Decay of References to Web sites in Articles Published in General Medical Journals: Mainstream vs Small Journals

P. Habibzadeh[1]

[1]Shiraz University of Medical Sciences, Student Research Committee, Shiraz University of Medical Sciences, Shiraz, Iran

Keywords

Internet, scientometrics, citation analysis, uniform resource locator, URL

Summary

Background: Over the last decade, Web sites (URLs) have been increasingly cited in scientific articles. However, the contents of the page of interest may change over the time.

Objective: To investigate the trend of citation to URLs in five general medical journals since January 2006 to June 2013 and to compare the trends in mainstream journals with small journals.

Methods: References of all original articles and review articles published between January 2006 and June 2013 in three regional journals – *Archives of Iranian Medicine (AIM), Eastern Mediterranean Health Journal (EMHJ),* and *Journal of Postgraduate Medical Institute (JPMI)* – and two mainstream journals – *The Lancet* and *British Medical Journal (BMJ)* – were reviewed. The references were checked to determine the frequency of citation to URLs as well as the rate of accessibility of the URLs cited.

Results: A total of 2822 articles was studied. Since January 2006 onward, the number of citations to URLs increased in the journals (doubling time ranged from 4.2 years in *EMHJ* to 13.9 years in *AIM*). Overall, the percentage of articles citing at least one URL has increased from 24% in 2006 to 48.5% in 2013. Accessibility to URLs decayed as the references got old (half life ranged from 2.2 years in *EMHJ* to 5.3 years in *BMJ*). The ratio of citation to URLs in the studied mainstream journals, as well as the ratio of URLs accessible were significantly ($p<0.001$) higher than the small medical journals.

Conclusion: URLs are increasingly cited, but their contents decay with time. The trend of citing and decaying URLs are different in mainstream journals compared to small medical journals. Decay of URL contents would jeopardize the accuracy of the references and thus, the body of evidence. One way to tackle this important obstacle is to archive URLs permanently.

Correspondence to:
Parham Habibzadeh
Student Research Committee
Shiraz University of Medical Sciences
Shiraz
Iran
Email: parham.habibzadeh@yahoo.com

Appl Clin Inform 2013; 4: 455–464
DOI: 10.4338/ACI-2013-07-RA-0055
received: August 2, 2013
accepted in revised form: September 14, 2013
published: October 2, 2013
Citation: Habibzadeh P. Decay of references to web sites in articles published in general medical journals: Mainstream vs small journals. Appl Clin Inf 2013; 4: 455–464
http://dx.doi.org/10.4338/ACI-2013-07-RA-0055

© Schattauer 2013

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

1. Introduction

Accuracy of the information referred to is one of the most important aspects of academic writing [1]. One of the examples would be the case of citing related or previously published articles. Undoubtedly, references are important in the flow of scientific knowledge. Correct citation of a reference let readers find further information on the subject of interest.

There have been a lot of concerns about the accuracy of references published [2]. An accurate citation includes accurate reference to all the identifying fields, namely the journal title, year or volume of publication, and the start page number of an article, the error in reporting each of which of these elements will certainly result in difficulties (even impossibility) in retrieving the original source. Previous investigations on reference accuracy in a wide variety of general and specialty biomedical journals reported error rates ranging from 10% to 70% [3-5].

With the development of World-Wide Web, the use of the Internet for identifying valuable information has become inevitable for many scientists round the globe, as many scientific fields like immunology and molecular biology are rapidly changing and many of the work is presented in digital format on the Internet every day. The biomedical community has therefore accepted the Internet as a readily available forum to share various file formats not suitable for print media like videos, sound, large databases, software programs, *etc*. With increasing availability of scientific materials on the Web, one type of references, which has become popular over recent years, is citing Web sites (URLs) so that currently, 1% to 19% of articles cite at least one URL [6]. Nonetheless, while print materials are archived in libraries, the contents of Web pages may not be permanent; URLs referenced within the scientific and medical literature become inaccessible over time [6]. There are also concerns about the quality and credibility of materials presented in the Web sites [7]. Nevertheless, many authors, though aware of all these limitations, cannot resist citing URLs in their articles. Scientific URLs also suffer from all these limitations. As citation to URLs is well accepted among all biomedical journals, the change in the cited URL contents would jeopardize their accuracy and thus the body of evidence [8, 9]. One study conducted in 2004 on five prestigious biomedical journals revealed that the rate of erroneous references to URLs ranged from 0% to 22% (mean = 8.7%, 95% CI: 2.8% to 20%) [10]. Many other studies came out to similar conclusions [8, 11–14].

