Subject analysis of cancer research: Procedure and identification of core literature

Chung-Yen Yu and Jiann-Cherng Shieh

Graduate Institute of Library and Information Studies,

National Taiwan Normal University, Taipei 10610, TAIWAN R.O.C

e-mail: jcshieh@ntnu.edu.tw (corresponding author)

ABSTRACT

Subject analysis is an approach for analyzing the development of academic research and paradigm shifts in Library and Information Science. Through content analysis and categorization of academic literatures, the focus in different development stages of research fields, such as research areas, issues of concern, and temporal and spatial shifts, is summarized and analyzed. This study proposes an analytical framework for the topics of electronic resources by adopting the Medical Subject Headings (MeSH) posted by the U.S. National Library of Medicine (NLM) as the basis for subject classification. The study covers cancer research as a subject case to verify and compare the databases of SCImago Journal and Country Rank and the bibliographies of PubMed biomedical literatures. Through systematic collection, collation, and analysis of raw data, variables such as specific key topics, number of publications, and year of publications are segmented for statistical analysis. The findings show that cancer research is a developing trend, and the compiled information provides a diversity of study topics for informatics and information behavior research.

Keywords: Subject analysis; Medical subject headings; MeSH, Electronic journals; Cancer research.

INTRODUCTION

In the field of library and information science, subject analysis serves as a way of analyzing academic research development and paradigm shifts. Such analyses chiefly include published works from the fields of academia and analyze, classify, and generalize their

contents to better understand the focus of development in various stages through the examination of the scopes of research in the fields of importance, issues of concern, and changes across time and space.

From the perspective of publication content, subject analysis can be defined as the process of identifying the intellectual content contained within a work in order to perform further analysis on the basis of the text's notable features, as well as the labeling of the work's content and the subject it describes using numbers, symbols, nouns, the combination of nouns and adjectives, or phrases as the basis for information searches (Chu and O'Brien 1993). Besides its traditional role in subject classification, subject analysis also plays an important role in document searching. In addition, analysis is performed on document content based on document storage and retrieval system demands according to fixed methods. Through this analysis, specific keyword concepts and topic categories are obtained, and structural analysis is conducted on keyword concepts, producing "subject terms" and "keywords." Typically, the results of subject analysis are presented in two formats: (a) numbers and symbols; and (b) vocabulary. The former is commonly seen in subject classification systems, whereas the latter is seen in subject/category labels and used in keyword searches (Lo, Chen and Lin 2001).

In contrast with subject heading lists that have been created by subject matter experts in the field of medicine and updated over time, because of the lack of commonly created subject heading lists updated over time in the past, subject analysis in library and information sciences was primarily conducted by scholars in one of two ways: (a) the creation of a complete subject list for the field in question; or (b) the creation of a relevant subject heading list after the literatures are already collated and the subjects have appeared in the literatures. The subject classification systems created or used in these two methods with a specific objective in mind are, for practical purposes, useful for studies with the appropriate time and scope in mind. In addition, over time, and as the environment changes, when analyzing development trends in research, there is a need to create a new classification system above current foundations as an incomplete subject analysis (in terms of concepts covered); an inaccurate reflection of existing literatures topics or other factors would have an impact on the accuracy of consequent analytic results and flexibility in conducting further research with these results (Tsay and Hsu 2009).

As such, this study does not attempt to create a subject heading list from scratch for the purposes of the subject analysis of electronic resources. Instead, a subject heading list created by subject area experts and recognized by third parties is used for the purposes of this study. This study demonstrates a procedure for the subject analysis of electronic resources and uses the field of cancer research to verify this proposal. As such, this paper applies Medical Subject Headings (MeSH) released by the U.S. National Library of Medicine as a basis for subject analysis. Publications listed in the PubMed database are used as the source of publication information, with the Cancer Research list in the SCImago Journal & Country Rank (SJR) journals database used as a case study. Simultaneously, the following were created by the author: (a) publication citations; (b) a MeSH Tree; and (c) a Document MeSH Tree Path database to aid the collection, organization, and analysis of literatures.

The limits of this study are as follows:

(a) the source of publication information for this study is the PubMed database, the content of which (publication list and keywords) is updated at irregular intervals. Hence, the researchers are required to conduct manual searches or deploy a crawler program to ensure that information obtained is up-to-date;

(b) journal information from the SJR and PubMed databases have to be compared manually by the author in order to maintain the consistency and completeness of journal information.

LITERATURE REVIEW

Some relevant works that support subject analysis on PubMed with MeSH are discussed in the following subsections.

a) Medical Subject Headings (MeSH)

Medical Subject Headings (MeSH) is a control vocabulary released in 1960 by the National Library of Medicine (NLM) (Medicine 2013). The list, which has its origins in the Subject Heading Authority List, is currently one of the most widely-used search tools for medical information (Medicine 2014b). MeSH aims to provide a subject catalog of a variety of biomedical literatures, books, etc., to facilitate the quick and precise location of literatures required by the user. It has also proven popular with medical libraries and search catalog

(Nelson, Johnston, and Humphreys 2001). As of 2014, MeSH contains 16 subject categories with a total of 26,582 headings listed (Medicine 2014a).

MeSH utilizes a hierarchical structure to show mutual relationships. Each heading that appears within the structure represents an independent concept. For example, the heading number for the term "Colorectal Neoplasms" is "C04.588.274.476.411.307"; if later we are to search for this term, it would fall under the "Diseases" category as a type of "Neoplasm." The heading will correspond with at least one heading number. The lower the level, the more generic the concept is. Conversely, the higher the level, the more specific the concept is (Figure 1).

