Scientometrics
DOI 10.1007/s11192-016-2007-1

Bibliometric analysis of Egyptian publications on Hepatitis C virus from PubMed using data mining of an in-house developed database (HCVDBegy)

Hanaa M. H. Alam El-Din[1,2] · Ahmed Sharaf Eldin[1] · Amro M. S. A. Hanora[3]

Received: 26 January 2016

Abstract Bibliometric analysis of Egyptian literature on HCV provides the intelligence needed for decision makers and gives an insight into research productivity in this area. We propose our database (HCVDBegy) on MS-SQL server by querying PubMed for "HCV and Egypt" with time limit till 31st March 2013. Fifty eight out of the 716 records were excluded and the rest 658 were divided into 22 domains. Analysis used data mining add-ins for Microsoft Excel, including association and regression algorithms. A fluctuation in numbers of papers was noticed from 2004 to 2009 with a steady increase onward. Eighty six percent of publications were the contribution of three or more authors. Top publishing bodies were Cairo and Ain Shams Universities, Faculty of Medicine, National Research Center and National Cancer Institute. Three Egyptian journals came on top, whereas other publishing journals were mainly from the USA. Few controlled clinical trials and meta-analyses were published. HCV epidemiology, review articles and sequence analysis domains were the most cited. Forecasting model showed a breakthrough in numbers of publications on 2013 and 2014 than those forecasted. Dependency network based on association rule model of MeSH topics was also extensively analyzed. Number of publications showed a promising increase which points to the better national awareness of HCV problem. Studying MeSH terms clustering showed some hot topics. We recommend that

✉ Hanaa M. H. Alam El-Din
halam63@hotmail.com

Ahmed Sharaf Eldin
Profase2000@yahoo.com

Amro M. S. A. Hanora
ahanora@yahoo.com

[1] Information Systems Department, Faculty of Computers and Information, Helwan University, Cairo, Egypt

[2] Virology and Immunology Unit, Cancer Biology Department, National Cancer Institute, Cairo University, 1st Kasr El-Aini St., Fom El-Khalig, PO Box 11796, Cairo, Egypt

[3] Microbiology and Immunology Department, Faculty of Pharmacy, Suez Canal University, Cairo, Egypt

Published online: 08 June 2016

the PubMed should alarm authors of the challenges of author affiliations. HCVDBegy availability opens the door for more drill down analysis for decision makers.

Keywords Bibliometrics · Data mining · Data base · Association rule · Regression analysis · Clustering · HCV and Egypt · PubMed

Abbreviations

HCV	Hepatitis C virus
XML	Extensible Markup Language
PMID	PubMed ID
BIDS	Business intelligence studio
MeSH	Medical subheadings

Introduction

Hepatitis C virus (HCV) is a major health problem worldwide, especially in Egypt. It has infected approximately 170 million people worldwide. HCV infection is cleared in about 25 % of cases (Alter et al. 1992; Hoofnagle 1997) leaving the rest to suffer chronic infection. Chronic HCV infection can lead to cirrhosis and liver cancer. In Egypt, however, HCV incidence rates have been estimated to be 2.4 per 1000 person-years (165,000 new infections annually) (Mostafa et al. 2010). In 2008, nearly 15 % of the population aged 15–59 years had antibodies to HCV (anti-HCV), and 10 % (approximately 5 million persons) had chronic HCV infection (El-Zanaty and Way 2009); overall, an estimated 6 million Egyptians had chronic HCV infection in 2008 (Centers for Disease Control and Prevention 2012).

Recently, the Egyptian Government has adopted a multidisciplinary national campaign to control HCV. Quantitative evaluation of scientific production in this field will thus be of a great asset for the decision makers in this country as well as for the scientific community concerned. Many methods are employed to satisfy this goal, the most important of which is bibliometric analysis.

Bibliometrics studies are performed to assess scientific literature in a given field. It was not until 1969 that the term *bibliometrics* first appeared in print (Pritchard 1969). It was defined as the 'application of mathematical and statistical methods to books and other media of communication', and the term was quickly adopted and used, particularly in North America (Wilson 1999). At almost the same time, Nalimov and Mulchenko (1969) mentioned the term *scientometrics* to refer to 'the application of quantitative methods which are dealing with the analysis of science viewed as an information process. However, this term was widely used in Europe (Wolfram 2003).

Many studies in different research domains were recently published to address knowledge production based on the MEDLINE database, among others (Alvi et al. 2014; Ghojazadeh et al. 2014; Thavamani 2013; Yao et al. 2012; Nichols 2008; Ramakrishnan and Babu 2007). They were mainly concerned with bibliometrics analysis because the recognition and exchange of research results are among the key forces for the advancement of science.

A group of research articles was concerned with bibliometrics and scientemtric of HCV. Thavamani (2013) published a bibliometric analysis of the literature output in the field of Hepatitis C covered in the Journal Gastroenterology from 2006 to 2010 (Thavamani 2013).

Another scientometric study on HCV world literature during 1999–2013 from Scopus was done by Alvi et al. (2014). On the other hand, another bibliometric study done by Ramakrishnan and Babu (2007), analyzing literature of hepatitis in Medline and two other databases. The latter focused on the degree of collaboration among authors (Ramakrishnan and Babu 2007). None of them however was specifically concerned with knowledge production of Egyptian literature in this field.

In this study, we analyzed Egyptian HCV-related literature from PubMed to evaluate their current status and provide an insight towards their productivity in one of the Egypt's national and major health problem. For this purpose, we developed a specialized database (HCVDBegy) for the ease of assessing related literature. We tried to use the powerful capabilities of relational database for this analysis instead of the traditional ways applied for bibliometric analysis. Thus we followed the lead and the same schema of this previous study (Oliver et al. 2004) in order to integrate the downloadable XML files available from PubMed on the records using the query "HCV and Egypt", up till the end of March 2013, into relational database using MS-SQL server 2008 R2 (Microsoft SQL). Analysis services of the previous program was used to create a data ware house that was suitable for data mining functionalities in Microsoft Excel data mining add-ins. Results of such analysis provide the intelligence needed by decision makers in this country and others.

Materials and methods

PubMed search

The query used was "HCV and Egypt", with time limits till the end of March 2013. On the 29th of November 2013, the query resulted in 716 publications. MEDLINE and Extensible Markup Language (XML) formats were the most conclusive, and we chose the latter for this task.

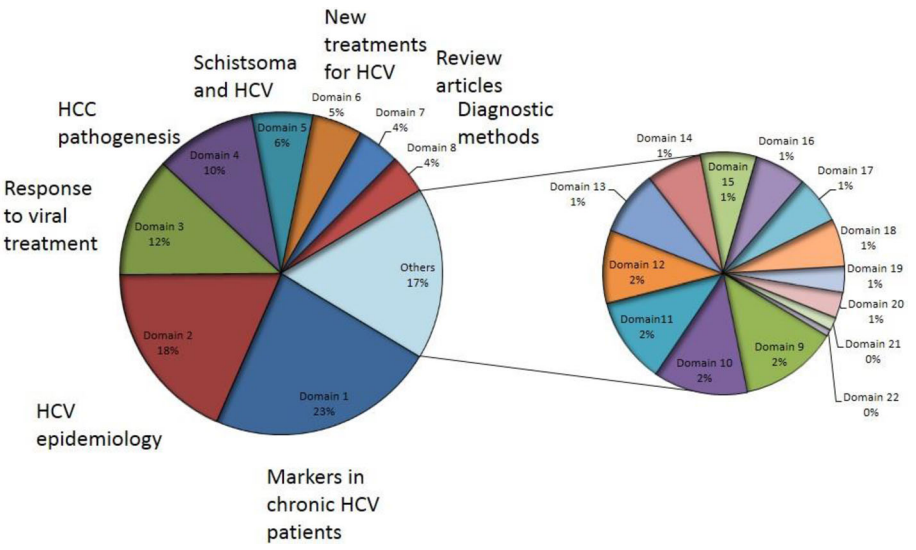


Fig. 1 Different domains (major and minor) of study in the present database

Inclusion and exclusion criteria

Inclusion criteria

(1) They must contain at least one Egyptian author, (2) and samples investigated should contain a group of Egyptian patients in reasonable numbers, (3) also HCV should be the main research area.

