Dynamic PubMed:
Self-updating software for PubMed bibliometric research

Kumara Mendis, Rasika Amarasiri
University of Sydney, Monash University
kmendis@med.usyd.edu.au

Abstract

Bibliometric analyses are increasingly being used to monitor research quality assessment and for synthesizing research evidence. The basic unit in bibliometric analysis is the journal article that is indexed in electronic bibliographic knowledge bases such as PubMed. Our aim is to demonstrate a methodology to auto-update PubMed bibliometric analysis that measures the quantity of publications and displays in a tabular and graphical format in a web interface.

PubMed e-Utilities along with a normal PubMed query was used to interrogate the database. A web-server based MySQL stored the queries and results. An online PC acted as a scheduler client also has several Java applications that are executed at scheduled time intervals to update the system. The Java applications query the MySQL database located at the web server to retrieve the queries that need to be executed. It then formulates the automated queries to be sent to the e-utilities interface with the three second time intervals. The resulting data are stored back in the MySQL database. Once all the queries have been executed, the Java application executes the scripts in the web server to update the graphs with the new information.

The system automatically updates many types of queries both yearly and monthly intervals. We have automated country specific total PubMed publications, discipline specific updates such as 'general practice' or 'health informatics' or a particular fast developing field such as 'biotechnology and genetics'. Comparison between countries can also be depicted as line charts.

The incorrect or incomplete use of the author, institution and country affiliation remains major problems for the accuracy. Errors on updating the PubMed queries were dependant on the time of the update. This was maximal at peak-times when the server load was highest.

1. Introduction

Economically developed nations are increasing their investments on health and medical research (HMR) at a rapid rate. Australia [1] and United Kingdom [2] are at the forefront of the Organization of Economic Co-operation and Development (OECD) countries which have almost doubled the HMR expenditure every few years. This huge increase in HMR funding makes tracking, assessment of funding imperative. Research quality assessment (RQA) processes has been under much discussion in the U.K. and Australia. Australia's RQA process is largely based on the UK model [3]. Traditional peer-review methods for RQA have come under scrutiny. Metric-based methods which include research income, publications, citations and the number of research students are increasingly being proposed for consideration as an alternative to the traditional peer-review processes [4] because they cost less and have been found to predict the same results as the peer-review methods.

The widening research-practice gap is a major concern for the practice of evidence based health care [5]. Traditional medical textbooks become out of date even before they are published. Investing on journals and electronic sources that synthesize evidence from the critically appraised top journals and is updated at least one a year is recommended for clinicians [6].

For synthesizing research evidence or for monitoring research funding using metric methods, the basic unit is the journal article that is published in electronic bibliographic knowledge bases such as PubMed [7], PsycInfo [8] or ISI Web of Science [9]. Publications are not the only, but certainly very important elements in the medical knowledge exchange process.

PubMed is the web interface of Medline in public domain that is commonly used by clinicians, academics and the public. Medline is the largest database that is accessed when using PubMed and currently contains more than 15 million bibliographic citations. PubMed uses the term 'citation' to refer to individual publications. To avoid confusion we will use the word 'publication' exclusively from this point. PubMed indexes more than 4,800 biomedical journals published more than 70 other countries. PubMed is essentially dynamic with an addition of between 1500-3500 completed references are added each day [10].

Bibliometrics is defined as 'the use of statistical methods in the analysis of a body of literature to reveal the historical development of subject fields and patterns of authorship, publication, and use. It was formerly called statistical bibliography' [11]. Bibliometric studies form a relatively minor group in PubMed publications with a total of approximately 1700 articles up to April 2006. It has however increased from 158 in 2000 to 217 in 2005. The types of bibliometric analysis are wide and cover many domains. It ranges from assessing what different countries get for their research spending [12], analysis of a specific domain like Parasitology [13] or to a single medical journal [14]. How PubMed citations can be used to track HMR funding in Australia was another publication using bibliometric methods to track expenditure [15]. In the debate over whether the academics are losing control over clinical research, the much needed empirical data is worked out using metric analyses [16]

However the continuous updating of bibliographic databases results in bibliometric studies that will be out of date no sooner the number of publications is retrieved to complete an article. Our aim is to demonstrate a methodology to auto-update PubMed bibliometric analysis that measures the quantity of publications in tabular and a graphical web-based interface.

