Davide Negrini, Andrea Padoan* and Mario Plebani

Between Web search engines and artificial intelligence: what side is shown in laboratory tests?

https://doi.org/10.1515/dx-2020-0022
Received February 5, 2020; accepted February 10, 2020

Abstract

Background: The number of websites providing laboratory test information is increasing fast, although the accuracy of reported resources is sometimes questionable. The aim of this study was to assess the quality of online retrievable information by Google Search engine.

Methods: Considering urinalysis, cholesterol and prostate-specific antigen (PSA) as keywords, the Google Search engine was queried. Using Google Trends, users' search trends (interest over time) were evaluated in a 5-year period. The first three or 10 retrieved hits were analysed in blind by two reviewers and classified according to the type of owner or publisher and for the quality of the reported Web content.

Results: The interest over time constantly increased for all the three considered tests. Most of the Web content owners were editorial and/or publishing groups (mean percentage 35.5% and 30.0% for the first three and 10 hits, respectively). Public and health agencies and scientific societies are less represented. Among the first three and 10 hits, cited sources were found to vary from 26.0% to 46.7% of Web page results, whilst for cholesterol, 60% of the retrieved Web contents reported only authors' signatures.

Conclusions: Our findings confirm those obtained in other studies in the literature, demonstrating that online Web searches can lead patients to inadequately written or reviewed health information.

Keywords: artificial intelligence; Google Trends; health information; laboratory tests; online information; Web search engine.

Introduction

The use of Internet as a source of health-related information has increased greatly in recent years, largely due to the improvement achieved in the performance of search engine service algorithms and the widespread availability of portable devices (e.g. smartphones) that allow access to search engines and Web pages [1]. Moreover, it is now possible to create new Web contents more easily and quickly, so not only large (private or public) institutions and editorial groups, but also single users, can generate professional-like websites, providing engaging contents. Interestingly, the rapid growth of online Web pages and resources providing health-related information has not gone hand in hand with regulatory updates, and the quality of their content is often controlled only by their authors. In 2010, Tonsaker and colleagues reported that for many of the 70% of Canadians who searched health-related information up online, the Internet, rather than their physician, was the first source of medical information [2]. In the United Kingdom, in 2015, more than 20% of people chose to diagnose their own diseases by using Internet resources [3]. Individuals who experience increased arousal when viewing online medical information may be at risk for developing cyberchondria, a phenomenon in which repeated Internet searches regarding medical information result in excessive concerns about physical health [4]. Interestingly, cyberchondria was found correlated with health anxiety and with a decreased quality of life [4].

Patients searching online for medical answers may improve their engagement in the management of their diseases in synergy with their physician [5], but sometimes it is difficult to check the quality and accuracy of the information they obtain. Any quality check of online health

***Corresponding author: Andrea Padoan,** Department of Laboratory Medicine, University-Hospital of Padova, via Giustiniani 2, Padova 35128, Italy; and Department of Medicine – DIMED, University of Padova, via Giustiniani 2, Padova 35128, Italy,
Phone: +390498212801, Fax: +390498211981, E-mail: andrea.padoan@unipd.it. https://orcid.org/0000-0003-1284-7885
Davide Negrini: Department of Laboratory Medicine, University-Hospital of Padova, Padova, Italy, E-mail: davidenegrini@outlook.com. https://orcid.org/0000-0002-8275-453X
Mario Plebani: Department of Laboratory Medicine, University-Hospital of Padova, Padova, Italy; and Department of Medicine – DIMED, University of Padova, Padova, Italy, E-mail: mario.plebani@unipd.it. https://orcid.org/0000-0002-0270-1711

Published online April 25, 2020

information is a challenge for Internet users, considering that results often include non-documented and unreliable information or, in the worst case scenario, are based on unscientifically founded, or harmful, health practices [2], especially nowadays, when "fake news" can become viral thanks to the Internet and its social media platforms [6].