Exclusion criteria

(1) Studies with no Egyptian authors (2) and no Egyptian samples investigated, (3) and those with samples selected to be negative for HCV (non-C). (4) Studies where HCV is not the major search theme based on personal judgment and experience were also excluded.

Table 1 Study domains and their citation indexes

No.	Search areas and domains	No. of researches	Range of citation index (average)	Publication years
Major domains				
1	Markers in chronic HCV patients	151 (23.3 %)	75–0 (10)	1980–2013
2	HCV epidemiology	120 (18.5 %)	998–1 (42)	1985–2013
3	Response to viral treatment	79 (12.2 %)	136–1 (18)	1996–2013
4	Hepatocellular carcinoma pathogenesis	67 (10.3 %)	182–1 (20)	1999–2013
5	Schistosoma and HCV	40 (6.17 %)	137–0 (23)	1994–2013
6	New treatments for HCV	33 (5.1 %)	174–0 (20)	2005–2013
7	Review articles	28 (4.3 %)	906–2 (90.5)	1998–2013
8	Diagnostic methods	26 (4 %)	72–0 (14)	1996–2012
	Total	545 (83.6 %)		
Rare or minor domains				
9	Transplantation studies	15 (2.3 %)	64–0 (15)	1995–2013
10	Noninvasive markers of fibrosis	14 (2.2 %)	27–1 (7)	2004–2013
11	Hepatocellular carcinoma positive for HCV	13 (2 %)	43–0 (14)	1993–2011
12	Sequence analysis	11 (1.7 %)	345–1 (73)	1997–2012
13	HCV and other diseases	10 (1.5 %)	32–1 (14)	1997–2011
14	Acute hepatitis	9 (1.3 %)	81–1 (21)	1997–2013
15	Occult HCV	9 (1.3 %)	24–0 (10)	1999–2012
16	Cloning and biotechnology	8 (1.2 %)		2005–2009
17	HCV genotyping	7 (1.1 %)	38–1 (9)	1994–2013
18	HCV and non-Hodgkin's lymphoma	7 (1.1 %)	48–0 (12)	2004–2012
19	HCV and hemodialysis	4 (0.6 %)	19–0 (6)	2007–2012
20	HCV and other organisms	4 (0.6 %)	6–0 (2)	2003–2006
21	HCV mathematical models	2 (0.3 %)	14–0	2010–2011
22	Infection control	1 (0.15 %)	10	2003
	Total	114 (17.4 %)		

Fifty eight studies were thus excluded. The rest 658 records were divided into major and minor areas or domains for building the database (Fig. 1; Table 1).

Citation index

Citations of all papers were collected from Google Scholar (http://scholar.google.com.eg/scholar) after 17 months from the last record (31st/3/213; i.e. on the 1st/10/2014) to allow time for others to cite the most recent records. Records with no citation index on google.scholars were searched for on Google and if not cited there too, they were scored with zero. No corrections for time were done. Chi square test (using SPSS 13.0 program for windows) was used to compare mean citation indexes between different domains.

Authors' affiliations

Analyzing author affiliations was a challenging task since it was subjective and depended on the authors' personal way of data entry. It was not curated in the PubMed nor were there any trials for unifying them for upcoming analysis. In the present study, they were manually curated from misspellings and such, and attempts for unifying them were made whenever feasible. Each field (Department, Center, University, and Governorate) was independently copied and pasted separately in Excel file. They were then integrated into the database for due analysis. On the other hand, 13 % of total citations had missing centers and no attempts were made to supply them for possibility of errors.

Database schema

A paper published back in 2004 showed Tools for loading MEDLINE into a local SQL relational database. We followed the lead of this paper in building up our HCVDBegy from PubMed data (Oliver et al. 2004). We used PubMed ID (PMID) as a primary key. A back up file of this database will be provided on demand.

Analysis of data and data mining functionalities

A data ware house of the present database was constructed for pivot chart building of the different descriptive parameters analyzed. Microsoft SQL Server 2008 R2-analysis services of the business intelligence studio (BIDS) data mining add-ins for Microsoft Excel, include many algorithm types. We used association and regression algorithms for data analysis. Association algorithms find correlations between different attributes in a dataset. Regression algorithms predict one or more continuous variables, based on other attributes in the dataset.

Microsoft regression algorithm (forecast tool)

The Microsoft regression algorithm includes two separate algorithms for analysing time series (https://msdn.microsoft.com/en-us/library/bb677216.aspx). Forecasting for number of publications excluded the first quarter of 2013 since it was inconclusive. Other data was used as a training set and the validation set was the numbers of citations for the years 2013 and 2014 (which are already available from PubMed web site). This analysis was applied to predict the numbers of citations every year till 2017.

Microsoft association algorithm (market basket analysis tool)

The Microsoft Association Rules algorithm is a straightforward implementation of the well-known Apriori algorithm (https://msdn.microsoft.com/en-us/library/ms174916.aspx). It was applied on medical subheadings (MeSH) and author names for each publication as one transaction. The algorithm uses two parameters, support and probability, to describe the item sets and rules that it generates. A dependency network appears which can investigate the interaction of the different items in the model. Each node in the viewer represents an item, while the lines between them represent rules. The selected node in the network highlights dependencies. By selecting a node (in cyan/turquoise), we can see which other nodes which predict the selected item (in Bronze), or which items the current item predicts (in steel blue). In some cases, there is a two-way association between items (nodes in dark magenta/violet), appearing in the same publication. The model could be as complex as the number of nodes they have.

Results

According to what PubMed made available and to what we thought analysis would be useful, we analysed authors and their affiliations, publishing journals, countries of publications, publication types. We chose citation indexes of publications as a sign of recognition of the work done and tried to analyse some merits of those highly cited publications. In additions, we tried to benefit from the powerful business intelligence provided by SQL database analysis service and see if we could predict number of upcoming publications, analyse networks that authors work in (results not shown), and networks of interesting medical subject headings under investigations.