2. Methods

Architecture of the Dynamic PubMed Graphing System

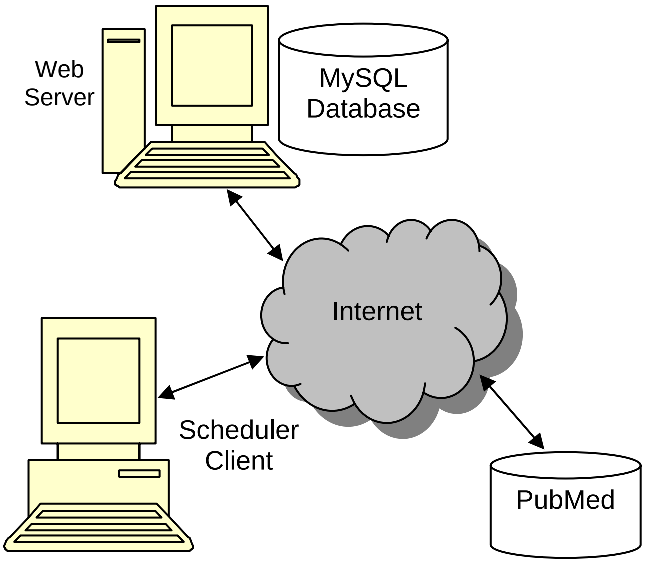


Figure 1 - Architecture of the Dynamic PubMed Graphing System

The basic architecture of the Dynamic PubMed Graphing System is illustrated in Figure 1. The system consists of three main components:

1. The PubMed online database.
2. A web server that is used to deliver the dynamic graphs over the web.
3. A scheduler client that updates the graphs at scheduled time intervals.

PubMed publication consists of many so-called ‘tags’ that are abbreviated names for the different fields. The tags are structured in a way that is similar to a structured abstract of a journal article but they are more comprehensive. The fields that were used in this analysis are: author affiliation (AD), publication type (PT), medical subject heading (MeSH), title (TI) and all fields [ALL].

The AD tag is the field that includes institutional affiliation and address of the first author. The information in this field was used to obtain the publications from Australia. We searched for the word ‘Australia’ and also for all state and territory names in the AD tag. The AD tag was also used to count the publications originating from the universities. The PT tag was used to pick out a specific publication type - ‘clinical trials’.

The MeSH tags are the main key words of the Medline database. Every year the MeSH words are revised - some are added and others deleted. Currently Medline has about 22,997 MeSH words. The specificity of PubMed can be increased when MeSH words are used to query Medline

The PubMed online database is the source for all the main information presented in the graphs generated by the Dynamic PubMed Graphing System. The e-utils service [17] provided by PubMed was used to query the database. The EGQuery option of the e-utils suite returns the count of records found in each and every database in the NCBI. A query sent to the EGQuery interface formulated as a HTTP post request results in an XML output with the following format:

```
<?xml version="1.0"?>
.......................
<Result>
  <Term>stem cells</Term>
  <eGQueryResult>
     <ResultItem>
        <DbName>pubmed</DbName>
        <MenuName>PubMed</MenuName>
        <Count>151496</Count>
        <Status>Ok</Status>
     </ResultItem>
.......................
  </eGQueryResult>
</Result>
```

An instance of <ResultItem> element is generated for each database entry. The <Count> element carries the number of citations in that database and the Status entity gives a feedback as to what resulted from the query of that database.

The queries that need to be executed are stored in a MySQL database located in the web server. This database also stores the queried results and other details for generating the graphs. Scripts in the web server are used to generate the dynamic graphs.

