Journal of Applied Pharmaceutical Science Vol. 7 (11), pp. 001-006, November, 2017
Available online at http://www.japsonline.com
DOI: 10.7324/JAPS.2017.71101
ISSN 2231-3354

Principal Component Analysis: A Tool for Identifying Web Document Characteristics Affecting Quality of Drug Information Websites

Lawan Sratthaphut[1*], Krittika Tanyasaensook[2], Piyaporn Tunneekul[3]

[1]Department of Health-Related Informatics, Faculty of Pharmacy, Silpakorn University, Nakhon Pathom, Thailand.
[2]Department of Pharmacy, Faculty of Pharmacy, Mahidol University, Bangkok, Thailand.
[3]Department of Law, Faculty of Humanities and Social Sciences, Nakhon Pathom Rajabhat University, Nakhon Pathom, Thailand.

ARTICLE INFO

Article history:
Received on: 29/08/2017
Accepted on: 14/10/2017
Available online: 30/11/2017

Key words:
Drug information, Principal component analysis, Web document, HTML document, Internet.

ABSTRACT

The objective of our study was to identify Web document characteristics affecting quality of drug information websites using principal component analysis (PCA) technique in order to assist consumers, patients, and Web developers for observing good designing aspects to achieve information quality. Internet websites were collected by using the 8 search terms and 3 mostly utilized search engines in Thailand. Thirty five drug information websites were assessed with two independent raters to find out the quality in drug information providing using DISCERN criteria. Sixteen characteristics of Web document were investigated. PCA was applied to the data and the principal components were plotted and visualized structurally for detection the most important characteristics which related to the quality of drug information websites. The six PCs accounted for 73.39% of the total variance. The overall mean of DISCERN score for quality was "fair" at 52.5 (range, 21–72; SD = 11.1). Four attributes then were chosen to be the factors which mostly influence quality status of drug information websites. These findings provide consumers and patients to observe the quality of sites that provide drug information as well as to support Web authors for improving the quality of drug information websites.

INTRODUCTION

Nowadays, looking through the Internet has turned into a typical device for people who wish to find out about their good beings and medical problems (Jansen and Spink, 2006). The amount of Internet websites offering health-related information increases quickly every day (NCCAM, 2006) including drug information websites. In the past seventy-two percent of online users surfed for health and medical information which described of one kind or another such as serious conditions of drug or diseases, general information, or minor health problems (Fox and Duggan, 2013). Moreover, seventy-seven percent of online health users started with a search engine such as Google, Yahoo, or Bing (Fox and Duggan, 2013). The patients used the World Wide Web for various health-related reasons. Fifty-eight percent mentioned employing the World Wide Web to review side effects of drug or complications of medical therapy (Joseph *et al*., 2002). Good drug information provided can be used to prevent medication errors and lead to enhanced quality of patient care. In the real world, websites can be developed by anyone. Thus, many websites provide significant information but others may provide information questioningly or deludingly. The consumers ought to consider online drug information since nobody has altered the tremendous amount of drug information on the web in order to guarantee its quality and accuracy. That is the motivation to have instruments for assessing the quality of the websites. There are several instruments for quality assessment of websites. Four instruments are broadly utilized today; the DISCERN tool, the HON Code, CyberGuide and the JAMA benchmarks. In this study, we were keen on inspecting the quality of accessible drug information to patients.

** Corresponding Author*
Department of Health-Related Informatics, Faculty of Pharmacy, Silpakorn University, 6 Rachamankhanai Rd., Sanam Chandra Palace Campus, Tam-bon Phra Pathom Chedi, Am-per Muang, Nakhon Pathom 73000, Thailand. Tel: +6634-253910 Fax +6634-255800
Email address: sratthaphut_l @ su.ac.th

