Journal of Cancer Education (2019) 34:435–440
https://doi.org/10.1007/s13187-017-1319-z

A Systematic Assessment of Google Search Queries and Readability of Online Gynecologic Oncology Patient Education Materials

Alexandra Martin [1] · J. Ryan Stewart [1] · Jeremy Gaskins [2] · Erin Medlin [1]

Published online: 20 January 2018

Abstract

The Internet is a major source of health information for gynecologic cancer patients. In this study, we systematically explore common Google search terms related to gynecologic cancer and calculate readability of top resulting websites. We used Google AdWords Keyword Planner to generate a list of commonly searched keywords related to gynecologic oncology, which were sorted into five groups (cervical cancer, ovarian cancer, uterine cancer, vulvar cancer, vaginal cancer) using five patient education websites from sgo.org. Each keyword was Google searched to create a list of top websites. The Python programming language (version 3.5.1) was used to describe frequencies of keywords, top-level domains (TLDs), domains, and readability of top websites using four validated formulae. Of the estimated 1,846,950 monthly searches resulting in 62,227 websites, the most common was cancer.org. The most common TLD was *.com. Most websites were above the eighth-grade reading level recommended by the American Medical Association (AMA) and the National Institute of Health (NIH). The SMOG Index was the most reliable formula. The mean grade level readability for all sites using SMOG was 9.4 ± 2.3, with 23.9% of sites falling at or below the eighth-grade reading level. The first ten results for each Google keyword were easiest to read with results beyond the first page of Google being consistently more difficult. Keywords related to gynecologic malignancies are Google-searched frequently. Most websites are difficult to read without a high school education. This knowledge may help gynecologic oncology providers adequately meet the needs of their patients.

Keywords Gynecologic cancers · Google search terms · Websites · Readability · Online patient education materials

Introduction

The Internet is a widely used source of health information. Approximately 84% of American adults use the Internet [1]. Of users, 80% look for health information online [2], 87% use a search engine [2], and 83% use Google most often [3]. The Internet can allow gynecologic oncology patients to proactively participate in their healthcare, supplement education provided by the physician and provide resources unavailable within the bounds of the clinical encounter [4].

Productive utilization of information is dependent on the reader's ability to understand it. Health literacy is the degree to which individuals have the capacity to obtain, process, and understand the basic health information and services they need to make appropriate health decisions [5]. Decreased health literacy is associated with a reduced understanding of cancer screening procedures, increased distress about developing cancer, and cancer diagnosis in later stages [6]. The American Medical Association (AMA) and the National Institute of Health (NIH) recommend that readability of patient education materials should not exceed a seventh- to eighth-grade reading level [7]. One study found that the majority of ovarian cancer websites were written at an eleventh grade level on average [8].

Though Internet use is ubiquitous and gynecologic malignancies are increasing [9], there is a paucity of objective data on the Internet search habits of the gynecologic oncology patient population and the readability of online patient information [8, 10–12]. Knowledge of this data could assist providers in empowering patients to become better-informed consumers of healthcare information.

✉ Alexandra Martin
almart15@louisville.edu

1 Department of Obstetrics, Gynecology and Women's Health, University of Louisville School of Medicine, 529 South Jackson Street, Louisville, KY 40202, USA

2 Department of Bioinformatics and Biostatistics, University of Louisville School of Public Health and Information Sciences, Louisville, KY, USA

Our objective is to systematically explore the most common Google search terms related to gynecologic oncology and to calculate the readability of the top websites resulting from these searches.

Methods

A list of keywords and their estimated average monthly search volumes was generated using the Google AdWords Keyword Planner [13]. Five patient education websites from sgo.org were utilized as "seed" pages to allow the keywords to be sorted into five groups (cervical cancer, ovarian cancer, uterine cancer, vaginal cancer, and vulvar cancer). The authors reviewed the keywords and excluded those that were irrelevant or had fewer than 100 searches per month to maximize the applicability of our results to a wide population.