An online PC is used as the scheduler client. This machine has several java applications that are executed at scheduled time intervals to update the system. The Java applications query the MySQL database located at the web server to retrieve the queries that need to be executed. It then formulates the automated queries to be sent to the e-utils interface with the correct time intervals. The resulting data are stored back in the MySQL database. Once all the queries have been executed, the Java application executes the scripts in the web server to update the graphs with the new information.

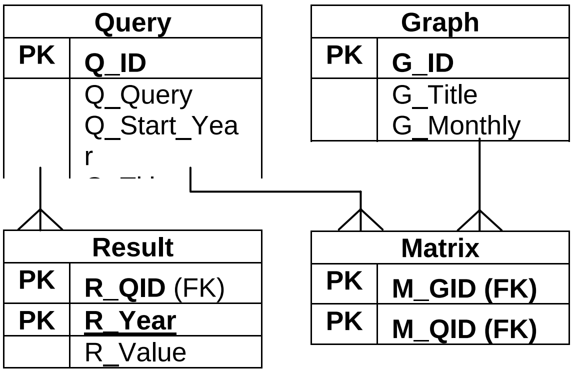
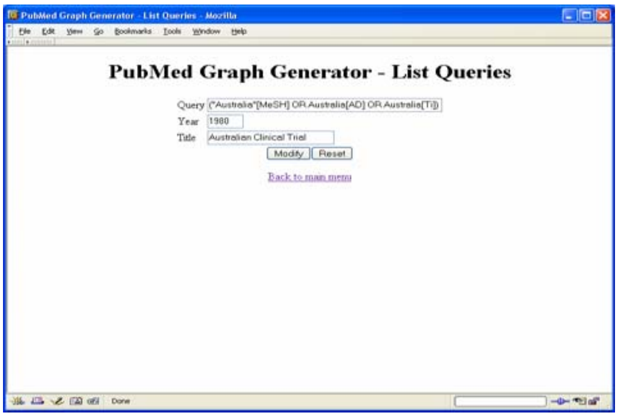


Figure 2. Database structure for the Dynamic PubMed Graphing System.

The system is designed to use the results from the same query in multiple graphs. To achieve this, the database was designed to have separate tables for the queries and the graphs. A linking table was used to connect the queries to the graphs. This architecture eliminates the requirement of updating the same query repeatedly from PubMed.


PubMed Graph Generator - List Queries - Mozilla
File Edit View Go Bookmarks Tools Window Help
Modify
Reset
Done

PubMed Graph Generator - List Queries

Query ("Australia"[MeSH] OR Australia[AD] OR Australia[TI])

Year 1980

Title Australian Clinical Trial

Back to main menu

Figure 3. Query Editor of the Dynamic PubMed Graph Graphing System

A screenshot of a typical dynamically update graph is illustrated in Figure 4. The generation of the graphs is done using a PHP script in the web server. In order to reduce the load on the server on generating the graphs for each and every request for a graph, the graphs and the relevant pages are generated only once after the database has been updated. Static images and HTML pages are created by a PHP script in the server that is called by the Java application that updates the database.

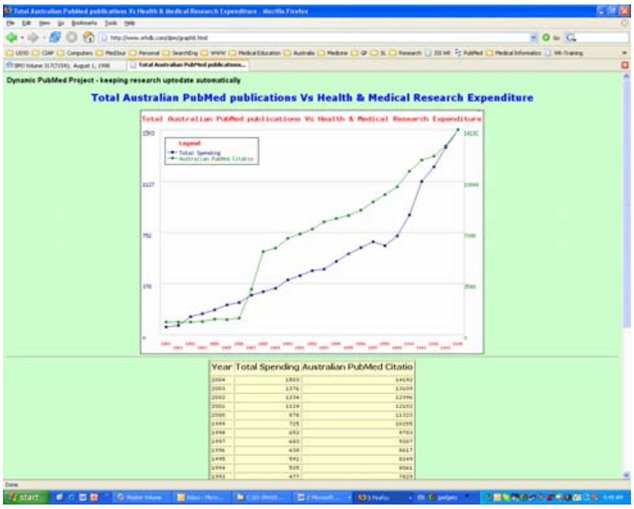
The structure of the database is illustrated in Figure 2. The query table stores the details of the queries that will be used to update the graphs. The query is stored in the format it is presented to the PubMed e-utils interface except for the date range. The starting year and a title for the query are also stored in the table.

