

Article

Not peer-reviewed version

---

# Evaluation of Replies to Voice Queries in Gynecologic Oncology by Virtual Assistants Siri, Alexa, Google, and Cortana

---

Jamie M Land , [Edward J. Pavlik](#) \* , [Elizabeth Ueland](#) † , Sara Ueland † , Nicholas Per , Kristen Quick , Justin W. Gorski , McKayla J. Riggs , [Megan L. Hutchcraft](#) , [Josie D. Llanora](#) , Do Hyun Yun

Posted Date: 6 May 2023

doi: [10.20944/preprints202305.0393.v1](https://doi.org/10.20944/preprints202305.0393.v1)

Keywords: Virtual assistants; Gynecologic; Oncology; Siri; Alexa; Google; Cortana; Validity



Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## Article

# Evaluation of Replies to Voice Queries in Gynecologic Oncology by Virtual Assistants Siri, Alexa, Google, and Cortana

Jamie M. Land <sup>1</sup>, Edward J. Pavlik <sup>2,\*</sup>, Elizabeth Ueland <sup>3,†</sup>, Sara Ueland <sup>4,†</sup>, Nicholas Per <sup>2</sup>, Kristen Quick <sup>2</sup>, Justin W. Gorski <sup>2</sup>, McKayla J. Riggs <sup>2</sup>, Megan L. Hutchcraft <sup>2</sup>, Josie D. Llanora <sup>1</sup> and Do Hyun Yun <sup>1</sup>

<sup>1</sup> University of Kentucky College of Medicine, Lexington, KY 40506, USA; jamie.land@uky.edu (J.M.L.); josie.llanora@uky.edu (J.D.L.); daniel.yun@uky.edu (D.H.Y.)

<sup>2</sup> Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, University of Kentucky Chandler Medical Center-Markey Cancer Center, Lexington, KY 40536-0293, USA; nicholas\_per@trihealth.com (N.P.); kequ228@luky.onmicrosoft.com (K.Q.); Justin.gorski@gmail.com (J.W.G.); zmbj6@goldmail.etsu.edu (M.J.R.); megan.hutchcraft@uky.edu (M.L.H.)

<sup>3</sup> Scripps College, Claremont, CA 91711, USA; epueland@gmail.com

<sup>4</sup> Claremont McKenna College, Claremont, CA 91711, USA; sueland22@students.claremontmckenna.edu

\* Correspondence: edward.pavlik@uky.edu; Tel.: +1-(859)-323-3830

† Equal Contribution: These authors contributed equally to this work.

**Abstract:** 1) Background: This study focused on the validity of audible replies to voice queries in gynecologic oncology by Siri, Alexa, Google, and Cortana virtual assistants (VAs).; (2) Methods: 21 evaluators analyzed VA audible answers to voice queries related to gynecologic oncology. A 3-tier template was utilized for each voice query: "X?" (A), "What is X?" (B), and "Define X?" (C) in a 24-question panel, allowing questions to be posed in different formats to assess the VAs at a greater depth. Responses were scored using a rubric designed to assess the validity and quality of each answer; (3) Results: For general queries, Google provided the most correct audible replies ( $n_{correct} = 20$ ; 83.3% correct), followed by Alexa ( $n_{correct} = 16$ ; 66.7% correct), Siri ( $n_{correct} = 11$ ; 45.8% correct), and Cortana ( $n_{correct} = 5$ ; 20.8% correct.); For Gynecologic Oncology related queries, Google also provided the most correct audible replies ( $n_{correct} = 222$ ; 18.14%), followed by Alexa ( $n_{correct} = 75$ ; 6.51%), Siri ( $n_{correct} = 55$ ; 5.46%), and Cortana ( $n_{correct} = 23$ ; 2.28%). (4) Conclusions: The audible replies to oral queries on gynecologic oncology topics varied between Siri, Alexa, Google, and Cortana. Google's audible reply accuracy was superior to audible replies by Siri, Alexa, and Cortana. Since there is room for considerable improvement in VA performance, the authors advise caution when using VAs for medical queries in gynecologic oncology.