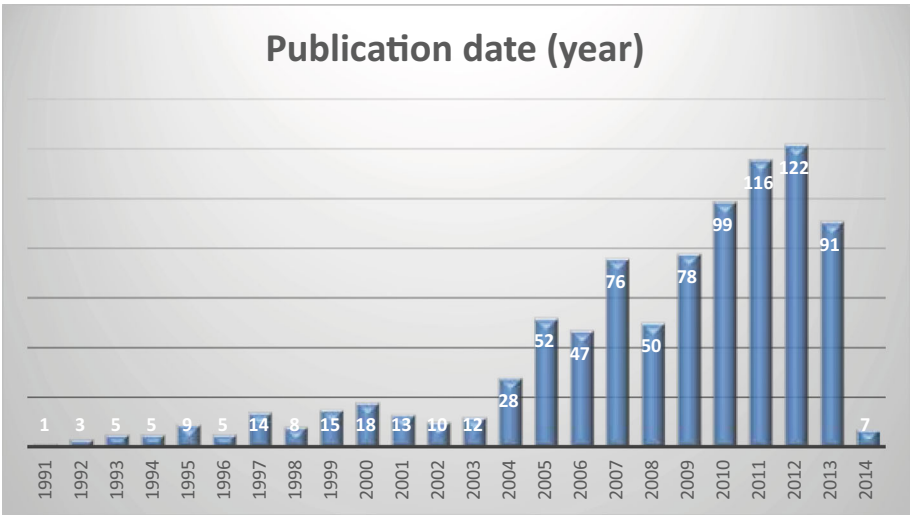


Fig. 2 Number of publications per year in the present database

Descriptive analysis

Total number of citations

Figure 2 shows total number of citations in the present database. The year 2012 showed a number of 122 citations, whereas in the few years right after HCV discovery the number of publications ranged from 1 to 5 citations and never exceeded twenty until 2004. A fluctuation in the number of publications occurred between the years 2004 and 2009. A steady increase was notice from 2009 till the end of study (Table 2).

Authors and their affiliations

Regarding the most influential four authors in the present study, Table 2 shows a comparison between their citation indexes and their types of studies they conducted. Esmat G came first among different authors with 52 journal articles, however, Mohamed MK was the most cited among them. The second and third author have statistically significant

Table 2 Google Scholar citation indexes and publication types of the four most influential authors in the present database

Parameter	Esmat G	Abdel-Hamid M	Mohamed MK	Zekri AR
Sum (SD) of citation indexes	1224 (35)	1949[a] (52)	2360[a] (157)	476[b] (13)
Average (max–min) citation index	23.5 (171–1)	49.8 (247–1)	73.75 (906–2)	18.3 (47–1)
Clinical trials	3	2	2	0
Comparative study	2	3	4	1
Evaluation study	0	0	0	1
Editorial	0	0	1	1
Journal article	52	40	34	27
Multicenter study	1	5	4	0
NIH grant	6	12	2	3
USA support	3	15	6	1
Review articles	3	0	0	0
Validation study	1	0	0	0
Publication years	1994–2013	1994–2013	1996–2013	1996–2013

T test proved a statistically significant difference between a and b ($p = 0.02$)

Table 3 Distribution of single and multiple authors for publications of the present database

No. of authors	No. of publications	% of publications	Rank
1 author	39	6	6
2 authors	53	8.2	4
3 authors	51	7.8	5
4 authors	104	16	3
5 authors	105	16.2	2
>5 authors	297	46	1
Total	649	100	

Fig. 3 Author affiliations. **A** Different Egyptian governorates and their percentages of sharing in the studies of the current database (courtesy of english.ahram.org.e.g.). **B** Percentage of sharing of the top ten universities in the database. **C** Top ten research centers sharing in the present database ▶

higher citation indexes than the fourth author ($p = 0.02$). The first three authors had more clinical trials, comparative studies, NIH grants, USA support, and multicenter studies, than the fourth author. Table 3 shows collaborative authorship of one or more authors and those with more than five authors. Percentages of single authored papers were less than that of multi-authored ones. Where 94 % of the papers were collaborative work of two or more authors and 86 % were papers with three or more contributing authors.

As shown in Fig. 3A, Cairo governorate was first among different governorates with 35.6 % of publications in the present database. Top ten Universities are shown in Fig. 3B, where Cairo and Ain Shams Universities were the most productive ones. The present database showed that Faculty of Medicine, followed by National Research Center and National Cancer Institute came first among the top ten research centers sharing in this database (Fig. 3C).

Collaboration with the USA was pointed top in author affiliations beside Egypt (32 publications), USA Navy Medical Research Unit; (NAMRU-3) in Egypt was affiliated in another 6 citations. Other Arab and non-Arab countries involved as author affiliations in the Egyptian studies for HCV are shown in Fig. 4. Saudia Arabia was the top Arab country in authors' affiliations in the present database.

All literature were in English language.

Publishing journals

Three Egyptian journals were among the top ten journals for publishing in the present database (Fig. 5). Liver international, The American journal of tropical medicine and hygiene, world journal of Gastroenterology were among the top ten journals.

The United States was the country of choice, followed by the United Kingdom and Egypt as places for publishing journals in the present database (Fig. 6).

Publication types

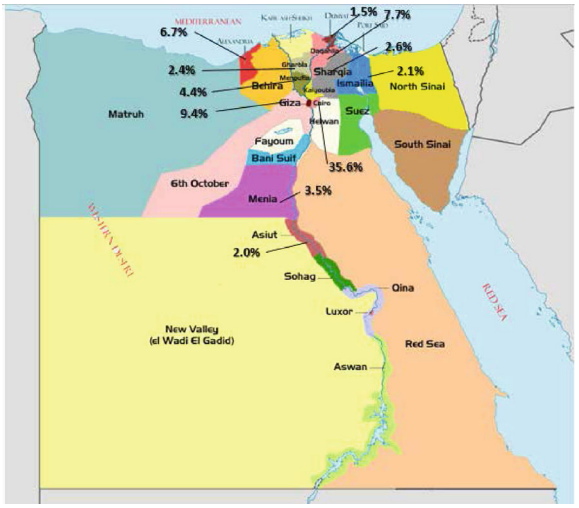
Both major and minor publication types are shown in Fig. 7A, B. Few review articles and meta- analysis were published whereas, citations were mainly journal articles.

Google Scholar citation index

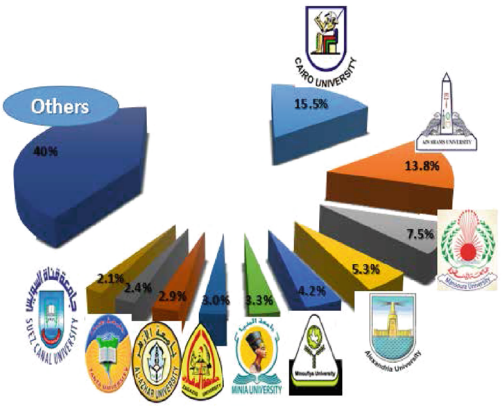
Figure 8 shows different categories of Google citation indexes. Fifty two percent (348/658) publications had impact factor of >10. Chi square test (using SPSS program for windows) did not show any statistically significant difference in citation indexes among the different domains ($p = 4.892$).On a closer look at the pivot table, the 12th domain (sequence analysis) had 8/11 citations with Google citation indexes above 20, one was 196 and the other was 345.

Table 1 shows that most of the top ten cited articles were either from domain 2 (HCV epidemiology), domain 7 (review articles) and HCV sequences (domain 12).