The graph table carries information about the graph. This includes the title of the graph and an indication whether it is a monthly graph or a yearly graph.

The matrix table is used to add multiple queries to a graph. The Dynamic PubMed Graphing System can display results from up to 5 queries in a single graph.

The results from the queries are stored in the result table. A similar table is used to store the results from the monthly queries.

A web based administration interface was also developed to add and modify the queries and the graphs that would be used in the system. A screenshot of adding a new query to the system is illustrated in Figure 3.


Total Australian PubMed publications Vs Health & Medical Research Expenditure - Mozilla Firefox
Legend
Total Spending
Australian PubMed Citatio
start

Dynamic PubMed Project - keeping research uptodate automatically

Total Australian PubMed publications Vs Health & Medical Research Expenditure

Year	Total Spending	Australian PubMed Citatio
2004	1503	14192
2003	1376	13109
2002	1234	12396
2001	1124	12102
2000	878	11323
1999	725	10255
1998	652	9702
1997	603	9207
1996	638	8617
1995	592	8249
1994	535	8061
1993	477	7829

Figure 4 A typical dynamically generated chart with part of the table

3. Results

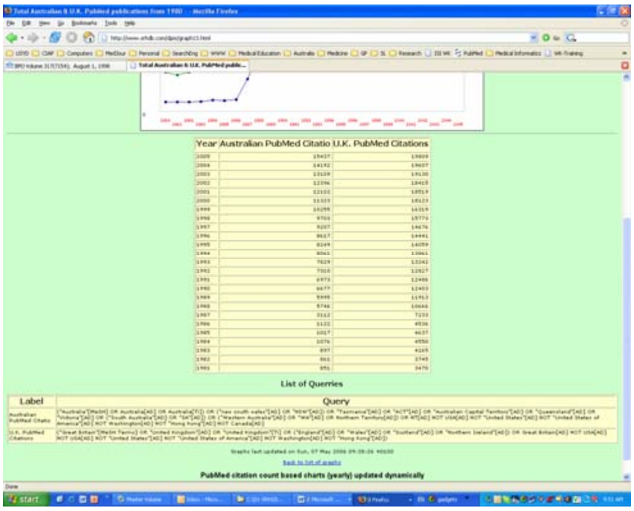
Typical Dynamic PubMed web pages include: (Fig 4)

a) The query or more than one query that was used to produce the results.

b) The results in Tabular and graphical form.

c) The date and time of last update.

Dynamic PubMed can also include data that do not get updated regularly - for e.g. the annual HMR non-governmental expenditure of Australia alongside dynamically updating data. These can be manually updated into the MySQL tables periodically.


Total Australian & U.K. PubMed publications from 1980 - Mozilla Firefox
Total Australian & U.K. PubMed publi...
Done
start

Year	Australian PubMed Citatio	U.K. PubMed Citations
2005	15437	19809
2004	14192	19607
2003	13109	19130
2002	12396	18415
2001	12102	18519
2000	11323	18123
1999	10255	16319
1998	9703	15771
1997	9207	14676
1996	8617	14441
1995	8249	14059
1994	8061	13861
1993	7629	13242
1992	7310	12827
1991	6973	12486
1990	6677	12403
1989	5995	11913
1988	5746	10666
1987	3112	7233
1986	1122	4536
1985	1017	4637
1984	1076	4550
1983	897	4165
1982	861	3745
1981	851	3470