As it is overwhelmingly the most popular search engine [3], Google searches were then performed on each of the resultant keywords. A list of websites resulting from each keyword search within each group was then generated. The Python programming language (version 3.5.1) was used to calculate frequencies of keywords, top-level domains (TLDs), and domains for each keyword group.

The readability of the top 20 websites resulting from each keyword search was calculated using four validated readability formulae. The Flesch-Kincaid formula (F-K) was originally formulated by John P. Kincaid for use in the US Navy, is now widely used in the field of education, and is often included as a function in word-processing packages, such as Microsoft Word [14]. It is calculated by using the total number of words, sentences, and syllables in a given text to generate a score as a US grade level [14]. Readability of the top websites was also determined using the Dale-Chall formula (D-C), which calculates the US grade level of a text using sentence length and number of "hard" words as determined from a list of 3000 words determined to be familiar to most 4th grade students. It is often used to evaluate health information [15]. The Gunning Fog Index (GFI), which was originally published by Robert Gunning in *The Technique of Clear Writing* in 1952, is used widely by the military, businessmen, engineers, and scientists to help decrease the complexity of their writing. This score is calculated by using the average sentence length (number of words divided by the number of sentences) and the number of "complex" words (those with three or more syllables) to estimate the number of years of formal education a person needs to understand a given text [16]. Lastly, the Simple Measure of Gobbledygook (SMOG) Index was determined by counting the number of polysyllables (words with three or more syllables) in three ten-sentence samples, calculating the count's square root, and adding three to estimate the years of education a person needs to understand a particular text. The SMOG Index developed by G. Harry McLaughlin in 1969 as a derivation of the GFI was recommended by the US National Cancer Institute for the evaluation of cancer pamphlets [6]. It is a more stringent measure of readability demonstrating a strong correlation with the required reading level in multiple validation studies ($r = 0.985$) [17]. As opposed to other measures that are based on the ability of readers to answer a certain portion of questions correctly (usually 50 or 75%), the SMOG Index produces reading grade levels for individuals with 100% comprehension of the text. [13].

A large systematic review of readability instruments used to assess web-based cancer information by Freidman et al. concluded it is preferable to use multiple tools to ensure the validity of the readability scores [6]. For these reasons, we calculated readability using the four abovementioned formulas, but the SMOG Index was taken as the preferred readability measure for this study [17].

Readability scores were described using eighth grade as the ideal reading level based on the recommendations of the AMA and the NIH [7]. Pages with fewer than 100 total words were excluded due to instability in the readability scores. The scores were summarized using means and standard deviations and compared using one-way ANOVA with a significance level of $\alpha = 0.05$. Relationships between the four readability formulas were analyzed using Spearman's correlation. The relationship between readability as calculated by the SMOG Index and Google ranking was then analyzed using a random-effects regression model. A first stage model was used with rank and group as categorical predictors and keyword as a random effect. This fit suggested a piecewise linear regression with a different slope for the Google search ranking for ranks 1 through 10 (page 1) versus ranks 11 through 20 (page 2). This piecewise regression was fit and analyzed using group as a confounder and the random effect for keyword. All statistical analyses were performed using the R statistical software [18].

Results

There were 3133 total keywords in the five groups. Of these, 749 were excluded for irrelevance and 1761 excluded for fewer than 100 estimated monthly searches. A total of 623 keywords (248 cervical, 202 ovarian, 140 uterine, 31 vulvar, and 2 vaginal) were included in the final analysis, resulting in an estimated 1,846,950 total monthly searches (918,220 cervical; 517,620 ovarian; 163,600 uterine; 233,810 vulvar; 13,700 vaginal). A Google search of each keyword yielded a list of 62,227 websites (24,776 cervical; 20,160 ovarian; 13,993 uterine; 3098 vulvar; and 200 vaginal). Of these websites, 135 had fewer than 100 words and were excluded from the readability calculations due to instability in the scores. The readability of the remaining sites was then calculated using

F-K, D-C, GFI, and SMOG. A schematic of the study design is seen in Fig. 1.