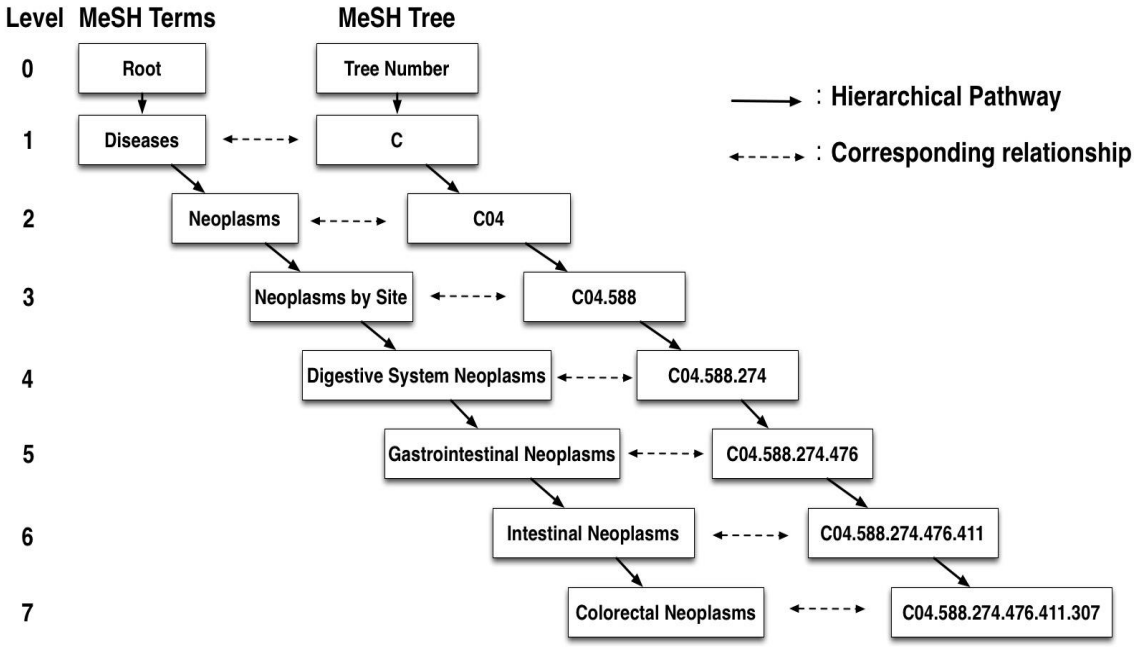

Figure 1: MeSH heading and qualifier classification numbers (Source: Created by the author based on the medical subject heading list from the National Library of Medicine

The composition rules for heading classification numbers consist of the following: the classification number is made up of letters and numbers, with the period (.) used as a separation mark. Each level is described by a combination of alphanumeric symbols in which a letter and numbers are used as the first symbol for the other symbols. The letter used for the first symbol corresponds with the 16 topic categories. Using Figure 1 as an example, the first symbol in the MeSH Tree Number is the letter "C," which corresponds to the topic category "Diseases." The Level 2 structure for the classification number is made

up of another three symbols, again with an alphabet as the first symbol and numerals as the second and third symbols, i.e., "C04," refers to the term "Neoplasms" in the heading. The subject in this example, "Colorectal Neoplasms," is situated on Level 7 of the classification. Therefore, as long as the user knows which particular heading to use in his search, he will be able to recreate the MeSH Tree path backward through an examination of the hierarchical relationship inherent in the heading classification number.

b) Related work that support subject analysis on PubMed with MeSH

Charidimou and Song (2015) performed an analysis to describe the cerebral amyloid angiopathy (CAA) research landscape from 1950 to 2015 and retrieved 2,153 CAA publications using a combination of search terms "amyloid amgiopathy" OR "congophilic angiopathy" OR "dysphoric angiopathy" OR "dysphoric angiopathy" on Pubmed. The authors discovered the following: 1) major topics relative to clinical practice included antithrombotic drugs and MRI; 2) capillary CAA and tangles, cognitive impairment/dementia, intracerebral haemorrhage, and subarachnoid haemorrhage were all covered for neuropathology; 3) other topics of major interest, including Alzheimer's disease, hypertension, and CAA-related ischaemic features (Charidimou and Song 2015).

Jia and colleagues used the MeSH term "urban health" to retrieve 5,579 MeSH terms and 11,299 publications between 1998 and 2007. MeSH term, co-word, and linear regression analyses were performed to discuss the thematic tendency of selected countries in urban health-related topics. Results revealed that air pollution, seasons, and smoke pollution were the environmental factors of greater concern in selected countries, including United States, India, China, South Africa, and Japan (Jia, Dai and Guo 2014).

Li, Pan and Ye (2013) conducted a search on Pubmed using the MeSH term "systemic lupus erythematosus" and identified 14,053 articles from the year 2002 to 2011. The authors figured out the following: a) these articles were spread across 1,627 different scientific journals, of which Lupus, Arthritis Rheum, and J Rheumatol were the most productive; b) English language publications were overwhelmingly predominant, whereas non-English articles accounted for only 11.492% of the total number of publications; c) major research areas were etiology and physiology, whereas transmission and veterinary were fields less studied.