**Keywords:** virtual assistants; gynecologic; oncology; Siri; Alexa; Google; Cortana; validity

## 1. Introduction

The internet is widely used by patients to obtain health information [1]. Voice technology allows internet searches through verbal queries which are answered by a virtual assistant. Virtual assistants (VAs), such as Siri (Apple), Alexa (Amazon), Cortana (Microsoft), and the Google Assistant, are used ubiquitously. Google Assistant deploys searches on Google, as does Siri, while Alexa and Cortana use Bing as their search engine [2,3]. In 2020, Google Assistant was available on >1 billion devices and was used by > 500 million users monthly [4] providing 27% of all global web searches. More than 500 million Apple customers use Siri as a virtual assistant [5]. The Amazon Echo Home Speaker alone has >40 million users in the United States, which uses Alexa VA [6]. Siri, Alexa, Google and Cortana have been voted in the top 10 best VAs, hence why we have chosen these VAs to investigate [7]. Since the

introduction of Siri as a feature of the iPhone 4S in 2011, VAs have become mainstream. Siri can make phone calls or send text messages when users cannot manually enter information while driving, biking, or walking [8]. VAs have utility for the elderly in situations that are simplified by speaking especially when there is physical impairment [9]. There have been steady declines in the personal computer market for the last eight years as people opt to go online with smartphones where voice queries are easier than keying in searches [10,11]. A surge in the use of VAs is underway in healthcare. VA usage increased to 21.0% of U.S. adults in 2021 with 54.4 million people using VAs for questions in healthcare about symptoms, medical information, and treatments [12]. VA responses to voice queries have been utilized for information on postpartum depression [12] healthy lifestyles [13], vaccinations [14], addiction [15], mental health, and interpersonal violence [16]. Several US hospitals have smart speakers installed in patients' rooms enabling them to make requests that can relieve the clinical staff [17]. Widespread access and utilization can be contrasted by the quality of the results received from VA's, especially in extending simple everyday voice queries to more complex questions related to GYN-ONC. The literature for VA queries in oncology or gynecologic oncology (GYN-ONC) does not exist, thus this paper investigates for the first time the degree to which voice queries related to GYN-ONC are accurately addressed by audible replies from Siri, Alexa, Google, and Cortana.

## 2. Materials and Methods

Power was calculated between the VAs using McNemar's test for comparing the percent correct between two sources of information. The McNemar's test was used because it was determined to be the best fit for our data set in which the data was paired and proportions were different. To calculate power, VA performance was estimated based on general queries made by two investigators as shown in the "Percent correct responses" line of Table 1 (bottom line). In this setting, Google was correct (83.3%) more often than Siri (45.8%: P value = 0.27, power = 0.90) and Cortana (20.8%: P value = 0.0003, power = 0.99) but not Alexa, and Alexa (66.7%) was correct more often than Cortana (20.8%: P value = 0.0023, power = 0.98). By enlarging the data set to additional rounds of 24 queries, assuming the results obtained in Table 1 would continue to hold, then to detect a difference between Google and Alexa, 90 queries would attain 90% power and this would hold for Alexa vs Siri, as well as Siri vs Cortana. To approach this power, 21 evaluators each presented a set of 24 questions specific to GYN ONC in querying the VAs. The general questions used for estimating power were selected from QUIZ Daily (1550 Larimer Street, Suite 431, Denver, CO 80202) and checked against Google and Bing for matching correct answers. Our baseline evaluation of VA performance was within 1-3% of the performance reported by others, except for Cortana for which our estimate was lower (20.8% vs 45%) [18,19]. We utilized a team of 21 evaluators making queries in anticipation of variability due to different devices and in the interpretation of VA replies by individuals that were provided with correct answers at the risk of the study being over-powered. Evaluators were chosen at random. Evaluators dialect, skills, and speech characteristics all varied, but we do not believe this alters the data because each VA is equipped to assess over 100 languages and dialects. Each evaluator accessed the most recent updated versions of Siri (iOS 14), Alexa (version 2.2.427375.0), and Google (version 1.9.28702) on their smartphones. To access Cortana (version 3.2106.14307.0), a Dell Latitude E6430 Windows 10 laptop was provided to each evaluator that did not have access to Windows 10, while evaluators who already had access completed the test on their own device. The panel of 24 questions related to GYN-ONC were posed in three different templates to all the VAs by 21 evaluators. A 3-tier template was used for each inquiry: "X?" (A), "What is X?" (B), and "Define X?" (C) in order to maximize the chances of a correct answer by each VA. The questions asked to each VA are listed in Table 2 in the 3-tier template. We weighted the 24 question panel toward ovarian cancer queries (16 questions) with the premise that the lower incidence of ovarian cancer might provide a more robust test of the VAs than more common gynecologic malignancies. In addition, six queries related to more prevalent cervical cancers were included. The remaining two questions were aimed at borderline epithelial tumors of the ovary and endometrial cancer. All questions are based on an expectation of correct knowledge by a fellow in training in GYN ONC.

