MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190233

Comparative Analysis of Topical Evolution Patterns and Temporal Trends of Hypertension Research

Yuxing Qian[a], Liqin Zhou[a], Rui Zhang[a], Zhiyuan Li[a], Chun Yan[a]

[a] *Center for Studies of Information Resources, Wuhan University, Wuhan, Hubei, China*

Abstract

Exploring the topical evolution patterns and temporal trends of hypertension can promote knowledge communication among experts, and is of great significance to understand the profile and frontiers of chronic disease. Current popular topic detection mainly focuses on two directions: one is based on social network analysis (SNA), the other is based on the topic models. Aiming at distinguishing their similarities and differences, this paper adopts the community detection method and expanded topic model Dirichlet-multinomial regression (DMR) respectively to detect the topic distribution and evolution trends of hypertension research. A total of 26,717 articles in the PubMed database were used as examples to construct the MeSH Terms co-occurrence matrix. It is found that hypertension literature is mainly concentrated on three communities and five research topics. MeSH Terms obtained from SNA are more specific and clearer, while the DMR has an advantage in exploring the evolution patterns of various themes.

Keywords:

Hypertension, Medical Informatics

Introduction

Hypertension is one of the most common chronic diseases and the main risk for cardio-cerebrovascular disease. Some 9.4 million people die of hypertension worldwide every year, in which the complications (such as stroke and heart disease) can lead to about half deaths. Thus, an increasing number of scholars have begun to devote themselves to the study of hypertension. As of May 12, 2017, 284322 articles related to hypertension were retrieved in the PubMed database. However, such large-scale medical literature has caused great trouble for researchers. It is difficult to present a comprehensive overview of the profile and frontiers of Hypertension. As the bibliometric method was widely used to analyze the medical literature, some scholars [1-3] have concentrated on the macro level (including major research countries, institutions, authors, etc.) of hypertension research, without in-depth analyzing their topic distribution and evolution patterns at the medium or micro levels.

Existing literature mainly focused on two directions to detect the topic distribution and evolution patterns: one is based on social network analysis (SNA), the other is based on the topic model [4]. The SNA is widely used in data mining, knowledge management, information dissemination, and knowledge network. It can show the relationship of subject terms clearly and visually and can provide effective support for analyzing the location and association of topic words in the network. M. L. Wallace et al.[5] utilized two case studies to demonstrate that the SNA method (such as community detection) helps in identifying research directions, and can reveal more structural details of the knowledge domain than traditional co-citation analysis.

Meanwhile, the topic model, which was widely used in natural language processing [6], information retrieval, text mining, and other fields, is based on the probability and statistics model in the machine learning field. D. Mimno and A. McCallum [7] proposed the Dirichlet-multinomial Regression (DMR) model extended and derived from the LDA model. M. Song et al. [8] utilized the DMR model to detect the topic distribution and evolution patterns of Alzheimer's disease and achieved good results. Compared with the co-word analysis and citation analysis, the topic model can better reflect the relationship between "words-topics-documents." It has great advantages in topic detection.

The purpose of this study is to detect the topic distribution and evolution patterns of hypertension research and distinguish the similarities and differences among topics obtained by the SNA method and Dirichlet-multinomial regression topic model. To illustrate the process of the methods, 26717 articles related to Hypertension in PubMed database are taken as an example to construct the co-occurrence knowledge network.

Methods

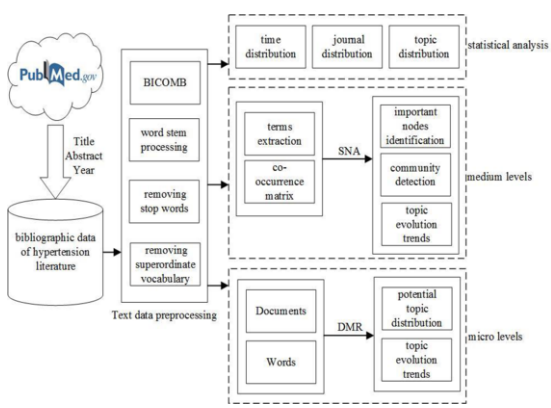

Figure 1– Research Framework

As shown in figure1, the process of our research includes data collection and preprocessing, basic statistical analysis, topic, and community detection by SNA and DMR, comparison of the analysis results.