For bibliometric analysis of publications on leishmaniasis, Ramos and colleagues consulted the Pubmed database using the MeSH terms “Leishmania” or “leishmaniasis” for the reference period of 1945 to 2010. The search yielded a total of 20,780 articles that were then grouped into five-year periods. Subsequent analysis revealed that for articles published on leishmaniasis, “animals” was the prevalent MeSH term with 8,564 (41.2%) publications, followed by “visceral leishmaniasis” with 6,216 (29.9%) entries. On the other hand, journal article was the preferred format of document with 17,982 publications accounting for 86.5 percent of the total number. Bibliometric analysis by geographical area revealed that Europe was the most productive region, delivering 31.7 percent of the articles, followed by Latin America and the Caribbean with 24.5 percent (Ramos, González-Alcaide, and Bolaños-Pizarro 2013).

These studies mentioned earlier reveal that past groups generally use a specific disease name as a MeSH term to retrieve a data set, which is then statistically evaluated in terms of MeSH term, subheading, publication year, publication region, publication language, growth of literature, and type of journal. On the other hand, our study is set apart from the aforementioned method studies in the following ways:

(a) instead of using disease name as the search term, record of inclusion in the SJR database was used as the key criterion in our study;
(b) our result was cross-referenced to PubMed prior to analysis;
(c) our study proposes a novel subject analysis method using cancer research as a case study to illustrate the process.

RESEARCH DESIGN

The convenience of using electronic resources (e-resources) lies in the fact that literatures may be accessed through the Internet. According to Yu and Hsieh (2014), the authors propose a method for the evaluation of e-resource usage using medical literatures found through e-resources as an example. In their study, the duo found that besides citation information found in e-resources databases or the web sites of the respective publishers, subject bibliographic databases can also be used to obtain bibliographic information on literatures found in journals. This information can then be used in the analysis of research subjects and related trends in the field alongside subject heading lists. Hence, for this

study, a web crawler program was designed to obtain bibliographic information through PubMed's Application Programming Interface (API). Then, the information obtained is used to create a repository of (a) Citation Information Documents, (b) MeSH Trees, and (c) Document MeSH Tree Paths to enable access to bibliographic information and statistics available in the shortest time possible with the help of automation. The information sources for this study, as well as study processes such as information analysis and system building, are described below.

a) Data sources

This study utilizes two information sources, the SCImago Journal & Country Rank (SJR) and the PubMed database (SCImago Lab 2012). From the SJR database, the subject area "Cancer Search" was selected to obtain the list of publications recorded in this area. The date of the search was May 3, 2013.

b) The Process

The information used in this study was obtained and processed in the following manner: (a) a list of journals in the area of cancer research was obtained from the SJR; (b) then, PubMed's official medical subject heading list was obtained, and the MeSH Tree repository was created; (c) thereafter, publication information on journals in the area of cancer research was obtained from PubMed. For more information on the data obtained from the steps described earlier, see Figure 2 and the description that follows.

i) A list of journals in the area of cancer research was obtained from the SJR:
   First, after connecting to the SJR database, journals were selected by journal ranking with subject category (Cancer Research) specified in ranking parameters. List information can be further sorted and selected by year of publication, with the SJR list containing journals published between 1999 and 2013. The data export function allows for the export of the list to Microsoft Excel format. This study takes the journal list from the SJR for cancer research as a case in performing verification.

