PubMed Grabber / Analyzer – a tool for off-line analysis of PubMed

Kumara Mendis, Indragith Solagarchchi, Chaminda Weerabaddana

Computer Centre, Faculty of Medicine, University of Kelaniya, Sri Lanka.

Abstract

MEDLINE is the largest bibliographic medical database available free of charges. PubMed is the Internet interface of MEDLINE.

Online searching is a luxury in most developing countries, even in academic settings because of ageing telecommunication infrastructure and prohibitive cost of the modern services like ISDN and ADSL.

Using PubMed, the Ceylon Medical Journal bibliography in MEDLINE format from 1965 – 2001 was downloaded. Important tagged fields - unique identifier (UI), publication ID (PMID), title (TI), authors (AU), source (SO), date of publication (DP), publication type (PT), country of origin (CY), Medical Subject Headings (MH), Abstract (AB) were written to a MS-Access database by PubMed Grabber- a Visual Basic program developed at the computer centre. The menu driven user interface to search the Access database was written using VB and SQL queries. Searching is possible by author, topic (MeSH), type of an article and date. Free text searching of the entire database is also possible.

PubMed Grabber/Analyzer will be a cost-effective tool for researchers in the developing countries who need to work with PubMed files off-line.

Keywords:

MEDLINE, PubMed, bibliography, Medical Informatics, off-line searching.

Introduction

MEDLINE[i] described as one of America's greatest gift to the World, is the largest medical bibliographic database. In an era where medical journals, databases and websites once available for free now demand credit card number for access, MEDLINE is free of charge. MEDLINE can be accessed on-line through the PubMed interface.

PubMed [ii] is mainly used for literature searches by medical personnel. Features like Automatic Term Mapping, Clinical Queries allow doctors with little knowledge of searching to get quick results for clinical questions. PubMed is yet to solve many issues of formulating complicated searches especially for doctors who have no training in bibliographic searches.

Online searching is a luxury in most developing countries, even in academic settings because of ageing telecommunication infrastructure or the prohibitive cost of the telecommunication services such as ISDN or ADSL. Still a properly working dial-up Internet connection of 64kbps would be a welcomed luxury in many developing countries. Furthermore, in developing countries, few medical librarians are trained search PubMed. In these circumstances minimum on-line time would be a welcome factor by many developing country doctors, if they can save and search their dataset in detail, offline.

Commercially available bibliographic management software like Reference Manager [iii] and Endnote [iv] can manage references efficiently but is not available as freeware. Tis kind of software is difficult to obtain at pirated software venders like the commonly used Windows/Office packages. Bibliographic management software use is limited to a very few licensed centres.

Other than using for searching the literature, PubMed can also be used for retrospective analysis of a topic e.g. electronic medical records [v], a Journal [vi] or an entire discipline.

We wanted to analyze Sri Lanka's oldest Medline Indexed Journal., The Ceylon Medical Journal which was established in 1887, using PubMed.

Our project had two objectives:

(a) To devise a program that can covert a MEDLINE format document to a relational database and provide menu-driven offline searching facilities.

(b) To analyze the Ceylon Medical Journal (CMJ) for the past 35 years using this application

The Analysis of the CMJ was presented elsewhere[vii] and has been submitted for publication.

This paper presents the methodology and functionality of our program PubMed grabber /viewer.

Materials and Methods

2.1 System requirements

IBM-PC with 128MB RAM, running Windows 2000. MS Access 2000 and Visual Basic 6 Professional

2.2 Medline and PubMeD

MEDLINE (Medical Literature, Analysis, and Retrieval System Online) contains 12 million references to journal articles in life sciences with a concentration on biomedicine, from 4,500 journals published in the United States and more than 70 other countries. Available for online searching since 1971, MEDLINE includes references to articles indexed from 1966 to the present. New citations are added weekly. All citations in MEDLINE are assigned MeSH Terms and Publication Types from NLM's controlled vocabulary. MEDLINE citations and abstracts are available as the primary component of NLM's PubMed database, which is searchable via the Internet.