Therefore, public institutions, scientific societies and large editorial groups, which usually implement a scientific revision process before publishing online, face a growing number of personal blogs, small private laboratories' websites and product sellers' online shops, in which information is not always properly controlled and reviewed.

One of the most important organisations aiming to help users identify the quality of information provided, the Health On the Net Foundation (Chêne-Bourg, Switzerland), a non-profit organisation, promotes transparent and reliable health information over the Internet [7]; many major/important health-related sites undergo the Health on the Net Foundation Code of Conduct (HONcode) certification process to prove that their content is reliable and validated by health professionals. Yet, not all Internet users understand the difference between a certified and a clickbait-driven website.

Most of the laboratory medicine sources of information available online are from public or private institutions, laboratories and publishing groups, while others are from patients' associations, foundations for research promotion or scientific societies. These resources provide a wide range of information, explaining, for example, the meaning of laboratory tests, how to get ready for them, how to interpret their results, etc.

It is now possible, by means of a Google Web Search (Alphabet Inc., Mountain View, CA, USA) flow, to identify about 75% of the total online searches [8], and users' behaviour data can be retrieved, aggregated and anonymised, using Google Trends public service tools, already employed in many different ways for biomedical research [9]. Google Web Search, as its market competitors, is constantly increasing its use of artificial intelligence and deep learning algorithms to achieve constant improvement in results and to "rank" them (to define their order of appearance in the search results page) while considering an ever increasing number of variables and conditions [10].

The aim of the present study was to assess the quality of online retrievable information in the Italian language by investigating the Google Web Search engine for laboratory medicine tests, and analysing Web search results of three example tests: urinalysis, cholesterol and prostate-specific antigen (PSA). These three tests were chosen because they are widely known by the general public, and commonly performed in clinical laboratories [11–13].

Materials and methods

Web search evaluation using Google Trends

The evaluation of the Web searches in the Italian language throughout a 5-year period (April 2014–April 2019: "interest over time" trend) was made using Google Trends tools. In particular, we evaluated which search keywords are most commonly adopted by users for the chosen laboratory tests.

As advised by Nuti and colleagues [9], we report that we accessed Google Trends in Italian on 15/04/2019 using the following settings: Italy as a region, last 5-year period, "Web search" and "all categories". We looked for: "Esame delle urine" as topic for urinalysis, "Colesterolo" for cholesterol and "Antigene prostatico specifico" for PSA. For each test, the first five "related queries" (with a meaning similar to that of the original query) were considered. These additional keywords, obtained from Google, are provided below.

Query execution in Google Web Search

We simulated user behaviour to record which Web pages are shown for each search phrase: the first 10 results and their order (corresponding to the first page of results) were recorded using Microsoft Excel 365 v.1909 (Microsoft Corporation, Redmond, WA, USA).

For each query to input on the Google Web Search, a clean browser session (with no previously saved cookies or other tracking methods enabled) was started, the website "https://www.google.it/" was opened directly, and then the chosen query was fed into the search field. The browser used was a Mozilla Firefox 66.0.3 (Mozilla Corporation, Mountain View, CA, USA) set in Italian and running on Microsoft Windows 10 Pro 1809 (Microsoft Corporation, Redmond, WA, USA), connected from the city centre of Padua, Italy (no device GPS position was given to the Web browser) using the mobile operator Vodafone (Vodafone Italia S.p.a., Milan, Italy) for the Internet connection. For each retrieved query, results were classified according to the type of owner or publisher. The evaluations of website owners were performed blindly and independently by two reviewers (D.N. and A.P.), after which the final classification was achieved mutually.

Website evaluation

Specific features in the Web pages found through the search engine were identified. In particular, for each website, we searched for: (1) the presence of cited sources (including references to papers published in journals, book chapters, encyclopaedias), (2) an indication that the content was written by health professionals (alone or in a group), (3) a statement that the text had undergone a scientific revision by a board of professionals, (4) a clear definition of the text's author(s)' name (irrespective of whether or not a professional) and (5) the presence of a valid HONcode certificate. For the above-described

five search queries, evaluations were performed considering the (a) first three results for each Web page (i.e. those deemed the most relevant by the search engine) and (b) the first 10 results (first page of results). Each Web page could include none to all five parameters simultaneously.