Hence, we selected a simple and common instruments - DISCERN. The DISCERN is a set of questionnaires which is divided into 3 parts: 1) reliability of the publication, 2) quality of information about treatment choices, and 3) the overall quality rating (Charnock D, 1998). DISCERN is easy to use and can be utilized not only patients but also pharmacists or authors of health information as a standard guide which users are qualified for anticipation. A variety of reports has characterized the manner in which consumers search for health-related information (Zhang and Dimitroff, 2005; Zhang and Dimitroff, 2005) and has been assessing the quality of health-related information on the websites (Woodruff, 1996; Jadad, 2006; McCool, *et al.*, 2015; Memon, *et al.*, 2016). There is no research study on the Web document characteristics of drug information websites and the quality of providing drug information. The purpose of this study is to apply principal component analysis (PCA) technique for identifying Web document characteristics that affect the designing drug information websites in order to achieve information quality. PCA is one of the statistical multivariate methods based on eigenvector decomposition. It was first presented by Pearson (1901) and developed separately by Hotelling (1933) (Jolliffe, 2002). PCA consolidates the majority of the variables in which there are interrelated into a smaller number of principal components (PCs) (Sratthaphut, *et al*., 2013). Those PCs then, are visualized structurally, while holding as much as possible of the variation exhibits in the data set.

MATERIALS AND METHODS

Sites identification and evaluation

Internet sites were identified using two general search terms (‘drug information’ and ‘medical information’) and six specific search terms (‘amoxicillin’, ‘celecoxib’, ‘hydrochlorothaizide’, lipitor’, ‘prozac’ and ‘spironolactone’) and three mostly utilized search engines in Thailand (Google, Yahoo!, and Bing, accessed on January, 2016). When the keywords were entered into these Web crawlers, the World Wide Web was filtered to discover websites related to these search keywords. A list of universal resource locators (URLs) was shown up in the search engine result pages (SERP) and arranged in decreasing order of relevance to the search keywords. Because drug information websites achieved through good quality information may not be good rank on SERPs. Thus, the top 30 consecutive URLs and the URLs in the range between 301^{th} and 330^{th} listed in SERP were collected in each search result. Once, one thousand four hundred and forty URLs (60x3x8) were returned by eight keyword searches conducted in each of the three search engines, the sites that met to the following criteria have been discarded.

- Non-English language sites
- Non-Drug information providing sites
- Illegal content or design sites
- Advertising purpose sites
- Duplicate sites
- Book review sites or journal abstract offering sites
- Non-operative sites or sites with required to apply for registration

The rest thirty five drug information websites were included. Then, each site was independently evaluated by two raters (registered pharmacists) according to DISCERN criteria. The DISCERN instrument comprises of 16 key inquiries with five-point Likert scale which provides users with a reliable framework for assessing the quality of health information. The raw scores are included toward the finish of the assessment. Along these lines, the greatest number of possible points is 80. The raw scores obtained were changed over the percentage scores. Class interval was calculated by the formula: (Smax-Smin)/3, where Smax is the overall maximum score (72), and Smin is the overall minimum score (21).

Class interval was employed to compute three quality levels (groups). All sites were classified into three groups according to their percentage scores and were interpreted as follows: >56% “good”, 39-56% “fair” and <39% “poor”.

Data collection

The twenty Web documents (also referred to as Web pages) of each evaluated drug information websites (35 sites) were randomly downloaded on March 2016 and stored for analysis. Hence, the data set used in this study composed of 700 Web documents (35x20). The observed Web document characteristics were document size, time used to download the document, image size, CSS size, the number of the broken links, the number of errors in HTML tags, and the number of HTML tags including <h1>, <h2>, <h3>, <h4>, <h5>, <h6>, <link>, <meta>, <p> and <img>. The values of these variables were obtained by using free tools. The document size, time used to download, image size, and CSS size were collected by Web Page Analyzer - 0.98 (http://www.websiteoptimization.com/services/analyze/). The number of the broken links was calculated by the W3C Link checker (http://validator.w3.org/checklink).

The errors in HTML tags related to the pages were found by the W3C Markup Validation Service (http://validator.w3.org/). And the last, the extraction of tagged text was done by using HTML Tag Count (http://redwriteblue.com/tags/htmlcount.html).