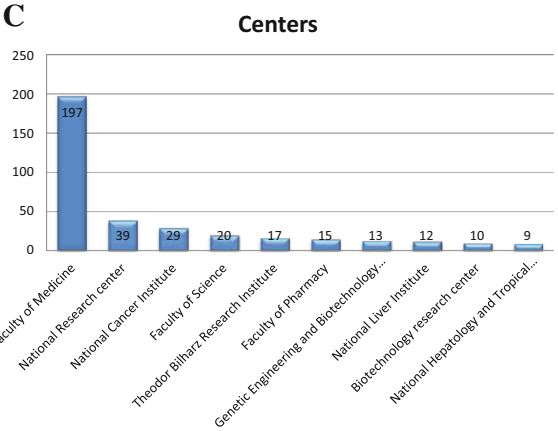
A


MEDITERRANEAN
Kafr el-Sheikh
Damietta
Port Said
1.5%
7.7%
6.7%
Alexandria
Dakahlia
2.6%
Gharbia
2.4%
Sharqia
2.1%
Menoufia
Ismailia
North Sinal
Behira
4.4%
Kalyoubia
Giza
Cairo
Matruh
9.4%
Suez
Helwan
South Sinal
Fayoum
Bani Suif
35.6%
6th October
Menia
3.5%
WESTERN DESERT
Asiut
2.0%
Sohag
Qina
Luxor
RED SEA
New Valley
(el Wadi El Gadid)
Red Sea
Aswan

B


Others
40%
CAIRO UNIVERSITY
15.5%
AIN SHAMS UNIVERSITY
13.8%
7.5%
5.3%
4.2%
3.3%
3.0%
2.9%
2.4%
2.1%
SUEZ CANAL UNIVERSITY
MINIA UNIVERSITY
Alexandria University

C


Centers
250
200
150
100
50
0
197
39
29
20
17
15
13
12
10
9
Faculty of Medicine
National Research center
National Cancer Institute
Faculty of Science
Theodor Bilharz Research Institute
Faculty of Pharmacy
Genetic Engineering and Biotechnology...
National Liver Institute
Biotechnology research center
National Hepatology and Tropical...

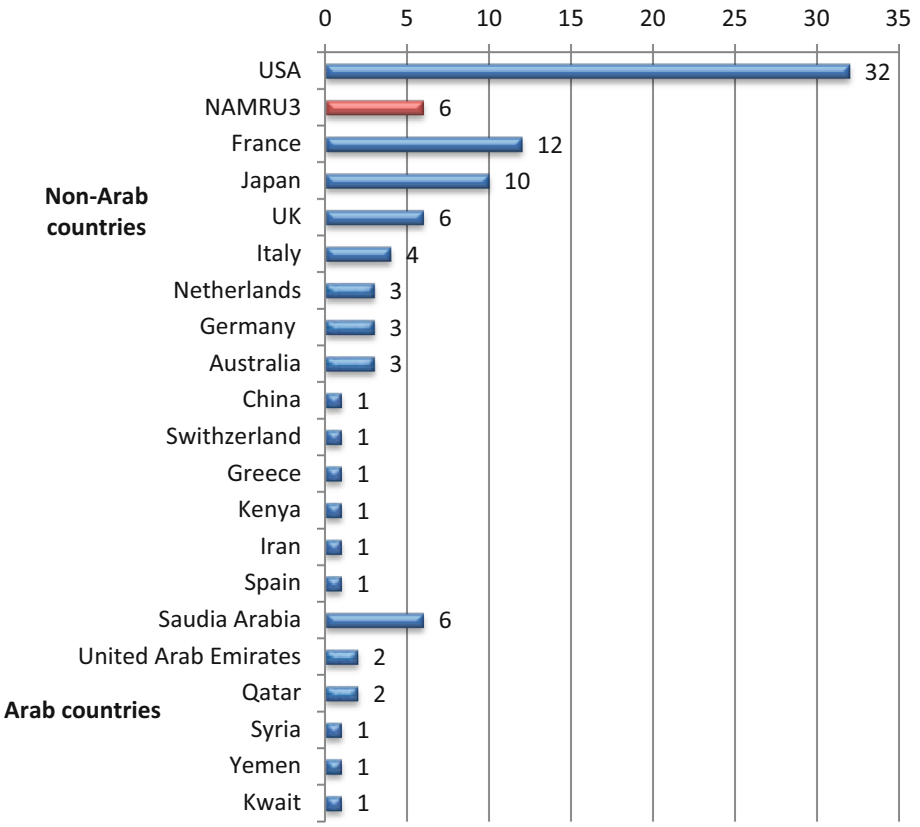


Fig. 4 Countries other than Egypt pointed as author affiliations in the present HCVDBegy

Data mining tools

Forecasting model

Figure 9 shows the number of forecasted publications till 2017 (using data till the end of 2012 as a training set). The actual numbers were retrieved from the PubMed for the years 2013 and 2014 (validation set). Among the 134 citations in the year 2013 and the 152 citations of 2014 from PubMed, 8 and 3 citations, respectively, were excluded according to our exclusion criteria (see previous).

The years 2013 and 2014 showed a breakthrough in the total number of citations than those forecasted, especially in the third domain (response to treatment), the Seventh (review articles), and the thirteenth domain (HCV and other diseases).

Association rules analysis

We tried to analyse citations' patterns regarding MeSH Terms, either as primary topics (most significant points in a record) or as other concepts discussed in the citations. Figure 10 shows the ten most common MESH terms authors used in this study. Hepatitis C

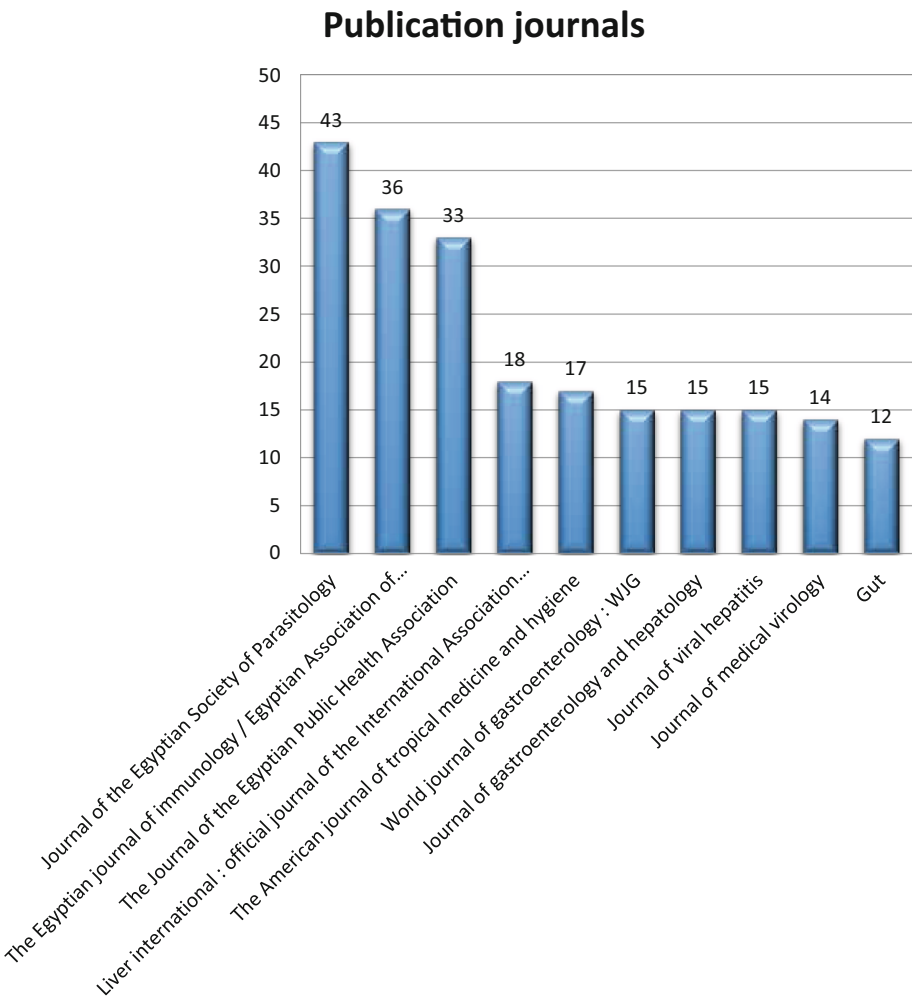


Fig. 5 Top ten journals for publishing citations appearing in the present database

and chronic HCV were among the top ten MeSH terms used. Egypt, Adults, middle aged, males and females, were also most commonly used followed by viral RNA.