+

2.3 Downloading the PubMed search file and MeSH

Entrez is the text-based search and retrieval system used in PubMed. Search results can be saved in many formats. We Used the MEDLINE format (.fcgi) which is a delimited text file. Citations are arranged according to the Medline 'Tags' or fields.

To down load all citations of the Ceylon Medical Journal (CMJ) from 1965 to 31st December 2001, we used the

PubMed query: ("Ceylon Med J"[Journal] AND ("1965"[PDat]: "2001/12/31"[PDat])).

The MEDLINE format file (two character tagged field format) was a 1402 KB with an extension of .fcgi

2.4 Writing PubMed MEDLINE format file to a RDBMS

Using the software application developed by us (PubMed Grabber), the important tagged fields in the MEDLINE format file was written to a relational database (MS Access 2000) **Figure 1.**

From all the different tags available in the MEDLINE format file, we picked only the relevant ones that will be required in searching a journal database. This was done after closely studying the PuMed help files that gives detail descriptions of the various citation tags and their functions. The citation tags that we captured were: unique identifier (UI), publication ID (PMID), title (TI), authors (AU), affiliation (AI) source (SO), date of publication (DP), publication type (PT), country of origin (CY), Medical Subject Headings (MH), Abstract (AB).

PubMed Grabber was written in Visual Basic 6. The Access database consists of multiple tables - CMJ, PT, AU, MeSH. The main CMJ table had one to many relationships with other four tables (PT, AD, AU, MeSH) with the UI field as the link field – **Figure 2**. When writing to the Access tables we encountered two main problems – the varying length of the fields and the number of repeat fields in one record. Dynamic arrays were used to overcome the number of repeated fields.

MEDLINE citations currently contain two unique identifying numbers, the MEDLINE Unique Identifier and the PMID (PubMed Identifier). Both numbers are found on PubMed. Because of the potential confusion in having two identifying numbers on the same citation the MEDLINE MESH update in 2004 the PMID will be the only unique number used in PubMed. We will be changing the linking filed from UI to PMID soon as this development was notified in August 2003.

MeSH is the National Library of Medicine's controlled vocabulary thesaurus. It consists of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity [viii]

The MeSH tree 2003 [ix] was downloaded in delimited format and then imported into the Access database.

2.5 PubMed - Viewer Analyzer interface

The PubMed analyzer interface has options to view and search the database. The database could be viewed in a form view (single record as in the MEDLINE file)-**Figure 4** or datasheet view (multiple records).

Menu driven searching using MeSH key words and also free text is possible from the search interface **Figure 3**. For those who need power and functionality of SQL queries, manual SQL queries are the limit.

3. Results

We converted the downloaded PubMed MEDLINE format file of 1402 KB to the relational database using 'PubMed Grabber' which took about 30 seconds. 1472 records (citations) were available.

Menu driven searching using keywords (MeSH), author name took less than a few seconds. Free text search including the title AND/OR abstract was also done. Combined searches were also possible which took less than a few seconds.

We cross checked the results using the on-line PubMed itself and our application for the type of queries given below.

1) Articles by an author, 2) Articles about a subject – (Using MeSH word), 3) Articles containing a given word in the Title or Abstract – Free text search; 4) Articles of Author XX for Subject YY;

PubMed Grabber/Analyzer was tested with large datasets. We downloaded a large single file of 8370 citations (PubMed allows only 10000 citations to be downloaded at a time) and converted it into Access database which took about 110 seconds. It is simple to merge the Access files to get one database using SQL queries.

Our detailed analysis of the Ceylon Medical Journal was presented earlier and has been submitted for publication.

4. Discussion

PubMed Grabber / Analyzer can convert a PubMed, MEDLINE format file to a relational database and provide an off-line menu driven interface for searching the dataset in detail. Formulating complex searches is possible with manual SQL queries.