The most commonly searched keywords in each group were "HPV" (368,000) for cervical cancer, "Ovarian cyst" and "Ovarian cancer" (135,000) for ovarian cancer, "Uterine cancer" (49,500) for uterine cancer, "Cancer" (201,000) for vulvar cancer, and "Vaginal cancer" (12,100) for vaginal cancer (Table 1).

Of the estimated 1,846,950 monthly searches resulting in 62,227 total websites, the most common TLD was *.com at 43%, and the most common domains were cancer.org (884), nih.gov (690), and cancerresearchuk.org (665) (Table 1).

We computed readability statistics for the most common pages within each keyword search. The mean grade level readability for all sites across all groups was 11.1 ± 6.4 (mean ± SD) using F-K, 8.5 ± 1.3 with D-C, 19.5 ± 6.6 with GFI, and 9.4 ± 2.3 using the SMOG Index. Few pages had readability at or below the recommended eighth grade level (F-K 15.8%; DC 35.7%; GFI 0%; SMOG 23.9%). When comparing the four readability formulas using Spearman's correlation, we find the SMOG Index to be the most reliable, due to its high correlation with the other formulae and the lower proportion of extreme outliers.

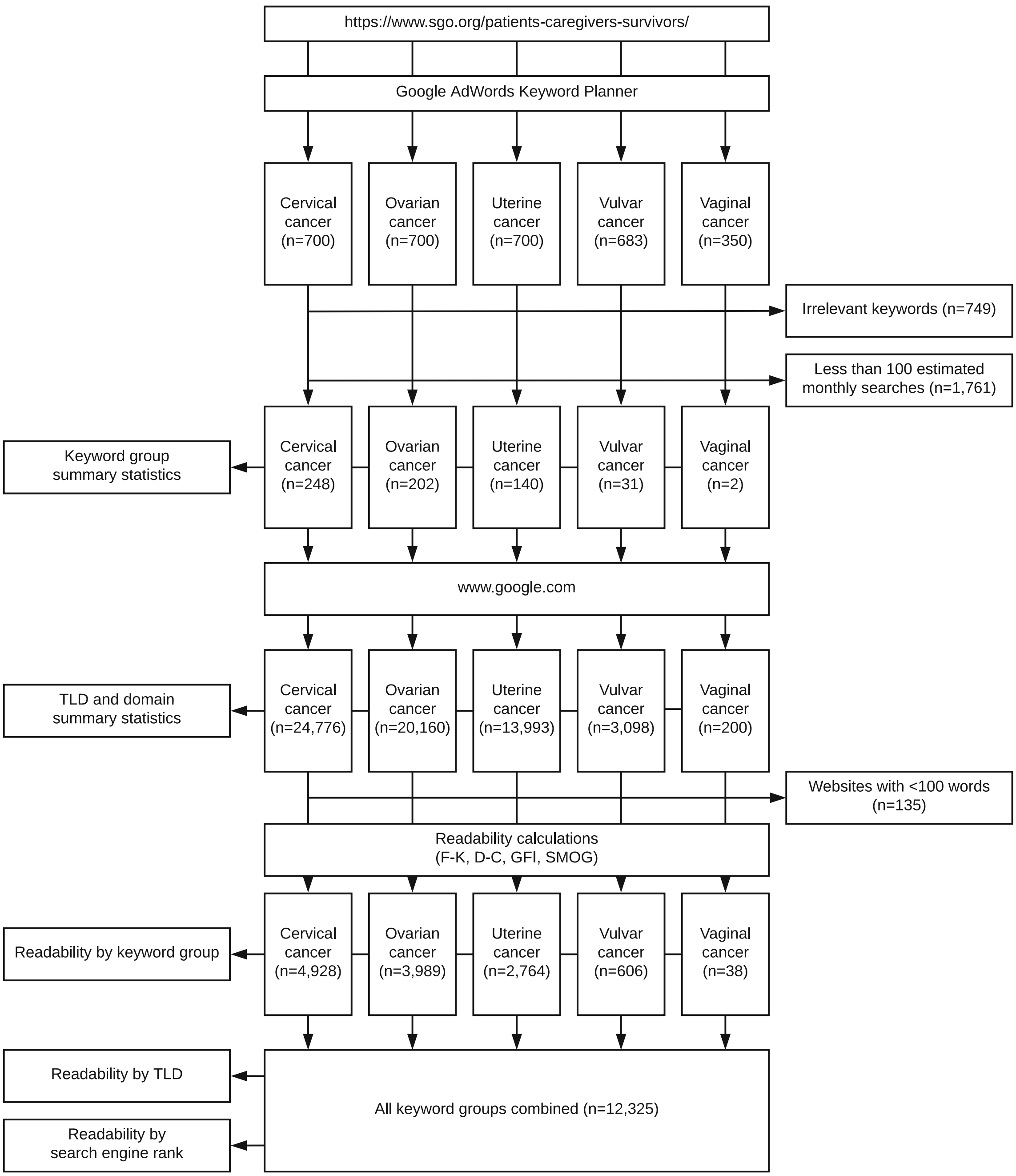

Fig. 1 Schematic of the study design

Table 1 Most common keywords, top-level domains (TLDs), domains, and their estimated frequencies

Keyword group	Total keywords	Total monthly searches	Most common keywords (# of monthly searches)	Most common TLDs (# of results)	Most common domains (# of results)
Cervical cancer	248	918,220	HPV (368,000) Cervical cancer (135,000) Pap smear (110,000)	.com (10,335) .org (7040) .gov (1650)	cancer.org (375) cancerresearchuk.org (291) webmd.com (275)
Ovarian cancer	202	517,620	Ovarian cyst (135,000) Ovarian cancer (135,000) Ovarian cancer symptoms (74,000)	.com (8605) .org (6948) .gov (863)	cancer.org (270) nih.gov (223) cancerresearchuk.org (216)