Data analysis

Cohen’s kappa statistic were applied to test inter-rater agreement and the kappa values were interpreted in agreement with Fleiss: 0.0 to 0.40 poor, 0.41 to 0.75 fair to good, and >0.76 excellent (Fleiss *et al.*, 2003; McCool *et al*., 2015). All data were processed by a Notebook equipped with Intel® Core™ i3 processor, 2GB for RAM, and Windows 7. The principal component analysis (PCA) and all statistical parameters were performed by SciCraft open source data analysis software 1.0.2 (Alsberg *et al.*, 2004) and PSPP 0.9.0 open source statistical software (Plaff and Darrington, 2015), respectively.

RESULTS AND DISCUSSION

A list of 35 websites, sorted alphabetically, is shown in Table 1. For sites evaluation, the overall mean of DISCERN score for quality was "fair" at 52.5 (range, 21–72; SD = 11.1). It was found that 37% of sites were "good", 46% of sites were "fair" and the remaining 17% of sites were "poor". The average kappa was 0.65. It notified 'fair to good' agreement conforming to Fleiss. In Web characteristics analysis, a variety of characteristics (variables) of 700 Web documents in 35 drug information websites has been analysed. The websites variables were presented below:

- The average of page size
- The average of downloadable time
- The average percentage of image size per page size
- The average percentage of CSS size per page size
- The average number of unique tags: <h1>, <h2>, <h3>, <h4>, <h5>, <h6>, <link>, <meta>, <p>, <img>
- The average percentage of dead link per total link
- The average number of HTML error per page

Table 1: A list of drug information websites.

Number	Drug Information Websites
1	www.allinahealth.org
2	www.arthritis.org
3	www.bccancer.bc.ca
4	www.canadadrugs.com
5	www.cancer.gov
6	www.centerwatch.com
7	www.chemocare.com
8	www.dailymed.nlm.nih.gov
9	www.drugguide.com
10	www.drugs.com
11	www.drugwatch.com
12	www.empr.com
13	www.familydoctor.org
14	www.fda.gov
15	www.healthline.com
16	www.iwk.nshealth.ca
17	www.library.everyonehealthy.com
18	www.mayoclinic.org
19	www.medbroadcast.com
20	www.medical-clinic.org
21	www.medicinenet.com
22	www.medindia.net
23	www.medinfo.co.uk
24	www.medlineplus.gov
25	www.nhs.uk
26	www.parkinsons.org.uk
27	www.patient.info
28	www.pdr.net
29	www.reference.medscape.com
30	www.riteaid.com
31	www.rxlist.com
32	www.uptodate.com
33	www.verywell.com
34	www.walgreens.com
35	www.webmd.com

Figure 1 presents a summary of the average number of unique tags and Table 2 reports a summary of the remaining characteristics. In Table 2, it was found that "good" and "fair" websites have the average of page size and the average percentage of image size per page size bigger than "poor" sites. On the other hand, "poor" sites have the average number of HTML error per page higher than "good" and "fair" sites. It is remarkable that the sites, which have good quality of drug information, enrich pages with information and image to make them more understanding and more attractive, respectively. This mention is confirmed by the results of the average number of unique tag <p> and <img> (Figure 1). Furthermore, we also notice that the correct coding has correlated with the quality of drug information websites.

Table 2: Summary table of Web document characteristics.

Variables	Group of Sites		
	Good	Fair	Poor
The average of page size (KB)	267	156	73
The average number of HTML error per page	130	87	269
The average of downloadable time (second)	4.1	3.4	3.3
The average percentage of image size per page size (%)	25.5	25.0	18.4
The average percentage of CSS size per page size (%)	14.9	9.4	15.1
The average percentage of dead link per total link (%)	0.1	0.4	0.1

In PCA study, PCA was applied to the data in order to detect the most important factors to describe the quality of drug information websites. The six PCs with eigen values greater than one, a general statistical cut-off level (Hamilton, 2010), were selected. The six PCs accounted for 73.39% of the total variance. Therefore, the number of variables were compressed from 16 variables to 6 uncorrelated PCs with 26.61% loss of variation. The first principal component (PC1) was the linear combination that best condenses varieties in the original data matrix (22.53% of the cumulative proportion of variation explained), while the others (PC2-PC6) outlined the rest of the variance (50.86%). In score plot Figure 2, websites were grouped according to their quality status. Most of the good quality sites were clustered on the top-right of the Figure 2. This indicates that PC1 and PC2 could be considered as a representative of quality status.