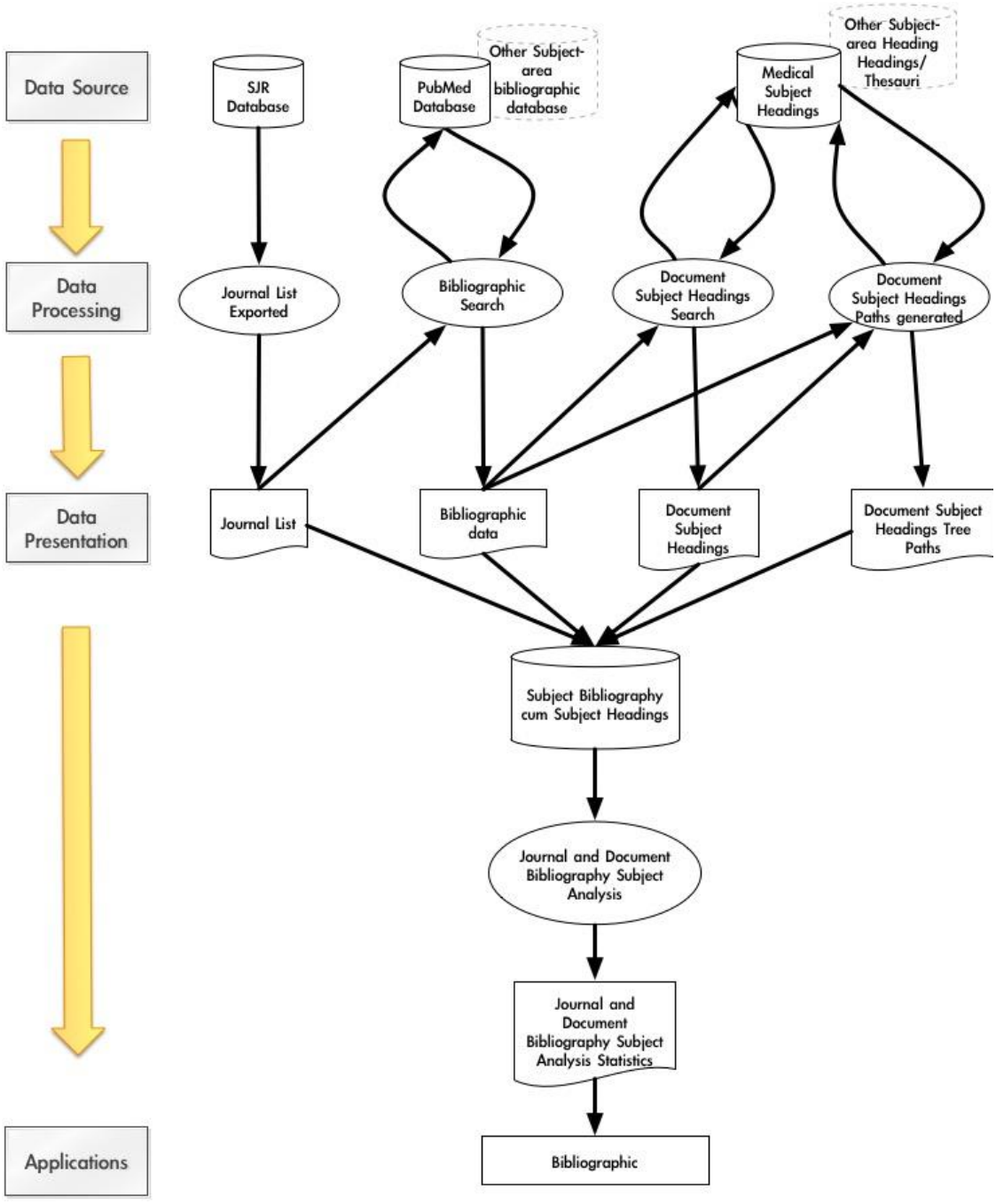

Figure 2: Subject structural diagram for journals in the area of cancer research (Source: As organized by the author)

ii) A MeSH Tree repository is created using medical subject headings released by PubMed, which releases an updated list on an annual basis:

For this study, a web crawler was deployed to obtain the current subject heading list for the year. From the information obtained, the following was used in creating

the yearly MeSH Tree repository: (a) heading, (b) tree number, (c) parent tree number, and (d) the specific level where the heading was found. This primarily allows for the creation of a heading path record with the MeSH Tree Database when analyzing the heading for each medical journal, which is beneficial to the efficacy of subject statistical analysis processing.

iii) Obtaining a publication list from the PubMed database:

The publication list obtained from the SJR must be verified for inclusion in PubMed using either the publication name or the ISSN as the search term. If the publication record exists in PubMed, then a list of all articles under the publication is obtained. To increase the efficiency of this process, the list obtained is exported in PMID List format.

iv) Publication summary information downloaded from PubMed:

Titles in the journal list obtained above are matched with the original PubMed URL and bibliographic content using the API and URL format provided by PubMed. PubMed provides a variety of bibliographic formats, including the MEDLINE format. For this study, the batch download method was used to download PubMed IDs for the journals listed in MEDLINE format using the PubMed API. Then, the publication information obtained was analyzed to obtain the required field information. As PubMed continually updates its information when obtaining publication information, care must be taken to perform batch downloads on a regular schedule to maintain data accuracy.

v) Obtaining bibliographic information and headings for each article:

Using the summary information obtained above, a bibliographic data analysis program was used to extract field information from the data, with key fields being the following: (a) PubMed ID; (b) article title; (c) article digital object identifier (DOI); (d) PubMed Main Headings; (e) major headings; (f) publication (journal); (g) year of publication; (h) journal ranking; (i) journal ISSN; and (j) author(s). Among these, major headings were obtained from PubMed Main Headings with further analysis and statistical processing to obtain heading count. Information for the field of journal ranking was obtained from the journal list earlier obtained from the SJR.

vi) The analysis of Medical Subject Heading format and the generation of MeSH Paths:

The major headings described above were processed using a lexical analysis

program written by the author to analyze the headings used within the MEDLINE format and to search for major topic headings in the MeSH Tree one-by-one to produce MeSH paths.

vii) The creation of a bibliographic and keyword database:

The aforementioned journal lists, bibliographic information, Headings, Heading Paths, etc., were imported into a bibliographic and keyword database on the academic subject built by the author.

Based on the aforementioned information analysis process and with cancer research journals recorded in the SJR used as a case study, a total of 402,495 articles using a total storage space of 2.1GB were found in PubMed and processed accordingly for the requisite field information.

RESULTS

On the basis of the data obtained from SJR with “cancer research” selected as the subject of focus, a total of 196 journals had been recorded according to journal information released in 2012. The publication information obtained for this study is used for subject headings that are then used in PubMed searches. Following the search conducted on May 3, 2013, a journal and article database was created by the author and analyzed for the purposes of this paper.

The ranking, count, and ratio of medical literatures on cancer research recorded in both the SJR and PubMed

On the basis of the information released in the SJR database for 2012, there are a total of 196 journals within the field of cancer research. However, if we use journal titles and ISSNs as search terms in PubMed, only 102 journals [of the 196] can be found. Simultaneously, a total of 402,795 articles—with headings provided by PubMed—were found in PubMed, with an average of 3,946 articles per journal. This shows that in the field of cancer research, over half of the journals recorded in SJR (102/196 = 52%) were also recorded in PubMed. Table 1 illustrates this in more detail.