Figure 11 shows dependency network based on association rule model of MeSH Terms that appear together in the same citation for detailed analysis of their relations with each other. The root nodes represent the hot topics authors are studying and publishing in this study. The most prominent example is the antiviral agents' node where it was central to many other MeSH terms satisfying the association rule. Also, middle aged was most prominent in association analysis and was associated with at least 30 other MeSH terms.

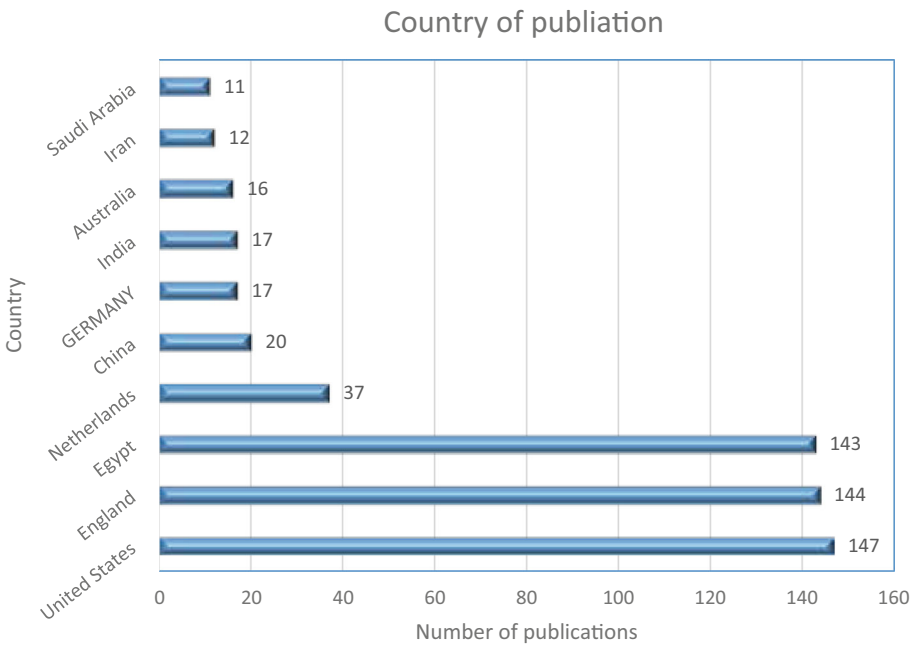


Fig. 6 Top ten countries as places for publishing journals in the present database

Discussion

In agreement with our study design, Yao et al. (2012) discussing bibliometric analysis and knowledge visualization of artemisinin research from SCI and Medline chose subject categories (domains), journals, countries, international collaboration, citations, authorship and co-authorship, author key word (MESH terms) among others. In addition we showed author's affiliations and citation indexes of published papers. Moreover, Gojazadeh et al. (2014), studying knowledge production status of Iranian researchers in areas of Gastric cancer, they studied publication date, author names, journal title, impact factor, and cooperation coefficient between researchers. Data mining in the database is a new in the field of bibliometrics that mainly depended on statistics in their analyses.

Descriptive analysis

Total number of citations

On 1991, when HCV was first discovered, there were 437 citations regarding HCV in PubMed and Egypt's share was 1. This was also the case in the following year (1992) where Egypt's share remained 3 citations out of total 627 PubMed citations. Alvi et al. (2014) studying world HCV literature from 1999 to 2013 showed that Egypt percentage overall share was 1.93 % and ranked the 15th among different countries studied. In accordance with the previous study, the present analysis showed that from the 3207 citations in the PubMed for the year 2012, Egypt's share was 122 (3.8 %). This rise might point to the increased national awareness of HCV problem and the recent national

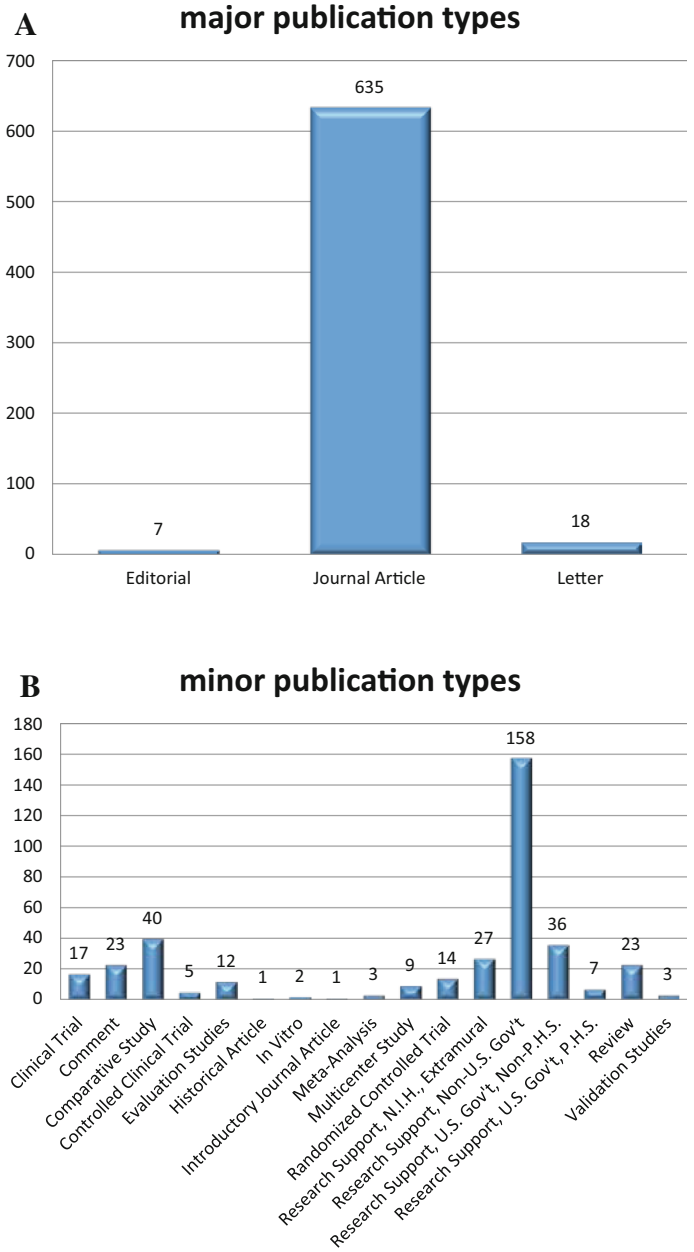


Fig. 7 **A** Major publication types in the present database, **B** minor publication types in the present database

campaign for its treatment as well as the improved calibre and number of researches and researchers in this area as well as the increased international collaborations.

