Indicators for websites: Particular reference to geriatric sites

Divya Srivastava[1] and Renu Bahadur[2]

[1]*Scientometrics Unit, Division of publications & Information,*
[2]*Indian Council of Medical Research (HQ), New Delhi 110029 (India)*

ABSTRACT

Several initiatives which could be applied at different levels to improve the average quality of web sites have been proposed. In the present ongoing study, efforts have been made to test the risk markers for disappearance of certain geriatric web sites with particular reference to India. The elderly population is increasing globally, so is in India. For this cross-section of society, Internet is a very useful communication media, so the sites catering to their needs must maintain their quality and should survive. Various webometric parameters are being employed to develop a model for calculation of survival chances of a web site.

Keywords: Webometrics Indicator, Web Ecology, Quality on the Web, Elderly Population, Link Analysis.

INTRODUCTION AND BACKGROUND

In 1990–91 India's population of the elderly was 56.68 million and is now around 75 million, which is expected to double in the next 20 years. The erosion of the joint family system has led the elderly to often being victims of emotional neglect and lacking physical, financial and other support. They need information on various issues. A proper solution can be an integrated approach involving information providers, providers of healthcare, etc. The Internet offers a great amount of health-related web sites, but after an early enthusiasm generated by the potential use of internet for medical issues, concerns have been raised about the content and quality of web sites.

In recent years, there has been intense debate concerning the definition of criteria for and markers of quality for medical web resources. The lack of universally accepted criteria of quality has prompted different organizations to propose their own criteria (Risk A 2001, Commission of the European Communities), or 'guidelines' (Health on the Net Foundation 2004, British Healthcare Internet Association 2004, Winker MA 2000). More and more data points to an 'innate' user ability to discern between poor- and high-quality medical web resources. Thus, usage analysis of medical web resources generates popularity indicators such as the number of inbound links, which correlate with their assessment by third parties and their degree of compliance with the above-mentioned quality criteria (Hernandez-Borges AA 1999).

It would seem that the best 'qualified' elements of any biosystem should have the greatest chances of survival. Similarly, the best resources on the web should also be the most likely to survive online. With this hypothesis, efforts are being made to determine whether online permanence of elements in a sample of geriatrics web pages correlated with indicators of quality such as their degree of compliance with criteria of quality and with a series of Webometric indicators.

In the present study we have focused on the disappearance of a subset of pages dealing with topics on health issues of the elderly population or geriatrics. The focus is on how they disappeared, how this disappearance correlated with certain quality markers and, finally, whether or not we could predict this event, which we named the 'death of a medical web page.' It could seem strange to

***Corresponding author.**

E-mail: drdivya.srivastava@gmail.com

DOI: 10.5530/jscires.2012.1.13

talk about web page disappearance in terms of death. However, some authors have previously considered the Internet as a biosystem in which its elements interact with each other to produce certain ecologic patterns. In addition, some authors have described this question in biological terms, for example, Koehler, in a work from 1999, found 8 to 20% of non-available pages, although some of them did come back from that state of non-availability that he called 'coma'. Another example is this study from Veronin, published in *J Med Inform Retrieval*, some years ago, which describes a rate of site attrition of 59% in 3 years.

In our study, we have considered that the best web pages, that is, the most qualified members of the biosystem 'Internet', should have more chance of permanence online. I have demonstrated this hypothesis.

OBJECTIVE

The objectives of the work were:

1. To describe the basic statistics regarding the disappearance of web pages in our sample. For the study we have defined the death of a web page as the lack of availability twice in a period of two weeks.
2. To study the associations, if any, of this event and certain variables, such as some webometric indexes and their degree of compliance with quality codes.
3. Finally, to design a predictive model of the disappearance of our pages.

Compared to other web methods such as a content-based analysis, the relative advantage of link analysis is that it is able to examine the way in which web sites form a certain kind of relation with others via hyperlinks. Given this interweaving hyperlink structure, it may be necessary to recognize individual web sites as mutually dependent entities, which constitute a web system. This provides the visibility index of web sites. We have made efforts to evaluate the quality of a web site by parameters culled out from Health on the Net Foundation (HON), British Healthcare Internet Association (BHIA) and American Medical Association (AMA). The reasoning behind our approach in calculating the relative importance of each criteria of quality in web sites is based on the assumption that some resources have higher visibility index because they comply with certain quality criteria. Determining the extent of compliance of each of these criteria by a given 'web site' enhanced its probability of 'online performance'. The variables which can predict 'online performance' of a 'web site' are: an increased number of 'inbound links' on follow up (IIL), the 'visibility index' of web pages and 'compliance of quality criteria'. The predictive power of this variable has been analyzed using contingency tables. Statistical significance was defined as P value less than or equal to 0.05 for 'link analysis' and 'visibility index'.