Uterine cancer	140	163,600	Uterine cancer (49,500) Endometrial cancer (33,100) Symptoms of uterine cancer (8100)	.com (6453) .org (4497) .gov (610)	cancer.org (192) nih.gov (168) cancer.gov (154)
Vulvar cancer	31	233,810	Cancer (201,000) Vulvar cancer (14,800) Cervical cancer treatment (4400)	.com (1220) .org (1043) .edu (207)	google.com (65) cancer.org (44) nih.gov (39)
Vaginal cancer	2	13,700	Vaginal cancer (12,100) Gynecological cancer (1600)	.org (97) .com (66) .gov (13)	google.com (4) [7 pages tied at 3]
Total	623	1,846,950	HPV (368,000) Cancer (201,000) Cervical cancer, ovarian cyst, and ovarian cancer (135,000)	.com (26,679) .org (19,623) .gov (3263)	cancer.org (884) nih.gov (690) cancerresearchuk.org (665)

Figure 2 displays the SMOG readability scores by keyword group. There were significant differences in readability across the groups (ANOVA, $p < 0.0001$) with uterine cancer websites being the most difficult to read (9.9 ± 2.5) and cervical cancer websites the easiest (8.9 ± 2.2). Comparing SMOG readability across the TLDs, we again find significant differences (ANOVA, $p < 0.0001$). Websites with the *.net TLD were easiest to read (8.8 ± 4.3), and *.gov were the most difficult (10.2 ± 2.4).

Finally, we sought to assess the relationship between the Google search ranking and readability. This piecewise linear regression model assumes the Google search rank may have a differing effect on SMOG score between the first ten and second ten ranks (first page vs. second page of Google). To account for differing difficulties between classes and keywords, we include class as a confounder and a random effect for keyword. The predicted best-fit line is shown in Fig. 3. The difficulty in readability rapidly increases over the first ten pages; each increase in rank corresponds to an increase in reading difficulty of 0.08 grade levels (95% CI 0.06–0.10, $p < 0.0001$). There is no appreciable difference in readability after page 10 ($p = 0.7534$).

