA)), Tier B included individuals/sites with medical expertise (e.g., WebMD/expert opinions) and Tier C included individuals/sites without medical expertise or sites with commercial backing. The presence of a disclaimer stating that the information should not be used to substitute medical advice was also assessed (disclaimer).

Comprehensiveness of the DVA responses were scored by calculating the proportion of points provided by the DVAs in relation to our compiled answer sheet (Appendix 1). The points provided by the DVAs were subsequently assessed for accuracy by comparing how exactly it matched the corresponding points in the answer sheet. The answer sheet was reviewed by one of the authors (JC) and then reviewed and discussed by the other authors (LLW, KY). A pilot test was conducted by 2 independent evaluators (WLC, AC) on a Samsung Bixby with at least 2 questions from each question category to ensure understandability of the rubric. Feedback was compiled from the pilot test, and questions and parts of the rubric that were unclear to the evaluators were discussed among the authors and rephrased for clarity. The final quality assessment was performed by 3 independent evaluators who were not involved in the pilot test (JC, AP, SN).



How many attempts do you need before the VA recognises the question and generates a successful response? (Comprehension ability)

3 points – 1 time
2 points – 2 times
1 point – 3 times

0 point – More than 3 times

Stop the evaluation

For the successful response/website provided, how many words are transcribed wrongly or are missing? (Transcribing ability)
2 points – 0 words
1 point – 1 to 2 words
0 point – 3 or more words

Is the VA response/website relevant to what is being asked? (Searching ability)

1 point – Yes

0 point – No

Stop the evaluation

What proportion of information in the answer sheet is provided by the VA response/website? (Comprehensiveness)
2 points – > 67%
1 point – 33 – 67%
0 point – <33%

Are the points in the VA response/website accurately matching the corresponding points in the answer sheet? (Accuracy)
2 points – Fully accurate (i.e. all points are correct)
1 point – Partially accurate (i.e. contains at least 1 incorrect point)
0 point – Totally inaccurate (i.e. all points are incorrect)

Is the content in the VA response/website presented in common, everyday language? (Ease of Understanding)
1 point – Absence of/minimal use of jargons OR Jargons present are explained in common, everyday language
0 point – Extensive use of jargons

When applicable, is it easy to obtain the answer? (Navigability; Only for Websites)
1 point – Yes (Clear headings/ organisation)
0 point – No (Lack of headings/ poor organisation)
NA – Not Applicable

Is the VA response/website updated? (Updatedness)

For questions NOT related to COVID-19:	For questions related to COVID-19:
2 points – Dated or last updated ≤ 1 year. 1 point – Dated or last updated >1 year but ≤ 5 years. 0 point – Dated or last updated > 5 years/ Uncertain/ No date was stated	2 points – Same date or a more recent date compared to the answer sheet. 1 point – An earlier date compared to the answer sheet. 0 point – Uncertain/ No date was stated

Is the view presented biased? (Biasness)
1 point – Not biased (Content is from an objective point of view/ does not sway or convince the user of the author's personal opinion)
0 point – Biased (Content is from the author's subjective point of view/ attempts to sway or convince the user of the author's personal opinion)

Are the authors/reviewers' name(s) and qualification(s) clearly stated? (Authorship)
2 points – Name(s) and qualification(s) explicitly stated AND author/ reviewer is qualified to write on the topic (i.e. expert in the field)
1 point – Either the name(s) or qualification(s) explicitly stated OR Both the name(s) and qualification(s) explicitly stated but the author/reviewer is not qualified to write on the topic (i.e. not an expert in the field)
0 point – Name(s) and qualification(s) not explicitly stated/ not stated

Are there any advertisements in the VA response/website provided? (Transparency)
1 point – No
0 point – Yes

How credible is the VA response/website and the reference citations? (Credibility)[a]

	Reference Cited		
Answer Source	Tier A	Tier B	Tier C or No Reference Present
Tier A	5	4	4
Tier B	3	2	1
Tier C or Answer Source not explicitly stated	2	1	0

When applicable, is there a disclaimer that the information provided is not a substitute for the advice of a healthcare professional? (Disclaimer)
2 points – Yes. It is stated explicitly (i.e. the information provided is not medical advice/ a substitute for medical advice)
1 point – Yes, but it is not stated explicitly (i.e. advised the user to seek healthcare professional advice)
0 point – No
NA – Not Applicable

[a] ***Tier A: Sites of Recognised Authorities***
- International or local government websites (e.g. International Diabetes Federation, World Health Organization, U.S. Centers for Disease Control and Prevention, U.K. National Health Service, Singapore Ministry of Health)
- Recognised authorities that develop health guidelines (e.g. American Diabetes Association, American Heart Association)
- Peer-reviewed scientific journals (e.g. Journal of the American Medical Association, The Lancet, Diabetes Care) or Research papers
- Websites of hospitals, medical centers, healthcare clusters or medical schools (e.g. National University Hospital, Mayo Clinic, Cleveland Clinic, SingHealth, Harvard Health)
- Sites that provide medical information to healthcare professionals (e.g. UpToDate, Medscape)
- Medical textbooks (e.g. Harrison's Principles of Internal Medicine)

*For blogs published in Tier A sources, if there is a disclaimer stating that the opinion of the author is representative of the organisation, consider it under this Tier. If there is no disclaimer or the opinion is solely representative of the author, consider it in Tier B if it's written by experts, and Tier C if it's written by non-experts.

Tier B: Individuals or sites with medical expertise
- Sites with medical expertise that provide health information (e.g. WebMD, Healthline)
- Expert Opinions
- Charity/ Non-Profit/ Non-Government organisations that provide health support or research findings (e.g. Diabetes.co.uk, Diabetes UK)

*For blogs published in Tier B sources written by non-experts: If there is a disclaimer stating that the opinion of the author is representative of the organisation, consider it under this Tier. If there is no disclaimer or the opinion is solely representative of the author, consider it under Tier C.
*For blogs published in Tier B sources written by experts: Regardless of whether there is a disclaimer stating that the opinion of the author is representative of the organisation, consider it under this Tier.

Tier C: Individuals or sites without medical expertise or with commercial backing
- Media sites (e.g. The Guardian, CNN)
- Pharmaceutical Companies (e.g. Dexcom, Accu-Chek)
- Charity/ Non-Profit/ Non-Government organisations that are not primarily known for providing health information
- Social media sites (e.g. Facebook)
- Computational knowledge engine (e.g. Wolfram alpha)
- Wikipedia

*For blogs published in Tier C sources, regardless of whether there is a disclaimer stating that the opinion of the author is representative of the organisation, consider it in Tier B if it's written by experts, and Tier C if it's written by non-experts.

Figure 1. Quality evaluation rubric for DVAs.

2.3. Compilation of questions on diabetes

Seventy-four questions were collated from Google Trends, AnswerThePublic.com, and the frequently-asked questions (FAQ) sections of government health websites [42,43] from 11 Jul 2021 to 30 Aug 2021. Google Trends and AnswerThePublic.com were used as the former website provided samples of actual Google searches [44], while the latter consolidated autocomplete data from Google and Bing [45]. The questions were organized into 6 categories: "General Diabetes Information", "Diabetes Diagnosis and Screening", "Diabetes Self-management", "Diabetes Treatment", "Diabetes Complications" and "COVID-19 and Diabetes". For each category, some search trend data phrases were rephrased to be presented as questions to the DVAs. Answers were compiled from health websites from the US [46-50], United Kingdom (UK) [51,52], Australia [53,54], and Singapore [55,56], together with their clinical guidelines [57–61]. The questions and answers were reviewed by a community pharmacist to ensure accuracy of the information and relevance to the public.

2.4. Evaluation of the DVAs

The smartphone DVAs evaluated were Apple Siri (accessed via iPhone 6S, iOS 14.7.1), Google Assistant and Samsung Bixby (both accessed via Samsung Galaxy Note 9, Android 10). The smart-speaker DVAs evaluated were Amazon Alexa and Sulli (both accessed via Amazon Echo Dot 2nd Generation) and Google Nest Mini. The laptop DVA evaluated was Microsoft Cortana (accessed via a Windows laptop).

Three evaluators (JC, female, AP, female and SN, male) independently assessed the DVAs using the same devices from 30 Aug 2021 to 22 Sep 2021 in Singapore. Factory reset of the smartphones and smart-speakers, creation of new accounts and reset of search history were conducted before each DVA evaluation to ensure depersonalization of search results. The device language was set as English (US) and the location function was disabled. For the smartphone and laptop DVAs, the verbal responses and first non-advertisement web result were evaluated. One of the evaluators (JC) conducted the Google Search as the comparison (https://www.google.com.sg/).