METHODOLOGY

To execute the objectives, endeavors were made to explore the availability of information on health problems in the field of geriatrics or elderly population. The starting point for this investigation was the entire population of Indian web pages dealing particularly with health & social problems of a cross-section of society ≥60 years of age. For comparison purposes initially, the entire gamut of web sites was navigated for culling out the requisite information. Popular search engines like Google were used. In this context 'web pages' refers to both text documents and hypertext documents. For limiting the time period 'year' has also been defined under the 'Advanced Search' protocol of the search engines.

One major drawback is that no search engine can index the whole web (Thelwell 2002b). If a commercial search engine is used, then clearly this is a limitation and one has to accept it. A second major problem was that the results returned by search engines fluctuated irregularly, sometimes drastically. To overcome this, multiple search rounds were carried out and in case of variations, an average was taken of the 'hits'.

Search Strategy:

The search commands that were used were:

'Health Problem' AND 'Elderly' OR 'Geriatrics'

Same query was repeated for each year individually, i.e., 2004, 2006 and 2008. Total hits thus obtained were tabulated. Search results were randomly screened to collect 'keywords' related to diseases and problems of elderly population. Each and every entry in the list was systematically screened for (1) provision of health information, social, psychological or economic issues (2) evaluated information likely to be accessed by the target population (3) the quality of information against certain criteria for example, by judging the authority of source, usefulness, readability or comprehensiveness of the provided information. After screening & filtering, a set of 35 websites was identified. To these 35 web pages the entire link from other sites was tabulated to see the popularity of the site. Number of inbounds links and number of pages

per web site were calculated using the syntax option for inbound links (link:http://www.nlai.ir/OR link:nlai.ir/) NOT (host:http://www.nlai.ir/OR host:nlai.ir). On the other hand, the number of pages in that site was calculated by the syntax (domain:www.nlai.ir OR domain:nlai.ir), wherever needed suitable statistical parameters were used for testing.

Quality criteria were taken from the Health on the Net Foundation (HON), British Healthcare Internet Association (BHIA) and the American Medical Association (AMA). Only those criteria were used which were identified as essential for the performance of pages online. These are as follows:

1. All the external site links work. **(AMA)**
2. All the internal site links are functional. **(AMA)**
3. Any claims about the benefits of a specific treatment, commercial product or service are supported with appropriate balanced evidence. **(HON)**
4. Any medical or health advice on the site is provided by medically trained and qualified professionals or otherwise by a clearly identified non-medically qualified individual or organization. **(HON)**
5. Clear references to source data are provided, and, where possible, specific links to those documents are included. **(HON)**
6. Contact addresses for visitors who seek further information or support or the Web editor's email are provided. **(HON)**
7. Features that facilitate the use of the site (site map, an FAQ section, customer service information, etc.) are provided. **(AMA)**
8. Site's viewers are allowed to return to a previous site, and Web designers avoid redirecting the viewer to a site the viewer did not intend to visit. **(AMA)**
9. The information is clearly provided. **(HON)**
10. The information provided on the site is designed to support, not replace, the relationship that exists between patients or site visitors and their existing physicians. **(HON)**
11. The intended audience is identified. (**BHIA**)
12. The last date of amendment or update is included. (**BHIA**)
13. The owners of the copyright are identified. (**AMA**)
14. The Web designers uphold the copyright and intellectual property laws. (**BHIA**)

High compliance to quality criteria could be not only a good marker of a web site's evolution, but its very cause. That is, certain web pages survive longer because they are better, and they are better because they comply with more quality criteria or with the most important ones.

OBSERVATIONS AND ANALYSIS

The number of inbound links has been proposed as a marker of relative quality for health-related web resources. Similarly, it has been shown that the number of inbound links to electronic editions of traditional journals correlates with the journal impact factor, a traditional marker of quality in printed publications. Furthermore, in our study, the number of inbound links and its increment showed a capacity for predicting the 'vital' evolution of pages. Whether this correlation has a causal significance or not is open to consideration. In other words, good indicators of popularity and use, such as visits or inbound links, could encourage web producers to maintain their online presence and the quality of their resources. From this point of view, similar to the increment of traffic that some sites experience in response to changes in their content or to links from other sites, web pages' popularity indexes would act not only as popularity markers, but as an incentive for web producers to reach certain quality standards in their web sites.