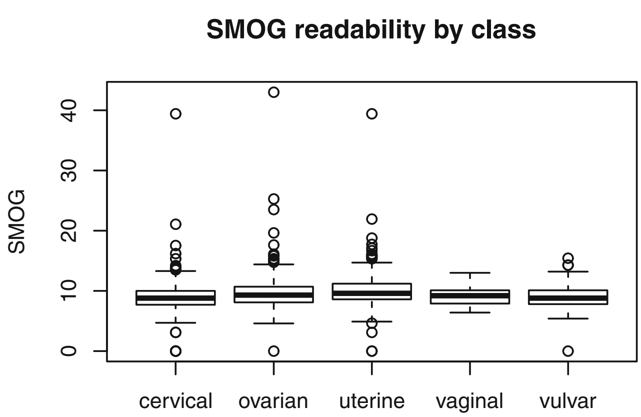

Fig. 2 Simple Measure of Gobbledygook (SMOG) Index readability scores by keyword group

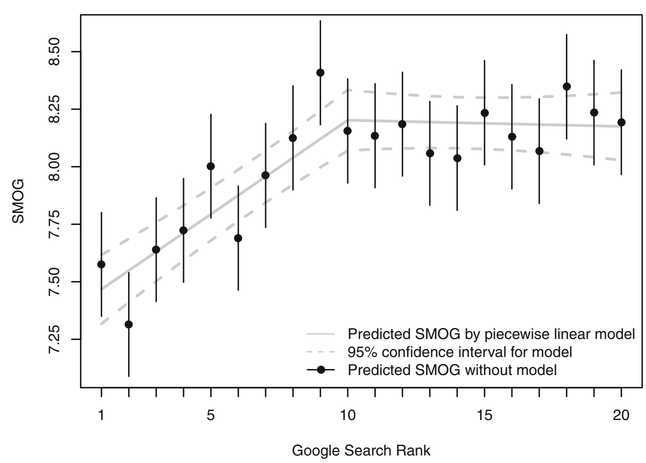

Fig. 3 Simple Measure of Gobbledygook (SMOG) Index readability scores by Google search rank. The gray line represents the predicted reading level by Google search ranking (for the reference class of cervical cancer). The dashed lines provide the 95% confidence interval. The black vertical lines represent the predicted difficulty by rank without using the piecewise linear model

Discussion

The Internet is undoubtedly a frequently consulted source for medical information and is gaining popularity [10]. Our study supports this phenomenon and makes it specific to the field of gynecologic oncology by demonstrating nearly 2 million monthly searches worldwide related to this topic. It also contributes to the body of evidence that much of online patient education material is difficult to read without a high school education.

There have been few attempts at characterizing the Internet search habits of the gynecologic oncology patient population. In a patient questionnaire study, 85% used the Internet as a gynecologic cancer resource with the most common search terms being cancer names (e.g., ovarian, endometrial, cervical) and specific treatments [12]. In 2013, keywords thought to be of interest to patients by two providers, and Google AdWords were used to show that "HPV" was searched approximately 305,400 per month; "ovarian cyst" 139,990; and "cervical cancer" 93,510 [10]. Our results using the same tool show that the number of monthly searches has increased in the last 3–4 years, but commonly searched keywords have remained similar.

A large systematic review of readability instruments used to assess web-based cancer information by Freidman, et al. concluded it is preferable to use multiple tools to ensure the validity of the readability scores [6]. For these reasons, we calculated readability using the four abovementioned formulas, but the SMOG Index was taken as the preferred readability measure for this study. Though all four readability formulae have good correlations, our data indicate that the SMOG Index is the most reliable, which is consistent with previous work comparing SMOG to other readability measures [19].