In loading plot Figure 3, the Web document variables are represented as a function of both PC1 and PC2. The PC2 is positive values on variable 3 (the average number of tag <h3>), 4 (the average number of tag <h4>), 5 (the average number of tag <h5>), 6 (the average the average number of tag <h6>), 7 (the average number of tag <link>), 8 (the average number of tag <meta>), 9 (the average number of tag <p>), 12 (the average percentage of image size per page size), 14 (the average number of <img>), and 16 (the average number of HTML error per page) and shows negative values on variable 1 (the average number of tag <h1>), 2 (the average number of tag <h2>), 10 (the average of download times), 11 (the average of page size), 13 (the average percentage of CSS size per page size) and 15 (the average percentage of dead link per total link). Moreover, the variable 3, 4, 5 and 9 have high positive values of PC1 and PC2. As seen from the Figure 2 and Figure 3, variables, appearing on top-right side of Figure 3, are about the same top-right quadrant of Figure 2. It illustrated that the tag <h3>, <h4>, <h5> and <p> were correlated with an increase in the quality of drug information websites.

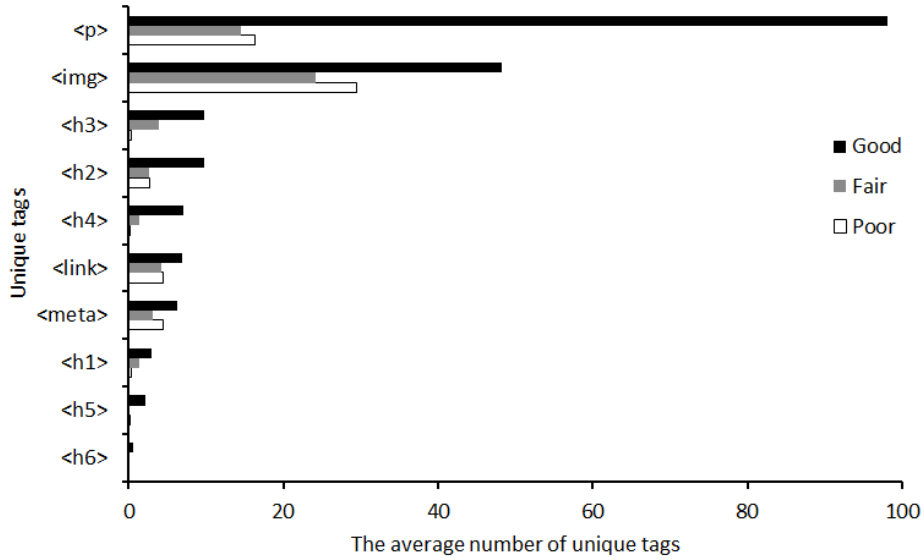

Fig. 1: Summary of the average number of unique tags.