Data Collection and Processing

The search strategy "Hypertension[MeSH Terms] AND ("2000/1/1"[PDat]: " 2017/5/1")" is used to retrieve the literature of hypertension in the PubMed database. 99252 articles were retrieved. After removing unrelated literature, we finally obtained 26717 articles containing both abstracts and full texts. Then this data was imported into the Bibliographic Items Co-occurrence Matrix Builder (BICOMB) to generate a co-occurrence matrix of bibliographic data. We can find that the 26717 articles were distributed in 1701 journals, involving 171637 authors and 9978 keywords through the extraction and statistics of journals, authors, and keywords.

In order to determine the high-frequency MeSH Terms, the frequency threshold of the subject words is obtained as 77 according to the formula of high and low-frequency word boundary [9]. The high-frequency MeSH subject terms are obtained from the bibliographic data, and the word co-occurrence matrix is constructed in the BICOMB.

Community Detection Based on Optimized Network Modularity

Social network refers to the collection of social actors and their relationships, mainly developed to measure, analyze, and predict the structure and attributes of relationships between various entities in the network [10]. In the academic literature, scholars always use the same or similar words to express the hot-spots in a certain field. The relationship between massive text data can be linked through the subject words of the text to form a huge network. The community is a common phenomenon in social networks composed of a group of highly aggregated and closely connected nodes. Nodes belonging to the same community are more likely to have similar functionality, and community structures can indicate the relationship between network structure and functionality. The most representative community identification algorithm is the optimized network modularity method proposed by M. E. J. Newman [11]. Module degree is an index that can measure the quality of network partitioning, also called Q value. In essence, the modularity-based algorithm performs community identification based on changes in the intermediaries and modularity of the edges.

Since the nodes in the word co-occurrence network are the topic words, the process of determining the community representative topics transformed into the process of finding the core nodes. A few core nodes represent the scientific research themes corresponding to the community. In complex networks, there are many important indicators of nodes such as centrality and PageRank values. These indicators consider the calculation of the number of edges, centrality, and connections with other nodes to determine the core nodes from the global level of the network.

Dirichlet-multinomial Regression Topic Model

The Dirichlet-multinomial regression topic model, as shown in figure 2, mainly obtains the distribution of topics under different conditions by adjusting the characteristics of the observed documents. This paper takes the time of publication of hypertension literature as a variable to explore the trend of the topic over time.

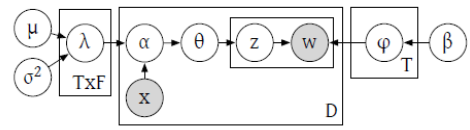

Figure 2– DMR topic model

In the document set D, for each document d, x_d represents the feature vector of the metadata, α is a function of the observable document feature, represents the prior probability distribution of the subject; given a prior probability distribution $\mathrm{N}(0,\Sigma)$, hyperparameters β , documents and words are generated as follows:

For each topic t, draw $\phi_t \sim Dir(\beta)$. $Dir(\beta)$ is a different topic-word distribution from the previous Dirichlet distribution.

For each document d, draw $\theta_d \sim Dir(\alpha_d) = \mathrm{Dir}(\exp(\tau_d))$, $\tau_d \in \tau$. For α_d of each document, the parameter of Dirichlet distribution and τ_d is covariance functions $\mathrm{f}(y_d, x_k)$, where y_d is the observed attribute vector of documents, and x_k is the vector of metadata.

For each word w, draw $z_{d,w} \sim Multi(\theta_d)$. $z_{d,w}$ is the subject allocation of the word t_w and θ_d is the proportion of the document d belonging to a certain topic, draw $T_{d,w} \sim Multi(\phi_{z_{d,w}})$. $T_{d,w}$ is the w-th word in document d, ϕ_t is the preference of topic t, $\Sigma_n \phi_{t,n} = 1$.

In the DMR topic model, three fixed parameters are set including the variance of the previously distributed parameter values σ^2, Dirichlet topic-word distribution β and the number of topics |T|.