Reference Manager (RM) and End Note (EN) are the leading personal bibliographic managing software in the market. There was a considerable difference in the time taken to convert a MEDLINE format PubMed search result to PubMed Grabber and Reference Manager. To convert the 1472 citation CMJ file (1404KB) took less than 30 seconds with our application while it took about 180 seconds to import it to the RM. The testing was done on a Pentium III 550 MHz PC with 128MB RAM, running Windows 2000 Professional operating system. RM and EN are not limited to MEDLINE format but can import numerous other bibliographic databases formats such as EMBASE. PubMed

Grabber /Analyzer deals with only MEDLINE format.

Searching bibliographic databases has been made very simple using commercial products like RM & EN. Nevertheless the formulating complex searches are also limited by this simplicity. Sometime even a relatively simple query of the frequency of all authors in the CMJ dataset cannot be done with RM. In our application, searches are limited only by the limitations of Access SQL.

PubMed citations have been used to retrospectively analyze up to 40 years of a particular subject[x]. Most of the papers do not give details as to how they analyzed the citations. We have yet to come across analyses that used tailor made off-line software systems. The reason may be that almost all of the papers about bibliographic analysis came from developed countries, where on-line PubMed searching and use of bibliographic reference manager software is the norm.

Our immediate plans are to use PubMed Grabber / Analyzer look at the discipline of Medical Informatics for the past three decades. The total citations involved are in the range of about 300000. PubMed has a limitation on retrieval of only 10000 records at one time. Our application will save considerable time when converting huge files. Citations of 300000 have to be downloaded as separate files and merged for the final analysis. These can be quite simple using SQL. Improvements will be made as we find problems in grabbing or analysis when handling large files.

The PubMed Grabber / Analyzer can be very useful to persons who wish to deal with large datasets and complex queries but knowledge of SQL is essential. The Access query grid makes SQL queries much easier for the non-programer. On the other extreme it can be used with small data sets by doctors without any knowledge of SQL needing simple answers using menu driven queries.

PubMed Grabber / Analyzer will be a handy tool for developing country researches (both Informatics experts with a programming background as well as non-programmers) interested in analyzing large datasets from PubMed, especially when Internet connections are expensive and slow.

Acknowledgments

Dr. Supun Wijesihnge, Chamara Wilasena, Saman Hettige of the computer centre.

References

1. MEDLINE Fact sheet: http://www.nlm.nih.gov/pubs/factsheets/medline.html (accessed on 11 September 2003)

2. PubMed Help: http://www.ncbi.nlm.nih.gov/entrez/query/static/help/pmhelp.html (accessed on 11 September 2003)

3. Reference Manager: http://www.refman.com/

4. Endnote www.endnote.com/

5. Moorman PW, van der Lei J. An inventory of publications on computer-based medical records: an update. Methods Inf Med. 2003;42(3):199-202

6. Simo Minana J, Gaztambide Ganuza M, Latour Perez J. Atencion Primaria journal in MEDLINE: an analysis of the first 7 years of indexing. Aten Primaria 1999 May;23 Suppl 1:5-13

7. Kumara Mendis, C.Weerabaddana, DIK Solangarachchi, CH Wanniarachchi. Three decades of Ceylon Medical Journal – an analysis using PubMed. Sri Lanka Medical association – 116th Anniversary academic sessions. March 26th -29th, 2003, Colombo Sri Lanka