2.5. Statistical analysis

The average score of the 3 evaluators was calculated and converted into percentages. Kolmogorov-Smirnov Test (for $n \geq 50$) or Shapiro-Wilk Test (for $n < 50$) was used to assess normality where applicable. Medians and interquartile ranges (IQRs) were used to describe the quality parameter scores and total quality scores of the DVAs. Kruskal-Wallis test was used to compare the quality parameter scores and total quality scores among the DVAs. Chi-Square and Wilcoxon Rank-Sum tests were used to compare the percentages of successful and relevant responses, quality parameters scores and total quality scores of COVID-19 and non-COVID-19 related questions. Wilcoxon Signed-Rank and Paired-Samples t-tests were used to compare the verbal-only responses of Google Assistant and Bixby with their verbal and web responses. Wilcoxon Rank-Sum test with Bonferroni adjustment was used for post-hoc analyses. Inter-rater reliability for each quality parameter and the total quality score was calculated using the Intraclass Correlation Coefficient (ICC). Data analysis was conducted using the Statistical Package for Social Sciences (SPSS) Version 27 and $p < 0.05$ was considered as statistically significant.

3. Results

On average, Siri returned mostly web results (n = 70/74, 94.6%), while Google Assistant (n = 66/74, 89.1%) and Bixby (n = 43/74, 58.1%) mostly returned verbal replies with web results (Table 1). Google Assistant (n = 69/74, 93.2%) had the highest average percentage of successful and relevant responses, followed by Siri and Nest Mini (n = 64/74, 86.5% each), while Cortana (n = 23/74, 31.1%) and Sulli (n = 10/74, 13.5%) had the lowest average percentage of successful and relevant responses. Cortana was unable to comprehend more than half of the questions (n = 43/74, 58.1%) and mostly responded with "Sorry I don't know the answer to this one", while Sulli provided irrelevant answers majority of the time (n = 62/74, 83.8%).

Table 1. Technical ability of the DVAs.

	Smartphone DVAs			Non-smartphone DVAs			
				Smart-speaker DVAs			Laptop DVA
	Google Assistant	Apple Siri	Samsung Bixby	Google Nest Mini	Amazon Alexa	Sulli the Diabetes Guru	Microsoft Cortana
Average number (%) of response types among the 3 evaluators (N = 74)							
Web-only response [a], n (%)	8 (10.8)	70 (94.6)	20 (27.0)	NA [d]	NA [d]	NA [d]	4 (5.4)
Verbal-only response [b], n (%)	0 (0)	1 (1.4)	11 (14.9)	74 (100)	74 (100)	74 (100)	63 (85.1)
Verbal + Web response [c], n (%)	66 (89.1)	4 (5.4)	43 (58.1)	NA [d]	NA [d]	NA [d]	7 (9.5)
Average number (%) of successful and relevant responses among the 3 evaluators (N = 74)							
Successful and relevant response, n (%)	69 (93.2)	64 (86.5)	50 (67.6)	64 (86.5)	44 (59.5)	10 (13.5)	23 (31.1)
Unsuccessful response, n (%)	5 (6.8)	10 (13.5)	24 (32.4)	10 (13.5)	30 (40.5)	64 (86.5)	51 (68.9)
Unable to comprehend, n (%)	0 (0)	0 (0)	8 (10.8)	3 (4.1)	15 (20.3)	2 (2.7)	43 (58.1)
Irrelevant, n (%)	5 (6.8)	10 (13.5)	16 (21.6)	7 (9.5)	15 (20.3)	62 (83.8)	8 (10.8)
Average number (%) of correct transcriptions among the 3 evaluators (N = 74)							
Correct transcription, n (%)	70 (95.0)	59 (79.3)	58 (77.9)	69 (92.8)	62 (83.3)	56 (75.5)	65 (87.4)

Continued on next page

	Smartphone DVAs			Non-smartphone DVAs			
				Smart-speaker DVAs			Laptop DVA
	Google Assistant	Apple Siri	Samsung Bixby	Google Nest Mini	Amazon Alexa	Sulli the Diabetes Guru	Microsoft Cortana
Average number (%) of responses that are the same for the 3 evaluators (N = 74)							
Same response provided to the evaluators, n (%)	56 (75.7)	49 (66.2)	33 (43.6)	55 (74.3)	39 (52.7)	60 (81.1)	39 (52.7)

Note: [a] Refers to instances when the DVA provided a web link to the query or verbally directed the users to the web link i.e. “Here’s what I found *[web link]”*.

[b] Refers to instances when the DVA provided a verbal response to the query.

[c] Refers to instances when the DVA provided both a verbal response and a web link to the query i.e. “The symptoms of diabetes include frequent urination, thirst, losing weight without trying and more *[web link]”*.

[d] NA: Not applicable to smart-speaker DVAs as they only provided a verbal response.

The median total quality score of the smartphone DVAs (73.3%) was not significantly different from Google Search (70.0%, p = 0.034) (Table 2). Among the smartphone DVAs, Bixby had the highest total quality score (median = 75.3%, IQR = 67.7–79.4%) followed by Google Assistant (median = 73.3%, IQR = 66.7–78.0%) and Siri (median = 72.0%, IQR = 65.3–77.3%). On the other hand, Google Nest Mini performed the best among the non-smartphone DVAs in terms of its total quality score (median = 56.9%, IQR = 50.0–65.3%), while Sulli performed the worst (median = 48.6%, IQR = 45.1–55.5%). The median total quality score of non-smartphone DVAs (54.2%) was significantly lower than those of Google Search (70.0%, p < 0.001) and smartphone DVAs (73.3%, p < 0.001).

All the DVAs scored highly for user-friendliness (median = 100.0% for all) and most of them were accurate (median > 80.0% for all except Sulli, 66.7%) (Table 2). User-friendliness and accuracy scores of all DVAs were comparable to Google Search. Google Assistant and Bixby scored the highest for comprehensiveness (median = 83.3% each, IQR = 50.0–100.0%), while Alexa scored the lowest (median = 0%, IQR = 0.0–31.2%). The median comprehensiveness score of the non-smartphone DVAs (16.7%) was significantly lower than those of Google Search (100.0%, p < 0.001) and the smartphone DVAs (83.3%, p < 0.001).

Bixby performed the best for reliability (median = 61.5%, IQR = 46.2–69.2%) while Sulli performed the worst (median = 23.2%, IQR = 18.2–33.4%). The median reliability score of non-smartphone DVAs (30.8%) was significantly lower than that of Google Search (61.5%, p < 0.001) and the smartphone DVAs (59.0%, p < 0.001). Sub-analysis of reliability scores revealed that the non-smartphone DVAs scored poorly for updatedness (median = 0% for all) and in terms of the presence of disclaimers (Table 3). In terms of the latter criterion, Sulli (median = 100.0%, IQR = 100.0–100.0%) scored the highest, while the rest of the other non-smartphone DVAs scored poorly instead (median = 0% for all). More than half of the responses by Bixby (63.3%), Google Search (61.6%), Google Assistant (56.0%) and Nest Mini (55.0%) were from Tier A sources, such as the Centre for Disease Control and Prevention (CDC), Mayo Clinic, and the National Health Service (NHS). The DVAs had a decreasing trend for the proportions of Tier A, Tier B and C sources except Alexa, Sulli and Cortana.

Table 2. Quality parameter scores and total quality scores of the DVAs.

Quality parameter		Scores of Smartphone DVAs [Median % (IQR)]			p-value between Internet Google Search and Smart phone DVAs [1]	Scores of Non-smartphone DVAs [Median % (IQR)]				p-value between Internet Google Search and Non-smart phone DVAs [2]	p-value between Internet Google Search and all DVAs [3]	Inter-rater reliability [4]
						Smart speaker DVAs			Laptop DVA			
	Internet Google Search	Google Assistant	Apple Siri	Samsung Bixby		Google Nest Mini	Amazon Alexa	Sulli the Diabetes Guru	Microsoft Cortana			
User-friendliness	100.0 (100.0–100.0)	100.0 (100.0–100.0)	100.0 (83.3–100.0)	100.0 (100.0–100.0)	0.132	100.0 (100.0–100.0)	100.0 (66.7–100.0)	100.0 (91.7–100.0)	100.0 (100.0–100.0)	0.233	0.284	0.570 (0.466–0.656)
Reliability	61.5 (53.9–69.2)	53.9 (46.2–66.7)	54.3 (46.2–66.7)	61.5 (46.2–69.2)	0.026	43.6 (23.1–46.2)	30.8 (21.2–46.2)	23.2 (18.2–33.4)	28.2 (15.4–46.2)	<0.001*	<0.001**^{a}	0.940 (0.926–0.951)
Comprehensive-ness	100.0 (50.0–100.0)	83.3 (50.0–100.0)	66.7 (50.0–100.0)	83.3 (50.0–100.0)	0.072	16.7 (0.0–50.0)	0.0 (0.0–31.2)	16.7 (0.0–50.0)	50.0 (8.4–75.0)	<0.001*	<0.001**^{b}	0.902 (0.880–0.920)
Accuracy	100.0 (50.0–100.0)	83.3 (66.7–100.0)	83.3 (50.0–100.0)	83.3 (62.5–100.0)	0.151	100.0 (66.7–100.0)	83.3 (50.0–100.0)	66.7 (50.0–87.5)	83.3 (50.0–100.0)	0.046	0.506	0.753 (0.696–0.801)

Continued on next page

Quality parameter		Scores of Smartphone DVAs [Median % (IQR)]			p-value between Internet Google Search and Smart phone DVAs[1]	Scores of Non-smartphone DVAs [Median % (IQR)]				p-value between Internet Google Search and Non-smart phone DVAs[2]	p-value between Internet Google Search and all DVAs[3]	Inter-rater reliability[4]
						Smart speaker DVAs			Laptop DVA			
	Internet Google Search	Google Assistant	Apple Siri	Samsung Bixby		Google Nest Mini	Amazon Alexa	Sulli the Diabetes Guru	Microsoft Cortana			
Total Quality Score	70.0 (60.0–75.0)	73.3 (66.7–78.0)	72.0 (65.3–77.3)	75.3 (67.7–79.4)	0.034	56.9 (50.0–65.3)	52.9 (48.1–58.3)	48.6 (45.1–55.5)	52.5 (41.7–64.7)	<0.001*	<0.001**[c]	0.931 (0.916–0.944)

Note: * $p < 0.0167$ based on Wilcoxon Rank-Sum test with Bonferroni adjustment ($\alpha = 0.05/3 = 0.0167$).