Mainstream journals and small medical journals are different in many aspects. Authors of articles in small medical journals are mostly from developing countries where they may not have readily access to subscription-based databases such as *Web of Science*, *Scopus*, and *EMBASE*. Therefore, they may have a different pattern of citing articles. Almost all of the studies so far conducted on the pattern of citations to URLs have studied published articles from developed countries. The current study was therefore conducted to investigate the trend in use of citation to URLs as well as the rate of accessibility to the citations over time in two leading general medical journals and three regional journals.

2. Methods

Two high-impact general medical journals, *The Lancet* (2012 journal impact factor [IF] = 39.06) and *British Medical Journal* (*BMJ*, 2012 IF = 17.215), both from the UK, and three general medical journals with lower profile published in the WHO Eastern Mediterranean Region, namely, *Archives of Iranian Medicine* (*AIM*, 2012 IF = 1.222) from Iran, *Eastern Mediterranean Health Journal* (*EMHJ*, no IF) from Egypt, and *Journal of Postgraduate Medicine Institute* (*JPMI*, no IF) from Pakistan, were included in this study. The two selected prestigious medical journals were arbitrary chosen from the top 10 general medical journals. The selected regional journals were chosen from the most important general medical journals published in the region, however, geographic distribution of the country of origin was also considered. The selected countries had the highest contribution to medical publications from the WHO Eastern Mediterranean Region. The study period was from January 2006 to June 2013.

The two British journals are weekly; the frequency of publication of the three regional journals varied from quarterly to monthly. All issues of the three regional journals were studied. Because of the high number of published articles in *The Lancet* and *BMJ*, using a systematic sampling method,

© Schattauer 2013

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

25% of the published issues, the second issue of each month, were included in this study. Only original articles and reviews published in English in the studied journals were analyzed. Original articles appeared under various titles in *The Lancet* and *BMJ.* In *The Lancet*, "Articles," "Seminars" and "Reviews" and in the *BMJ*, "Research" articles and "Clinical Reviews" were included in this study. In the regional journals distinguishing original articles and reviews was simple as they appeared under the same headings.

The total references cited in each article and the number of citations to URLs were counted. Then, each of the URLs was copied and pasted in a Web browser (Google Chrome®) to determine the accessibility of the URL and the accuracy of the page contents retrieved. In evaluating URLs, any white spaces (space [ASCII 32], and horizontal tab [ASCII 9]) in the URL string were not considered an error, and the URL was corrected by concatenating the next consecutive string after the URL. A URL was considered inaccessible if the site could not be gotten through within 60 seconds or encountering an error indicating that the site could not be found (*e.g.*, error "404 not found," *etc.*). If error "404 not found" was encountered, the site was rechecked almost 24 hours later to see if it is temporarily down or not. Furthermore, referencing to generic Web sites (*e.g.*, CDC or WHO) where the landed page did not contain exactly the information referred to in the article reference was also considered "inaccessible."

The ratio of references to URLs as well as the ratio of accessible URLs were then calculated for each volume (year) of the journals. This was done to abolish the seasonal variation that might occur in the calculated ratios. Then, the trend of the ratios over the study period was determined using an exponential curve fit model.

3. Results

Over the studied period, *The Lancet* and *BMJ* published 874 eligible articles. In these articles 39429 references were cited, 1468 (3.7%) of which referred to URLs out of which 901 (61%) of citations were accessible (▶Table 1). The number of articles found eligible in the three studied regional journals were 1948; the total number of references were 50369, out of which 1240 (2.5%) referred to URLs of which 524 (42.3%) were accessible (▶Table 1). The ratio of citation to URLs in the two studied prestigious journals was significantly (χ^2=120.303, df = 1, p<0.001) higher than that in the low-profile journals. The ratio of URLs accessible was also significantly (χ^2 = 98.546, df = 1, p<0.001) higher in the prestigious journals compared to the low-profile journals.