8. MeSH Fact Sheet http://www.nlm.nih.gov/mesh/filelist.html

9. MeSH tree: http://www.nlm.nih.gov/mesh/filelist.html

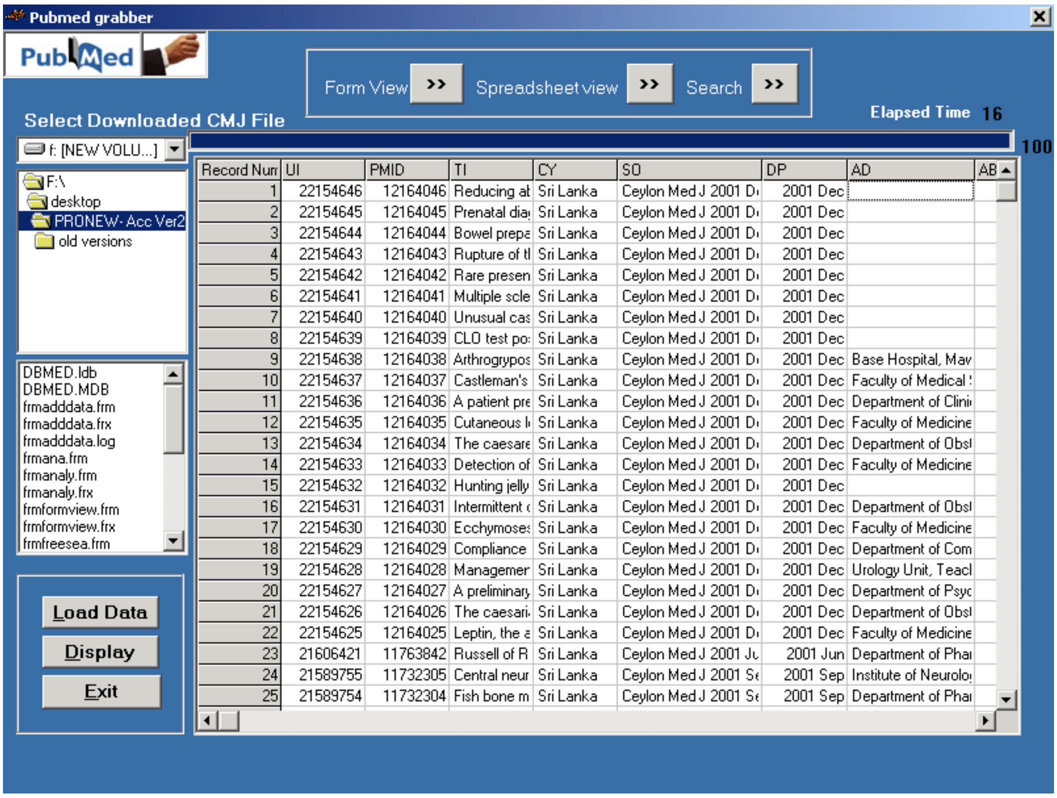
10. Banos JE, Ruiz G, Guardiola E. An analysis of articles on neonatal pain published from 1965 to 1999. Pain Res Manag 2001 Spring;6(1):45-50

Address for Correspondence

Kumara Mendis MBBS, MD (Family Medicine), MSc (Medical Informatics) is the Head of the Computer Centre, Faculty of Medicine, University of Kelaniya, Sri Lanka. Graduated in Medicine from University of Colombo and specialized in Family Medicine. Interest in Electronic Medical Records lead to the sub-specialty of Medical Informatics with a Masters degree from Erasmus University, The Netherlands. Pioneered the starting of a Computer Centre in a Sri Lankan Medical Faculty. Interested in using ICT to facilitate medical teaching/learning with Intranets and dynamic enabled-web sites. Pragmatic Medical Informatics that helps to increase the quality of care of the frontline doctors is his main focus. Currently involved in developing an EPR for a developing country primary care doctor with prescription decision support to be system used at the time of the consultation. Member of the Classification and Medical Informatics committees of the World Organization of Family Doctors - WONCA. President elect of Health Informatics Society of Sri Lanka -HISSL.