** $p < 0.05$ based on Kruskal-Wallis test.

[1] p-value was based on post-hoc analysis of Kruskal-Wallis test comparing Internet Google Search, smartphone DVAs and non-smartphone DVAs. Wilcoxon Rank-Sum test between smartphone DVAs and Internet Google Search was conducted with Bonferroni adjustment ($\alpha = 0.05/3 = 0.0167$).

[2] p-value was based on post-hoc analysis of Kruskal-Wallis test comparing Internet Google Search, smartphone DVAs and non-smartphone DVAs. Wilcoxon Rank-Sum test between non-smartphone DVAs and Internet Google Search was conducted with Bonferroni Adjustment ($\alpha = 0.05/3 = 0.0167$).

[3] p-value was based on Kruskal-Wallis test comparing Internet Google Search, Google Assistant, Siri, Bixby, Nest Mini, Alexa, Sulli the Diabetes Guru and Cortana. Post-hoc analyses were conducted using Wilcoxon Rank-Sum test with Bonferroni adjustment ($\alpha = 0.05/28 = 0.00179$).

[4] Intraclass Correlation Coefficients was calculated based on the mean-rating (3 evaluators), absolute-agreement, and a 2-way mixed-effects model. Reported as ICC estimate (95% confidence interval).

[a] Reliability scores for non-smartphone DVAs were significantly lower than Internet Google Search, Google Assistant, Siri and Bixby ($p < 0.001$ for each pairwise comparison).

[b] Comprehensiveness scores for non-smartphone DVAs were significantly lower than Internet Google Search, Google Assistant and Bixby ($p < 0.001$ for each pairwise comparison). Non-smartphone DVAs ($p < 0.001$ for each pairwise comparison) scored significantly lower than Siri except Cortana ($p = 0.005$). Alexa scored significantly lower than Cortana ($p < 0.001$).

[c] Total quality scores for non-smartphone DVAs were significantly lower than Internet Google Search, Google Assistant, Siri and Bixby ($p < 0.001$ for each pairwise comparison).

Table 3. Breakdown of the reliability and credibility scores of the DVAs.

		Smartphone DVAs			p-value between Internet Google Search and Smart phone DVAs [1]	Non-smartphone DVAs				p-value between Internet Google Search and Non-smart phone VAs [2]	p-value between Internet Google Search and all DVAs [3]
						Smart-speaker DVAs			Laptop DVA		
	Internet Google Search	Google Assistant	Apple Siri	Samsung Bixby		Google Nest Mini	Amazon Alexa	Sulli the Diabetes Guru	Microsoft Cortana		
Breakdown of Reliability [Scores in Median % (IQR)]											
Updatedness	50.0 (0.0–100.0)	50.0 (0.0–83.3)	50.0 (16.7–100.0)	50.0 (33.3–100.0)	0.562	0.0 (0.0–0.0)	0.0 (0.0–0.0)	0.0 (0.0–0.0)	0.0 (0.0–29.2)	<0.001*	<0.001**[a]
Absence of bias	100.0 (100.0–100.0)	100.0 (100.0–100.0)	100.0 (100.0–100.0)	100.0 (100.0–100.0)	-	100.0 (100.0–100.0)	100.0 (100.0–100.0)	100.0 (100.0–100.0)	100.0 (100.0–100.0)	-	0.221
Transparency	33.3 (33.3–66.7)	33.3 (22.2–66.7)	33.3 (33.3–66.7)	33.3 (33.3–66.7)	0.773	33.3 (33.3–33.3)	33.3 (33.3–33.3)	33.3 (33.3–33.3)	33.3 (27.8–33.3)	<0.001*	0.002**[b]
Credibility	80.0 (60.0–80.0)	70.0 (33.3–80.0)	60.0 (33.3–80.0)	80.0 (40.0–80.0)	0.151	70.0 (20.0–80.0)	40.0 (15.0–80.0)	0.0 (0.0–6.7)	20.0 (0.0–70.0)	<0.001*	<0.001**[c]
Disclaimer	50.0 (0.0–50.0)	16.7 (0.0–50.0)	33.3 (0.0–75.0)	25.0 (0.0–79.2)	0.116	0.0 (0.0–0.0)	0.0 (0.0–0.0)	100.0 (100.0–100.0)	0.0 (0.0–0.0)	<0.001*	<0.001**[d]
Proportion of answer sources classified in the different tiers (%)											
Tier A sources (%)	61.6	56.0	48.2	63.3	-	55.0	43.8	0.0	35.3	-	-
Tier B sources (%)	34.2	29.5	33.0	30.0	-	34.0	19.2	0.0	8.8	-	-

Continued on next page

		Smartphone DVAs			p-value between Internet Google Search and Smart phone DVAs [1]	Non-smartphone DVAs				p-value between Internet Google Search and Non-smart phone VAs [2]	p-value between Internet Google Search and all DVAs [3]
						Smart-speaker DVAs			Laptop DVA		
	Internet Google Search	Google Assistant	Apple Siri	Samsung Bixby		Google Nest Mini	Amazon Alexa	Sulli the Diabetes Guru	Microsoft Cortana		
Proportion of answer sources classified in the different tiers (%)											
Tier C sources/ Answer source not explicitly stated (%)	4.1	14.5	18.8	6.7	-	11.0	36.9	100.0	55.9	-	-

Note: * p < 0.0167 based on Wilcoxon Rank-Sum test with Bonferroni adjustment (α = 0.05/3 = 0.0167).

** p < 0.05 based on Kruskal-Wallis test.

[1] p-value was based on post-hoc analysis of Kruskal-Wallis test comparing Internet Google Search, smartphone DVAs and non-smartphone DVAs. Wilcoxon Rank-Sum test between smartphone DVAs and Internet Google Search was conducted with Bonferroni adjustment (α = 0.05/3 = 0.0167).

[2] p-value was based on post-hoc analysis of Kruskal-Wallis test comparing Internet Google Search, smartphone DVAs and non-smartphone DVAs. Wilcoxon Rank-Sum test between non-smartphone DVAs and Internet Google Search was conducted with Bonferroni adjustment (α = 0.05/3 = 0.0167).

[3] p-value was based on Kruskal-Wallis test comparing Internet Google Search, Google Assistant, Siri, Bixby, Nest Mini, Alexa, Sulli the Diabetes Guru and Cortana. Post-hoc analyses were conducted using Wilcoxon Rank-Sum Test with Bonferroni adjustment (α = 0.05/28 = 0.00179).

[a] Updatedness scores for non-smartphone VAs were significantly lower than Internet Google Search, Google Assistant, Siri and Bixby (p < 0.001 for each pairwise comparison). Cortana scored significantly higher than Nest Mini and Alexa (p < 0.001 for both).

[b] Transparency scores for Nest Mini were significantly lower than Bixby (p < 0.001).

[c] Credibility scores for Alexa were significantly lower than Internet Google Search and Bixby (p < 0.001 for both). Cortana scored significantly lower than Internet Google Search, Google Assistant, Siri and Bixby (p < 0.001 for each pairwise comparison). Sulli scored significantly lower than each of the DVAs (p < 0.001 for each pairwise comparison) except Cortana (p = 0.145).

[d] Disclaimer scores for Sulli the Diabetes Guru was significantly higher than the other DVAs (p < 0.001 or p = 0.001 for each pairwise comparison). Nest Mini, Alexa and Cortana scored significantly lower than Internet Google Search, Google Assistant, Siri and Bixby (p < 0.001 for each pairwise comparison).

Since the responses of Google Assistant and Bixby consisted mostly of a verbal response with a web result, a sub-analysis was conducted for their verbal responses alone (Figure 2). The verbal responses of both the DVAs had much lower median comprehensiveness and reliability scores than when both the verbal and web responses were taken into consideration.