Results

Web search evaluation using Google Trends

The first parameter considered in Google Trends results was "interest over time" in the chosen laboratory tests. As reported by Google, the numbers obtained represent search interest relative to the highest point (always set to 100) on the chart for the given region and time (a value of 100 represents peak popularity for the term, and a value of 50 means that the term is half as popular) [14].

In Figure 1, which reports Google Trends data, an increase in the mean "interest over time" value is identifiable for all the considered laboratory tests. Comparing the first 3 months of 2014 and first 3 months of 2019, PSA-related searches had the highest increase in "interest over time", with an increase in the number of searches of 25%; cholesterol had an increase of 19% and urinalysis, an increase of 15%. The number of searches showed a seasonal trend, with minimums in the middle of August and in the last week of December (Christmas period), for all the tests considered.

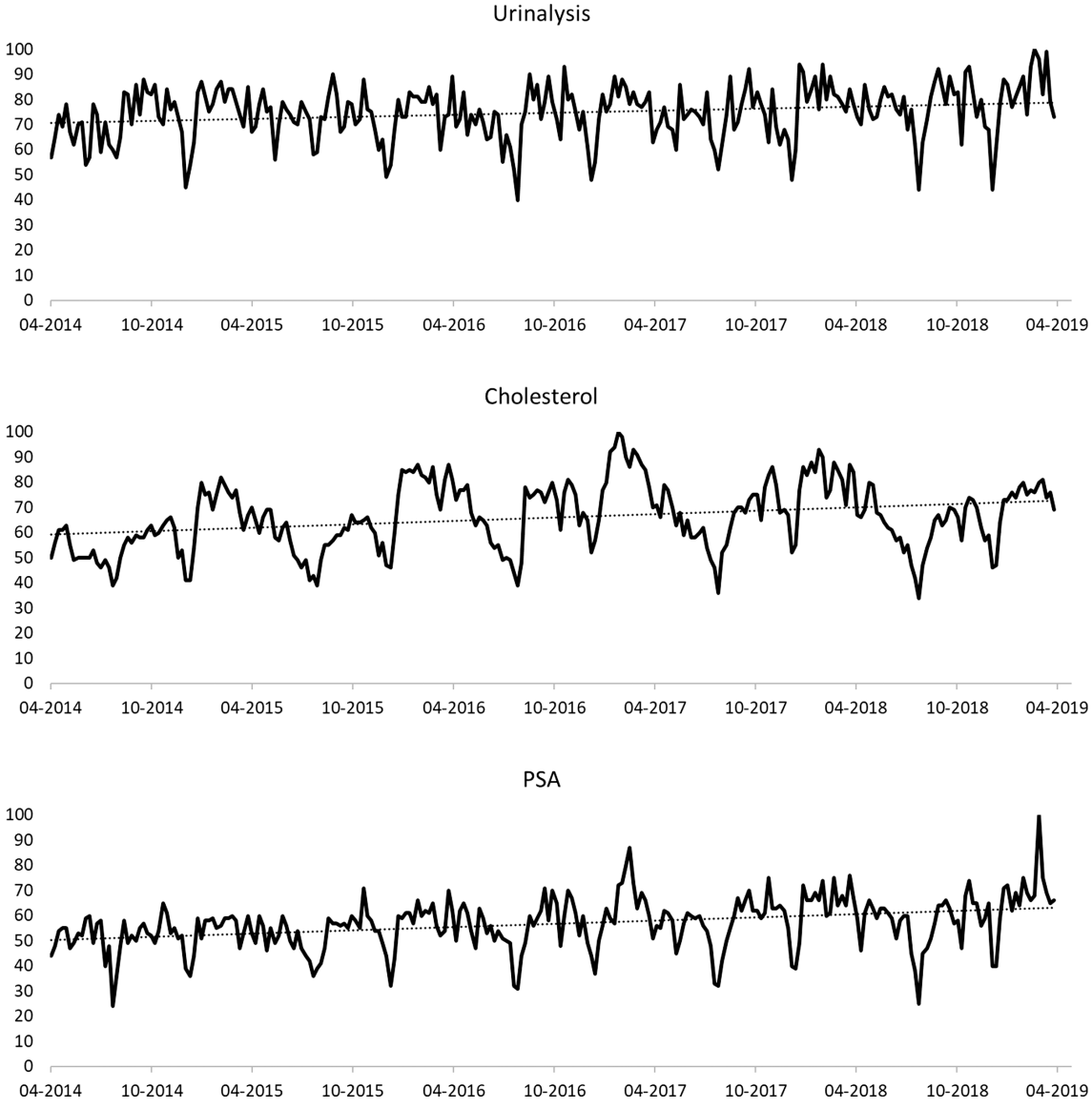

Figure 1: Google Trends "interest over time", from April 16, 2014 to April 15, 2019.

The first five "related queries" found for urinalysis were (in brackets the English translation): "analisi delle urine" ("urinalysis"), "analisi urine" ("urine analysis"), "esame delle urine" ("examination of urine"), "esami delle urine" ("examinations of urine"), "esami urine" ("urine tests"). For cholesterol, related queries were: "colesterolo" ("cholesterol"), "colesterolo alto" ("high/elevated cholesterol"), "colesterolo hdl" ("hdl cholesterol"), "hdl" ("hdl"), "il colesterolo" ("the cholesterol"). For PSA, related queries were: "prostata psa" ("prostate psa"), "psa" ("psa"), "psa libero" ("free psa"), "psa totale" ("total psa"), "valori psa" ("psa values").

Query execution in Google Web Search

Table 1 shows aggregated information on Web search results, based on the type of owner or publisher of the contents, divided into the following groups: patients' associations or foundations for research promotion, personal blogs, editorial or publishing groups, online encyclopaedias, private hospitals and laboratories, public hospitals and laboratories, public health agencies, scientific societies and product sellers. Overall, results showed that editorial or publishing groups, private hospitals and laboratories are the most commonly represented type of owner of the Web contents, displayed in the Web search results for all the tests considered. These categories accounted for more than a half of the total results. On the other hand, public health agencies and scientific societies were not well represented.