Email: kmendis@sltnet.lk kmendis@mfac.kln.ac.lk

Figure 1. The PubMed Grabber-Analyzer: Main Interface.
After importing the CMJ MEDLINE format file



Elapsed Time 16

100

Record Num	UI	PMID	TI	CY	SO	DP	AD	AB
1	22154646	12164046	Reducing at	Sri Lanka	Ceylon Med J 2001 D	2001 Dec		
2	22154645	12164045	Prenatal dia	Sri Lanka	Ceylon Med J 2001 D	2001 Dec		
3	22154644	12164044	Bowel prepa	Sri Lanka	Ceylon Med J 2001 D	2001 Dec		
4	22154643	12164043	Rupture of tl	Sri Lanka	Ceylon Med J 2001 D	2001 Dec		
5	22154642	12164042	Rare presen	Sri Lanka	Ceylon Med J 2001 D	2001 Dec		
6	22154641	12164041	Multiple scle	Sri Lanka	Ceylon Med J 2001 D	2001 Dec		
7	22154640	12164040	Unusual cas	Sri Lanka	Ceylon Med J 2001 D	2001 Dec		
8	22154639	12164039	CLO test po:	Sri Lanka	Ceylon Med J 2001 D	2001 Dec		
9	22154638	12164038	Arthrogrypos	Sri Lanka	Ceylon Med J 2001 D	2001 Dec	Base Hospital, Mav	
10	22154637	12164037	Castleman's	Sri Lanka	Ceylon Med J 2001 D	2001 Dec	Faculty of Medical !	
11	22154636	12164036	A patient pre	Sri Lanka	Ceylon Med J 2001 D	2001 Dec	Department of Clini	
12	22154635	12164035	Cutaneous l	Sri Lanka	Ceylon Med J 2001 D	2001 Dec	Faculty of Medicine	
13	22154634	12164034	The caesare	Sri Lanka	Ceylon Med J 2001 D	2001 Dec	Department of Obsl	
14	22154633	12164033	Detection of	Sri Lanka	Ceylon Med J 2001 D	2001 Dec	Faculty of Medicine	
15	22154632	12164032	Hunting jelly	Sri Lanka	Ceylon Med J 2001 D	2001 Dec		
16	22154631	12164031	Intermittent (	Sri Lanka	Ceylon Med J 2001 D	2001 Dec	Department of Obsl	
17	22154630	12164030	Ecchymoses	Sri Lanka	Ceylon Med J 2001 D	2001 Dec	Faculty of Medicine	
18	22154629	12164029	Compliance	Sri Lanka	Ceylon Med J 2001 D	2001 Dec	Department of Com	
19	22154628	12164028	Managemer	Sri Lanka	Ceylon Med J 2001 D	2001 Dec	Urology Unit, Teacl	
20	22154627	12164027	A preliminary	Sri Lanka	Ceylon Med J 2001 D	2001 Dec	Department of Psyc	
21	22154626	12164026	The caesari	Sri Lanka	Ceylon Med J 2001 D	2001 Dec	Department of Obsl	
22	22154625	12164025	Leptin, the a	Sri Lanka	Ceylon Med J 2001 D	2001 Dec	Faculty of Medicine	
23	21606421	11763842	Russell of R	Sri Lanka	Ceylon Med J 2001 Ju	2001 Jun	Department of Phai	
24	21589755	11732305	Central neur	Sri Lanka	Ceylon Med J 2001 Se	2001 Sep	Institute of Neurolo	
25	21589754	11732304	Fish bone m	Sri Lanka	Ceylon Med J 2001 Se	2001 Sep	Department of Phai	