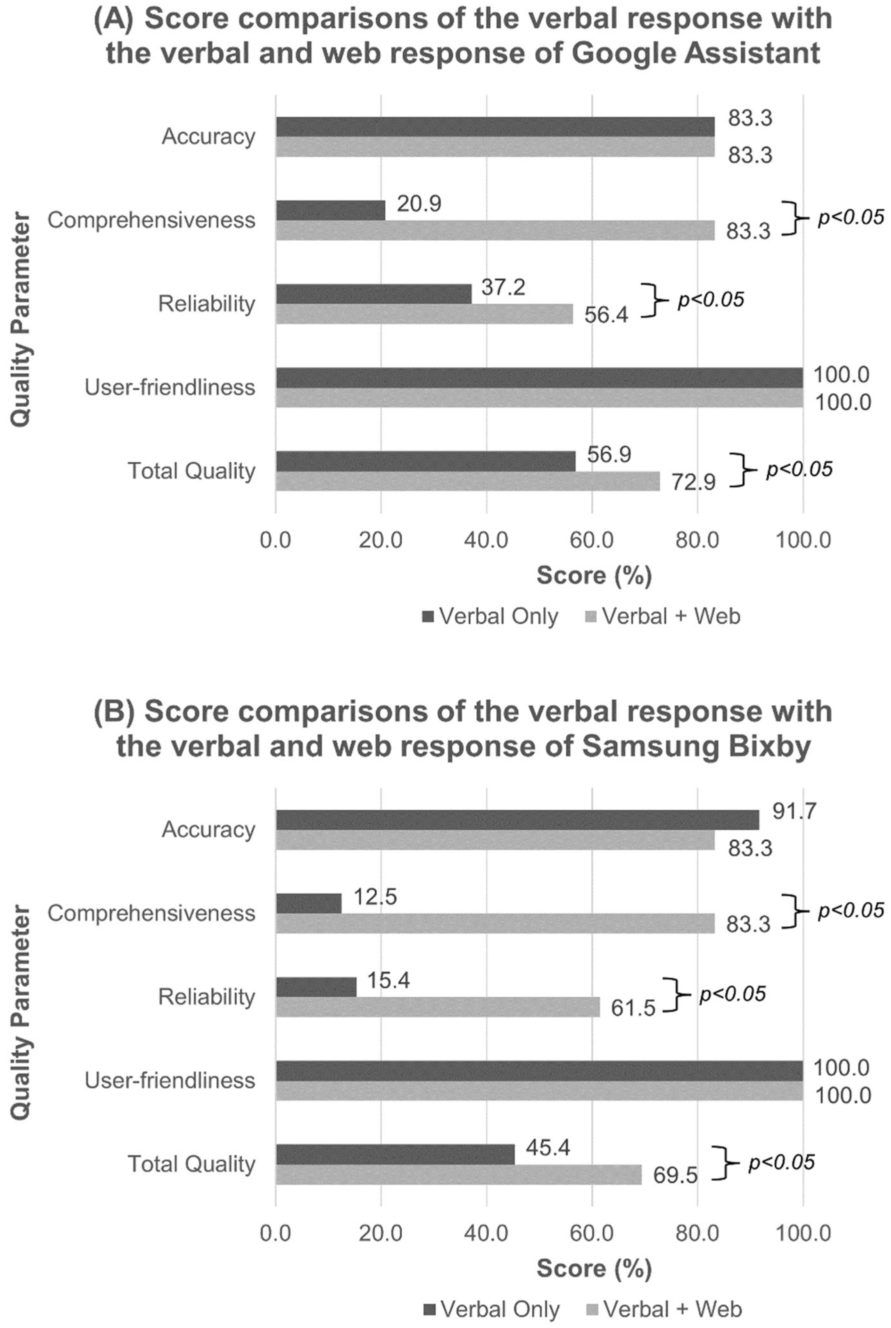


Figure 2. Sub-analysis of the quality parameter scores and total quality scores of the verbal response with verbal and web response of Google Assistant and Samsung Bixby.

When the proportion of successful and relevant responses were compared among the DVAs, 3 of the non-smartphone DVAs (Alexa, Sulli and Cortana) did not have any successful and relevant responses for COVID-19 related questions (Table 4). Among the non-COVID-19 related questions, there was a decreasing trend of successful and relevant responses, with Nest Mini being the best, followed by Alexa, Sulli and Cortana. Nest Mini had a significantly higher proportion of successful and relevant responses for non-COVID-19 related questions (93.9% versus 20.8%, $p < 0.001$). Similarly, among the smartphone DVAs, only Google Assistant (96.5% versus 66.7%, $p < 0.001$) and Bixby (72.2% versus 29.2%, $p < 0.001$) had higher proportions of successful and relevant responses for non-COVID-19 related questions compared to COVID-19 related questions.

Table 4. Proportion of successful and relevant responses for COVID-19 and non-COVID-19 related questions by DVAs.

DVA/Internet Google Search	Proportion of successful and relevant responses (%)			p-value between COVID-19 and non-COVID-19 related questions
	Overall (COVID-19 and non-COVID-19 related questions combined)	COVID-19 related questions (N = 66)	Non-COVID-19 related questions (N = 8)	
Internet Google Search (%)	98.7	87.5	100	<0.001*
Smartphone DVAs				
Google Assistant (%)	93.2	66.7	96.5	<0.001*
Apple Siri (%)	86.5	91.7	84.5	0.127
Samsung Bixby (%)	67.6	29.2	72.2	<0.001*
Non-smartphone DVAs				
Google Nest Mini (%)	86.5	20.8	93.9	<0.001*
Amazon Alexa (%)	59.5	0	66.2	<0.001*
Sulli the Diabetes Guru (%)	13.5	0	14.6	<0.001*
Microsoft Cortana (%)	31.1	0	10.6	<0.001*

Note: * $p < 0.05$ based on chi-square test.

For DVAs with successful and relevant responses (Google Assistant, Siri, Bixby and Nest Mini), there were no significant differences between how the DVAs performed for COVID-19 and non-COVID-19 related questions, except for Google Assistant, which performed better in terms of accuracy for COVID-19 related questions (100% versus 83.3%, $p = 0.009$) (Table 5). In terms of overall quality (total quality score), in general, Google Assistant and Siri seemed to fare equally well for COVID-19 and non-COVID-19 related questions, compared to Bixby and Nest Mini, which seemed to fare a little better for non-COVID-19 related questions and COVID-19 related questions respectively.

There was moderate–excellent inter-rater reliability among the 3 evaluators for each quality parameter and for the total quality score (Table 2). The quality parameters with the highest and lowest ICC values were reliability [0.940 (0.926–0.951)] and user-friendliness [0.570 (0.466–0.656)] respectively.

Table 5. Comparison of the quality parameter scores and total quality scores of DVAs between non-COVID-19 related and COVID-19 related questions.

Voice Assistants	Quality parameter scores of DVAs [Median% (IQR)]		p-value
	Non-COVID-19 related questions (N = 66)	COVID-19 related questions (N = 8)	
Accuracy [a]			
Internet Google Search	100.0 (50.0–100.0)	100.0 (100.0–100.0)	0.089
Google Assistant	83.3 (66.7–100.0)	100.0 (100.0–100.0)	0.009*
Apple Siri	83.3 (50.0–100.0)	100.0 (70.9–100.0)	0.230
Samsung Bixby	83.3 (66.7–100.0)	50.0 (25.0–87.5)	0.088
Google Nest Mini	83.3 (58.4–100.0)	100.0 (100.0–100.0)	0.196
Comprehensiveness [b]			
Internet Google Search	100.0 (50.0–100.0)	100.0 (50.0–100.0)	1.000
Google Assistant	83.3 (50.0–100.0)	91.7 (75.0–100.0)	0.221
Apple Siri	66.7 (50.0–100.0)	83.3 (54.2–95.8)	0.449
Samsung Bixby	83.3 (54.2–95.8)	25.0 (0.0–100.0)	0.230
Google Nest Mini	16.7 (0.0–50.0)	0.0 (0.0–25.0)	0.598
Reliability [c]			
Internet Google Search	61.5 (53.9–69.2)	46.2 (46.2–61.5)	0.112
Google Assistant	56.4 (43.6–66.7)	48.1 (45.6–66.4)	0.789
Apple Siri	57.7 (46.2–66.7)	50.0 (44.9–59.6)	0.429
Samsung Bixby	63.6 (52.5–69.2)	46.2 (46.2–65.4)	0.300
Google Nest Mini	38.5 (23.1–46.2)	46.2 (46.2–50.1)	0.111
User-Friendliness [d]			
Internet Google Search	100.0 (100.0–100.0)	100.0 (50.0–100.0)	0.205
Google Assistant	100.0 (100.0–100.0)	100.0 (83.3–100.0)	0.584
Apple Siri	100.0 (83.3–100.0)	100.0 (100.0–100.0)	0.272
Samsung Bixby	100.0 (100.0–100.0)	100.0 (87.5–100.0)	0.759
Google Nest Mini	100.0 (100.0–100.0)	100.0 (100.0–100.0)	0.644

Continued on next page

Voice Assistants	Quality parameter scores of DVAs [Median% (IQR)]		p-value
	Non-COVID-19 related questions (N = 66)	COVID-19 related questions (N = 8)	
Total Score [e]			
Internet Google Search	70.0 (63.8–75.0)	65.0 (55.0–75.0)	0.525
Google Assistant	73.3 (65.7–77.7)	71.7 (68.7–82.5)	0.617
Apple Siri	71.0 (64.7–77.8)	73.3 (69.0–77.0)	0.570
Samsung Bixby	76.0 (68.0–80.0)	68.0 (59.2–73.0)	0.075
Google Nest Mini	56.9 (50.0–65.3)	62.5 (60.4–64.6)	0.305

Note: * $p < 0.005$ based on Wilcoxon Rank-Sum Test.