As shown in ▶Figure 1, citing URLs has had an increasing trend in the studied journals with a doubling time ranging from 4.2 years for *EMHJ* to 13.9 years for *AIM* (▶Table 1). After *EMHJ*, *The Lancet* had the steepest trend. As ▶Figure 2 depicts, the accessibility to URLs, however, has decreased over time with a half life of URLs varying from 2.2 years for *EMHJ* to 5.3 years for *BMJ* (▶Table 1). The average half life of URLs for the two prestigious journals was higher than that of regional journals (4.7 *vs.* 2.6 years).

Out of 874 articles studied in the two prestigious journals, 424 (48.5%) articles had at least one citation to a URL; the ratio in the three studied regional journals was significantly (χ^2 = 127.000, df = 1, p<0.001) lower (26.8%; 523/1948).

4. Discussion

It was found that citing URLs has become more popular in scientific articles so that almost half of the articles published in the two mainstream journals and around a quarter of articles published in the regional journals had at least one citation to a URL. Likewise, in a study on three leading US scientific journals, the *New England Journal of Medicine, JAMA,* and *Science*, it was found that 30% of articles cited at least one URL [15].

The increasing trend in citing URLs was also reported in other scientific disciplines. For example, a study conducted in 2000 on almost 271000 computer science articles revealed a dramatic rise in the number of URLs per article since foundation of the Web [16]. The frequency of citing URLs was higher in high-impact journals compared with that in low-profile journals. Most of the authors of

© Schattauer 2013

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

articles published in regional journals are from the region, and have somewhat limited access to the Internet comparing to the authors publishing in leading medical journals who are generally from developed countries with readily access to the Internet [17]. That would explain the higher prevalence of citing URLs in high-impact journals compared to regional journals.

It was found that a considerable proportion of the URLs cited was not accessible. The half life of URLs ranged from 2.2 years for *EMHJ* to 5.3 years for *BMJ*. This is in accord with the findings of other studies. A study conducted in April 2004 on five mainstream medical journals reported an inaccessible or inaccurate ratio of 11.8% for URLs cited in original articles published in January 2004 in the studied journals [10]. In another study performed between 2003 and 2006, 2.5% and 3.9% of the reviewed references in the *New England Journal of Medicine* and *The Lancet*, respectively, referred to URLs, out of which 14.6% and 17.9%, respectively, were not accessible. Over the study period, the rate of inaccessibility increased as the references got old [13]. The reported rate of inaccessible URLs of 17.9% (over four years) for *The Lancet* reported in that article is comparable with that of the findings in the current study (35.4% over 7.5 years). In another study conducted in 2003, Wren reported that almost 37% of 1630 unique URLs identified in MEDLINE records published between 1966 and April 2003 were either non-accessible or intermittently accessible [18].

Inaccessibility of the URLs cited may be due to either inaccurate referencing to or change of the contents of the URLs. A study investigating the prevalence of inaccessible URLs in biomedical publications reported that 11.9% of the URLs cited were already inaccessible within two days after an article's release to the public [14].

We cannot be sure what we see at the time of visiting a Web page is exactly what the author had observed at the time of his research. Several studies reported various life spans for URLs ranging from 44 to 100 days [19, 20]. There are many reasons for ending the life of a Web page including closure of an organization or merger of the organization with other ones causing removal of the materials of the Web page. The second reason would be the failure in maintaining the old links. In all these situations the visitor will get a "404 error not found" message instead of the content of interest.

▶Figure 2 shows that the accessibility rate to the URLs has sharply declined after a few years. However, the rate of decay in high-impact journals is lower than that in low-profile journals (▶Table 1). The URLs cited in regional journals had generally shorter half life (2.6 years) compared to those cited in prestigious journals (4.7 years). This might be due to a better infrastructure existing in high-impact journals which are mostly based in developed countries. At these journals URLs might be checked before publication either manually or automatically while such infrastructures are generally not available to the regional journals.