A nationwide movement to address health literacy has resulted in federal policies including the Affordable Care Act, The National Action Plan to Improve Health Literacy, and The Plain Writing Act of 2010, which aim to improve healthcare access and quality [5]. Providing easy-to-read patient materials is an important component of this initiative. In a Committee Opinion on Health Literacy, The American College of Obstetrics and Gynecology places the onus on providers to provide simple, clear patient information [5]. Despite this, there have been few reports regarding readability of gynecologic or gynecologic oncology patient information since 1999 [20].

Readability studies from other oncologic subspecialties show that difficult-to-read online patient information is not an issue specific to gynecologic oncology. The average readability of websites about colorectal cancer screening is 11.0 ± 2.2 [21], 13.36 (95% CI 12.83–13.89) for radiation oncology sites [22], and 11.2 ± 2.2 for kidney and bladder cancers sites [23].

Our data also indicates that the search results on the first page of Google are the easiest to read with increasing difficulty up to the top 10. Search results beyond the first page are consistently more challenging. This information can be used to counsel patients that they should limit themselves to the top 10 results when "Google-ing" information.

This study demonstrates a novel approach to the objective description of common Google searches related to gynecologic oncology by creating a curated list of keywords from a list generated by the most commonly used search engine. It is a robust study, evaluating more sites than done previously [8, 10–12]. It is the first study of readability in gynecologic oncology literature to evaluate websites for all five gynecologic malignancies and the first to use the SMOG formula [8].

It is limited in that we identified websites using Google exclusively and cannot draw conclusions regarding other search engines. However, since Americans use Google more than any other search engine [3], our results are likely applicable to most Internet users. Though higher-ranked results are viewed more commonly than lower-ranked ones [24], we cannot conclude that the "most common" sites are necessarily the most commonly viewed. We also cannot determine who is performing the searches or potentially viewing the websites (i.e., patients, families, friends, healthcare professionals). The readability calculations were done programmatically, and errors in calculation (likely based on text parsing) may go unnoticed, but our results are comparable to those found in similar studies [11, 19]. Lastly, though readability formulas provide a general idea of how difficult a document is to read, they do not measure the quality of information, its effectiveness as a communication tool, or the reader's comprehension. Comprehension levels are often at least two grades below reading level and drop even further in stressful situations [25], such as new cancer diagnoses. Future studies should evaluate the quality of online patient information and comprehension levels of readers.

Knowledge of common keywords, websites, and readability is important for gynecologic oncology providers. It can provide better insight into the needs and apprehensions of patients, allowing providers to tailor patient counseling and suggest Google search strategies that may provide more readable information. It could also help refine the provider vocabulary to match patient search language in an effort to minimize asymmetry of information. Creators of patient education materials should be more cognizant of readability, introduce guidelines based on the NIH and AMA recommendations, and use the SMOG Index to evaluate their text.

Acknowledgements The following abstract was displayed as a poster during the Western Association of Gynecologic Oncologists (WAGO) 2017 Annual Meeting. All presented WAGO abstracts are published in the October issue of *Gynecologic Oncology.*

Compliance with Ethical Standards

Conflict of Interest The authors have no conflicts of interest to disclose.

Ethics Approval and Consent to participate The study was reviewed by the University of Louisville Institutional Review Board and deemed exempt from approval, as it does not meet the "Common Rule" definition of human subjects' research.