A fluctuation in number of HCV research from Gastroenterology (a journal) was noticed. There was a decrease in 2008, and then an increase in 2009 followed by a decrease in 2010 (Thavamani 2013). In agreement with those findings, the present study showed a

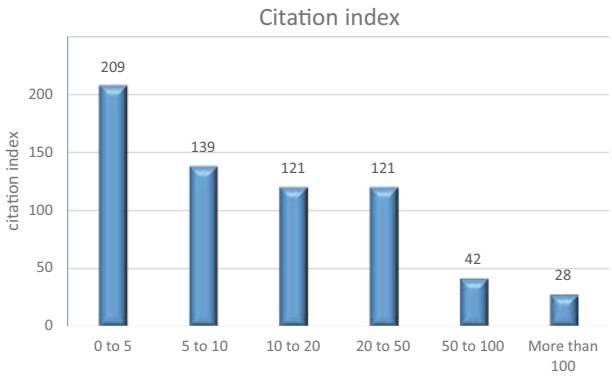


Fig. 8 Google citation indexes categories of different citations

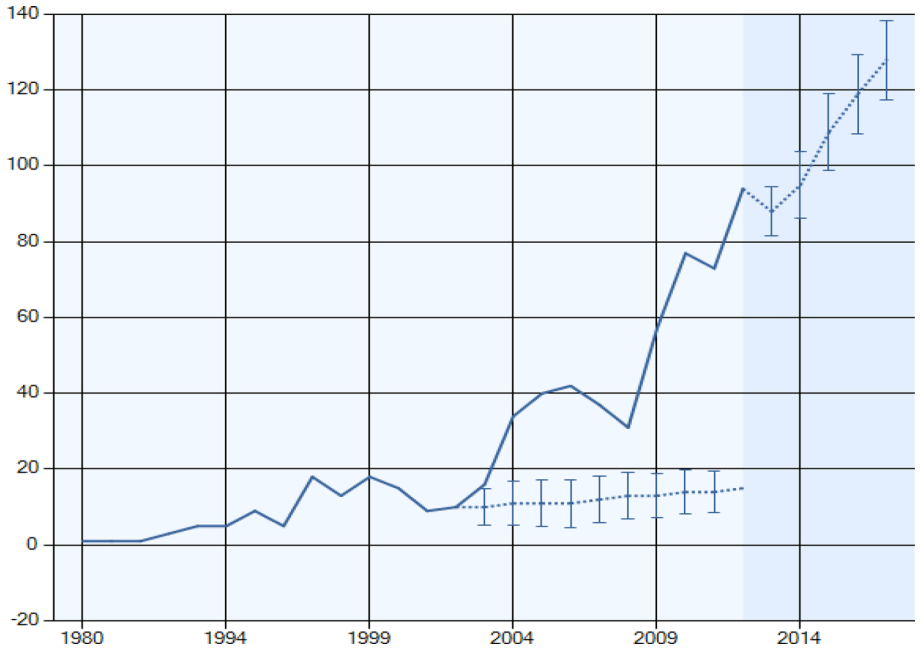


Fig. 9 The numbers of forecasted citations according to the forecasting citation model

decrease in number of publications in the year 2008, followed by a constant increase to present, with some fluctuations from the years 2004 to the year 2009. Moreover, studying world literature on HCV research from 1999 to 2013 revealed that HCV showed upward trend and that the highest quantity was produced during the block 2011–2013, a number similar to our forecasted and verification sets (Alvi et al. 2014).

Authors and their affiliations

Cooperation among authors was very clear since only 6 % of articles were done by single authors, whereas 86 % were done by 3 or more authors. A study of hepatitis literature from

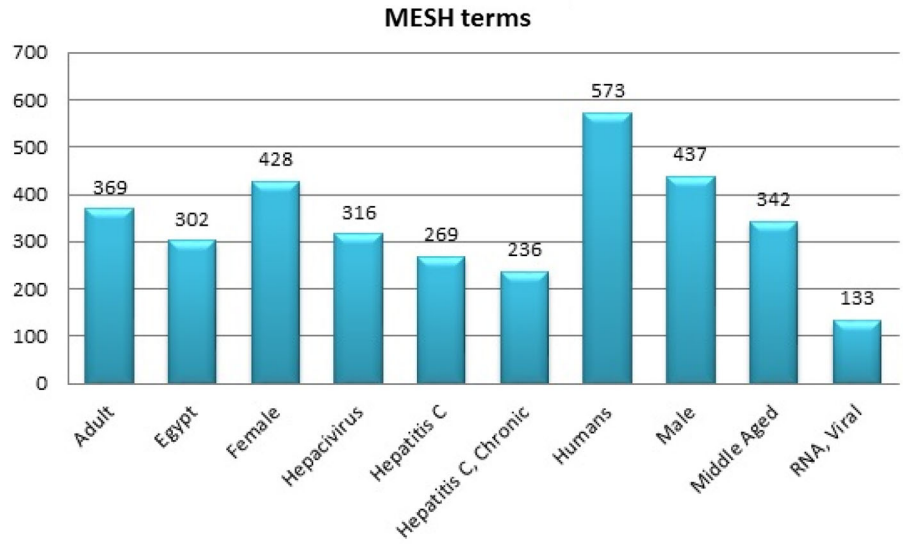


Fig. 10 Most commonly used MESH terms in the present study

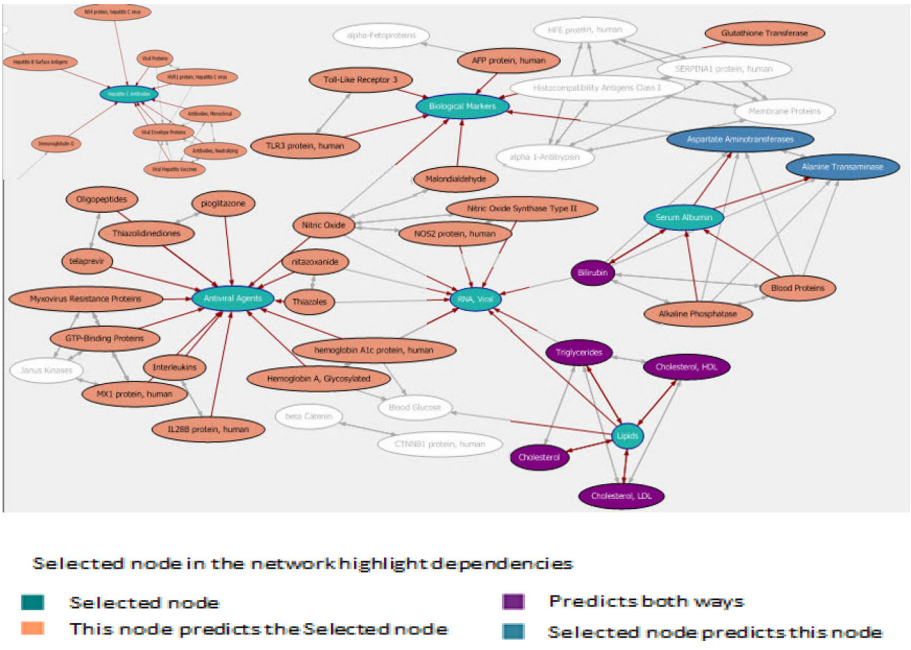


Fig. 11 Dependency network based on association rule model of MeSH primary topics that appear together in the same citation

MEDLINE and two other databases from 1984 to 2003 also showed that the percentages of single authored papers were less than that of multiple-authored papers. Where 85 % of their total contributions were collaborative research (Ramakrishnan and Babu 2007). This shows that research in the field of hepatitis tends to be more collaborative.