Individuals making queries were provided answers to queries against which responses returned by VAs were graded. Correct answers were determined from consensus recommendations published by the Society of GYN-ONC, the GYN-ONC Group, and the American College of Obstetrics and Gynecology. Audible answers from the VAs were scored as incorrect = 0; does not understand or know or returns only web-links = 1, <40% correct = 2, 40-50% correct = 3, 50-85% correct = 4, 100% correct = 5.

**Table 1. Responses to General Questions Presented to Virtual Assistants.** General Questions that were verbally asked to the VAs are identified. Chi square and Fisher's Exact Probability tests showed that Siri and Cortana underperformed Google ( $p < .05$ ), and that Cortana underperformed Alexa ( $p < .05$ ). The performance of Google and Alexa were not significantly different.

Question: YES = Correct Answer, NO = No audible answer or incorrect	Google	Alexa	Siri	Cortana
What is the current temperature?	YES	YES	YES	NO
What is the weather forecast?	YES	YES	YES	YES
What is the weather tomorrow?	YES	YES	YES	YES
What is \$5 US in Euros?	YES	YES	YES	YES
What is a flock of flamingos known as?	YES	YES	NO	NO
Which actress has the most Oscar nominations without a win?	YES	YES	NO	NO
What is the world's largest body of water?	YES	YES	YES	NO
What river feeds the Dead Sea?	YES	YES	YES	NO
What is the minimum age for a U.S. President?	YES	YES	YES	NO
What popular vegetable is poisonous if eaten raw?	NO	NO	NO	NO
Which cathedral has more statues than any other in the world?	YES	NO	NO	NO
What President appointed a former President to the Supreme Court?	YES	YES	NO	NO
Who was awarded the highest-ever rank in the U.S. armed forces?	NO	YES	NO	NO
What is the smallest planet in our solar system?	YES	YES	YES	NO
Who was the drummer for the Beatles?	YES	YES	YES	NO
Which state produces nearly half of all U.S. mushrooms?	YES	YES	NO	NO
Which President owned a haberdashery?	YES	NO	NO	NO
What is the name of the land bridge that once connected Asia and North America?	YES	NO	YES	NO
What is the only state with an official jelly?	YES	NO	NO	NO
Which artist is best known for their work in Cubism?	YES	YES	NO	NO
In which state is it illegal to serve margarine instead of butter in restaurants?	NO	NO	NO	YES
Which state's official dance is the polka?	NO	NO	NO	NO
Where is the longest fence in the world?	YES	YES	NO	YES
What does "Häagen-Dazs" mean?	YES	NO	YES	NO
<i>Total number of correct responses</i>	20	16	11	5
<i>Percent correct responses</i>	83.3%	66.7%	45.8%	20.8%

**Table 2. GYN-ONC Questions Asked to Virtual Assistants.** This table shows the questions that were asked to each VA in the three different formats, "X?" (A), "What is X?" (B), and "Define X?" (C). Sources of correct answers are hyperlinked to each query.

Question #	Question:	Correct Answer Link
1	A. Stage I ovarian cancer? B. What is stage I ovarian cancer? C. Define stage I ovarian cancer?	<a href="#">1</a>
2	A. Stage II ovarian cancer? B. What is stage II ovarian cancer? C. Define stage II ovarian cancer?	<a href="#">2</a>
3	A. Stage III ovarian cancer? B. What is stage III ovarian cancer? C. Define stage III ovarian cancer?	<a href="#">3</a>