Results

Topic Detection and Evolution Trend Analysis Based on SNA

In order to achieve better visualization, the nodes with the frequency no higher than 77, and the isolated nodes are deleted. 632 nodes are obtained to construct a co-occurrence matrix of high-frequency words. To reduce its complexity, the top 100 nodes are selected, and 4950 edges are included. A node is a biological entity derived from articles. The edge represents the relationship between the entities. The weight of the edge indicates the frequency at which the two entities co-occur in the specific sentence of the article. The data is imported into Gephi, and the community detection algorithm [12] is used for visualization, as shown in Figure 3.

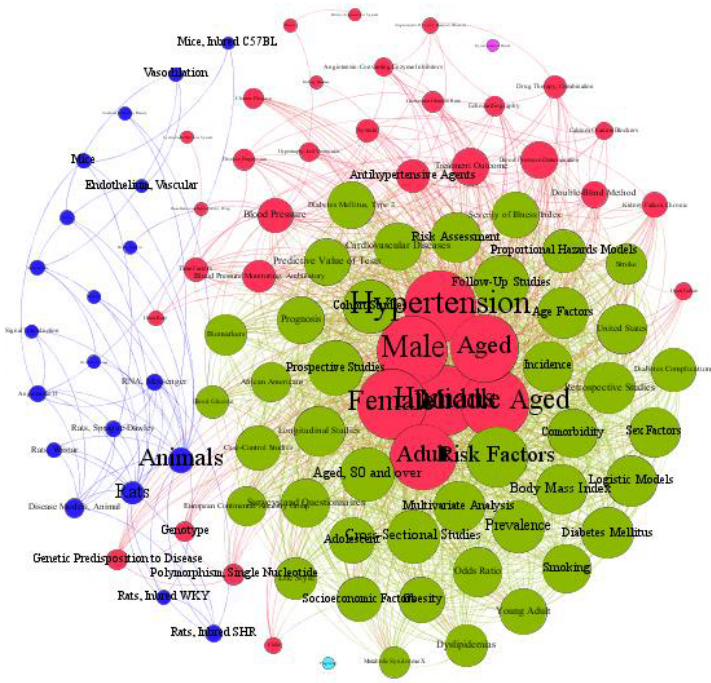

Figure 3–community detection (SNA)

To identify the most important biological entities in the literature of hypertension, four well-known centrality indicators including PageRank value, weighted centrality, closeness centrality, and betweenness centrality are calculated. Detailed centrality measurements are made by S. Wasserman [13] and S. Brin [14]. The top-ranked biological entities with PageRank values are similar to the top entities with weighted centrality rankings. Moreover, PageRank has unique biological entities, such as rates that only appear in the PageRank values top 10. The risk factor, prospective studies, and follow-up studies, only appear in the weighted centrality top 10, while Blood Pressure, Animals, and Body Weight only appear in the betweenness centrality top 10. The top 10 words in closeness centrality are exactly the same as the weighted centrality. In general, the nodes that are at a critical position in the entire network are mainly Hypertension, Male, Female, Age, Adult, Human, Risk factors, Body Weight, Blood Pressure, Animals, Prospective Studies, etc. These nodes are also at the core of the community distribution map.

The modular algorithm proposed by V. D. Blondel [12] is used for community detection and the resolution is set to 1 [15]. 5 modules are detected including 3 main modules (shown in red, blue, and green in Figure 3). The modular value Q is 0.187. The average path length is 2.645, and the diameter is 6. Choosing the Fruchterman Reingold layout, the two communities of light blue and purple in the picture account for only 1% of the total communities and are negligible. Therefore, communities of red, blue and green are mainly considered. The largest green community accounts for 42% of the network, including Risk Factors, Aged, Prevalence, Sex Factors, Prospective Studies, Follow-Up Studies. The second largest community (red part) accounts for 36% of the network, including Hypertension, humans, female, middle-aged, Treatment Outcome, Antihypertensive Agents, etc. The third largest community (blue part) accounts for 20% of the network, including animals, Rats, Inbred SHR, Messenger, Mice, etc. Through manual judgment and expert identification, the three communities can be initially divided into five topics.