It is expected that there would be no difference between prestigious and regional journals in terms of the pattern of citation to URLs and their decay over time, as inaccessibility of the Internet references is presumably an inherent problem with the Internet and not be related to the journal type, reader base, or impact factor. However, as it is shown in the current research, mainstream journals are different from low-profile journals in terms of both the pattern of citing Internet references and decay of URLs cited. That could probably be explained by the fact that the authors publishing in these two types of journals are in fact not similar. The authors of the studied regional journals are mainly from developing countries and may not have readily access to subscription-based databases such as *Web of Science*, *Scopus*, *EMBASE*, *etc*. The credibility and quality of the Web sites these authors have access and refer to may not be as high as those generally cited by the authors publishing in the prestigious medical journals investigated in this study; these authors are usually working in large research centers based in developed countries with readily access to many databases, IT consultants, and software programs (*e.g.*, reference managers).

Some researchers suggest that publishers should be advised to discourage authors from citing URLs in their manuscripts [15, 21]. However, every day, tens of thousands of pages, many of which are not available elsewhere, are published online. Examples are articles published on electronic-only journals. Considering the rapid penetration of the Internet in societies and data explosion, it seems that using citations to URLs is inevitable.

One option to tackle the problem of loosing URLs would be to archive the cited URL contents on the journals servers. This approach, however, has some serious limitations including problems associated with managing of the available resources including hardware upgrade, increasing demand for better archiving (size, accuracy and retrieval/storage speed), installing the necessary software

© Schattauer 2013

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

programs that demand RAM and CPU time, and their maintenance [9]. It would be much difficult for small journals to setup such archiving systems due to their limited access to the necessary resources.

Another way to deal with this problem is the use of persistent URLs (PURLs). PURLs are Web addresses that locate the resource itself rather than addressing where the resource resides in the Internet. Technically, a PURL is an URL that takes the Web page visitor to an intermediary PURL Resolution Service, instead of directly taking them to the exact location of the specified data. The Resolution Service translates each PURL to the pertinent URL where the data of interest are resided. PURLs exist permanently, however, they have some disadvantages as well; they are not location-independent, *i.e.*, a PURL address contains the address of a PURL Resolution Service, without which the redirection cannot be done correctly. Therefore, a PURL would not work properly if that certain PURL Service did not exist anymore [16, 22].

Another alternative for facing this issue is using Internet archiving systems. One example is WebCite® which was developed to prevent "link rot" in scholarly publications. WebCite® archives permanently the URLs. WebCite® is available to authors, editors, readers, *etc* [23]. Editors and publishers of journals are encouraged to ask the authors to use WebCite®-enhanced references, *i.e.*, references which contain not only a link to an archived copy of the document, but also the original URL (which will probably become inaccessible or its content may change in the future). A recent study investigating the effectiveness of archiving Web sites revealed significant results. Although, 46% of 144 URLs cited in 55 articles published from March 2009 to June 2010 in *Annals of Emergency Medicine* were lost over 18 months after their publication, none of the archived articles did [6].

Decay of URLs over time is an important issue to scientists. Although, we did not study the association between the length of the URLs and their rate of decay, previous studies showed that longer URLs are more likely to be inaccessible. This might be obvious, as the longer the URLs, the more likely that an error will occur in their handling. Digital Object Identifier (DOI) and WebCite® links are usually less than 20 characters in length and are thus less likely to cause errors [24].

It is reasonable to suggest authors to use DOIs of the cited articles whenever possible. They also should be advised to use WebCite® (which is currently free for authors) or similar archiving systems. It is also better to use reference managers like EndNote® that guaranty error-free transfer of citation information from databases.

Despite many authors suggesting ways to preserve URLs, only few journals have so far mentioned in their "instructions for authors" a statement asking authors for archiving the cited URLs in their submissions. One reason for this might be that there is no universal consensus on how to preserve the cited URLs. Another reason might be that medical journal editors, who are mostly physicians, do not care enough to references as much as they do to other parts of a submitted manuscript. Even in conferences on biomedical journalism, the topic of preserving URLs is not considered as important as other issues like peer review, journal metrics, indexing systems, statistical presentation of data, *etc*. They do not know that after a while, with increasing the number of broken links, even citation analysis and correct assessment of the impact of a journal would be very difficult, if not impossible at all.