4	A. Stage IV ovarian cancer? B. What is stage IV ovarian cancer? C. Define stage IV ovarian cancer?	<a href="#">4</a>
5	A. Stage IC1 ovarian cancer? B. What is stage IC1 ovarian cancer? C. Define stage IC1 ovarian cancer?	<a href="#">5</a>
6	A. Stage IIIA1 ovarian cancer? B. What is stage IIIA1 ovarian cancer? C. Define stage IIIA1 ovarian cancer?	<a href="#">6</a>
7	A. Stage IVB ovarian cancer? B. What is stage IVB ovarian cancer? C. Define stage IVB ovarian cancer?	<a href="#">7</a>
8	A. Subtypes of epithelial ovarian cancer? B. What are the subtypes of epithelial ovarian cancer? C. Define the subtypes of epithelial ovarian cancer?	<a href="#">8</a>
9	A. Screening for ovarian cancer? B. What is screening for ovarian cancer? C. Define screening for ovarian cancer?	<a href="#">9</a>
10	A. Screening recommendations for ovarian cancer? B. What are the screening recommendations for ovarian cancer? C. Define the screening recommendations for ovarian cancer?	<a href="#">10</a>
11	A. Ways to prevent ovarian cancer? B. What are ways to prevent ovarian cancer? C. Define the ways to prevent ovarian cancer?	<a href="#">11</a>
12	A. Symptoms of ovarian cancer? B. What are the symptoms of ovarian cancer? C. Define the symptoms of ovarian cancer?	<a href="#">12</a>
13	A. Hereditary ovarian cancer? B. What is hereditary ovarian cancer? C. Define hereditary ovarian cancer?	<a href="#">13</a>
14	A. Ovarian cancer risk reduction? B. What is ovarian cancer risk reduction? C. Define ovarian cancer risk reduction?	<a href="#">14</a>
15	A. Screening for cervical cancer? B. What is screening for cervical cancer? C. Define screening for cervical cancer?	<a href="#">15</a>
16	A. Screening recommendations for cervical cancer?	<a href="#">16</a>

	B. What are the screening recommendations for cervical cancer? C. Define the screening recommendations for cervical cancer?	
17	A. Options for a 20-year-old sexually active woman who requests a Pap smear? B. What are the options for a 20-year-old sexually active woman who requests a Pap smear? C. Define the options for a 20-year-old sexually active woman who requests a Pap smear?	<a href="#">17</a>
18	A. HPV vaccine? B. What is the HPV vaccine? C. Define the HPV vaccine?	<a href="#">18</a>
19	A. Ages for HPV vaccination? B. What are the ages for HPV vaccination? C. Define the ages for HPV vaccination?	<a href="#">19</a>
20	A. Three dose HPV vaccine recommendations? B. What are the three dose HPV vaccine recommendations? C. Define the three dose HPV vaccine recommendations?	<a href="#">20</a>
21	A. Borderline epithelial tumors of the ovary? B. What are borderline epithelial tumors of the ovary? C. Define borderline epithelial tumors of the ovary?	<a href="#">21</a>
22	A. Carcinosarcoma of the ovary? B. What is carcinosarcoma of the ovary? C. Define carcinosarcoma of the ovary?	<a href="#">22</a>
23	A. High grade serous tumors of the ovary? B. What are high-grade serous tumors of the ovary? C. Define high-grade serous tumors of the ovary?	<a href="#">23</a>
24	A. Stage IB endometrial cancer? B. What is stage IB endometrial cancer? C. Define stage IB endometrial cancer?	<a href="#">24</a>

### 3. Statistical Analysis

We calculated the overall intra-class correlation coefficient (ICC) using the results from a one-way ANOVA on Winstat (version 2012.1) and calculated individual coefficients of variation for each question, as well as results across VAs and templates. Finally, we quantified the responses across the VAs by aggregating the frequency of each score and have reported the respective results in percentages. Graded scores were averaged across all graders and expressed with the standard error of the mean (SEM). An average graded score  $\pm$  SEM was calculated across individuals that evaluated

responses by each VA. An average % score was determined as a percentage of the total score possible. Median scores and 75th percentiles were determined as a function of the total score possible. Minimum and maximum scores were determined as a function of the total score possible and the difference between maximum and minimum as a function of the total possible score. Counts of total correct answers to VA queries are expressed as a percentage of total queries. Significant differences were determined as  $p < .05$ .

Chi-square and Fisher's exact probability tests were used for non-parametric analyses, and Student's t-test was performed to compare means in the time-series study.