[a] Based on Kruskal-Wallis Test, accuracy scores of Internet Google Search, Google Assistant, Siri, Bixby and Nest Mini were not significantly different for COVID-19 related questions ($p = 0.003$, post-hoc analyses with Wilcoxon Rank-Sum Test revealed no significant difference) and non-COVID-19 related questions ($p = 0.812$).

[b] Based on Kruskal-Wallis Test, comprehensiveness scores of Internet Google Search, Google Assistant, Siri, Bixby and Nest Mini were not significantly different for COVID-19 related questions ($p = 0.090$). Comprehensiveness score of Nest Mini was significantly lower than Internet Google Search, Google Assistant, Siri and Bixby for non-COVID-19 related questions ($p < 0.001$ for each pairwise comparison).

[c] Based on Kruskal-Wallis Test, reliability scores of Internet Google Search, Google Assistant, Siri, Bixby and Nest Mini were not significantly different for COVID-19 related questions ($p = 0.971$). Reliability score of Nest Mini was significantly lower than Internet Google Search, Google Assistant, Siri and Bixby for non-COVID-19 related questions ($p < 0.001$ for each pairwise comparison).

[d] Based on Kruskal-Wallis Test, user-friendliness scores of Internet Google Search, Google Assistant, Siri, Bixby and Nest Mini were not significantly different for COVID-19 related questions ($p = 0.779$) and non-COVID-19 related questions ($p = 0.208$).

[e] Based on Kruskal-Wallis Test, total scores of Internet Google Search, Google Assistant, Siri, Bixby and Nest Mini were not significantly different for COVID-19 related questions ($p = 0.154$). Total score of Nest Mini was significantly lower than Internet Google Search, Google Assistant, Siri and Bixby for non-COVID-19 related questions ($p < 0.001$ for each pairwise comparison).

4. Discussion

This study evaluated the quality of DVAs in terms of their technical ability, user-friendliness, reliability, comprehensiveness and accuracy. Overall, the total quality scores of the smartphone DVAs (Google Assistant, Bixby and Siri) were high and comparable to Google Search. Despite having the highest total quality score (Table 2), Bixby had a much lower percentage of successful and relevant responses compared to the other smartphone DVAs (Table 4). Non-smartphone DVAs (Nest Mini, Alexa, Sulli and Cortana) had lower total quality scores compared to the smartphone DVAs, which is consistent with other studies that evaluated the quality of VA responses to health and COVID-19 queries [8,9].

User-friendliness was the highest scored quality parameter, suggesting that information provided by the DVAs was easy to navigate and understand. Accuracy was the next highest scored quality parameter. All DVAs were penalized for accuracy mainly because their points were not specific enough when compared to our compiled answers. In addition, incorrect information was provided on two occasions. The web responses of the smartphone DVAs, Nest Mini and Cortana stated that the ADA's fasting blood glucose recommendation was 70–130 mg/dL. However, this information was outdated as ADA had changed its recommendation from 70–130 mg/dL to 80–130 mg/dL in 2015 [62]. The evaluators also observed that the web responses of Google Assistant and Nest Mini stated that people with diabetes were at a higher risk of getting an influenza virus infection, which was in contrast to the information provided by ADA and CDC, even though people with diabetes were at higher risks of serious flu complications [63,64]. These two examples suggested that DVAs could potentially provide information that could be misinterpreted by users with diabetes, who might be predisposed to harm if they unknowingly act on the inaccurate information [11].

Most of the responses of Nest Mini and Alexa began with a web resource from the internet (e.g., "According to cdc.gov…"). However, only a subset of information from the web resource was verbalized, resulting in low comprehensiveness scores. In response to "How is diabetes diagnosed?", Nest Mini replied with "On the website clevelandclinic.org, they say: "Diabetes is diagnosed and managed by checking your glucose level in a blood test." It did not include the next sentence from the web resource, which provided information on the types of tests (i.e., fasting glucose, random glucose and HbA1c tests). On the other hand, even though Alexa was described in several studies to provide long verbal responses [7,8], this feature did not translate to a more comprehensive response in our study.

In our study, reliability was observed to be the lowest scored quality parameter. This could be attributed to the fact that a substantial proportion of health websites providing diabetes information did not provide the date of update regarding their information and details on authorship [65,66]. In cases where the websites had actually contained this information, the smart-speaker DVAs did not provide it to the evaluators. Hence, smart-speaker DVAs scored much lower for reliability. The recommendation by the US National Library of medicine is that such information should be included in websites providing health information online [67]. In this sense, we also advocate that the recommendation by the US National Library of Medicine also be followed for the responses by DVAs, so that users can evaluate the quality and trustworthiness of the responses.

The credibility was higher for smartphone DVAs and Nest Mini as most of their responses were from sites of recognized authorities or sites/individuals with medical expertise. This was in contrast to Alexa, which scored lower for credibility as it referred users to Wikipedia for questions on definition (average n = 12/74, 12.2%). Our results were similar to the findings by Alagha and Helbing [7], who evaluated the quality of vaccine-related information provided by DVAs. However, in recent years, there is an upward trend in the number of studies discussing the use of Wikipedia's health information, which hints at the growth in acceptance of health information provided by it [68]. We encourage users to supplement the information from Wikipedia with other sources by recognized authorities or with medical expertise wherever possible. On the other hand, majority of the answer sources of Cortana were classified as not explicitly stated because the information provided in the application did not match the Bing Search which it directed its users to, which could be a potential cause of confusion.

All the DVAs except Sulli lacked a medical disclaimer in their responses, which was contrary to a previous study which found that Alexa provided clear disclaimers in its verbal responses [8]. As Amazon and Google required health-specific skills/actions to include such disclaimers [69,70], users of Sulli would be informed of the disclaimer when they first enabled the skill. However, for Sulli, it is not a requirement for subsequent responses to have a disclaimer. Users may forget that the information provided by Sulli is not a substitute for medical advice. Thus, it is important for DVAs to include disclaimers in their responses since the management of diabetes is individualized [71] and users should seek professional medical advice before following the information provided by the DVAs.

The verbal-only responses of Google Assistant and Bixby were not comprehensive and lacked details on reliability, which resulted in occasions whereby the evaluators were misled. For example, in response to "Who should get tested for diabetes?", Bixby responded with "Here's what the internet says: ADA recommends…", which led to the evaluators perceiving ADA to be the answer source. However, the information was taken from a Healthline website which cited ADA instead. Users of these smartphone DVAs who rely solely on their verbal responses might perceive the source to be more credible than the actual website source. Hence, we encourage users of these smartphone DVAs to refer to the original source of information provided by the DVAs to clarify their doubts whenever possible.