Two kinds of calculations were made on the sample:

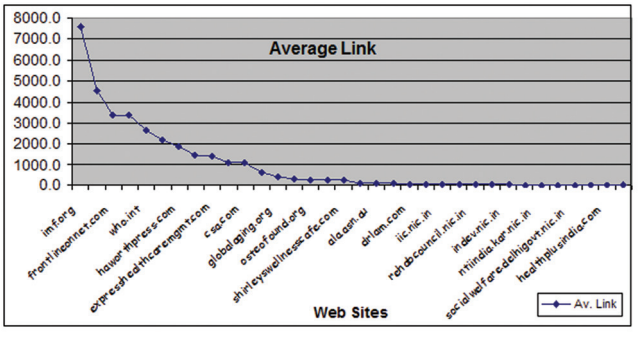
First, in 2004, 2006 and 2008, the number of inbound links was calculated and other webometric indexes by using the search engine Google. In 2008 we calculated how much the pages and sites in our sample conformed to the quality criteria of HON, BHIA and AMA. During follow-up some pages of our sample disappeared, with the result 88% of the original sample was available at the end of the follow-up.

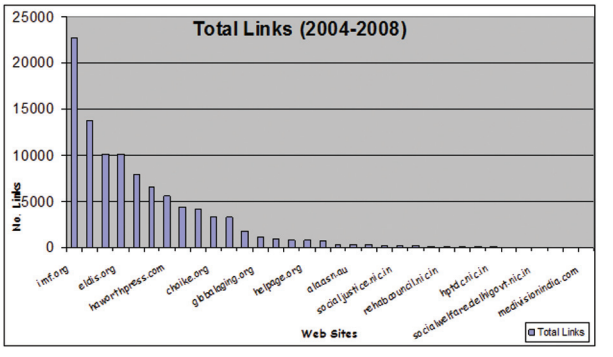
We also wanted to evaluate whether or not the disappearance of web pages was associated to some variables. Also, what about the pages disappearance with regard to certain quality markers? In the last few years we have done some research on the association between some webometric indexes and the quality of medical web pages. Specifically, we have focused on the number of inbound links. We wanted to test whether or not the survival of the pages correlated to those indexes. The web sites from the sample, received 2,54,480 inbound links during 2008, this figure was 37,006 during 2004 and 37,578 during 2006. The site having the highest link during 2006 was www.imf.org with a broad subject area. Some of the sites with more than 5000 total inbound links for all the three blocks were www.imf.org, www.aafp.org, www.frontlineonnet.com, www.eldis.org, www.who.int, www.flonnet.com, www.haworthpress.com. The follow-up indicated that some of the sites disappeared from online presence during 2006 and 2008.

Table 1. Link analysis for time periods (2004*, 2006, 2008***)**

S No.	WEB ADD	LINK (google), 2004	LINK (google), 2006	LINK (google), 2008
1	www.aafp.org	5460	5450	2790
2	www.abbott.co.in	1	5	7
3	www.ala.asn.au	129	63	92
4	www.buddhistinformation.com	58	46	**Died**
5	www.choike.org	760	1350	1210
6	www.continenceworldwide.org	19	18	53
7	www.csa.com	1140	1470	656
8	www.csupomona.edu	1700	1980	680
9	www.drlam.com	98	45	78
10	www.eldis.org	2620	5980	1510
11	www.expresshealthcaremgmt.com	1590	1010	1560
12	www.flonnet.com	1870	2450	2250
13	www.frontlineonnet.com	5440	2450	2250
14	www.globalaging.org	480	473	232
15	www.haworthpress.com	3570	1620	425
16	www.healthplusindia.com	3	3	**Died**
17	www.helpage.org	194	260	300
18	www.hptdc.nic.in	14	31	25
19	www.iic.nic.in	62	50	72
20	www.imf.org	7470	7490	7730
21	www.indev.nic.in	84	**Died**	**Died**
22	www.indiatogether.org	458	911	448
23	www.iussp.org	133	503	251
24	www.medivisionindia.com	4	**Died**	**Died**
25	www.ntiindia.kar.nic.in	11	9	9
26	www.osteofound.org	206	233	383
27	www.ota.org	81	86	92
28	www.populationcommission.nic.in	8	6	7
29	www.rehabcouncil.nic.in	14	42	35
30	www.shirleyswellnesscafe.com	391	219	106
31	www.socialjustice.nic.in	57	90	70
32	www.socialwelfare.delhigovt.nic.in	1	1	12
33	www.who.int	2850	3100	1960
34	www.whoindia.org	30	134	155

** Data captured on 6 September, 2004*
*** Data captured on 10 July, 2006*
**** Data captured on 14 August, 2008*





In this graph one can see how the average number of inbound links calculated on a yearly basis was much higher among the pages which survived than among the pages which disappeared.