List of Querries

Label	Query
Australian PubMed Citatio	("Australia"[MeSH] OR Australia[AD] OR Australia[TI]) OR ("new south wales"[AD] OR "NSW"[AD]) OR "Tasmania"[AD] OR "ACT"[AD] OR "Australian Capital Territory"[AD] OR "Queensland"[AD] OR "Victoria"[AD] OR ("South Australia"[AD] OR "SA"[AD]) OR ("Western Australia"[AD] OR "WA"[AD] OR Northern Territory[AD]) OR NT[AD] NOT USA[AD] NOT "United States"[AD] NOT "United States of America"[AD] NOT Washington[AD] NOT "Hong Kong"[AD] NOT Canada[AD]
U.K. PubMed Citations	("Great Britain"[MeSH Terms] OR "United Kingdom"[AD] OR "United Kingdom"[TI] OR ("England"[AD] OR "Wales"[AD] OR "Scotland"[AD] OR "Northern Ireland"[AD]) OR Great Britain[AD] NOT USA[AD] NOT USA[AD] NOT "United States"[AD] NOT "United States of America"[AD] NOT Washington[AD] NOT "Hong Kong"[AD])

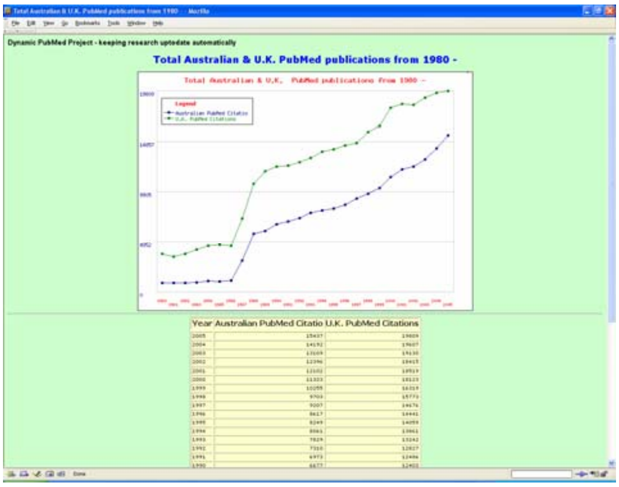
Graphs last updated on Sun, 07 May 2006 09:38:26 +00:00

Back to list of graphs

PubMed citation count based charts (yearly) updated dynamically

Figure 5 A typical dynamically generated table, related query and time of last update (with part of the chart)

We have also designed to have two types of charts - yearly and monthly. The annual update (for e.g. between two countries publications) gives a long-term trends (Fig 6) and the monthly update show the current trends (Fig. 7).

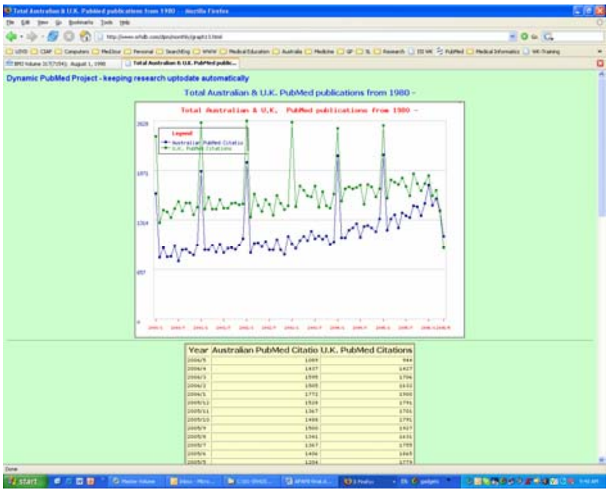

Total Australian & U.K. PubMed publications from 1980 - Mozilla
Dynamic PubMed Project - keeping research uptodate automatically
Total Australian & U.K. PubMed publications from 1980 -
Total Australian & U.K. PubMed publications from 1980 -
Legend
Australian PubMed Citatio
U.K. PubMed Citations
Done