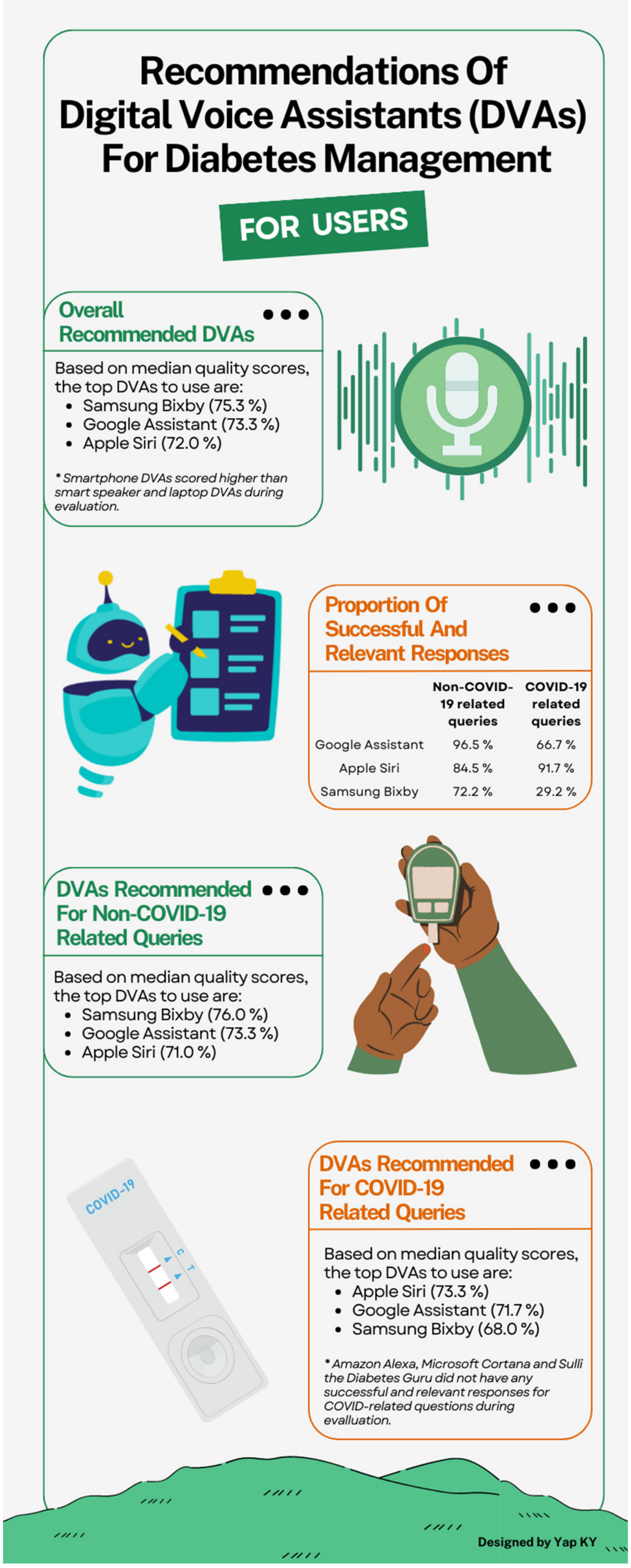
Some phrases taken from search trend data were rephrased as questions to reflect typical voice searches [72]. For example, "diabetes symptoms" was rephrased as "What are the symptoms of diabetes?" Despite rephrasing, Sulli and Bixby could not really recognize users' intent. When a range of insulin-related queries were posed to Sulli, it responded in the exact same way with all the queries by providing a general description of insulin. Bixby also provided generic answers in response to COVID-19 related questions. We postulate that these DVAs might have recognized the queries based on keywords instead of the user's intent. In addition, the evaluators noted that some DVAs were unable to recognize critical situations. Our results concurred with findings by Kocaballi et al. [9]. For example, in response to "How do I treat hypoglycemia?", Alexa replied that "treatment may include dietary, medical, and/or surgical therapies", while Siri responded with "the answer I found is diazoxide". A user asking this question might be experiencing hypoglycemia symptoms, which is dangerous if it is left untreated [73]. The more appropriate recommendation would be for users to consume 15–20 g of fast-acting carbohydrate such as 3–4 glucose tablets immediately [74,75]. In this regard, we suggest that the DVAs could potentially improve in terms of their natural language understanding algorithms to better recognize users' intents and adapt to different types of questions [76].

5. Limitations

The answers to our compiled questions on diabetes were crafted with reference to the US, UK, Australia and Singapore government health websites and clinical guidelines. Since recommendations for screening and diagnosis, and target blood glucose levels can vary depending on each country's standard of practice, the results of this study might not be able to be directly extrapolated to other countries. The accuracy of our compiled answers is also limited to the period of study as clinical guidelines for diabetes management and the algorithms of the DVAs are subjected to updates and/or changes. However, this study still adds to the literature by providing a current snapshot of the quality of DVAs for the management of chronic conditions, such as diabetes, especially during this period of time when the world is just transitioning and adjusting to the post-pandemic era. For patients with diabetes, it is important for them to know which DVAs are able to address their general queries and concerns regarding their diabetes and medications, as well as the impacts of COVID-19 and vaccinations on their medical condition. Furthermore, our evaluation process might not have mimicked how users interact with DVAs as the evaluators did not continue the conversation with the DVAs when prompted. Furthermore, Sulli was only accessed through Alexa since the evaluators were unable to access it through Google Assistant in Singapore. It is unknown how this diabetes-specific DVA will perform through a different platform. Lastly, the evaluators in this study were between the ages of 18 to 25 years old. The interpretation of user-friendliness might not be extrapolatable to the older population, especially those above 45 years old who are increasingly moving towards owning and using DVAs [77], yet might have a higher probability of being diagnosed with diabetes [78].

6. Conclusions

Google Assistant, Siri and Bixby were the best performing DVAs overall when presented with a range of diabetes-related questions. In contrast, the diabetes-specific DVA (Sulli) performed the worst. Although Bixby had the highest median quality score, its lack of successful and relevant responses may frustrate users, especially if they have COVID-19 related queries. A list of recommendations is provided in Figure 3 for users who use DVAs for their diabetes-related queries. In general, smartphone DVAs scored higher in terms of quality than non-smartphone DVAs. The responses of smartphone DVAs were comparable to Google Search in terms of user-friendliness, reliability, comprehensiveness and accuracy, while the non-smartphone DVAs scored much lower for reliability and comprehensiveness. In general, there is still room for improvement for all the DVAs in terms of reliability, comprehensiveness and accuracy. As such, DVA designers are encouraged to consider features related to these quality parameters, in addition to user-friendliness (Figure 4), when developing their products for medical purposes, such as diabetes management.



Recommendations Of Digital Voice Assistants (DVAs) For Diabetes Management

FOR USERS

Overall Recommended DVAs

Based on median quality scores, the top DVAs to use are:

- Samsung Bixby (75.3 %)
- Google Assistant (73.3 %)
- Apple Siri (72.0 %)

** Smartphone DVAs scored higher than smart speaker and laptop DVAs during evaluation.*

Proportion Of Successful And Relevant Responses

	Non-COVID-19 related queries	COVID-19 related queries
Google Assistant	96.5 %	66.7 %
Apple Siri	84.5 %	91.7 %
Samsung Bixby	72.2 %	29.2 %

DVAs Recommended For Non-COVID-19 Related Queries

Based on median quality scores, the top DVAs to use are:

- Samsung Bixby (76.0 %)
- Google Assistant (73.3 %)
- Apple Siri (71.0 %)

DVAs Recommended For COVID-19 Related Queries

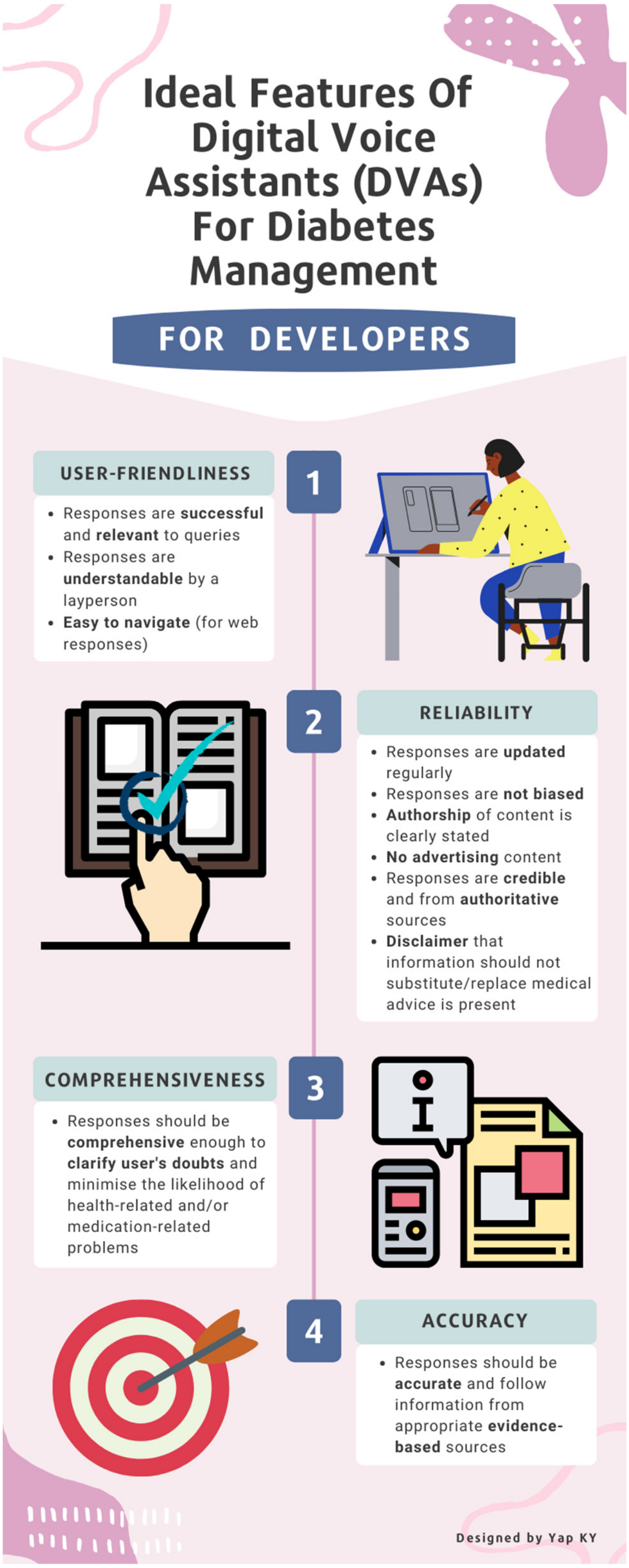
Based on median quality scores, the top DVAs to use are:

- Apple Siri (73.3 %)
- Google Assistant (71.7 %)
- Samsung Bixby (68.0 %)

** Amazon Alexa, Microsoft Cortana and Sulli the Diabetes Guru did not have any successful and relevant responses for COVID-related questions during evalluation.*

Designed by Yap KY

Figure 3. Recommendations for users on which DVAs to use for their diabetes-related queries.