Similarly, the increment of inbound links from 2004 to 2008 was much higher on average for the survivors than for the dead ones. It may also be noted that the dead pages even lost inbound links during the follow-up. This is important for an understanding for 'predicting' the survival of a site.

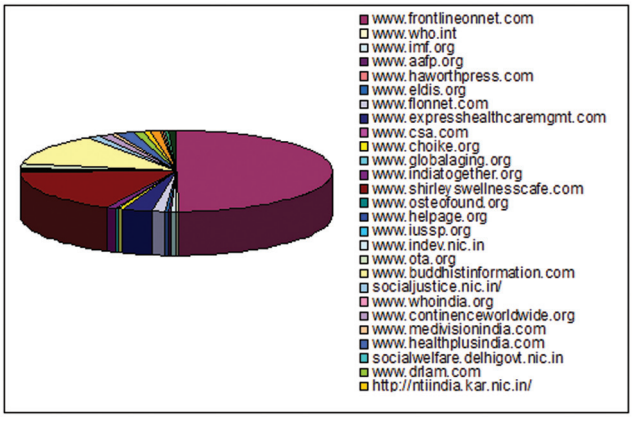


Figure. Visibility index of web sites.
Data captured on 10 July, 2006.

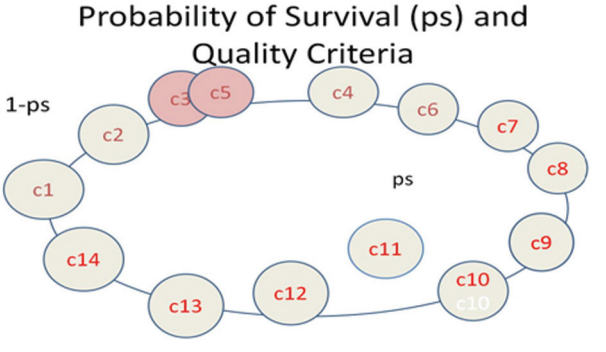
Link counts also indicate visibility of any web site, the more the link count the more site reliability. Apart from link analysis we are doing 'self link analyses' that indicates that information and pages within a web site are well connected. It is worth stating that a higher 'self link' indicates the related resources are available in one site. Search engines track self links and create precise indexes from an indicator of real routing of a web site. But the indicator of the real ranking of a web site is through the 'visibility index'. For computing this, self links were subtracted to get the actual 'link' for individual web sites. The ratio of pages in each web site and actual link gave the values for generating the visibility index.

As can be seen from the figure that the site www.frontlineonnet.com had the highest rate of visibility index (162.77) and www.globalaging.org has the least value. From the table, it is clear that only having more inbound links and self links is not important. The web pages being indexed by search engines are also an important factor that can affect the visibility index of any site and indirectly the chances of survival of that web site also.

But the ultimate causes for the success of a given page should be implicit in the essence of the pages, that is, the way they provide the information, the information itself and its sources, its navigability, the way it respects users' confidentiality, etc. So how does the compliance of quality criteria correlate to the disappearance of pages? Again we found an association between the survival of pages and their rate of compliance with the quality codes. But, were all the quality criteria equally important in producing that association? In this somewhat complex scheme we have tried to explain our reasoning. We believe that different criteria could have different importance with regards to the survival of a page. In other words, some quality criteria may be more clearly associated than others with the survival of a page.

Let's call pS the probability of survival for two years of a page due to its compliance with quality criteria. Let's look at the criterion 1. The probability that C1 is met by the pages which survived is p(C1/S) but in fact is lesser than the probability of the same criterion among the pages which didn't survive (in dark shade in the graph). In contrast, the quality criteria C11, C12 and C10 are completely related to the survival of a page. It becomes complicated when two criteria express very similar concepts (for example, C3 and C5 in this graph), so that we should remove one of them to avoid duplicating their importance. So, how could we calculate the relative importance of each quality criteria, that is, p(Ci/S)? This is a problem of conditional probability and can be solved by using Bayes' theorem. The values have been calculated by the formula

$$p\left(\frac{Ci}{S}\right) = \frac{p\left(\frac{S}{Ci}\right) \bullet p(Ci)}{\sum_{n=1}^{i} p\left(\frac{S}{Ci}\right) \bullet p(Ci)}$$



All the quality codes of the three organizations were compiled in a checklist of more than fifty items after removing the duplicated criteria. We found that 14 criteria accumulated 90% of survival probability, and that the top 10 criteria accumulated almost 80%.