Table 1: The ranking, count, and ratio of medical literatures on cancer research recorded in both SJR and PubMed

Journal rank	PQn	PQn/P	JQn	JQn/J
Q1	251,357	62.45%	39	38.24%
Q2	62,592	15.55%	30	29.41%
Q3	70,317	17.47%	29	28.43%
Q4	18,229	4.53%	4	3.92%
(P = 402,495, J = 102, P/J = 3,946)				

When we further analyze the 102 journals in the set, we find that there are 39 journals, or 38 percent of the total, that fall into the Q1 category with a total of 251,357 articles or about 62 percent of the total recorded. This shows that the 39 journals ranked Q1 by the SJR comprised a larger proportion of articles recorded (62.45%) compared with the proportion of Q1 journals in total journal count (38.24%). In the next category, Q2, there are 30 journals with 62,592 articles recorded (15.55% of total). The 29 Q3 journals have a total of 70,317 articles (or 17.47% of total) recorded. These numbers show that the difference in the articles recorded from Q2 and Q3 journals is minuscule. Q4 journals, however, have far fewer articles recorded in PubMed, with only 18,229 articles (4.53% of total) from four journals found. However, of the journals recorded in both the SJR and PubMed, 39 or 38 percent of the total are Q1 journals, whereas the numbers of Q2 and Q3 journals found are 30 and 29 (both around 30% of total each), respectively. Together, Q1 and Q2 journals comprise almost 70 percent of journals recorded in both the SJR and PubMed.

Top 10 cancer research journals in the Q1 category

A further analysis of the 39 journals ranked Q1 shows that in order of the number of articles recorded, *Cancer Research* tops the top 10 publications with a total of 47,099 articles recorded and is followed by *Cancer* with 37,055 articles. Other publications were associated with less than 20,000 articles each. In terms of the country of origin of each publication, apart from *Cancer Letters,* which is published in Ireland, the vast majority of publications are published in the U.K. and the U.S.A. Furthermore, journals with a higher

number of articles recorded may not necessarily have a higher Q1 ranking. For instance, *Cancer Research* has the highest number of articles among Q1 journals but was ranked only eighth in 2012. The *European Journal of Cancer* (Oxford, England: 1990), which is ranked 10th, also has an article count of only 9,653. This shows that while the ranking of the journal is related to its number of citations and the reputations of its authors, its article count must still reach a certain scale. For more details, see Table 2.

Table 2: Top 10 Cancer Research Journals in the Q1 Category

Journal Title	**ISSN**	**No of articles**	**Q1 Rank (2012)**	**Country**
Cancer Research	0008-5472	47,099	8	USA
Cancer	1097-0142	37,055	20	USA
Journal of the National Cancer Institute	0027-8874	19,707	6	UK
British Journal of Cancer	1532-1827	19,558	24	UK
**Journal of Clinical Oncology*	1527-7755	18,603	3	USA
Oncogene	0950-9232	15,669	11	UK
***Clinical Cancer Research*	1078-0432	12,594	10	USA
Carcinogenesis	1460-2180	10,809	21	UK
Cancer Letters	1872-7980	10,539	44	Ireland
European Journal of Cancer (Oxford, England: 1990)	0959-8049	9,653	22	UK

* *Official Journal of the American Society of Clinical Oncology*

*** An Official Journal of the American Association for Cancer Research*

During the processing of raw data, although journal names and ISSNs were the most discernible fields in the cross-referencing between the SJR and PubMed, there were still instances wherein matches could not be obtained. This is because, although journal names and ISSNs had been recorded at the time of entry, PubMed also recorded information such as the publisher of the journal, location and year of publication, and the names of any supplements, together with the journal title, as in the case of the *European journal of cancer* (Oxford, England: 1990); conversely, the same publication would simply be recorded as the *European Journal of Cancer* in the SJR. In addition, if there are any changes in journal ISSN, information data reconciliation is even more challenging during data cleaning.

Distribution by year of publication for journal articles in the area of cancer research: The example of Q1 journals

Regarding the distribution by year of publication, considering the 251,357 Q1 articles published in 39 journals, we see that the earliest texts recorded in PubMed were published in 1945. To divide publication dates into periods of 10 years each, we can see that the period from 1945 to 1949 can be described as the infancy of cancer research with only 593 articles published. However, the field grew rapidly since the 1950s, with 4,618 articles published in the 1950s and 7,706 articles—a growth of almost 100 percent over the last decade—recorded in the 1960s. In the next stage, beginning with the 1970s, the number of articles published in each decade passed the 10,000 mark. Compared with the 1960s, in the 1970s, the number of articles published increased by almost 12,000. From the 1970s to the 1990s, the number of articles published rose from 19,726 to 62,388, while the number of articles published between 2000 and 2009 peaked at 91,339. Between 2010 and 2013, 30,168 articles were published—figure that almost equals the 34,819 articles published between 1980 and 1989. It is extremely likely that the figure for the 2010s would exceed the figures of the 1990s and the 2000s. It is worth noting that the number of articles published during the period from 1990 to 2009 already comprises 38.19 percent (153,727/402,495) of the total number of articles published. These data show that of the many subject areas within the field of medical research, cancer research has grown in scale to a certain degree and that cancer research is an area of focus for researchers. For more details, see Figure 3.