Year	Australian PubMed Citatio	U.K. PubMed Citations
2005	15437	19809
2004	14192	19607
2003	13109	19130
2002	12396	18415
2001	12102	18519
2000	11323	18123
1999	10255	16319
1998	9703	15771
1997	9207	14676
1996	8617	14441
1995	8249	14059
1994	8061	13861
1993	7629	13242
1992	7310	12827
1991	6973	12486
1990	6677	12403

Figure 6 Total annual Australian and U.K. PubMed publications from 1960 - 2006

4. Discussion

Dynamic-PubMed software can automatically run a PubMed query at a given time and frequency and write the publication counts to a relational database. This data is displayed as a line charts on a web page. In addition the results are also shown in tabular form. The results of a query can be displayed in two modes – a long-term summary over more than a decade and a more detailed monthly summary to view the current variation in trends.


Total Australian & U.K. PubMed publications from 1980 - Mozilla Firefox
Total Australian & U.K. PubMed publi...
Dynamic PubMed Project - keeping research uptodate automatically
Total Australian & U.K. PubMed publications from 1980 -
Total Australian & U.K. PubMed publications from 1980 -
Legend
Australian PubMed Citatio
U.K. PubMed Citations
start

Year	Australian PubMed Citatio	U.K. PubMed Citations
2006/5	1069	944
2006/4	1437	1627
2006/3	1595	1706
2006/2	1505	1632
2006/1	1772	1900
2005/12	1519	1791
2005/11	1367	1701
2005/10	1488	1791
2005/9	1500	1927
2005/8	1341	1631
2005/7	1367	1759
2005/6	1406	1865
2005/5	1234	1779

Figure 7 Total monthly Australian and U.K. PubMed publications from 2000 - 2006

Problems encountered during implementation

During the implementation of the system several problems were encountered. These included problems associated with access to the Internet and problems in the PubMed server. Some of these problems and the suggested remedies are as follows:

1) Accessing the e-utilities interface

Since the web service provider did not allow accessing the Internet from the web server, the system had to be broken down into two parts. One to actually display the graphs on the web server and the other to dynamically query the PubMed database via the e-utilities interface.

The solution adopted was to develop the part that updated the graphs as a separate Java application. The Java application would access the e-utilities interface and update the MySQL database in the web server. The Java application was installed on a PC and scheduled to run daily at a pre-determined time interval that was recommended by PubMed.

The Java application was developed in such a way that the queries to the e-utilities interface were done at intervals of three seconds as recommended by PubMed. This was to avoid the application from flooding the e-utilities interface with queries.

2) HTTP errors from e-utilities

While accessing the e-utilities interface it was noted that at times the result from the EGQuery interface did not work. A detailed analysis showed that the interface was returning a HTTP 500 error. The code corresponds to an internal error in the system where the software responsible for generating the dynamic results has generated an error.

As there was no valid count that could be generated from a page containing an HTTP 500 error, the Java application was modified to detect these errors and retry the query until a valid result was returned.

Recently several of the graphs started showing incomplete results and a further investigation revealed that the EGQuery interface was now generating some HTTP 502 errors. These errors occur when the server gets overloaded. The error is reported from the proxy or a load balancer that may be sitting in front of the web server.

Similar to the HTTP 500 error, the Java application was modified to detect this error and retry the query after a short time interval.

3) Reducing the number of queries

It was notes that citations that were older than about five years were rarely getting updated. Therefore, it was unnecessary to update these values daily. Some of the graphs had queries dating back to 1961 and these queries seldom returned a modified value. In order to reduce the number of queries sent to e-utilities it was decided to update the values of queries related to citations older than year 2000 only once a month.

A new field was added to the query table to record the last full update. If the last full update was done more than a month ago, the query will be executed for the full date range and from year 2000 onwards otherwise.

The above modification resulted in a significant reduction in the number of queries sent to PubMed and an increase in the performance.

4) Port blockings

Due to the security setups in the academic network it was not possible to implement the system initially. The web access had to be done through a proxy server and access to the MySQL database on the web server was not possible at all. These had to be resolved with the assistance from the network administrators.