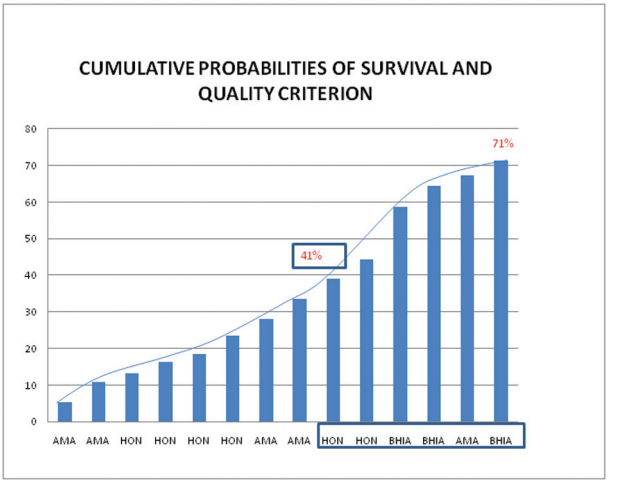


Figure. Compliance of different quality criteria.

Among the 14 'core criteria', as we called them, we could find those four quality criteria proposed by Silberg et al in 1997 in an editorial from JAMA, that is, the presence of the web author's name, the date of the last update, a disclaimer and references to the sources of information. Importantly, there are no criteria on confidentiality or privacy policies among those core criteria. On the other hand, we did not study some criteria on electronic commerce due to the low amount of pages doing e-commerce in our sample

Finally, using the associations we found a strong relation between the web page survival and inbound links and compliance of quality criteria. Thus, we designed tests to predict the disappearance of web pages. We also tried to evaluate how accurate those tests could be. Firstly, we checked the accuracy of the test, 'To be conformed with the Top 10 quality criteria', as a predictor of the survival of a page. This test had a very low sensitivity and negative predictive value, that is, we could say little if a given page did not meet the Top 10 criteria. However, if a given page met those Top 10 criteria, it would survive on-line for two years with a probability of almost 90%.

SURVIVAL ONLINE

increased inbound links	yes	no	total
yes	24	2	26
no	9	3	12
total	33	5	38

POWER OF THE IL TEST :SENSITIVITY = 72%, SPECIFICITY = 60%, POSITIVE PREDICTIVE VALUE=92.31%, NEGATIVE PREDICTIVE VALUE=25%

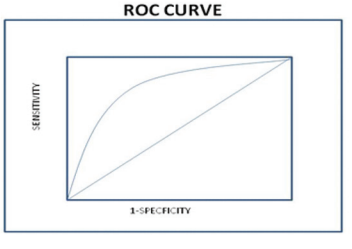


Figure. ROC curve.

In this figure the parameters are as follows:

Sensitivity = .72, at 95%, Confidence Interval is 15.40% to 93.51%
Specificity = .6, at 95% Confidence Interval is 15.40% to 93.51%
Prevalence = 86.84% at 95% Confidence Interval is 71.90% to 95.54%
PPV = 92.31%, CI 74.83% to 98.83%
NPV= 25% CI 5.78% to57.16%

With regard to the inbound links a new test was designed and checked: 'To have a loss or no-increment of inbound links for the two years before'. We found that this test had a higher, although not excellent, sensitivity to detect pages in risk of disappearance in the next two year block. So if a given web page had an increment of inbound links, it would not be in risk of disappearance with a probability of almost 90%. So, unlike the previous test, this one could have some value as a screening test.