Regarding first author names, analysis of the present database proved that it was a bad indicator of the overall authors' contributions and efficiencies in general and would have given us misleading results. Analysis thus must include all other authors. PubMed's redundant, subjective, missing, and unclean authors affiliations' data posed a challenge during analysis. A manual curation of the authors' affiliations and division into separate fields have been tried in this study to enable proper analysis. A system for unifying spelling and editing of governorates, Universities, and centers have been manually tried (since the number of citations was in the order of 658). This problem arises as publishing journals left author's affiliations to be edited by the authors themselves and was integrated in PubMed data base as one entry. A step that is lastly and hastily been taken care of after a tiresome research and a long editing process. Thus, we recommend that the PubMed should put a note for all the journals to alarm authors of such a problem and recommend a unifying system for all affiliations, especially in Egypt. We also recommend that every researcher should stick to one name spelling throughout his/her career.

Missing data for research centers were around 13 % of the total citations. This deprives them of their rights as research entities of this kind of work.

Cairo governorate was the most productive among the different governorates in Egypt, with 35.6 % of the publications in the present database. Cairo metropolitan area has four high caliber Governmental Universities with major contributions in the present database; namely, Cairo, Ain Shams, Al-Azhar, and Helwan Universities. Other Non-Governmental Universities (such as The German University in Cairo) also contributed to this situation.

Appearance of Faculty of Medicine as the first center who shared in the present database was misleading, since this research entity is shared in all universities. National Research Center and the National Cancer Institute, are both two high calibre research entities. The latter belongs to Cairo University with more than one specialized team in the field of HCV (as top ten authors analysis also revealed). Three of the top ten first authors were from National Cancer Institute, Cairo University.

Yao et al. (2012) studying cooperation network maps showed that some Universities were in the core of the map whereas others were at the edge of the network due to their less cooperation. The reason behind these findings might be attributed to a concentration of budget and facilities in the capital which provide better conditions for knowledge production in these centers and vice versa (as stated by Gojazadeh et al. 2014). Thus, our decision makers should be considering what they indicated regarding empowering universities with small share to become in the lead for this kind of research. Where Ghojazadeh et al. (2014) stated that it seems necessary to fairly distribute the resources and provide more proper access of researches from universities of small cities to research facilities in order to promote ever-increasing process of scientific productions and provide required conditions to equal attendance of all universities and researchers of the country (Ghojazadeh et al. 2014).

USA followed by France and Japan were the top three foreign countries appearing as authors' affiliations in the present database. This shows the trend of cooperation between Egyptian researchers and foreign countries in this special field of research. Saudi Arabia also came first as an Arab non-Egyptian country in author affiliations.

One of the limitations in this study was that PubMed made available only the first authors' affiliations at the time this work was conducted, thus depriving other authors' centers from taking a chance to be counted in this analysis.

All published articles were in English language. In agreement with our findings, Alvi et al. (2014) studying the world literature on HCV research stated that any author despite his country of origin or language, publishes his research mainly in English rather than their

own mother language. English was the language of publication in 86.6 % of the world literature on HCV (Alvi et al. 2014).

Publishing journals

Despite the low impact factor of such journals, three of the top ten publishing journals were from Egypt. An easy to reach (especially before the internet period and the electronic on line submission), easy processing, fast editing process, low fees with the local currencies, and backdates of such journals might be the reasons for this situation. Encouraging authors to go through the peer review process of the high caliber journals is thus recommended. The presence of many Egyptian journals with the same mission might in fact dilute our efforts in maintaining better standard for them.

The impact factors of the other non-Egyptian top ten journals were between 2.347 (Journal of Medical Virology) and 14.66 (Gut). Publishing journals were mainly from USA and United Kingdom, beside Egypt. None of them however, was among the top 100 journals in Biology and Medicine (http://dbiosla.org/publications/resources/dbio100.html).

Alvi et al. (2014) studying world HCV literature showed that the highest number of articles in HCV was in the Journal of Hepatology, followed by Journal of Viral Hepatitis, Journal of Medical Virology and World Journal of Gastroenterology. This agrees more or less with our present study where Journal of Viral Hepatitis, Journal of Medical Virology and World Journal of Gastroenterology were among the top ten journals for publishing in the present database chosen by Egyptian authors.

Publication types

Regarding types of publications, there were few review articles and randomized controlled trials. Four percent were review articles and 13 % were in PubMed Central (the free database of PubMed). Fifty two percent (285/543) publications of the major domains had impact factor of >10. A figure that is comparable to that in the minor domains (54 %, 63/117).

In the present study, few controlled clinical trials and meta- analyses were published whereas, citations were mainly journal articles (original research articles). In agreement with our findings, Alvi et al. (2014) found that world HCV studies were mainly articles then reviews followed by letters.

Google Scholar citation index

Alvi et al. (2014) found out that the most important subject (of total 134) was Medicine. Next comes immunology and microbiology followed by biochemistry, genetics and molecular biology, pharmacology, toxicology and pharmaceutics, agricultural and biological sciences, chemistry, nursing, social sciences and others (Alvi et al. 2014). Our study was organized differently where publications were divided from the start into 22 domains or area of research according to the judgement and experience of the first author in the present work.

Based on Google Scholar, we were able to retrieve the citation indexes of all authors. Also we computed all citations and found no statistically significant difference among the different domains. As for the 4 most influential authors in the present database, there was a statistically significant difference regarding citation indexes ($p = 0.02$) between top two

authors and the fourth one. The higher citation indexes in their cases were associated with more comparative studies, clinical trials, multicenter studies, NIH grants, and USA grants than that of the fourth author.

We agree with what Tousoulis and Stefanadis (2014) have stated regarding the high citation indexes and its association with total recognition/influence, relative impact and efficiency. Also in the present work, it was associated with more foreign cooperation and grants and more efforts regarding summarizing data and writing review articles.

Genotyping domain, on the other hand, had high citation indexes, this might be attributed to the fact that sequences' GenBank accession numbers associated with these studies must be cited every time they are being used by other authors. This is also because our sequences are of a special genotype (Genotype 4) which is present mainly in Egypt and thus it is considered unique among other countries. It is worth mentioning that all accession numbers associated with these studies are present in a separate table in this database for further analysis.

In agreement with another article, our study showed high citation indexes for studies of HCV epidemiology, review articles and sequence analysis domains (Tsafrir and Reis 1990). We cannot however recommend for authors to stick to those three domains only if they were to be recognized by others, because this will minimize the importance of the original researches even if they have less citation indexes. A draw back in this analysis was that we did no time correction for this kind of analysis.

Data mining tools

Forecasting model

Forecasting model showed a lag in the model (training set) than the number of citations for the years 2013 and 2014 (validation set) retrieved from the PubMed. Thus the model was not accurate enough to predict the actual situation. The fact that there is a national campaign for treating HCV and the government supportive policy might be the cause of the increase in citations in all fields and especially in the field of the third domain (response to treatment). More and more authors are trying to do research in this important and up to date field.

Other least studied domains might continue and prove strong in the future as shown for the thirteenth domain (HCV and other diseases) other than HCC and chronic HCV. The third domain (response to treatment) showed a breakthrough in number of publications in the years 2013–2014. A phenomena to be verified in the years to follow.

Few articles however became oriented towards genotyping and sequencing of the virus in the past 2 years (2013 and 2014). This does not coincide with the ease of sequencing that Egypt encounter nowadays.