PubMed and Bibliometrics

PubMed was used as it is one of the largest biomedical bibliographic databases in the public domain and facilitates access to the ‘raw’ data of indexed biomedical publications. PubMed improved access to the database recently by enhancing the functionality of e-utilities with code examples to the ENTEZ engine [18]. This makes it easier when designing software.

Bibliometric assessment of research performance is based on one central assumption - ‘scientists who have to something worthwhile to contribute publish their findings in the international journals’. One of the largest biomedical knowledgebase in public domain is PubMed and the high visibility of Australian [19] publications in international biomedical literature make this combination suitable for Australia. Furthermore bibliometric analyses performed at the macro-level (e.g., a whole country) yield at best general assessments of fields as a whole [20], for instance, how good a country's performance is in physics, chemistry, psychology or immunology, without a reliable breakdown to the individual research groups or programs.

Querying bibliographic databases is a reliable way of retrieving information. More than three quarters of the studies included in systematic reviews are identified through electronic bibliographic databases that are searched using Boolean logic [21]. However, the limitations imposed by the absence of important journals and papers from a single database analysis such as PubMed may be of concern.

Most health-related bibliometric analyses use the ISI databases and particularly Impact Factor related metrics. However, the impact factor has its own critics [22]. Prevention Research Centre is one of the Centers for Disease Control (CDC) largest funded programs [23]. The PRC’s influence factor (the number of PRC peer-reviewed publications that considered a journal highly influential) was only weakly correlated with the ISI impact factor. This suggested that the ISI metric and the PRC judgment are not closely aligned.

It has also been pointed out that counting publications (without citations) would not give a good indication of the quality or the impact. It has been difficult anyhow to ascertain the real impact on clinical practice by even tracking highly promising articles. One bibliometric study concluded that [24] ‘one in four promising technologies resulted in a published randomized control trial and fewer than one in ten entered routine clinical use within 20 years of the index basic science publication’.

Evaluation of scientific research is crucial with the increase amounts of funding that goes into HMR research. Peer review is typically a predominantly qualitative assessment of research performance and can be influenced by personal prejudices and preferences. Bibliometric indicators discussed here represent the quantitative measurements. But quantitative elements may be also present in peer review, e.g., number of publications in high prestige scientific journals. Conversely, citations given to research work can be seen as judgments, "votes" of colleague-scientists in favour of the work cited.

Bibliometric analyses with PubMed data is dependant on the accuracy provided in the relevant tags. The accuracy of indexing for authorship and country is dependant on the information included on the [AD] tag. This includes the first authors information. The similar state names (Victoria in Australia and Canada) and abbreviations (WA – Western Australia, Washington) can give erroneous counts even at country level. Similarly the [AU] tag has the author name information. The inconsistency of the name submitted to journal article may adversely affect the individual publication counts.

There have been other bibliographic software utilities like BIRS [25] that assists the end users to get better understanding of their search domain, formulate and expand their search queries, and visualize the bibliographic research results. The Dynamic-PubMed will be just one step the beginning of automating bibliometric analysis.

Dynamic PubMed may be used as a tool that automatically keeps track of the number of PubMed research publications to ascertain both long and short term trends. It will be a good tool to track the results of short term HMR funding at country, institution or specific specialty or domain level.

References

[1] The Hon Tony Abbott MHR - Media Releases. $10 million boost for health research- 21/03/2006. http://www.health.gov.au/internet/ministers/publishing.nsf/Content/health-mediarel-yr2006-ta-abb033.htm?OpenDocument&yr=2006&mth=3 (Accessed 14-05-2006)

[2] Improving the nation's health, increasing the nation's wealth: A new deal for medical research HM Treasury. http://www.hmtreasury.gov.uk/newsroom_and_speeches/press/2005/press_100_05.cfm
(Accessed 14-05-2006)

[3] Shewan LG, Coats AJ. The Research Quality Framework and its implications for health and medical research: time to take stock? Med J Aust. 2006 May 1;184(9):463-6

[4] Hobbs FDR, Stewart PM. How should we rate research? BMJ 2006;332:983-984

[5] Haynes RB, Hayward RS, Lomas J. Bridges between health care research evidence and clinical practice. J Am Med Inform Assoc. 1995 Nov-Dec;2 (6):342-50.