Quality of information reported on websites

The results of the five-feature investigation conducted to evaluate the quality and the usage of scientific sources and revision are shown in Table 2, which shows the percentages of Web pages of the search engine results that fulfil the requested features. Cited sources were found in a percentage of Web page results ranging from 26% to 46.7%, proving that more than 50% of the Web search results do not report them. Considering the first three cholesterol results, author signature was the only source guaranteeing the Web page contents in the 60% of cases. In some Web pages, the authors were content writers rather than health professionals. For the PSA test, on the other hand, we found the lowest percentages of Web results that provide only the author signature (26.7% in the first three results) without any additional statements, but the percentage of Web pages written by health professionals was higher than that for the other two tests.

Table 1: Web search results by type of owner or publisher for the five search queries of the three laboratory tests.

Tests	Cholesterol		PSA		Urinalysis	
	Considering the first 3 results, %	Considering the first 10 results, %	Considering the first 3 results, %	Considering the first 10 results, %	Considering the first 3 results, %	Considering the first 10 results, %
Patients' associations or foundations for research promotion	0.0	0.0	13.4	12.0	0.0	0.0
Personal blogs	0.0	0.0	33.3	12.0	0.0	0.0
Editorial or publishing groups	40.0	32.0	13.3	28.0	53.3	30.0
Online encyclopaedias	0.0	6.0	6.7	4.0	26.7	12.0
Private hospitals and laboratories	40.0	26.0	33.3	34.0	13.3	44.0
Public hospitals and laboratories	0.0	0.0	0.0	2.0	0.0	0.0
Public health agencies	0.0	6.0	0.0	2.0	6.7	10.0
Scientific societies	0.0	2.0	0.0	6.0	0.0	4.0
Product sellers	20.0	28.0	0.0	0.0	0.0	0.0
All	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**

All values are in % and are calculated considering a total of 15 queries for the first three Web search results or a total of 50 queries for the first 10 Web search results.