Topic 1 mainly includes words related to risk factors for hypertension, such as age, pregnancy, gender, smoking, obesity, myocardial infarction, etc. Topic 2 mainly includes research methods and models related to hypertension, such as prospective studies, follow-up studies, Lateral studies, cohort studies, retrospective studies, etc., also includes some indicator parameters for hypertension research, such as morbidity, disease severity index, experimental expectations, and so on. Topic 3 mainly includes the basic elements of hypertension, such as gender, age, blood pressure, heart rate, renin, glomerular filtration rate, etc. Topic 4 mainly includes disease diagnosis and treatment, such as diagnostic effects, antihypertensive drugs, blood pressure measurement, drug dose response relationship and so on. Topic 5 mainly includes animal, rat, RNA, and Inbred SHR. That is to verify the indicators of hypertension through animal experiments. In general, the biological entities within each community are closely linked to form a variety of specific research topics relevant to the hypertension literature.

According to the time distribution, the hypertension literature was divided into three stages: 2000-2005, 2006-2010, 2011-2017. After data processing, they are imported into Gephi and visualized by a community detection algorithm. According to the visual distribution map and various parameters, the three stages of hypertension topics are distributed in three communities, and the proportion of each community is relatively average. The proportion of communities in the period from 2000 to 2005 was 38%, 33%, and 29 %, the proportion of each community in the period of 2006-2010 was 40%, 38%, 22%. In 2011-2017, the proportion of each community was 42%, 37%, 21%; compared with the MeSH Terms obtained in the three stages, the most frequent occurrences of the three stages are hypertension, male, and female. Most of the terms appear in three stages at the same time, but the distribution of the theme communities in each stage is slightly different. The topic community distribution parameters at each stage are shown in Table 1. The average degree, graph density, and modularity in these three stages are constantly increasing,

indicating that the number of MeSH erms in each stage is increasing with time and the distribution of subject communities is constantly changing. However, this method is very labor intensive, and it is difficult to accurately detect the proportion of each topic in each time period and the path that the theme evolves over time.

Table 1–Parameters at Each Stage

	2000-2005	2006-2010	2011-2017
Average degree	13.94	16.88	20.4
Graph density	0.141	0.171	0.202
Modularity	0.126	0.158	0.175
Average clustering coefficient	0.891	0.438	0.869

Topic Detection and Evolution Trend Analysis Based on DMR

The bibliographic data of 26717 articles were processed as follows: word stem processing; removing stop words, words of length 1 and words with a frequency less than 5 times; removing the superordinate vocabulary in hypertension field. Each bibliographic data generates a text file as a document for the DMR topic model. The data is then processed through an open source machine learning language processing package Mallet [16], according to the DMR model and algorithm. In order to contrast with the topics detected by SNA, the number of topics |T| is set to 5, while adjusting the σ^2 and β values, the resulting five topics and related subject terms. In order to make the identified topics and terms more meaningful, the numbers of terms in the topic-vocabulary distribution of each topic are set to 10, 20, and 30. 10 words that appear more frequent and meaningful are selected, as shown in Table 2.

Topic 1 contains mice, angiotensin, renin, vascular, response, effects, receptor, rats, etc., mainly to describe animal experiments related to hypertension; topic 2 contains risk, factors, age, obesity, gene, women and other words Mainly used to describe the risk factors of hypertension, including age, diabetes, obesity, gender, genes, etc. Topic 3 includes systolic, diastolic, group, compare, rate, invalid, significant, etc., mainly used to describe and hypertension related research methods. Topic 4 contains patients, blood, gene, treatment, results, and other words, mainly used to describe basic elements of hypertension. Topic 5 contains treatment, antihypertensive, therapy, coronary, mortgage, medication, care, control, etc., mainly used to describe the diagnosis and treatment of hypertension.

Table 2–Topic Distribution (DMR)

Topic 1 animal experiments	**Topic 2** risk factors	**Topic 3** research methods	**Topic 4** Basic elements	**Topic 5** Diagnosis Treatments
Mice	Risk	Systolic	Patient	Patient
Renal	Age	Group	Blood	Antihypertensive
Vascular	Prevalence	Significant	Treatment	Coronary
Effects	Blood	Rate	Results	medication
Expression	Disease	Pressure	Finding	clinical
Angiotensin	Factors	Diastolic	cardiovascular	Treatment
Proteins	Diabetes	Compare	Gene	Therapy
Response	Obesity	Invalid	Association	Mortality
Receptor	Gene	Term	Evidence	care