[6] Sharon E. Straus, W. Scott Richardson, Paul Glasziou, R. Brian Haynes. Evidence Based Medicine 3rd Edition. Churchill Livingstone. 2005

[7] PubMed. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi (Accessed on 15-05-2006)

[8] PsycINFO
http://www.apa.org/psycinfo/ (Accessed on 15-5-2006)

[9] Institute of Scientific Information http://www.isinet.com/ (Accessed on 15-5-2006)

[10] MEDLINE – Fact sheet
http://www.nlm.nih.gov/pubs/factsheets/medline.html
(Accessed on 15-05-2006)

[11] National Library of Medicine. Medical Subject Headings. http://www.nlm.nih.gov/cgi/mesh/2006/MB_cgi?mode=&index=15026&field=all&HM=&II=&PA=&form=&input=
(Accessed on 15-05-2006)
[12] King D. Scientific impact of nations. Nature. VOL 430 15 JULY 2004

[13] Falagas MF, Papastamataki PA, Bliziotis IA. A bibliometric analysis of research productivity in Parasitology by different world regions during a 9-year period (1995–2003). BMC Infectious Diseases 2006, 6:56

[14] Mendis K, Solagarchchi I, Weerabaddana C.
Three decades of the Ceylon Medical Journal— analysis using MEDLINE (PubMed). Ceylon Medical Journal. Vol. 50. No. 1, March 2005

[15] Mendis K, McLean R. PubMed citations as an Index of Outcome for Research: Monitoring the Response to Australian Health and Medical Research Expenditure. (Accepted for publication in the Medical Journal of Australia)

[16] Patsopoulos NA, Ioannidis JPA, Analatos AA. Origin and funding of the most frequently cited papers in medicine: database analysis. BMJ 2006;332:1061-1064

[17] Sayers E, Wheeler D. Building Customized Data Pipelines Using the Entrez Programming Utilities (eUtils). http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=coursework.chapter.eutils (Accessed 14-05-2006)

[18] ENTEZ The life sciences search engine.
http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi
(Accessed on 15-05-2006)

[19] Butler L. Explaining Australia's increased share of ISI publications-the effects of a funding formula based on publication counts. Research Policy 2003; 32: 143-155

[20] van Raan AFJ. The use of bibliometric analysis in research performance assessment and monitoring of interdisciplinary scientific
developments. Technikfolgenabschätzung -Theorie und Praxis Jahrgang - März 2003, S. 20-29.
http://www.itas.fzk.de/tatup/031/raan03a.htm
(Accessed on 15-05-2006)

[21] Sampson M, Barrowman NJ, Moher D, Clifford TJ, Platt RW, Morrison A, Klassen TP, Zhang L. Can electronic search engines optimize screening of search results in systematic reviews: an empirical study.
BMC Med Res Methodol. 2006 Feb 24;6:7.

[22] Lundberg G. The "omnipotent" Science Citation Index Impact Factor Med J Aust 2003 178: 253-254

[23] Franks AL, Simoes EJ, Singh R, Sajor Gray B. Assessing prevention research impact a bibliometric analysis. Am J Prev Med. 2006 Mar;30(3):211-6.

[24] Contopoulos-Ioannidis DG, Ntzani E, Loannidis JP. Translation of highly promising basic science research into clinical applications, Am J Med 2003; 114: 477–484

[25] Ding Y, Chowdhury CG, Foo S, Qian W. Bibliometric Information Retrieval System (BIRS): A Web Search Interface Utilizing Bibliometric. Journal of the american society for information science. 51(13):1190–1204, 2000