## 4. Results

Google provided the most correct audible replies with the general questions. ( $n_{correct} = 20$ ; 83.3% correct), followed by Alexa ( $n_{correct} = 16$ ; 66.7% correct), Siri ( $n_{correct} = 11$ ; 45.8% correct), and Cortana ( $n_{correct} = 5$ ; 20.8% correct). Differences in correct audible responses to the general voice queries occurred with Google > Alexa > Siri > Cortana (Table 1). Regarding queries related to GYN-ONC, Google's average graded score ( $2.88 \pm 0.04$ , Table 3(B) mean  $\pm$  SEM) was significantly higher ( $p < 0.01$ ) than the average graded scores for Alexa, Siri, and Cortana ( $1.60 \pm 0.04$ ,  $1.52 \pm 0.04$ ,  $1.28 \pm 0.03$ , Table 3(B)). The graded score for Cortana was statistically lower than the other VAs. When the average graded scores were expressed as a percent of the highest possible score, Google graded almost twice as high as the other VAs ( $57.6\% \pm 0.9$  vs.  $32\% \pm 0.8$ ,  $30.5\% \pm 0.8$ ,  $25.7\% \pm 0.6\%$ , Table 3(C)). Google's performance was mirrored by comparing scores at the median (Table 3(D)) and at the 75th percentile (Table 3(E)). Although the range displayed between the lowest and highest maximum scores was wide (Table 3(H)), the agreement between graders showed acceptable reliability with an ICC score of 0.525. Totally correct replies to Google queries yielded a significantly higher number of responses graded as totally correct ( $n = 222$ , 18.2%) than the other VAs (2.3-6.5%), Table 3(I). Examination of the query formats in terms of totally correct responses showed that there was no significant difference in the Google responses to the three formats, while Alexa, Siri, and Cortana had more correct responses to the "What is X" format,  $p < .05$  (Table 4). Consequently, query formats can influence VA responses to queries. The number of queries varied due to dissimilar access devices for VA applications. The total number of evaluators for Google were 17, 16 for Alexa, 14 for Siri, and 14 for Cortana. Nevertheless, the data still proves that Google provided the most accurate responses for both general queries and Gynecologic Oncology related questions.

**Table 3. Response Summary to Questions Related to GYN-ONC.** Graded scores were averaged across all graders and expressed with the standard error of the mean (SEM). The number of queries (A) varied due to dissimilar access devices for VA applications by different graders. Average Graded Score  $\pm$  standard error (SEM) (B) was calculated across individuals that evaluated responses by each VA. Average % Score was determined as a percentage of the total score possible (C). Medians (D) and 75th percentiles (E) were determined as a function of the total score possible. Minimum (F) and maximum (G) scores were determined as a function of the total score possible and the difference between maximum and minimum as a function of the total possible score (H). Count of total correct answers to VA queries for each VA and percentages are expressed as a percentage of total queries (I). Significantly different  $p < .05$  ANOVA<sup>a</sup> or Chi square<sup>b</sup>.

Google	1224	2.88 ±0.04 <sup>a</sup>	57.6±0.9%	60%	80%	23.6%	75.6%	51.9%	222 (18.2%) <sup>b</sup>
Alexa	1152	1.60±0.04	32.0±0.8%	20%	40%	24.7%	58.3%	33.6%	75 (6.5%)
Siri	1008	1.52±0.04	30.5±0.8%	20%	20%	23.6%	80.3%	57.2%	55 (5.5%)
Cortana	1008	1.28±0.03	25.7±0.6%	20%	20%	16.4%	43.3%	26.9%	23 (2.3%)

**Table 4. Analyses of Query Format Presented to Google, Alexa, Siri and Cortana.** Number of totally correct responses by each VA. Each percentage is based on the number of totally correct answers by each VA for the template queries: "X?", "What is X?", "Define X." N= total number of correct responses to Gynecologic Oncology related queries, responses with a grade of 5 (meaning 100% correct). Significance was determined both by the Chi square test and Fisher's Exact Probability test.

	Google	Alexa	Siri	Cortana
"X?"	74, 33.3%	20, 26.7%	12, 21.8%	7, 30.4%
"What is X?"	67, 30.2%	35, 46.7%	30, 54.5%	10, 43.5%
"Define X"	81, 36.5%	20, 26.7%	13, 23.6%	6, 26.1%
n =	222, 100%	75, 100%	55, 100%	23, 100%
	NS	P<.05	P<.05	P<.05