Association rules analysis

Out of the three studies concerning bibliometrics and scientometric analysis of hepatitis and HCV (Alvi et al. 2014; Thavamani 2013; Ramakrishnan and Babu 2007) only Alvi et al. (2014) was concerned with subjects and key words that authors studied. The latter study showed that out of 134 subjects, the most important subjects were Medicine then Immunology and Microbiology followed by Biochemistry, Genetics and Molecular Biology, Pharmacology, Toxicology and Pharmaceutics, Agricultural and Biological Sciences, Chemistry, Nursing, Social Sciences and others. Our present study showed that Virology,

pathology, immunology and genetics were most commonly used secondary MeSH terms. Yao et al. (2012) showed that the artemisinin researches on the subjects of pharmacology and pharmacy and chemistry occupy a dominant position from the outset (other top subjects such as biochemistry and molecular biology, plant sciences, microbiology and infectious diseases are also involved). They stated that since artemisinin was proven to treat malaria well in 1971, scientists tend to focus on the research of chemical composition and pharmacology for drug development, which can benefit malarial patients all over the world. In agreement with the findings of Yao et al. (2012), our study revealed that pathology and genetics were extensively studied since chronic hepatitis must be confirmed by pathology, and genetics was used for studies of the genes involved in the malignant transformation by HCV.

In the present study, market basket and association data mining analysis was tried on MeSH terms. This was in attempt to make analysis of such terms useful for upcoming authors as these research fields will be the next potential technology applications and are worth more attention. Yao et al. (2012) also studied key words clustering and showed that some of them were hot topics leaving the chance for other authors in the future to work on.

Schistosoma was also mentioned in the previous analysis, because it was one of the alleged factors for transmitting HCV to our population due to the anti-Bilharzia parenteral treatment program (Rao et al. 2002). Also, in Egypt, viral hepatitis along with infection with Schistosoma mansoni were among the major causes of chronic liver disease and liver cirrhosis (Halim et al. 1999). Also the fifth domain in this study was Schistosoma and HCV accounting for 40 citations (6.17 %).

Conclusions

To the best of our knowledge, this is the first paper to analyze the quantity and quality of research productivity in the field of HCV in Egypt. It directs attention and initiates discussions among those interested in this field and could be considered as a decision support system for the decision makers in this country. Number of publications showed a promising increase than the original forecasting model. This rise might points to the increased national awareness of HCV problem and the recent national campaign for its treatment among others. Research in the HCV field tends to be collaborative. Analysis should include all authors, since analysis of the first author only could miss other highly productive ones. Empowering small-town universities to become in the lead for this kind of research was found to be mandatory and should be planned for by the governmental bodies concerned. Encouraging authors to go through the peer review process of the high caliber journals is highly recommended. High citation indexes was associated with more foreign cooperation and grants and more efforts regarding summarizing data and writing review articles.

The third domain (response to treatment) showed a breakthrough in number of publications, which might be due to the national campaign for treatment of HCV and the national awareness to combat it. Studying MeSH terms clustering showed some hot topics leaving the chance for other authors in the future to work on.

Limitations

PubMed made available only the first authors' affiliations at the time this search was conducted, thus depriving other authors from taking a chance for their affiliations to be counted in this analysis. No corrections for time was done during analysis of citation indexes in this study.

Recommendations

We recommend that the PubMed should put a note for all the journals to alarm authors of affiliations problem and recommend a unifying system for them, especially in Egypt. We also recommend that all researchers should stick to one name spelling for future and career recognition.

Author contributions Alam El-Din HM, did the work, and edited the manuscript. Ahmed AS, shared in the idea, and revised the manuscript. Hanora A, put the idea and revised the manuscript.

References

Alter, M. J., Margolis, H. S., Krawczynski, K., Judson, F. N., Mares, A., Alexander, W. J., et al. (1992). The natural history of community-acquired Hepatitis C in the United States. The Sentinel Counties Chronic non-A, non-B Hepatitis Study Team. *New England Journal of Medicine, 327*, 1899–1905.

Alvi, K. S., Vinitha, K., & Ravanan, C. (2014). World literature on Hepatitis C virus research: A scientometric study. *Journal of Advances in Library and Information Science, 3*(4), 361–368.

Centers for Disease Control and Prevention. (2012). (CDC). Progress toward prevention and control of Hepatitis C virus infection–Egypt, 2001–2012. *MMWR. Morbidity and Mortality Weekly Report, 61*(29), 545–549.

El-Zanaty, F., & Way, A. (2009). *Egypt demographic and health survey 2008*. Cairo, Egypt: Ministry of Health, El-Zanaty and Associates, and Macro International. http://www.measuredhs.com/pubs/pdf/fr220/fr220.pdf. Accessed July 18, 2012.

Ghojazadeh, M., Naghavi-Behzad, M., Nasrolah-Zadeh, R., Bayat-Khajeh, P., Piri, R., Mirnia, K., & Azami-Aghdash, S. (2014). Knowledge production status of Iranian researchers in the gastric cancer area: Based on the medline database. *Asian Pacific Journal of Cancer Prevention, 15*(12), 5083–5088.

Halim, A. B., Garry, R. F., Dash, S., & Gerber, M. A. (1999). Effect of schistosomiasis and hepatitis on liver disease. *American Journal of Tropical Medicine and Hygeine, 60*(6), 915–920.

Hoofnagle, J. H. (1997). Hepatitis C: The clinical spectrum of disease. *Hepatology, 26*(Suppl), 15S–20S.

Mostafa, A., Taylor, S., El-Daly, M., el-Hoseiny, M., Bakr, I., Arafa, N., et al. (2010). Is the Hepatitis C virus epidemic over in Egypt? Incidence and risk factors of new Hepatitis C virus infections. *Liver International, 31*, 560–566.

Nalimov, V. V., & Mulchenko, B. M. (1969). *Scientometrics*. Moscow: Nauca.

Nichols, A. W. (2008). Sports medicine clinical trial research publications in academic medical journals between 1996 and 2005: An audit of the PubMed MEDLINE database. *British Journal of Sports Medicine, 42*(11), 909–921.

Oliver, D. E., Bhalotia, G., Schwartz, A. S., Altman, R. B., & Hearst, M. A. (2004). Tools for loading MEDLINE into a local relational database. *BMC Bioinformatics, 5*, 146.

Pritchard, A. (1969). Statistical bibliography or bibliometrics? *Journal of Documentation, 24*, 348–349.

Ramakrishnan, J., & Babu, B. (2007). Ramesh. Literature on hepatitis (1984–2003): A bibliometric analysis. *Annals of Library and Information Studies, 54*(4), 195–200.

Rao, M. R., Naficy, A. B., Darwish, M. A., Darwish, N. M., Schisterman, E., Clemens, J. D., et al. (2002). Further evidence for association of Hepatitis C infection with parenteral schistosomiasis treatment in Egypt. *BMC Infectious Diseases, 2*, 29.

Thavamani, K. (2013). *Growth of literature in the field of Hepatitis-C*. Library Philosophy and Practice (e-journal). Paper 944. http://digitalcommons.unl.edu/libphilprac/944.

Tousoulis, D., & Stefanadis, C. (2014). How can we assess scientific quality? Citation index only for original research and/or for authorship in the guidelines? *Hellenic Journal of Cardiology, 55*(5), 353–354.

Tsafrir, J. S., & Reis, T. (1990). Using the citation index to assess performance. *BMJ, 301*(6764), 1333–1334.

Wilson, C. S. (1999). Informetrics. In M. Williams (Ed.), *Annual review of information science and technology*. NJ: Information Today, Medford.

Wolfram, D. (2003). *Applied informetrics for information retrieval research*. Westport, CT: Libraries Unlimited.

Yao, Q., Chen, J., Lyu, P. H., Zhang, S. J., Ma, F. C., & Fang, J. G. (2012). Knowledge map of artemisinin research in SCI and Medline database. *Journal of Vector Borne Diseases, 49*(4), 205–216.