Figure 2: Relationships of the Access tables

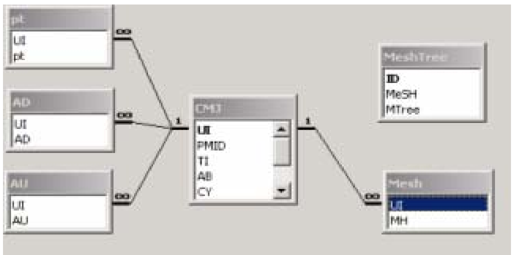


Figure 3: PubMed Analyzer – The Search interface

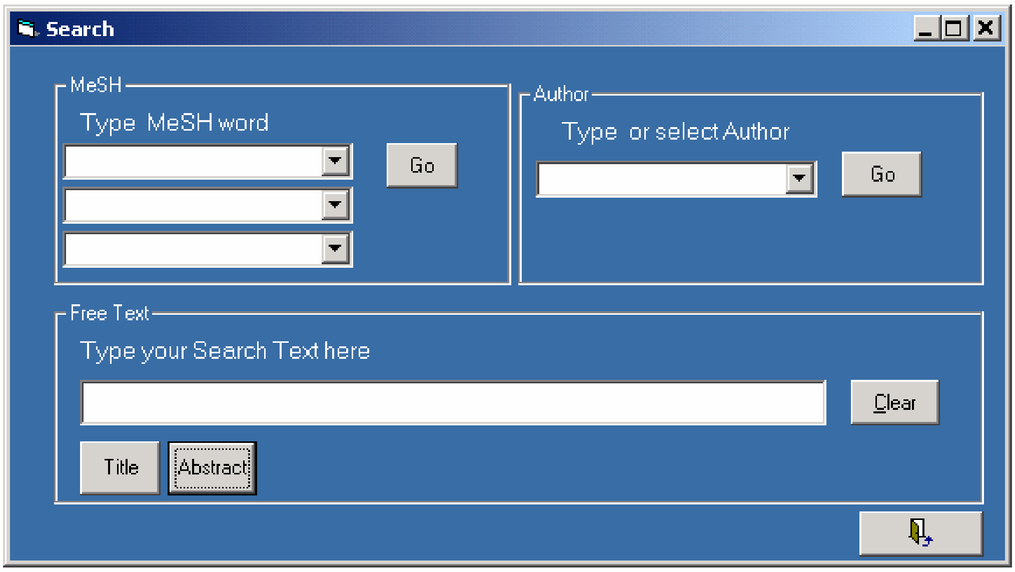
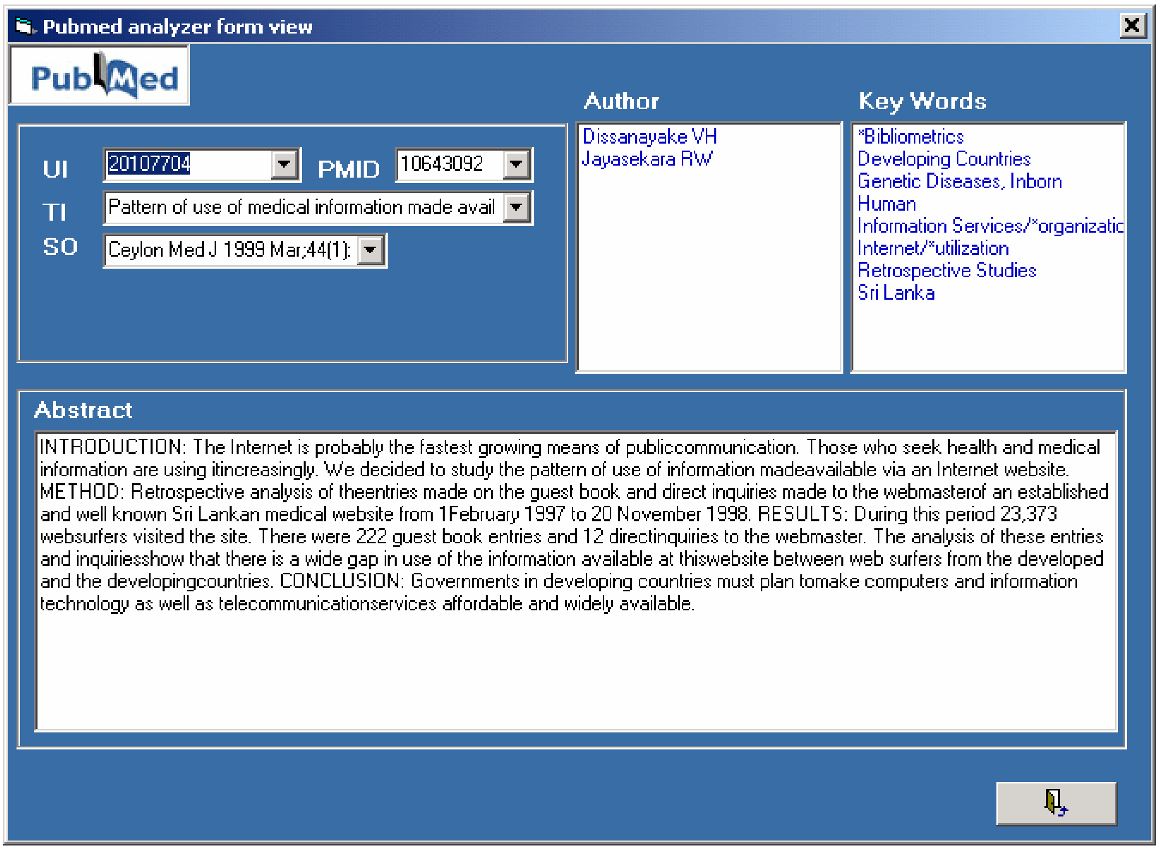