## 5. Discussion

In summary, audible replies by VAs to queries related to GYN-ONC have room for improved accuracy. Our findings support those of Bickmore et al. that patients should not rely solely on VAs to answer medical questions [20]. For the queries evaluated, a well-trained gynecologic oncologist would answer the queries with more accuracy than the VAs. Our results are consistent with surveys showing that the Google Assistant performed better than Siri or Alexa in addressing non-clinical queries [19]. VA responses related to family medicine revealed room for improvement [21], and did not provide addiction help [22]. We have not found reports on VA queries by GYN-ONC patients; however, virtual visits to gynecologic oncologists have increased during the pandemic [23] and should continue in the future [24]. A proof-of-concept utilization of a virtual assistant has been reported for creating a molecular tumor board in GYN-ONC to integrate automated methods in collaborative treatment decisions [25]. Certain GYN-ONC providers have already introduced "live chat" [26,27] which can be readily augmented with "voice chat" through specialized VAs. A data-driven approach, substituting virtual visits for in-person visits, has been suggested for the identification of symptoms related to ovarian cancer recurrence [28]. While not specific to GYN-ONC alone, the Alberta Health Service is the first public health care organization in Canada to offer health care information via voice queries on Google and Amazon devices [29]. In addition, the United Kingdom's National Health Service is partnering with Amazon Alexa, to answer health-related questions [30]. Amazon is also partnering with a telemedicine provider to start a voice-activated virtual program that prompts a call back from a telemedicine physician [30]. Our view is that the availability and utilization of VAs for medical information is increasing. This paper indicates that improvements are needed in the accuracy of GYN-ONC information provided by the VAs considered here. Differences in VA performance may be related to the search engine employed by the VA, as well as the proprietary artificial intelligence employed by each. The degree to which physical devices can clearly interpret speech without interference from background sounds is also a factor in VA performance. In the present study, 21 evaluators made 24 queries in three different formats, using different access devices and hence should mirror performance expected in the real world. The present

work provides an empirical standard for evaluating the reliability of information obtained from virtual assistants. Inaccuracy may be reduced by improvements at the search engine and VA artificial intelligence levels. Utilization of VAs internationally is a function of language and dialect [31]. Alexa supports 8 languages and 10 dialects, while Google Assistant supports 12 languages and 13 dialects and has been working on more than 115 languages capable of speech recognition and natural language understanding [32]. By early 2022, Google became conversant in 30 languages in 80 countries, with Siri supporting 21 languages in 36 countries, and Cortana supporting 8 languages in 16 countries [32]. VA availability in an expanding number of languages attests to the international relevance of the topic presented here.

## 6. Conclusions

In summary, audible replies by VAs to oral queries related to gynecologic oncology have considerable room for improved accuracy. The findings of this study support those of Bickmore *et al.*, showing that patients should not rely solely on VAs to answer medical questions [20]. Overall, we recommend caution when using VAs to obtain information in gynecologic oncology.

In recent news, Microsoft has announced that it will launch an AI-powered Bing search engine on the Edge browser to “deliver better search, more complete answers, a new chat experience and the ability to generate content” [33]. Google plans a rollout of its AI chatbot named BARD which is intended to enhance Google Search; however in a preview demonstration Google Bard provided erroneous information about discoveries made by the James Webb Space [34,35], while Microsoft’s AI-powered Bing has also generated false information [36,37]. At present neither Microsoft nor Google has associated these AI-technologies with receiving and responding to voice instructions. With new technology like these, further analysis of such VAs may be required when the technology has been improved. The new AI-powered Bing search Engine and Google BARD will need to be examined for their validity in answering healthcare questions, specifically as related to gynecologic oncology.

In addition to Microsoft’s AI-powered Bing search engine, OpenAI has released an AI system called Generative Pretrained Transformer 4 (GPT-4) that has a chat interface [38]. The chatbot gives a natural-language “response”, normally within 1 second that is relevant to the prompt [38]. This opens up a new conversation for how the chatbot could pertain to the medical field. Just like with the VAs, there is concern for how accurate the information that the chatbot gives is. According to an article in the New England Journal of Medicine, a false response given by GPT-4 is referred to as a “hallucination” [38]. These errors can be dangerous in medical scenarios because they can often be subtle and stated in a manner in which the chatbot is very convincing. These concerns are the same with the VAs, and thus it is extremely important to verify the information from both the VA and GPT-4.