One of the limitations of this study would be the limited number of journals analyzed. Only two leading journals published in the UK and three general journals from the WHO Eastern Mediterranean Region were studied. These might not represent all regional journals published; however, they are for sure among the most prominent medical journals published in the region. The current study was conducted by a human who could analyze various problems in accessing a URL, and identify various presentations of URLs (particularly in the regional journals where standard referencing was not practiced well). Due to the fact that all the data collection was made by a human no further expansion of the sample size was possible. Nevertheless, the number of references studied was much higher than many previous studies, even some of those which were conducted by software programs. In this study, only general medical journals were investigated and the results may not be generalized to specialty medical journals.

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

Conflicts of Interest
The author has no conflicts of interest to declare.

Financial Support
None.

Authorship
The author involved in the (1) conception and design, acquisition of data and/or analysis/interpretation of data and (2) drafting and/or critical revision of the article for important intellectual content. The author approved the final version to be published.

Protection of Human Subjects and Animals in Research
Not applicable.

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

© Schattauer 2013

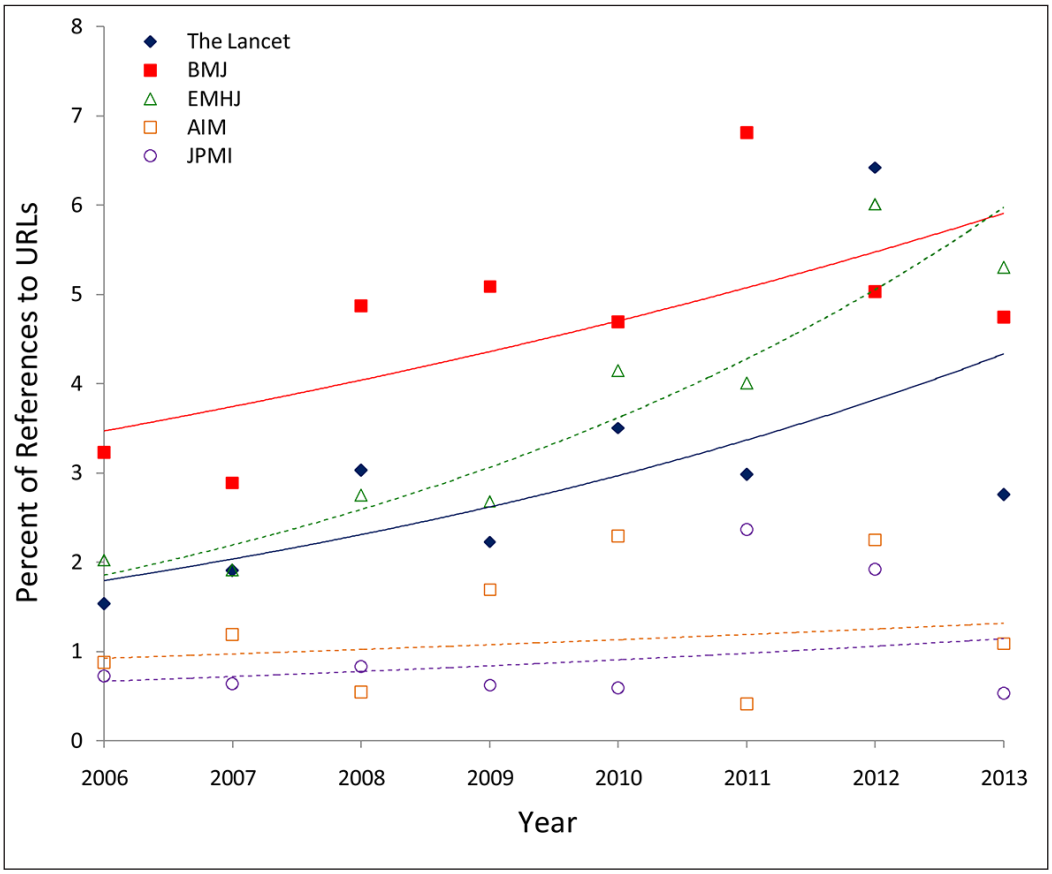

Fig. 1 The trend in the relative frequency of references to URLs for the studied journals