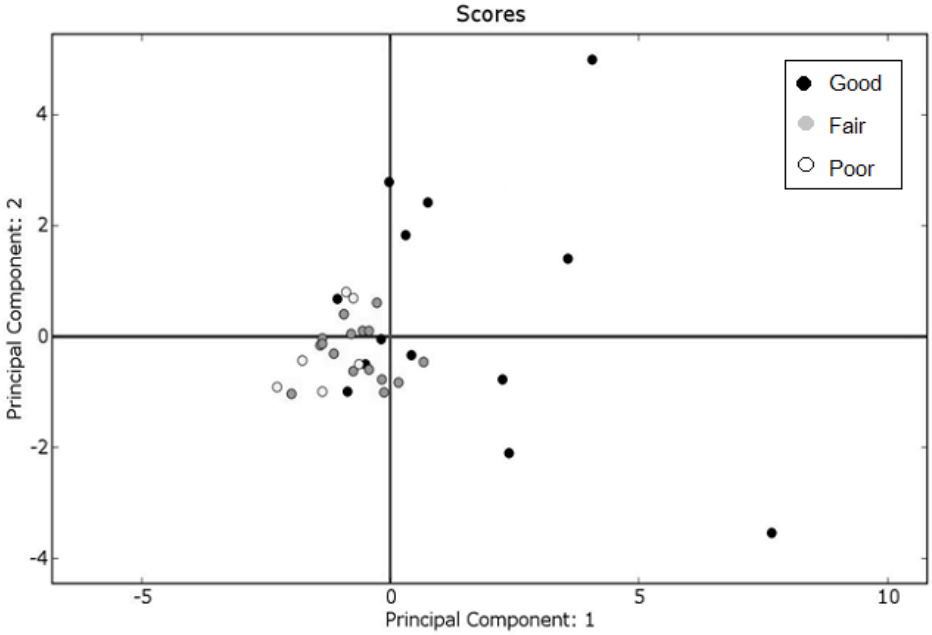

Fig. 2: Quality status of drug information websites on PC1 vs PC2.

Thus, those variables should be the potential factors responsible for quality of drug information websites. In general, tag <h1> is used for the most significant headline on the page then tag <h2>, <h3> and so on.

Web authors use the headings tag to isolate subjects related to the importance of the information. Additionally, heading tags are often followed by a short paragraph which represents in tag <p>. Headings and paragraph structure make content more understandable. For this reason, the good quality drug information websites have high average number of tag <h3>, <h4>, <h5> and <p>. The previous mention that the good quality of drug information sites enriched pages with information to make them more understanding was confirmed by these PCA results.

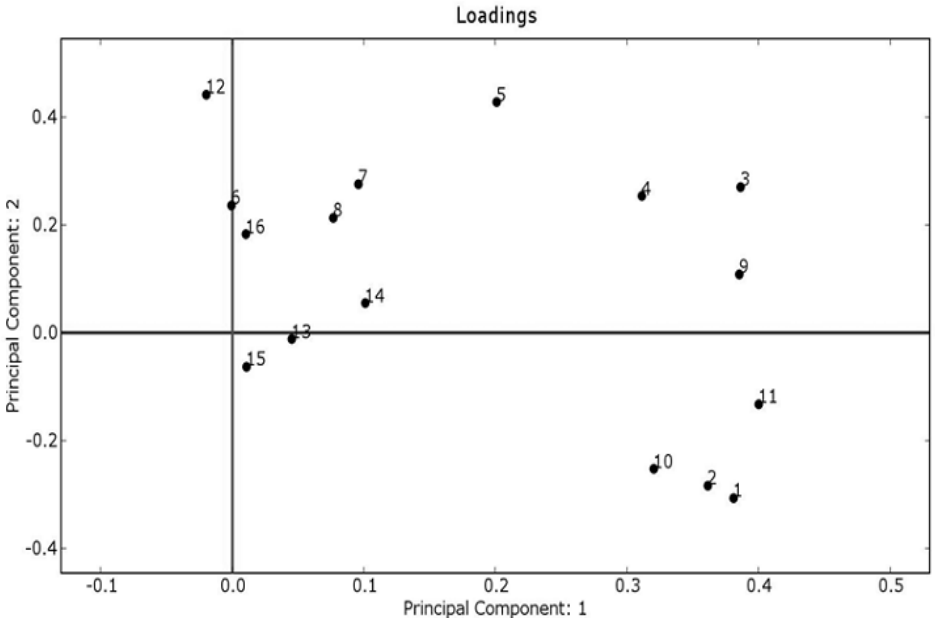

Fig. 3: Web document variables as a function of PC1 and PC2.

CONCLUSIONS

This paper aims to investigate into various Web document characteristics related to quality drug information providing of websites. PCA technique is the powerful choice for the determination of most significant factors on the quality. It was found that websites parameters mostly influence quality status were tag <h3>, <h4>, <h5>, and <p>. The explanation of these results is the more number of paragraphs are added, the more number of qualities of sites are increased. This conclusion allows health information consumers and patients to make primary decisions about the reliance on drug information websites. In addition, this conclusion is also useful for Web developers in highlighting these attributes to refine good designing aspects in order to achieve information quality. We can extend this work to investigate whether these characteristics are correlated with quality assessment by using HON Code, CyberGuide and JAMA benchmarks.