**Author Contributions:** Conceptualization, E.P. and J.M.L.; methodology, J.M.L.; software, E.P. and J.M.L.; validation, E.P. and J.M.L.; formal analysis, E.P. and J.M.L.; investigation, J.M.L., E.U. and S.U.; resources, E.P.; data curation, J.M.L., E.U., S.U., N.P., K.Q., J.G., M.R., M.H., J.D.L., D.H.Y., E.P.; writing—original draft preparation, E.P., J.M.L.; writing—review and editing, J.M.L., E.U., S.U., N.P., K.Q., J.G., M.R., M.H., J.D.L., D.H.Y., E.P.; visualization, J.M.L.; supervision, E.P.; project administration, J.M.L.; funding acquisition, E.P. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** All data will be made available if requested.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Fox S, Duggan M. Health online 2013. *Pew Research Center* 2013. <https://www.pewresearch.org/internet/2013/01/15/health-online-2013/> (2013, accessed 1 November 2022).
2. Shah P. How to Change Siri's Search Engine (And Other Tricks). *Guiding Tech*. 2019. <https://www.guidingtech.com/change-siri-search-engine-tricks/> (2019, accessed 25 May 2022).
3. Snead A. What Search Engine Does Alexa Use? And Can I Use Google To... *Smarter Home Guide* 2020. <https://smarterhomeguide.com/alexa-search-engine/#:~:text=Alexa%20utilizes%20Bing's%20search%20engine%20for%20all%20of%20her%20search%20queries>. (2020, accessed 25 May 2022).
4. DBS Interactive. Voice Search Statistics and Emerging Trends- Voice Search Statistics and Emerging Trends. <https://www.dbswebsite.com/blog/trends-in-voice-search/> (Accessed 25 May 2022).
5. Georgiev D. 2022's Voice Search Statistics- Is Voice Search Growing? *Review* 42. <https://serpwatch.io/blog/voice-searchstatistics/#:~:text=Around%20500%20million%20people%20are,assistants%20worldwide%2C%20alongside%20Google%20Assistant>. (2022, accessed 1 November 2022).
6. SafeAtLast. Intriguing Amazon Alexa Statistics You Need to Know in 2022. <https://safeatlast.co/blog/amazon-alexa-statistics/#gref> (2022, accessed 25 May 2022).
7. McFarland A. 10 Best AI Assistants (November 2022). *Unite. AI*. <https://www.unite.ai/10-best-ai-assistants/>. (2022, accessed 1 November 2022).
8. CNBC. Here's how Siri made it onto your iPhone. <https://www.cnbc.com/2017/06/29/how-siri-got-on-the-iphone.html>. (2022, accessed 25 May 2022).
9. Yaghoubzadeh R, Kramer M, Pitsch K, et al. Virtual agents as daily assistants for elderly or cognitively impaired people. *Proceedings of the 13th International Conference on Intelligent Virtual Agents* 2013; 79–91. DOI:10.1007/978-3-642-40415-3\_7].
10. Phys Org. Personal computer sales fall for fifth year in a row. <https://phys.org/news/2017-01-personal-sales-fall-year-row.html> (2017, accessed 25 May 2022).
11. Roopinder T. The Desktop Computer Was in Decline. The Pandemic Made It Worse. *Engineering*. <https://www.engineering.com/story/the-desktop-computer-was-in-decline-the-pandemic-made-it-worse> (2020, 25 May 2022).
12. Yang S, Lee J, Sezgin E, et al. Clinical advice by voice assistants on postpartum depression: cross-sectional investigation using Apple Siri, Amazon Alexa, Google Assistant, and Microsoft Cortana. *JMIR Mhealth Uhealth* 2021. DOI: 10.2196/24045.
13. Kocaballi AB, Quiroz JC, Rezazadegan D, et al. Responses of conversational agents to health and lifestyle prompts: investigation of appropriateness and presentation structures. *J Med Internet Res* 2020.Feb 09;22(2):e15823.DOI: 10.2196/15823.Medline: 32039810.
14. 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. DOI: 10.1136/bmjhci-2019-100075. Medline: 31767629.
15. Nobles AL, Leas EC, Caputi TL, et al. Responses to addiction help-seeking from Alexa, Siri, Google Assistant, Cortana, and Bixby intelligent virtual assistants. *NPJ Digit Med* 2020;3:11. DOI:10.1038/s41746-019-0215-9. Medline: 32025572.
16. Miner AS, Milstein A, Schueller S, et al. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA Intern Med* 2016;176(5):619-625. DOI: 10.1001/jamainternmed.2016.0400. Medline: 26974260.
17. Medscape. Amazon's Alexa Is Now a Healthcare Provider. <https://www.medscape.com/viewarticle/968719> (2022, accessed 25 May 2022).
18. Hong G, Folcarelli A, Less J, et al. Voice Assistants and Cancer Screening: A Comparison of Alexa, Siri, Google Assistant, and Cortana. *Ann Fam Med* 2021;447-449. DOI: 10.1370/afm.2713. PMID: 34546951; PMCID: PMC8437561.
19. Laricchia, L. Share of questions answered correctly by selected digital assistants as of 2019, by category. *Statista*. <https://www.statista.com/statistics/1040539/digital-assistant-performance-comparison/> (2019, accessed 25 May 2022).
20. Bickmore TW, Trinh H, Olafsson S, 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.
21. Hong G, Folcarelli A, Less J, et al. Voice Assistants and Cancer Screening: A Comparison of Alexa, Siri, Google Assistant, and Cortana. *The Annals of Family Medicine* 2021;19(5)447-449; DOI: 10.1370/afm.2713.
22. Nobles, A.L., Leas, E.C., Caputi, T.L. et al. Responses to addiction help-seeking from Alexa, Siri, Google Assistant, Cortana, and Bixby intelligent virtual assistants. *npj Digit. Med* 2020;3:11; <https://doi.org/10.1038/s41746-019-0215-9>