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

© Schattauer 2013

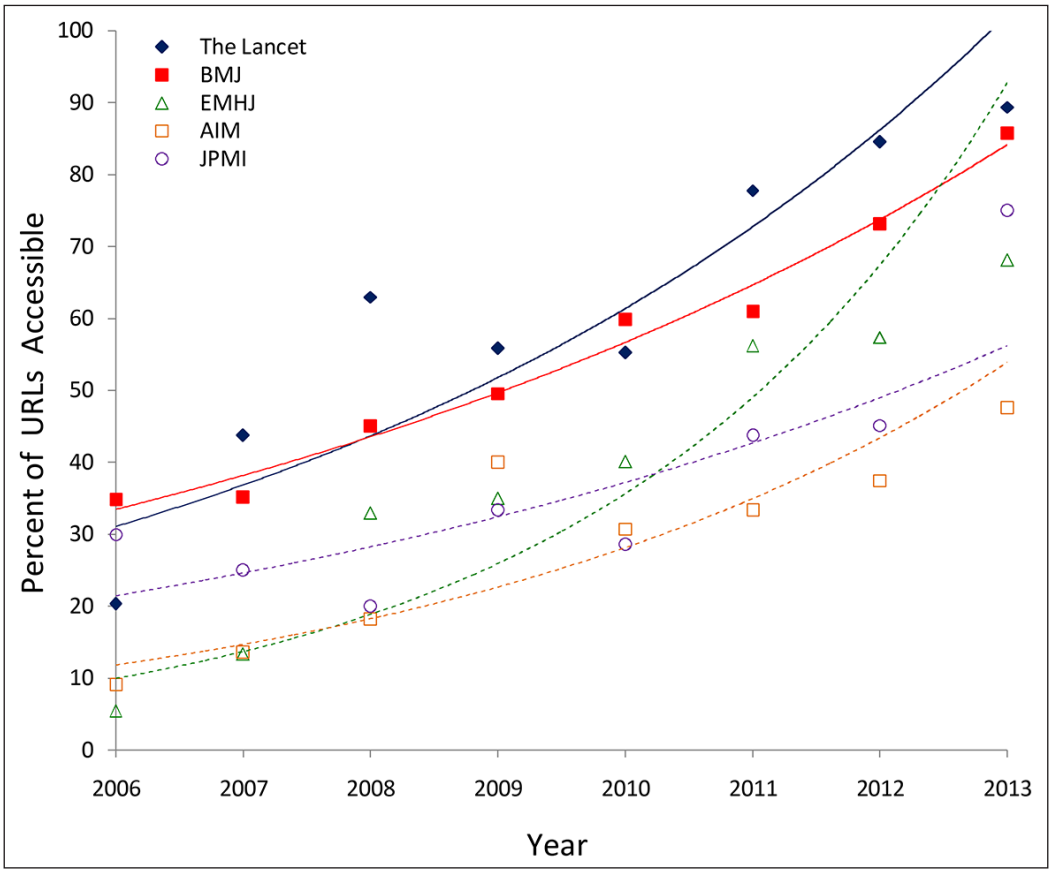

Fig. 2 Decay in the rate of accessibility of the cited URLs over time in the studied journals

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

© Schattauer 2013

Table 1 Parameters measured in the studied journals

Journal	Articles	References	References to URLs (%)	URLs accessible (%)	Doubling-time (yrs)	Half-life (yrs)	Articles with at least one citation to a URL (%)
Lancet	407	23477	689 (2.9)	445 (64.6)	5.5	4.1	191 (46.9)
BMJ	467	15952	779 (4.9)	456 (58.5)	9.1	5.3	233 (49.9)
EMHJ	1016	25157	917 (3.7)	414 (45.2)	4.2	2.2	369 (36.3)
AIM	487	15005	212 (1.4)	67 (31.6)	13.9	3.2	91 (18.7)
JPMI	445	10207	111 (1.1)	43 (38.7)	9.0	5.1	63 (14.2)