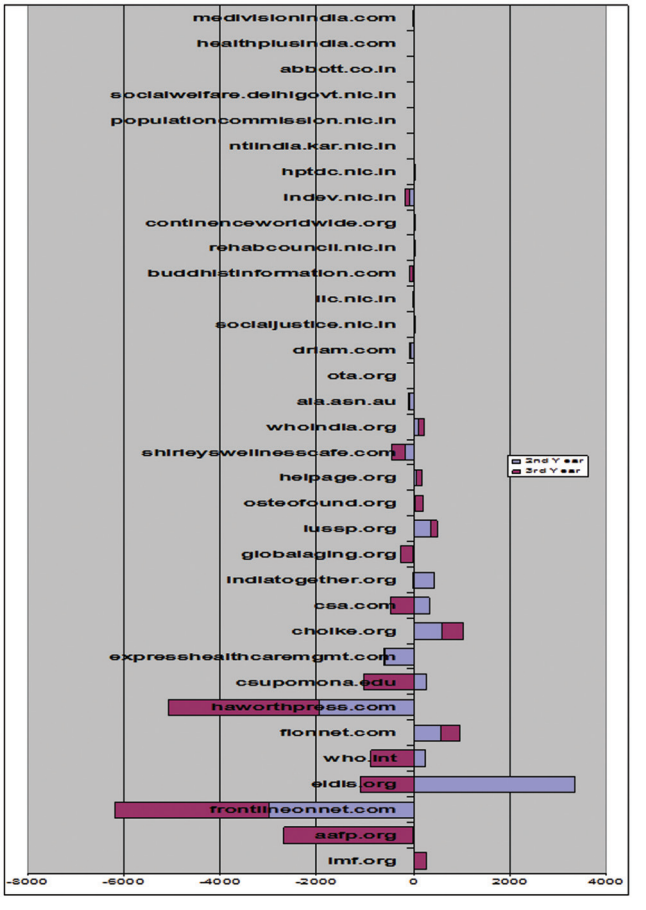


Figure. Increment of links (2004–2008).

KEY FINDINGS OF THE STUDY

Web site disappearance seems to be not just by natural wastage or a natural death affecting the oldest ones. It seems that the compliance of some quality criteria (core criteria) is especially important to protect a given web resource from an early disappearance. Its compliance could serve as a diagnostic test of the pages' good state of health. Certain webometric indexes could have a role as a screening test to detect web pages in risk of early disappearance. Further studies, i.e., prospective studies, could allow us to confirm these associations and the accuracy of these diagnostic tests. Sensitivity gives the true +ve rate of prevalence (criteria). A high PPV shows an increase of inbound link and the survival data indicates that the compliance of criteria has a positive impact on the status of web site and the assumption that the criteria

adherence coupled with increment data on inbound link can predict the survival of a site.

Various authors have shown that the web has a certain organization and stable behaviour with respect to some of its indexes of popularity. Likewise, if the survival of web resources shows the same regularity as other characteristics of the web, it should be possible to create a mathematical model that explains the site's online permanence according to certain variables. As in epidemiological studies in medicine, analysis of the variables correlated with survival of medical web resources would allow the identification of factors associated with their disappearance. Future studies using larger and more homogeneous samples could clarify the utility of webometric indicators resulting from the analysis of inbound links as markers of the online survival of resources. These studies might be incorporated into a new discipline that we could call 'Weboepidemiology'. Since the study is ongoing the observations made pertaining to the death and survival of web sites will also be tested for their statistical significance using appropriate statistical tests at both 5% and 1% level of significance.

ACKNOWLEDGMENT

We are grateful to Dr. VM Katoch, Secretary DHR & Director General, ICMR and Dr. VK Srivastava Head of the Division of Publication & Information, Indian Council of Medical Research for their encouragement and guidance. We also wish to acknowledge the other collogues especially Ajit Mathur at the council's headquarters for their valuable suggestions and constructive comments towards the analysis of the data and the project staff Arvind Singh Kushwah and Mona Gupta for processing the data.

REFERENCES

Angel A. *et al*: Is it possible to predict the death of a medical website? (accessed on 6th August, 2006 http://www.mednet2002.org/).

Wyatt Jeremy C.: Commentary: Measuring quality and impact of the world wide web. (accessed on 6th August, 2006 http://www.bmj.com/).

Eysenbach G, Powell J, Kuss Sa. Ryoung, Empirical studies assessing the quality of health information for consumers on the world wide web. JAMA 2002;287:2691–2700.

Synder H, Rosenbaum H. (1999). Can search engines be used as tool for web-link analysis? A critical view, *Journal Of Documentation*, 55:375–84.

Ingwersen P and Hjortgaard Christensen F. (1997). Data set isolation for bibliometric online analysis of research publications. Fundamental Methodological Issues. *Journal of The American Society For Information Science*, 4893), 205–17.

Osareh F. Scientometric: Aspects, methods and applications. In: Proceedings of 'Conference held by Iranian Library Association'. Eds: Mohsen Haji Zeinolabedini, (Tehran: Iranian national library, Iranian library association), 2005;2:271.