Table 2: Quality of information reported on Web search results by the investigated features for the five search queries, for the three laboratory tests.

Test	Cholesterol		PSA		Urinalysis	
	Considering the first 3 results, %	Considering the first 10 results, %	Considering the first 3 results, %	Considering the first 10 results, %	Considering the first 3 results, %	Considering the first 10 results, %
Presence of cited sources	33.3	46.0	33.3	34.0	46.7	26.0
Written by health professionals	46.7	48.0	80.0	60.0	66.7	64.0
Scientific revision by a board of professionals	20.0	18.0	13.3	22.0	53.3	34.0
Author's signature only	60.0	44.0	26.7	38.0	80.0	40.0
HONcode certificate	0.0	6.0	13.3	22.0	20.0	14.0

All values are in % and are calculated considering a total of 15 queries for the first three or a total of 50 queries for the first 10 results. Each Web page can be considered for more parameters at the same time.

Discussion

The fact that nowadays patients look for health information online is well established [2–3], and the trend has increased worldwide, thanks also to groups of patients' forums, and enthusiasts' blogs, which are easily accessible through Web search engines, addressing the user searches on the Web. Every Web search engine crawls all the pages it receives (or finds by surfing the Web on following links) and creates an index [15]. To provide results for each query (the text corresponding to the user's question), search engine uses different algorithms in order to rank all possible results available in its index, and then displays a list of Web pages ordered by "relevance", the first result being considered the one most relevant for the user, and the query made [16]. Then the user can click (or tap, depending on the type of device used) on the results to directly navigate to the Web page. In view of the variety of search engines and their back-end algorithms, we chose the Google Web Search, the engine most widely used by the population in Italy and in many other countries [8].

The first element we analysed was the "interest over time" trend. The overall increase of "interest" in these arguments can be considered a confirmation of the growing number of individuals searching for health-related information on the Web, probably linked to the increase in health-related searches by already connected users, but also the spread of Internet access to new users, who now find it easier than ever to surf the Web: there is greater availability of mobile devices with fast connections and more user-friendly interfaces, where Web search engines have a strong presence. The PSA test had the greatest growth in trend in the latest 5 years, perhaps because the test itself is younger than the other two tests, which are well established in laboratories. Regarding seasonality in the "interest over time", the reduction in the number of searches in August and in the last week of December appears to correspond to a reduction in the number of requests for the three tests in laboratory routine.

In the second part of the investigation, we identified the owners or publishers of the Web contents in the search results (Table 1). In particular, from Table 1, the mean percentages of results (considering all three tests) from Web search of pages owned by editorial and publishing groups were 35.5% and 30.0% of the first three and the first 10 results, respectively; the mean percentages of results (considering all three tests) from Web search owned by private hospitals and laboratories were 28.9% and 34.7% of the first three and the first 10 results, respectively. Public health agencies and scientific societies were displayed only rarely when the first three results were considered, and accounting for only 2–10% on considering the first 10 results.

Furthermore, we looked at the Web pages in the search engine results in order to ascertain how many fulfil some basic quality features, such as source citation or scientific revision processes of the contents (Table 2). Most of the Web contents were written by health professionals and included cited sources, while HONcode certificate was not frequently used. Interestingly, HONcode represents a certification that focuses on the reliability and credibility of health and medical online information and sets rules for website to (a) hold basic ethical standard in the presentation of information and (b) facilitate readers to know the source of the information [7].

Finally, this study presents some specific limitations, such as the possibility that patients use keywords different from those initially queried and the utilization of a single Web search engine for obtaining information on online laboratory tests.