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

© Schattauer 2013

References

1. Brender J, Talmon J. On using references as evidence. Methods Inf Med 2009; 48: 503-507.
2. Browne RF, Logan PM, Lee MJ, Torreggiani WC. The accuracy of references in manuscripts submitted for publication. Can Assoc Radiol J 2004; 55: 170-173.
3. Siebers R, Holt S. Accuracy of references in five leading medical journals. Lancet 2000; 356: 1445.
4. Goldberg R, Newton E, Cameron J, Jacobson R, Chan L, Bukata WR, et al. Reference accuracy in the emergency medicine literature. Ann Emerg Med 1993; 22: 1450-1454.
5. Jackson K, Porrino JA, Jr., Tan V, Daluiski A. Reference accuracy in the Journal of Hand Surgery. J Hand Surg Am 2003; 28: 377-380.
6. Thorp AW, Schriger DL. Citations to Web pages in scientific articles: the permanence of archived references. Ann Emerg Med 2011; 57: 165-168.
7. Adelhard K, Obst O. Evaluation of medical internet sites. Methods Inf Med 1999; 38: 75-79.
8. Carnevale RJ, Aronsky D. The life and death of URLs in five biomedical informatics journals. Int J Med Inform 2007; 76: 269-273.
9. Wren JD. URL decay in MEDLINE–a 4-year follow-up study. Bioinformatics 2008; 24: 1381-1385.
10. Crichlow R, Winbush N, Davies S. Accessibility and accuracy of web page references in 5 major medical journals. JAMA 2004; 292: 2723-2724.
11. Wagner C, Gebremichael MD, Taylor MK, Soltys MJ. Disappearing act: decay of uniform resource locators in health care management journals. J Med Libr Assoc 2009; 97: 122-130.
12. Wren JD, Johnson KR, Crockett DM, Heilig LF, Schilling LM, Dellavalle RP. Uniform resource locator decay in dermatology journals: author attitudes and preservation practices. Arch Dermatol 2006; 142: 1147-1152.
13. Falagas ME, Karveli EA, Tritsaroli VI. The risk of using the Internet as reference resource: a comparative study. Int J Med Inform 2008; 77: 280-286.
14. Aronsky D, Madani S, Carnevale RJ, Duda S, Feyder MT. The prevalence and inaccessibility of Internet references in the biomedical literature at the time of publication. J Am Med Inform Assoc 2007; 14: 232-234.
15. Dellavalle RP, Hester EJ, Heilig LF, Drake AL, Kuntzman JW, Graber M, et al. Information science. Going, going, gone: lost Internet references. Science 2003; 302: 787-788.
16. Lawrence S, Pennock DM, Flake GW, Krovetz R, Coetzee FM, Glover E, et al. Persistence of Web References in Scientific Research. Computer 2001; 34: 26–31.
17. Internet users per 100 inhabitants, 2001–2011: ITU Statistics. Available from: http://www.itu.int/ITU-D/ict/statistics/material/excel/2011/Internet_users_01–11.xls and http://www.webcitation.org/6IV6Z70cn (Archived by WebCite). [accessed July 30, 2013]
18. Wren JD. 404 not found: the stability and persistence of URLs published in MEDLINE. Bioinformatics. 2004; 20: 668-672.
19. Kahle B. Preserving the Internet 1997. Available from: http://web.archive.org/web/19970504212157/http://www.sciam.com/0397issue/0397kahle.html (Archived by Internet Archive). [accessed July 30, 2013]
20. Weiss R. On the Web, Research Work Proves Ephemeral 2003. Available from: http://stevereads.com/cache/ephemeral_web_pages.html and http://www.webcitation.org/6IV8qT1QC (Archived by WebCite). [accessed July 30, 2013]
21. Ducut E, Liu F, Fontelo P. An update on Uniform Resource Locator (URL) decay in MEDLINE abstracts and measures for its mitigation. BMC Med Inform Decis Mak 2008; 8: 23.
22. Persistent Uniform Resource Locator (PURL). Available from: http://purl.oclc.org/docs/index.html and http://www.webcitation.org/6IVBrfh3B (Archived by WebCite). [accessed july 30, 2013]
23. WebCite. Available from: http://www.webcitation.org/. [accessed July 30, 2013]
24. DiCarlo JV, Pastor X, Markovitz BP. The shadow uniform resource locator: standardizing citations of electronically published materials. J Am Med Inform Assoc 2000; 7: 149-151.

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.