Then the time of publication of the hypertension literature is taken as a variable to explore the trend of the topic over time. The relative distribution of each topic from 2000 to 2017 is shown in Figure 4 below. In general, each topic is constantly evolving over time. In 2000, topic 1 (animal experiment) and topic 4 (basic elements) accounted for a relatively large proportion, and topic 5 (diagnosis and treatment) research was relatively weak; over time, topic 1 (animal experiment) The trend of gradual decline, the topic 4 (basic elements) first decline and then rise, but it has been in an important position; and the proportion of topic 5 (diagnosis and treatment) has increased year by year. It had a relatively important proportion in 2017. Topic 2 (risk factors) has developed relatively stable and has been in a relatively important position; topic 3 (research method) has slightly fluctuated and its proportion has increased year by year since 2007.

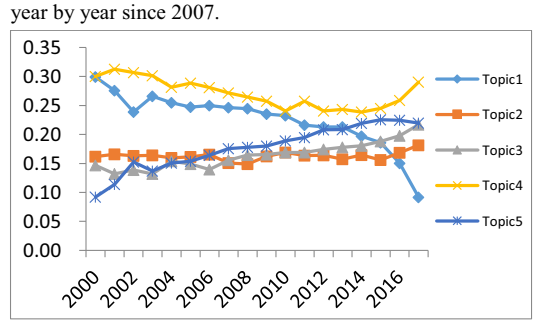

Figure 4– Trend of the Topic Evolution (DMR)

Discussion

The results show that the topics obtained by the SNA are almost the same with those detected by the DMR, both including risk factors, research methods, basic elements, diagnosis, and treatment as well as animal experiments. From the macroscopic view, the MeSH Terms obtained using the SNA method are more specific and more precise, while the DMR are broader. For example, in the topic risk factors, the MeSH Terms determined by SNA, including age factors, Diabetes Mellitus, sex factors, smoking, cardiovascular disease, obesity, and lifestyle can represent relatively concrete risk factors. On the contrary, the term identified by DMR, including age, Diabetes, obesity, and gene, are wider, only to illustrate each general categories of risk factors. In addition, in research methods part, SNA can not only recognize the terms covering prospective studies, logistic models, surveys and questionnaires which stand for research methods, but also include targets such as risk assessment, odds ratio, severity of illness index while DMR recognizes the terms including group, compare, rate and significant. This is because the dataset is too large. Those objects selected by SNA are the top 100 highest-frequency MeSH Terms, while DMR focuses on the entire document. Different research objects thus result in the dissimilarity in the outcome. The number of community topics is also different. SNA sets the number subjectively, and DMR adjusts the number of topics and terms in advance as well. Besides, the relation among 'terms-topics-documents' can be better displayed by DMR so that the probability distribution of the topics can be figured out in any document.

In the comparison in evolution trends, SNA can merely detect the distribution of topic communities for a specific period, hard to compare the evolution trends happening in every duration. However, DMR can detect the ratios of different topics in every period and the evolution trends of topics depending on time,

and it has advantages in exploring the process of topics evolution trends.

In general, the SNA method and DMR method pay different attention to detecting community and evolution trends. The MeSH Terms obtained by SNA are more concrete and precise while terms obtained by DMR are more widely, which need interpretations but are more advantageous in exploring the evolution trends of every topic. Given that combining these two methods together, exploring the community and evolution trends of knowledge network from both the medium and micro view. As a result, they can supplement each other.

Conclusions

Firstly, it is found that the hypertension literature is mainly concentrated on three communities, which can be divided into five research topics, such as risk factors, research methods, basic elements, diagnosis and treatment, and animal experiments. Secondly, the topic changes constantly with time going by. The basic situation of patients has always occupied a high proportion of research. Researches of animal experiments have decreased yearly. Development of risk factors analysis has accounted for a relatively important ratio steadily. The percentage of research topics have been increasing since 2007. Third, it is also found that the topic obtained from SNA and DMR are basically similar. But the MeSH Terms obtained using the SNA method are more specific and precise, while the DMR are broader and have an advantage in exploring the evolution of various themes. If they are applied jointly, the result will be better.