The top 20 literature headings in the area of cancer research

Regarding heading distribution, as publication information found for publications recorded in both PubMed and MEDLINE contain an unspecified number of headings or keywords based on the nature of the publication concerned, if the keyword used is considered to be one that can be used as a major topic keyword for searching other publications, the asterisk symbol (*) is used to distinguish it. Analysis shows that there are a total of 17,061 major topic keywords across 402,495 articles, and the top 20 keywords among these appear 10,000 times or more. Based on the frequency of their appearance, they are further divided into three groups that can be regarded as core topics in the field of cancer research (see Table 3 for more details).

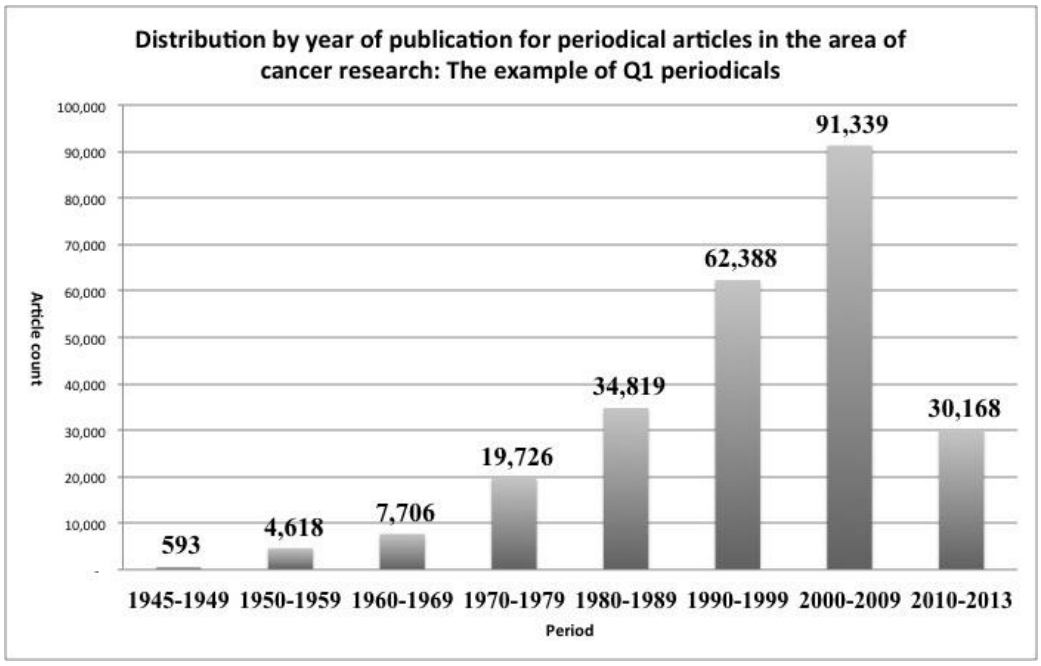

Figure 3: Distribution by year of publication for journal articles in the area of cancer research: the example of Q1 journals

Regarding distribution and the sorting of term incidences, we can see that of 17,061 major topic keywords, there are 20 headings that appear 10,000 times or more. These 20 keywords can be categorized into three groups. (1) The first group comprises of terms that have appeared 50,000 times or more, with the term "Breast Neoplasms" topping this list at 63,767 incidences and followed by "Neoplasms" at 50,680 incidences. These two terms appear most frequently and can be regarded as the most representative and most commonly used major topic keywords in cancer research. (2) The second group comprises of terms that have appeared 20,000–30,000 times: "Antineoplastic agents," (Antineoplastic Combined Chemotherapy Protocols), "Lung Neoplasms," and "Adenocarcinoma." There four terms can be regarded as Tier One terms. (3) The third group comprises terms that have appeared 10,000 times but under 20,000 times: the 14 terms are "Prostatic Neoplasms," "Carcinoma, Squamous Cell," "Liver Neoplasms," "Colorectal Neoplasms," "Brain Neoplasms," "Ovarian Neoplasms," "Melanoma," "Stomach Neoplasms," "Tumor Markers, Biological," "Apoptosis," "Carcinoma," "Colonic Neoplasms," "Skin Neoplasms," and "Carcinoma, Non-Small-Cell Lung." The aforementioned 20 key major topic headings are sufficient for the classification of key topics in cancer research.

Table 3: The top 20 literatures headings in the area of cancer research

Rank	Major Topic	(Total no. of incidences)
1	Breast Neoplasms	63,767
2	Neoplasms	50,680
3	Antineoplastic Agents	36,570
4	Antineoplastic Combined Chemotherapy Protocols	33,743
5	Lung Neoplasms	30,632
6	Adenocarcinoma	21,150
7	Prostatic Neoplasms	17,465
8	Carcinoma, Squamous Cell	16,558
9	Liver Neoplasms	16,171
10	Colorectal Neoplasms	14,972
11	Brain Neoplasms	14,890
12	Ovarian Neoplasms	14,274
13	Melanoma	13,528
14	Stomach Neoplasms	13,389
15	Tumor Markers, Biological	13,213
16	Apoptosis	12,042
17	Carcinoma	11,987
18	Colonic Neoplasms	11,413
19	Skin Neoplasms	11,072
20	Carcinoma, Non-Small-Cell Lung	10,899

Source: As organized by the author

Note: Sorted and based on the number of appearance, from highest to lowest

The distribution of specific major topic headings: The example of the term "Breast Neoplasms"

Narrowly and broadly speaking, the heading distribution within the literatures surveyed can be divided into two categories: (a) texts with the term "Breast Neoplasms" included in the heading field; and (b) other texts with MeSH Tree heading numbers that also encompass the topic. For example, the term "Breast Neoplasms" has two possible MeSH

paths: “C04.588.180” and “C17.800.090.500.” In other words, texts with heading numbers that cover the abovementioned two can also be regarded as being related to the subject of “Breast Neoplasms.” For the correspondence between heading structure and Tree number, please see Figure 4.