References

1. Perrin A, Duggan M (2015) Americans' Internet access: 2000–2015. Pew Research Center Numbers, Facts and Trends Shaping the World:1–12
2. Fox S (2011) Health topics. Pew Research Center Pew Internet and American Life Project
3. Purcell K, Brenner J, Lee R (2012) Search engine use 2012. Pew Research Center Pew Internet and American Life Project (http://pewinternet.org/Reports/2012/Search-Engine-Use-2012.aspx)
4. Lee K, Kreshnik H, Jeffery H, Lynne E (2014) Dr. Google and the consumer: a qualitative study exploring the navigational needs and online health information-seeking behaviors of consumers with chronic health conditions. J Med Internet Res 16(12):1
5. Gynecologists, American College of Obstetricians and (2016) Health literacy to promote quality of care. Committee Opinion No. 676. Obstet Gynecol 128:183–186
6. Friedman DB, Hoffman-Goetz L (2006) A systematic review of readability and comprehension instruments used for print and web-based cancer information. Heal Educ Behav 33(3):352–373. https://doi.org/10.1177/1090198105277329
7. National Institutes of Health (2007) How to write easy to read health materials. http://www.nlm.nih.gov/medlineplus/etr.html. Accessed 26 March 2017
8. Ingledew P-A, El-Zammar D, Scali E, Brar B, Lin J (2014) Caught in the web: the quality of online resources for cancer patients. Int J Radiat Oncol 90(1):S604. https://doi.org/10.1016/j.ijrobp.2014.05.1808
9. Society, American Cancer (2016) Cancer Facts and Figures. http://www.cancer.org/research/cancerfactsstatistics/allcancerfactsfigures/index. Accessed
10. Baazeem M, Abenhaim H (2014) Google and women's health-related issues: what does the search engine data reveal? Online J Public Health Informatics 6(2):187
11. Fu LY, Zook K, Spoehr-Labutta Z, Hu P, Joseph JG (2016) Search engine ranking, quality, and content of web pages that are critical versus noncritical of human papillomavirus vaccine. J Adolesc Health 58(1):33–39. https://doi.org/10.1016/j.jadohealth.2015.09.016
12. McLeod J, Yu I, Ingledew PA (2016) Peering into the deep: characterizing the Internet search patterns of patients with gynecologic cancers. J Canc Educ 32:85–90
13. https://adwords.google.com/home/tools/keyword-planner/. Accessed
14. Kincaid JP, Fishburne RP Jr, Rogers RL, Chissom BS (1975) Derivation of new readability formulas (automated readability index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. Naval Technical Training Command Millington TN No. RBR-8-75 (Research Branch)
15. Dale E, Chall JS (1948) A formula for predicting readability: instructions. Educ Res Bull 27(2):37–54
16. Gunning R (1969) The fog index after twenty years. Int J Bus Commun 6(2):3–13. https://doi.org/10.1177/002194366900600202
17. Laughlin GHM (1969) SMOG grading-a new readability formula. J Read 12(8):639–646
18. Team, R Core (2016) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna URL https://www.R-project.org/
19. Fitzsimmons PR, Michael BD, Hulley JL, Scott GO (2010) A readability assessment of online Parkinson's disease information. R Coll Physicians Edinburgh 40(4):292–296. https://doi.org/10.4997/JRCPE.2010.401
20. Freda MC, Damus K, Merkatz IR (1999) Evaluation of the readability of ACOG patient education pamphlets. Obstet Gynecol 93(5):771–774
21. Schreuders EH, Grobbee EJ, Kuipers EJ, Spaander MCW, Veldhuyzen van Zanten SJO (2017) Variable quality and readability of patient-oriented websites on colorectal cancer screening. Clin Gastroenterol Hepatol 15(1):79–85. https://doi.org/10.1016/j.cgh.2016.06.029
22. Rosenberg SA, Francis DM, Hullet CR, Morris ZS, Brower JV, Anderson BM, Bradley KA, Bassetti MF, Kimple RJ (2017) Online patient information from radiation oncology departments is too complex for the general population. Pract Radiat Oncol 7: 57–62
23. Azer SA, Alghofaili MM, Alsultan RM, Alrumaih NS (2017) Accuracy and readability of websites on kidney and bladder cancers. J Cancer Educ:1–19
24. Kamvar M, Baluja S (2006) A large scale study of wireless search behavior: Google Mobile Search. CHI '06 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: 701–709
25. Prevention, United States Department of Health and Human Services. Centers for Disease Control and. 2010. Simply put a guide for creating easy-to-understand materials. Strategic and Proactive Communication Branch 3