These investigations can help researchers who just start to explore hypertension study to understand the field overview, discover research hot-spots and predictive research frontiers in the field, and promote knowledge exchange within and between domains among experts to help decision makers follow up the flow of knowledge in the field of hypertension. At the same time, the analysis methods of community detection and topic evolution trends in this paper can be extended to other areas of chronic diseases, such as diabetes, coronary heart disease, etc.

Based on the results of this study, PubMed and other databases can provide such information services, helping researchers to obtain a global understanding of relevant fields in the literature search, and improve the relevance and efficiency of the literature search.

This paper also has some limitations. Some Mesh Terms are wildly used in the PubMed publications and are not specific to hypertension, such as “humans”, “female”, “middle-aged”. The accuracy of this research will be further improved by means of filtrating these terms by normalizing their frequency in the target corpus with their overall frequency in PubMed. The DMR topic model needs to preset the number of topics when detecting potential topics. But the number of topics identified by DMR in this paper is subjectively determined; In addition, “a direction for future work is analyzing the internal structures and correlations of communities and topics, which is the next step of our study.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Nos. 71661167007).

References

[1] Y.S. Oh and Z.S. Galis, Anatomy of success: the top 100 cited scientific reports focused on hypertension research, *Hypertension* **63** (2014), 641-647.
[2] C. Schreiber, C. Edlinger, S. Eder, M. Ausserwinkler, B. Wernly, I. Pretsch, C. Jung, U.C. Hoppe, and M. Lichtenauer, Global research trends in the medical therapy of pulmonary arterial hypertension 2000–2014, *Pulmonary Pharmacology & Therapeutics* **39** (2016), 21-27.
[3] M. Götting, M. Schwarzer, A. Gerber, D. Klingelhöfer, and D.A. Groneberg, Pulmonary hypertension: scientometric analysis and density-equalizing mapping, *PloS One* **12** (2017), e0169238.
[4] Y. Ding, Community detection: topological vs. topical, *Journal of Informetrics* **5** (2011), 498-514.
[5] M.L. Wallace, Y. Gingras, and R. Duhon, A new approach for detecting scientific specialties from raw cocitation networks, *Journal of the American Society for Information Science and Technology* **60** (2009), 240-246.
[6] A. Gross and D. Murthy, Modeling virtual organizations with Latent Dirichlet Allocation: a case for natural language processing, *Neural networks* **58** (2014), 38-49.
[7] D. Mimno and A. Mccallum, Topic models conditioned on arbitrary features with Dirichlet-multinomial regression, in: *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI2008)*, 2008, pp. 411-418
[8] M. Song, G.E. Heo, and D. Lee, Identifying the landscape of Alzheimer’s disease research with network and content analysis, *Scientometrics* **102** (2015), 905-927.
[9] Y. Yang, M. Wu, and L. Cui, Integration of three visualization methods based on co-word analysis, *Scientometrics* **90** (2012), 659-673.
[10] C.T. Butts, Social network analysis: A methodological introduction, *Asian Journal of Social Psychology* **11** (2008), 13-41.
[11] M.E. Newman and M. Girvan, Finding and evaluating community structure in networks, *Physical review E* **69** (2004), 026113.
[12] V.D. Blondel, J.L. Guillaume, R. Lambiotte, and E. Lefebvre, Fast unfolding of communities in large networks, *Journal of Statistical Mechanics* **2008** (2008), 155-168.
[13] A.W. Wolfe, Social network analysis : methods and applications, *American Ethnologist* **24** (1997), 219-220.
[14] S. Brin and L. Page, The anatomy of a large-scale hypertextual web search engine, *Computer networks and ISDN Systems* **30** (1998), 107-117.
[15] R. Lambiotte, J.-C. Delvenne, and M. Barahona, Laplacian Dynamics and multiscale modular structure in networks, *Physics* (2008), 1-29.
[16] H.M. Wallach, D.M. Mimno, and A. McCallum, Rethinking LDA: why priors matter, in: *Advances in Neural Information Processing Systems*, 2009, pp. 1973-1981.

Address for correspondence

Corresponding author: Liqin Zhou, PhD student. Center for Studies of Information Resources, Wuhan University, Wuhan, Hubei, China
E-mail: zhoulq92@163.com