In conclusion, although we evaluated a small number of clinical laboratory tests, and did not analyse trends and results on social media or video sharing platforms, our findings confirm those made in other studies in the literature, demonstrating that online Web searches can lead patients to inadequately written or reviewed health information. Web search engines are providing results to an ever increasing number of users, but it is hoped that artificial intelligence and deep learning applications in Web search algorithms will give ever better and ever more accurate results [17], higher ranking being conferred to contents with cited sources, and those which undergo scientific revision by professionals.

Author contributions: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.
Research funding: None declared.
Employment or leadership: None declared.
Honorarium: None declared.
Competing interests: The funding organisation(s) played no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the report for publication.

References

1. Lippi G, Mattiuzzi C, Cervellin G. Is digital epidemiology the future of clinical epidemiology? J Epidemiol Glob Health 2019;9:146.
2. Tonsaker T, Bartlett G, Trpkov C. Health information on the Internet: gold mine or minefield? Can Fam Physician 2014;60:407–8.
3. Kirk A. One in four self-diagnose on the internet instead of visiting the doctor. The Telegraph; 2015. Available at: https://www.telegraph.co.uk/news/health/news/11760658/One-in-four-self-diagnose-on-the-internet-instead-of-visiting-the-doctor.html. Accessed 30 Dec 2019.
4. Mathes BM, Norr AM, Allan NP, Albanese BJ, Schmidt NB. Cyberchondria: overlap with health anxiety and unique relations with impairment, quality of life, and service utilization. Psychiatry Res 2018;261:204–11.
5. Iverson SA, Howard KB, Penney BK. Impact of internet use on health-related behaviors and the patient-physician relationship: a survey-based study and review. J Am Osteopath Assoc 2008;108:699–711.
6. Wang Y, McKee M, Torbica A, Stuckler D. Systematic literature review on the spread of health-related misinformation on social media. Soc Sci Med 2019;240:112552.
7. Health On the Net – Non-Governmental Organisation (NGO). Available at: https://www.hon.ch/en/. Accessed 30 Dec 2019.
8. Search Engine Market Share, NetApplications.com. Available at: https://bit.ly/2Mavvaj. Accessed 30 Dec 2019.
9. Nuti SV, Wayda B, Ranasinghe I, Wang S, Dreyer RP, Chen SI, et al. The use of Google trends in health care research: a systematic review. PLoS One 2014;9:e109583.
10. Metz C. AI is transforming Google Search. The rest of the Web is next. Wired. 2016. Available at: https://www.wired.com/2016/02/ai-is-changing-the-technology-behind-google-searches/. Accessed 30 Dec 2019.
11. Stanford Health Care. Different types of lab tests. Available at: https://stanfordhealthcare.org/medical-tests/l/lab-tests/types.html. Accessed 30 Dec 2019.
12. Lin K, Lipsitz R, Miller T, Janakiraman S. Benefits and harms of prostate-specific antigen screening for prostate cancer: an evidence update for the U.S. Preventive services task force. Ann Intern Med 2008;149:192–9.
13. Ilic D, Djulbegovic M, Jung JH, Hwang EC, Zhou Q, Cleves A, et al. Prostate cancer screening with prostate-specific antigen (PSA) test: a systematic review and meta-analysis. BMJ 2018;362:k3519.
14. Google. FAQ about Google Trends data. Trends Help. Available at: https://support.google.com/trends/answer/4365533?hl=en. Accessed 30 Dec 2019.
15. Marsden S. How do search engines work? Deep crawl. 2019. Available at: https://www.deepcrawl.com/knowledge/technical-seo-library/how-do-search-engines-work/. Accessed 30 Dec 2019.
16. Turnbull D. What is search relevance? OpenSource Connections. 2014. Available at: https://opensourceconnections.com/blog/2014/06/10/what-is-search-relevancy/. Accessed 30 Dec 2019.
17. Lippi G. Machine learning in laboratory diagnostics: valuable resources or a big hoax? Diagnosis 14 Sep 2019. DOI: https://doi.org/10.1515/dx-2019-0060 [Epub ahead of print].