In Figure 4, we can see that the heading “Breast Neoplasms,” by tracing the relevant MeSH paths and Tree numbers, appears in a total of 63,767 instances with 43,836 articles containing “Breast Neoplasms” as a keyword. An additional 107 articles were located using the Tree numbers “C04.588.180” and “C17.800.090.500” for a total of 43,943 articles. In other words, when searching for literatures related to the topic “Breast Neoplasms,” 43,943 articles can be located with a recall of 100 percent. However, precision would still need to be determined according to the specific needs of the user.

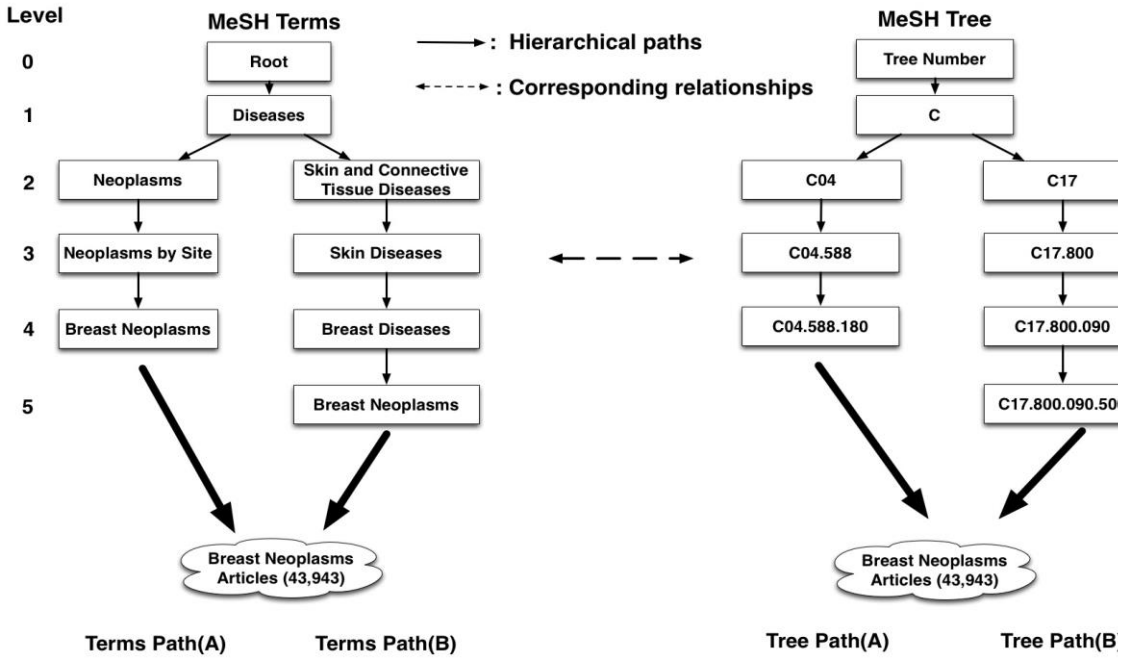

Figure 4: MeSH Paths and Tree Numbers for the heading “Breast Neoplasms”

Source: As organized by the author

Analysis by SJR rank

Using the ranking in the SJR database, 99 journals with 43,943 articles with the keyword “Breast Neoplasms” (P = 43,943, J = 99, P/J = 443.86) were found. Regarding ranking

distribution, Q1 journals constituted the bulk of the set, with 30,839 articles published in 38 journals for an article ratio of 70.18 percent (30,839/43,943) and an average article count of 811.55 per journal. This suggests that the term “Breast Neoplasms” is a rather significant and commonly-seen term in the 38 Q1 journals. Q2 and Q3 journals, at 29 and 28, respectively, have a more similar article count at 4,685 (Q2) and 6,829 (Q3). It would appear that the subject is more salient in Q3 journals than in Q2 journals, since the average article count per journal for Q3 is also higher than for Q2. Relatively fewer Q4 journals have been included in the PubMed database, which is why the average article count per journal is also lower compared with Q1, Q2, and Q3 journals.

Analysis by year of publication

Regarding the year of publication of the literatures surveyed, the first article with the keyword “Breast Neoplasms” was published in 1945. Since then, there has been a clear increasing trend for the number of articles on the subject. Through an examination of developments in 10-year periods, we find that in the two decades between 1990 and 2009, as many as 30,008 articles, or 68.29 percent of the total, were published on “Breast Neoplasms.” Specifically, between 2000 and 2009, the number of articles published peaked at 19,645, while 10,363 articles were published between 1990 and 1999. Simultaneously, the growth in the number of articles published was also highest between the 1990s and 2000s at 9,282. For more details, please see Figure 5.