Ideal Features Of Digital Voice Assistants (DVAs) For Diabetes Management

FOR DEVELOPERS

1 USER-FRIENDLINESS

- Responses are **successful** and **relevant** to queries
- Responses are **understandable** by a layperson
- **Easy to navigate** (for web responses)

2 RELIABILITY

- Responses are **updated** regularly
- Responses are **not biased**
- **Authorship** of content is clearly stated
- **No advertising** content
- Responses are **credible** and from **authoritative** sources
- **Disclaimer** that information should not substitute/replace medical advice is present

3 COMPREHENSIVENESS

- Responses should be **comprehensive** enough to **clarify user's doubts** and minimise the likelihood of health-related and/or medication-related problems

4 ACCURACY

- Responses should be **accurate** and follow information from appropriate **evidence-based** sources

Designed by Yap KY

Figure 4. Recommendations for developers on the quality considerations when designing DVA products for diabetes management.

Acknowledgments

The authors would like to thank Ms Wing Lam Chung and Ms Amanda Shue Qi Chia for conducting the pilot test; and Ms Alyssa Elyn Pua and Mr Shandon Shi Kai Ng for performing the quality evaluations. The authors would also like to thank Mr Jovan Kris Kamadjaja, Mr Francis Tee Heng Chia, Mr Shandon Shi Kai Ng and Mdm Betty Gan for lending their smart devices for this study (Google Nest Mini, Samsung Galaxy Note 9, Echo Dot and iPhone 6S respectively).

Conflict of interest

No competing financial interests exist. All authors have no conflicts of interest with any of the evaluated apps or their 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. Kinsella B, Mutchler A. Voice assistant consumer adoption report. Voicebot.AI, 2018. Available from: https://voicebot.ai/wp-content/uploads/2018/11/voice-assistant-consumer-adoption-report-2018-voicebot.pdf. Accessed 25 Mar 2022.
2. National Public Media, Edison Research. The Smart Audio Report. National Public Media, 2020. Available from: https://www.nationalpublicmedia.com/insights/reports/smart-audio-report/. Accessed 25 Mar 2022.
3. Laricchia F. Number of voice assistants in use worldwide from 2019 to 2024 (in billions). Statista, 2022. Available from: https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/. Accessed 25 Mar 2022.
4. Laricchia F. Factors surrounding preference of voice assistants over websites and applications, worldwide, as of 2017. Statista, 2022. Available from: https://www.statista.com/statistics/801980/worldwide-preference-voice-assistant-websites-app/. Accessed 25 Mar 2022.
5. Kinsella B, Mutchler A. Voice assistant consumer adoption in healthcare. Voicebot.AI, 2019. Available from: https://voicebot.ai/wp-content/uploads/2019/10/voice_assistant_consumer_adoption_in_healthcare_report_voicebot.pdf. Accessed 25 Mar 2022.
6. Ferrand J, Hockensmith R, Houghton RF, et al. (2020) Evaluating smart assistant responses for accuracy and misinformation regarding human papillomavirus vaccination: Content analysis study. *J Med Internet Res* 22: e19018. https://doi.org/10.2196/19018
7. 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
8. Goh ASY, Wong LL, Yap KYL (2021) Evaluation of COVID-19 information provided by digital voice assistants. *Int J Digit Health* 1: 3. https://doi.org/10.29337/ijdh.25
9. 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

10. 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
11. Bickmore TW, Trinh H, Olafsson S, et al. (2018) 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* 20: e11510. https://doi.org/10.2196/11510
12. Bérubé C, Schachner T, Keller R, et al. (2021) Voice-based conversational agents for the prevention and management of chronic and mental health conditions: Systematic literature review. *J Med Internet Res* 23: e25933. https://doi.org/10.2196/25933
13. World Health Organization. Global report on diabetes. World Health Organization, 2016. Available from: https://www.who.int/publications/i/item/9789241565257. Accessed 25 Mar 2022.
14. World Health Organization. New WHO Global Compact to speed up action to tackle diabetes. World Health Organization, 2021. Available from: https://www.who.int/news/item/14-04-2021-new-who-global-compact-to-speed-up-action-to-tackle-diabetes. Accessed 25 Mar 2022.
15. Ow Yong LM, Koe LWP (2021) War on diabetes in Singapore: A policy analysis. *Health Res Policy Syst* 19: 15. https://doi.org/10.1186/s12961-021-00678-1
16. Amazon. Alexa developer document—Design your skill. Amazon, (n.d.). Available from: https://developer.amazon.com/en-US/docs/alexa/design/design-your-skill.html. Accessed 25 Mar 2022.
17. Google Developers. Integrate with Google Assistant. Google, (n.d.). Available from: https://developers.google.com/assistant. Accessed 25 Mar 2022.
18. Martineau P. Alexa, What's my blood-sugar level? Wired, 2019. Available from: https://www.wired.com/story/alexa-whats-my-blood-sugar-level/. Accessed 27 Mar 2023.
19. One Drop. Alexa skills: One Drop. Amazon, (n.d.). Available from: https://www.amazon.com/One-Drop/dp/B072QDCSQH. Accessed 25 Mar 2022.
20. DataMystic. Alexa skills: My Sugar by Jade Diabetes. Amazon, (n.d.). Available from: https://www.amazon.com/My-Sugar-by-Jade-Diabetes/dp/B0874717X2/ref=sr_1_3?dchild=1&keywords=diabetes&qid=1627894652&s=digital-skills&sr=1-3. Accessed 25 Mar 2022.
21. Epad Inc. Google Assistant: Diabetes Tips. Google, (n.d.). Available from: https://assistant.google.com/services/a/uid/00000063a0945bf1?hl=en-US. Accessed 25 Mar 2022.
22. Epad Incc. Google Assistant: Diabetes Checkup. Google, (n.d.). Available from: https://assistant.google.com/services/a/uid/000000e2388921d2?hl=en-US. Accessed 25 Mar 2022.
23. DietLabs. Google Assistant: Well With Diabetes. Google, (n.d.). Available from: https://assistant.google.com/services/a/uid/0000000a181531b7?hl=en-US. Accessed 25 Mar 2022.
24. Roche. Alexa skills: Sulli the Diabetes Guru. Amazon, (n.d.). Available from: https://www.amazon.com/Roche-Sulli-the-Diabetes-Guru/dp/B08BLTFY75. Accessed 25 Mar 2022.
25. eHealth Support Networks. Alexa skills: My Diabetes Lifestyle. Amazon, (n.d.). Available from: https://www.amazon.com/eHealth-Support-Networks-Diabetes-Lifestyle/dp/B08V8DGL69/ref=sr_1_1?dchild=1&keywords=diabetes&qid=1625367230&s=digital-skills&sr=1-1. Accessed 25 Mar 2022.

26. Heifner M. Sulli The Diabetes Guru: Your diabetes voice assistant, 2020. Available from: https://beyondtype2.org/sulli-the-diabetes-guru/. Accessed 25 Mar 2022.
27. Schachner T, Keller R, Wangenheim FV (2020) Artificial intelligence-based conversational agents for chronic conditions: Systematic literature review. *J Med Internet Res* 22: e20701. https://doi.org/10.2196/20701
28. 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
29. Cheng A, Raghavaraju V, Kanugo J, et al. (2018) Development and evaluation of a healthy coping voice interface application using the Google home for elderly patients with type 2 diabetes. *2018—15th IEEE Annual Consumer Communications and Networking Conference (CCNC)*. https://doi.org/10.1109/CCNC.2018.8319283
30. Akturk HK, Snell-Bergeon JK, Shah VN (2021) Continuous glucose monitor with Siri integration improves glycemic control in legally blind patients with diabetes. *Diabetes Technol Ther* 23: 81–83. https://doi.org/10.1089/dia.2020.0320
31. Maharjan B, Li J, Kong J, et al. (2019) Alexa, what should I eat?: A personalized virtual nutrition coach for native American diabetes patients using Amazon's smart speaker technology. *2019—IEEE International Conference on E-Health Networking, Application and Services, (HealthCom)*. https://doi.org/10.1109/HealthCom46333.2019.9009613
32. Statista Research Department. Worldwide desktop market share of leading search engines from January 2010 to July 2022. Statista, 2022. Available from: https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/. Accessed 3 Sep 2022.
33. Yap KYL, 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
34. Health On The Net (HON) Foundation. HONcode certification. Health On The Net, 2020. Available from: https://myhon.ch/en/certification.html. Accessed 3 Sep 2022.
35. 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://a-f-r.org/wp-content/uploads/sites/3/2016/01/1998-Radcliffe-Medical-Press-Quality-criteria-for-consumer-health-information-on-treatment-choices.pdf. Accessed 27 Mar 2023.
36. Charnock D, Shepperd S, Needham G, et al. (1999) DISCERN: An instrument for judging the quality of written consumer health information on treatment choices. *J Epidemiol Community Health* 53: 105–111. https://doi.org/10.1136/jech.53.2.105
37. 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
38. Shoemaker SJ, Wolf MS, Brach C (2014) Development of the Patient Education Materials Assessment Tool (PEMAT): A new measure of understandability and actionability for print and audiovisual patient information. *Patient Educ Couns* 96: 395–403. https://doi.org/10.1016/j.pec.2014.05.027
39. Shoemaker SJ, Wolf MS, Brach C (2013) The Patient Education Materials Assessment Tool (PEMAT) and User's Guide. Available from: https://www.ahrq.gov/health-literacy/patient-education/pemat.html