ACKNOWLEDGEMENTS

The authors would like to thank Wantip Saengthinthong, a pharmacist of Ban Pong Hospital, Thailand, for her kind assistance in website assessments. The authors also gratefully acknowledge the Silpakorn University Research and Development Institute, (Grant no. SURDI 52/01/23), Thailand, for financial support.

CONFLICT OF INTEREST

The authors declare that no conflict of interest is associated with this work.

REFERENCES

Alsberg B, Kirkhus L, Tangstad T, Andressen E. 2004. Data analysis of microarrays using SciCraft. Knowledge exploration in life science informatics, proceedings lecture notes in artificial intelligence.

Charnock D. 1998. The DISCERN Handbook: quality criteria for consumer health information on treatment choices. University of Oxford, Division of Public Health and Primary Health Care. Oxon, UK: Radcliffe Medical Press. Available at: http://www.discern.org.uk/discern.pdf. [Accessed on 11 November 2015].

Fleiss JL, Levin B, Paik MC. 2003. Statistical-Methods for Rates and Proportions, 3nd Edition. New Jersey, USA: John Wiley & Sons.

Fox S, Duggan M. 2013. Health Online 2013. Pew Internet and American Life Project. [ONLINE] Available at http://pewinternet.org/Reports/2011/Health Topics.aspx/. [Accessed 2 July 2017].

Hamilton LC. 2010. Regression with graphics: a second course in applied statistics. Michigan, USA: Brooks/Cole Publishing.

Jadad AR, Deshpande A. Healia, a search engine for finding high quality and personalized health information. J Mens Health Gend. 2013; 3(4): 418–419.

Jansen BJ, Spink A. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. Inf Process Manag, 2006; 42(1): 248–263.

Joseph A, Diaz MD, Rebecca A, Griffith MD, James J Ng, Steven ER, *et al.* Patients' Use of the Internet for Medical Information. J Gen Intern Med, 2002; 17(3): 180-185.

Jolliffe IT. 2002. Pricipal Component Analysis 2nd ed. New York, USA: Springer-Verlag.

McCool ME, Wahl J, Schlecht I, Apfelbacher C. Evaluating Written Patient Information for Eczema in German: Comparing the Reliability of Two Instruments, DISCERN and EQIP. PLOS ONE. 2015; October 6. doi:10.1371/journal.pone.0139895

Memon M, Ginsberg L, Simunovic N, Ristevski B, Bhandari M, Kleinlugtenbelt YV. Quality of Web-based Information for the 10 Most Common Fractures. Interact J Med Res 2016; 5(2): e19. doi:10.2196/ijmr.5767

NCCAM 2006. 10 Things to Know About Evaluating Medical Resources on the Web. National Cancer Institute [ONLINE] Available at http://wellnessproposals.com/health-care/complimentary-and-alternative-medicine/10-things-to-know-when-evaluating-online-health-resources.pdf. [Accessed 2 July 2017].

Plaff B, Darrington J, *et al*. 2011. GNU PSPP. Version 0.7.8. Boston: Free Software Foundation.

Sratthaphut L, Jamrus S, Woothianusorn S, Toyama O. Principal Component Analysis Coupled with Artificial Neural Networks for Therapeutic Indication Prediction of Thai Herbal Formulae. Silpakorn U Science & Tech J, 2013; 7(1): 41-48.

Woodruff A, Aoki PM, Brewer E, Gauthier P, Rowe LA. An investigation of documents from the World Wide Web. Comput Netw ISDN Syst. 1996; 28: 963-980.

Zhang J, Dimitroff A. The impact of webpage content characteristics on webpage visibility in search engine results (Part I). Inf Process Manag. 2005; 41: 665–690.

Zhang J, Dimitroff A. The impact of metadata implementation on webpage visibility in search engine results (Part II). Inf Process Manag. 2005; 41: 691–715.

How to cite this article:

Sratthaphut L, Tanyasaensook K, Tunneekul P. Principal Component Analysis: A Tool for Identifying Web Document Characteristics Affecting Quality of Drug Information Websites. J App Pharm Sci, 2017; 7 (11): 001-006.