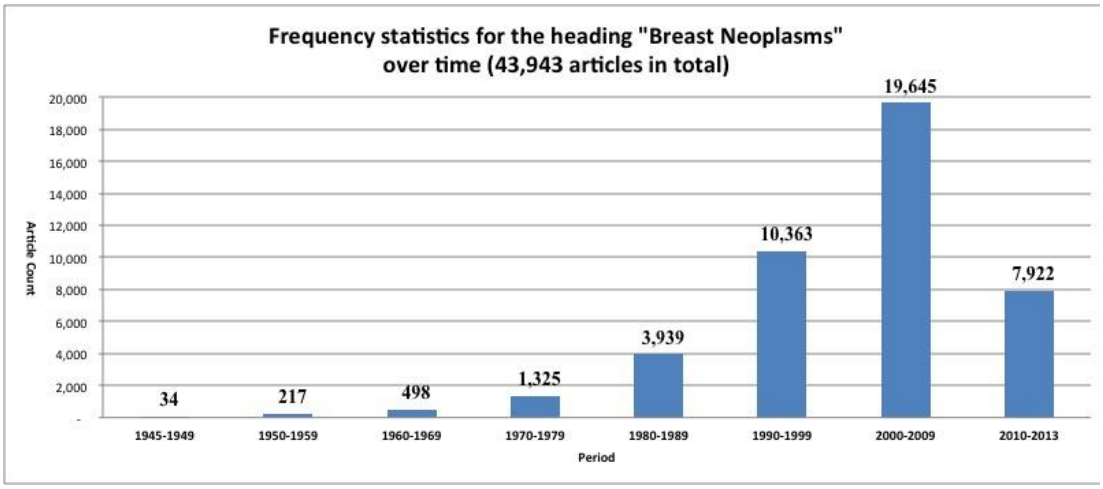

Figure 5: Frequency statistics for the heading “Breast Neoplasms” (by year of publication in 10-year periods)

Notes:

1. The first article with the keyword “Breast Neoplasms” was published in the year 1945;

2. Statistics for the 2010s collected till 2013

CONCLUSIONS

This study proposes a structure for the subject analysis of e-resources. Using medical literatures as the subject of the study, the information sources include the following: (a) a journal list from the SJR database; and (b) the PubMed database from where bibliographic information, including journal type, ranking, and headings used, and year of publication was extracted and organized with reference to the Medical Subject Heading (MeSH) list released by the U.S. National Library of Medicine. The data was statistically analyzed, and a number of tables were produced, to better understand development trends in the field of cancer research.

In this study, subject analysis was conducted on electronic resources in the field of medical research. As there was an authoritative medical subject heading list released by the U.S. National Library of Medicine, the study did not have to build a subject classification system from scratch. For specific fields of research, a third-party database journal list can also be used. Together with the aforementioned two types of materials, the use of such a list

would minimize researcher subjectivity during subject analysis. Simultaneously, medical librarians may seek to improve the service level of their reference services by creating resources such as their own subject heading trees, bibliographies, and subject heading path records for the purpose of automatic analysis of subject trends in large volumes of literatures (and indexing).

ACKNOWLEDGMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

Charidimou, A. and Song, M. 2015. Evolving trends in cerebral amyloid angiopathy research themes: insights from medical subject heading analysis. *Journal of the Neurological Sciences,* Vol. 357, no. 1–2: 341-42.

Chu, C.M. and O'Brien, A. 1993. Subject analysis: the critical first stage in indexing. *Journal of Information Science*, Vol.19, no. 6: 439-454.

Jia, X, Dai, T. and Guo, X. 2014. Comprehensive exploration of urban health by bibliometric analysis: 35 years and 11,299 articles. *Scientometrics*, Vol. 99, no. 3: 881-94.

Li, B., Pan, H. and Ye, D. 2013. A bibliometric study of literature on SLE research in PubMed (2002–2011). *Lupus*, Vol. 22, no. 8: 772-777.

Lo, S-C., Chen, K-H. and Lin, C-J. 2001. The study of framework of subject classification for journal articles in library and information science. *Journal of Library & Information Studies,* Vol. 16: 185-207.

Medicine, U.S. National Library of. 2013. *Introduction to MeSH - 2014.* Available at: http://www.nlm.nih.gov/mesh/introduction.html.

Medicine, U.S. National Library of. 2014a. *MeSH Tree Structures — 2014*. Available at: http://www.nlm.nih.gov/mesh/2014/mesh_trees/trees.html

Medicine, U.S. National Library of. 2014b. *MeSH-Preface.* Available at: http://www.nlm.nih.gov/mesh/intro_preface.html - pref_rem.

Nelson, S.J., Johnston, W.D. and Humphreys, B.L. 2001. Relationships in medical subject headings (MeSH). In *Relationships in the Organization of Knowledge*, 171-184.

Springer.

Ramos, J. M., González-Alcaide, G. and Bolaños-Pizarro, M. 2013. Bibliometric analysis of leishmaniasis research in Medline (1945-2010). *Parasites & vectors*, Vol. 6, no.1: 55.

Scimago Lab. 2012. *Journal rankings on cancer research*. Available at: http://www.scimagojr.com/journalrank.php?area=0&category=1306&country=all&year=2012&order=sjr&min=0&min_type=cd.

Tsay, M-Y, and Hsu Y-T. 2009. An analysis of subjects and citations in core journals of library and information science in Taiwan and China from 1997 to 2006. *Journal of Library & Information Studies,* Vol. 70: 17-38.

Yu, C-Y, and Shieh, J-C. 2014. The study of analytical model of library electronic resources usage-a case of medical electronic resources. *Journal of Educational Media & Library Sciences*, Vol. 51, no. 5: 57-89.