40. National Library of Medicine. Evaluating Internet Health Information Tutorial. MedlinePlus, 2020. Available from: https://medlineplus.gov/webeval/intro1.html. Accessed 3 Sep 2022.
41. Tan RY, Pua AE, Wong LL, et al. (2021) Assessing the quality of COVID-19 vaccine videos on video-sharing platforms. *Explor Res Clin Soc Pharm* 2: 100035. https://doi.org/10.1016/j.rcsop.2021.100035
42. Ministry of Health Singapore. Your Diabetes Questions Answered. HealthHub, 2020. Available from: https://www.healthhub.sg/live-healthy/1392/your-diabetes-questions-answered. Accessed 9 Aug 2021.
43. American Diabetes Association. Frequently Asked Questions: COVID-19 and Diabetes—How COVID-19 impacts people with diabetes. American Diabetes Association. Available from: https://www.diabetes.org/coronavirus-covid-19/how-coronavirus-impacts-people-with-diabetes. Accessed 3 Seo 2022.
44. Google. FAQ about Google Trends data—Trends Help. Available from: https://support.google.com/trends/answer/4365533?hl=en&ref_topic=6248052. Accessed 7 Aug 2021.
45. AnswerThePublic. Search listening tool for market, customer & content research. Available from: https://answerthepublic.com/. Accessed 7 Aug 2021.
46. American Diabetes Association. American Diabetes Association—Connected for Life. Available from: https://diabetes.org/. Accessed 20 Oct 2021.
47. Centers for Disease Control and Prevention. Diabetes. Centers for Disease Control and Prevention, 2021. Available from: https://www.cdc.gov/diabetes/index.html. Accessed 20 Oct 2021.
48. National Institute of Diabetes and Digestive and Kidney Diseases. Diabetes. National Institute of Diabetes and Digestive and Kidney diseases (NIDDK). Available from: https://www.niddk.nih.gov/health-information/diabetes. Accessed 20 Oct 2021.
49. Cleveland Clinic. Diabetes: An overview. Cleveland Clinic, 2021. Available from: https://my.clevelandclinic.org/health/diseases/7104-diabetes-mellitus-an-overview. Accessed 20 Oct 2021.
50. US Food and Drug Administration. Food and Drug Administration homepage. US FDA. Available from: https://www.fda.gov/. Accessed 20 Oct 2021.
51. Diabetes UK. Diabetes UK—Know diabetes. Fight diabetes. Diabetes UK. Available from: https://www.diabetes.org.uk/. Accessed 20 Oct 2021.
52. UK National Health Service. Diabetes—NHS. National Health Service. Available from: https://www.nhs.uk/conditions/diabetes/. Accessed 20 Oct 2021.
53. Victoria State Government. Better Health Channel—Diabetes. Available from: https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/diabetes. Accessed 20 Oct 2021.
54. Diabetes Australia. Diabetes Australia homepage. Available from: https://www.diabetesaustralia.com.au/. Accessed 20 Oct 2021.
55. Ministry of Health Singapore. HealthHub Health Services. Available from: https://www.healthhub.sg/. Accessed 20 Oct 2021.
56. National University Health System. National University Hospital: Home. National University Hospital. Available from: https://www.nuh.com.sg/Pages/Home.aspx. Accessed 20 Oct 2021.

57. Royal Australian College of General Practitioners, Diabetes Australia. Management of type 2 diabetes: A handbook for general practice. Royal Australian College of General Practitioners, 2020. Available from: https://www.diabetesaustralia.com.au/wp-content/uploads/Available-here.pdf. Accessed 20 Oct 2021.
58. Ministry of Health Singapore. Diabetes Mellitus—MOH Clinical Practice Guidelines 1/2014. Ministry of Health, 2014. Available from: https://www.moh.gov.sg/docs/librariesprovider4/guidelines/cpg_diabetes-mellitus-booklet---jul-2014.pdf. Accessed 20 Oct 2021.
59. Diabetes Care. Introduction: Standards of Medical Care in Diabetes—2021. Diabetes Care, 2021. Available from: https://care.diabetesjournals.org/content/44/Supplement_1. Accessed 20 Oct 2021.
60. National Institute for Health and Care Excellence. Type 1 diabetes in adults: Diagnosis and management—NICE guideline. National Institute for Health and Care Excellence, 2021. Available from: https://www.nice.org.uk/guidance/ng17. Accessed 20 Oct 2021.
61. Cosentino F, Grant PJ, Aboyans V, et al. (2020) 2019 ESC Guidelines on diabetes, pre-diabetes, and cardiovascular diseases developed in collaboration with the EASD: The Task Force for diabetes, pre-diabetes, and cardiovascular diseases of the European Society of Cardiology (ESC) and the European Association for the Study of Diabetes (EASD). *Eur Heart J* 41: 255–323. https://doi.org/10.1093/eurheartj/ehz486
62. American Diabetes Association (2021) 6. Glycemic targets: Standards of medical care in diabetes—2021. *Diabetes Care* 44: S73–S84. https://doi.org/10.2337/dc21-S006
63. American Diabetes Association. Myths about Diabetes. American Diabetes Association. Available from: https://www.diabetes.org/diabetes-risk/prediabetes/myths-about-diabetes. Accessed 3 Sep 2022.
64. Centers for Disease Control and Prevention. Flu & people with diabetes. Centers for Disease Control and Prevention, 2022. Available from: https://www.cdc.gov/flu/highrisk/diabetes.htm. Accessed 3 Sep 2022.
65. Rahmatizadeh S, Valizadeh-Haghi S (2018) Evaluating the trustworthiness of consumer-oriented health websites on diabetes. *Libr Philos Pract,* 1786.
66. Keselman A, Arnott Smith C, Murcko AC, et al. (2019) Evaluating the quality of health information in a changing digital ecosystem. *J Med Internet Res* 21: e11129. https://doi.org/10.2196/11129
67. National Library of Medicine. MedlinePlus evaluating internet health information: A tutorial National Library of Medicine, 2018. Available from: https://medlineplus.gov/webeval/EvaluatingInternetHealthInformationTutorial.pdf. Accessed 3 Sep 2022.
68. Smith DA (2020) Situating Wikipedia as a health information resource in various contexts: A scoping review. *PLoS One* 15: e0228786. https://doi.org/10.1371/journal.pone.0228786
69. Google Developers. Policies for actions on Google. Google Developers. Available from: https://developers.google.com/assistant/console/policies/general-policies. Accessed 3 Sep 2022.
70. Amazon.com Inc. Alexa developer documentation: Policy requirements. Amazon. Available from: https://developer.amazon.com/en-US/docs/alexa/custom-skills/policy-testing-for-an-alexa-skill.html. Accessed 27 Mar 2023.

71. Savitha S, Hirsch IB (2014) Personalized diabetes management: Moving from algorithmic to individualized therapy. *Diabetes Spectrum* 27: 87–91. https://doi.org/10.2337/diaspect.27.2.87
72. Patel N. 6 timely SEO strategies and resources for voice search. Available from: https://neilpatel.com/blog/seo-for-voice-search/. Accessed 3 Sep 2022.
73. Shafiee G, Mohajeri-Tehrani M, Pajouhi M, et al. (2012) The importance of hypoglycemia in diabetic patients. *J Diabetes Metab Disord* 11: 17. https://doi.org/10.1186/2251-6581-11-17
74. Diabetes.co.uk. Hypoglycemia (low blood glucose levels). Diabetes.co.uk, 2022. Available from: https://www.diabetes.co.uk/Diabetes-and-Hypoglycaemia.html#:~:text=Diabetes%20UK%20recommend%20that%20you,a%20non%2Ddiet%20soft%20drink. Accessed 27 Mar 2023.
75. Centers for Disease Control and Prevention. How to treat low blood sugar (Hypoglycemia). Centers for Disease Control and Prevention, 2021. Available from: https://www.cdc.gov/diabetes/basics/low-blood-sugar-treatment.html. Accessed 3 Sep 2022.
76. Amazon.com Inc. What is natural language understanding? Amazon. Available from: https://developer.amazon.com/en-US/alexa/alexa-skills-kit/nlu. Accessed 3 Sep 2022.
77. Kinsella B. Voice assistant demographic data—Young consumers more likely to own smart speakers while over 60 bias toward Alexa and Siri. Voicebot.ai, 2019. 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 3 Sep 2022.
78. Cherney K, Wood K. Age of onset for type 2 diabetes: Know your risk. Healthline, 2022. Available from: https://www.healthline.com/health/type-2-diabetes-age-of-onset#age-at-diagnosis. Accessed 3 Sep 2022.