23. McAlarnen A, Tsaih SW, Aliani R, et al. Virtual visits among gynecologic oncology patients during the COVID-19 pandemic are accessible across the social vulnerability spectrum. *Gynecol Oncol* 2021;162(1):4-11. DOI: 10.1016/j.ygyno.2021.04.037.
24. Mancebo G, Solé-Sedeño J, Membrive I, et al. Gynecologic cancer surveillance in the era of SARS-CoV-2 (COVID-19). *International Journal of Gynecologic Cancer* 2021; 31:914-919. <http://dx.doi.org/10.1136/ijgc-2020-001942>.
25. Macchia G, Ferrandina G, Patarnello S, et al. Multidisciplinary Tumor Board Smart Virtual Assistant in Locally Advanced Cervical Cancer: A Proof of Concept. *Front Oncol.* 2022; 11:797454. DOI: 10.3389/fonc.2021.797454.
26. Virtua Health. Gynecologic Oncology. <https://www.virtua.org/services/cancer-treatment/gynecologic-oncology> (accessed 25 May 2022).
27. Dignity Health. Treating Gynecologic Cancers. <https://www.dignityhealth.org/campaign-landers/gyn-oncology-surgery> (accessed 25 May 2022).
28. Feinberg J, Carthew K, Webster E, et al. Ovarian cancer recurrence detection may not require in-person physical examination: an MSK team ovary study. *Int J Gynecol Cancer* 2022; 32(2):159-164. DOI: 10.1136/ijgc-2021-002885.
29. Brown S. Partnerships between health authorities and Amazon Alexa raise many possibilities — and just as many questions. *CMAJ* 2019; 191(41)E1141-E1142; DOI: 10.1503/cmaj.1095799.
30. Associated Press. New York Post. You can now ask Amazon's Alexa to call you a doctor. <https://nypost.com/2022/02/28/amazons-voice-assistant-alexa-to-start-seeking-doctor-help/> (2022, 25 May 2022).
31. Summa Linguae. Language Support in Voice Assistants Compared. <https://summalinguae.com/language-technology/language-support-voice-assistants-compared/> (accessed 25 May 2022).
32. Wiggers K. Which voice assistant speaks the most languages, and why? *The Machine.* <https://venturebeat.com/2019/02/02/which-voice-assistant-speaks-the-most-languages-and-why/> (2019, accessed 25 May 2022).
33. Mehdi Y. Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web. *Official Microsoft Blog.* <https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/> (2023, accessed 14 February 2023).
34. Miao H. Alphabet Stock Drops 8% After Google Rollout of AI Search Features. *The Wall Street Journal* February 8, 2023 <https://www.wsj.com/livecoverage/stock-market-news-today-02-08-2023/card/alphabet-stock-drops-after-google-parent-introduces-ai-search-features-wgCJG3IDoSbfL3SgyrNI>
35. Martindale J. How to use Google Bard, the latest AI chatbot service. *DigitalTrends.* <https://www.digitaltrends.com/computing/how-to-use-google-bard/#dt-heading-what-question-did-google-bard-get-wrong> (2023, accessed 16 February 2023).
36. Hao K. What Is ChatGPT? What to Know About the AI Chatbot That Will Power Microsoft Bing. *The Wall Street Journal.* February 10, 2023 [https://www.wsj.com/articles/chatgpt-ai-chatbot-app-explained-11675865177?st=q4wbp2ercfh1zo3&reflink=share\\_mobilewebshare](https://www.wsj.com/articles/chatgpt-ai-chatbot-app-explained-11675865177?st=q4wbp2ercfh1zo3&reflink=share_mobilewebshare) (2023, accessed 16 February 2023).
37. Quach K. Microsoft's AI Bing also factually wrong, fabricated text during launch demo. *The Register.* [https://www.theregister.com/2023/02/14/microsoft\\_ai\\_bing\\_error/](https://www.theregister.com/2023/02/14/microsoft_ai_bing_error/) (2023, accessed 16 February 2023).
38. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *The New England Journal of Medicine.* 2023; 388:1233-1239. DOI: 10.1056/NEJMsr2214184.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.