Figure 4: PubMed Analyzer – The Form view
Displaying a single MEDLINE format citation with abstract



UI 20107704 PMID 10643092

TI Pattern of use of medical information made avail

SO Ceylon Med J 1999 Mar;44(1):

Author

Dissanayake VH
Jayasekara RW

Key Words

*Bibliometrics
Developing Countries
Genetic Diseases, Inborn
Human
Information Services/*organizatic
Internet/*utilization
Retrospective Studies
Sri Lanka

Abstract

INTRODUCTION: The Internet is probably the fastest growing means of publiccommunication. Those who seek health and medical information are using itincreasingly. We decided to study the pattern of use of information madeavailable via an Internet website. METHOD: Retrospective analysis of theentries made on the guest book and direct inquiries made to the webmasterof an established and well known Sri Lankan medical website from 1February 1997 to 20 November 1998. RESULTS: During this period 23,373 websurfers visited the site. There were 222 guest book entries and 12 directinquiries to the webmaster. The analysis of these entries and inquiriesshow that there is a wide gap in use of the information available at thiswebsite between web surfers from the developed and the developingcountries. CONCLUSION: Governments in developing countries must plan tomake computers and information technology as well as telecommunicationservices affordable and widely available.

References

[i] MEDLINE Fact sheet: http://www.nlm.nih.gov/pubs/factsheets/medline.html (accessed on 11 September 2003)

[ii] PuibMed Help: http://www.ncbi.nlm.nih.gov/entrez/query/static/help/pmhelp.html (accessed on 11 September 2003)

iii Reference Manager: http://www.refman.com/

[iv] Endnote www.endnote.com/

[v] Moorman PW, van der Lei J. An inventory of publications on computer-based medical records: an update. Methods Inf Med. 2003;42(3):199-202

[vi] Simo Minana J, Gaztambide Ganuza M, Latour Perez J. Atencion Primaria journal in MEDLINE: an analysis of the first 7 years of indexing. Aten Primaria 1999 May;23 Suppl 1:5-13

[vii] Kumara Mendis, C.Weerabaddana, DIK Solangarachchi, CH Wanniarachchi. Three decades of Ceylon Medical Journal – an analysis using PubMed. Sri Lanka Medical association – 116th Anniversary academic sessions. March 26th -29th, 2003, Colombo Sri Lanka

[viii] MeSH Fact Sheet http://www.nlm.nih.gov/mesh/filelist.html

[ix] MeSH tree: http://www.nlm.nih.gov/mesh/filelist.html

[x] Banos JE, Ruiz G, Guardiola E. An analysis of articles on neonatal pain published from 1965 to 1999. Pain Res Manag 2001 Spring;6(1):45-50