Comparing Keywords Plus of WOS and Author Keywords: A Case Study of Patient Adherence Research

Juan Zhang  
School of Public Health, Shanxi Medical University, No. 56, Xin-Jian S. Road, Taiyuan 030001, China. E-mail: lizhao2000849@126.com

Qi Yu  
Department of Information Management, Shanxi Medical University, No. 56, Xin-Jian S. Road, Taiyuan 030001, China. E-mail: yuqi351@gmail.com

Fashan Zheng  
The Robert Darbishire Practice, Rusholme Health Centre, Walmer Street, Rusholme, Manchester, England M14 5NP, United Kingdom. E-mail: fashan.zheng@nhs.net

Chao Long  
Stanford University School of Medicine, 871 Warren Way, Palo Alto, CA 94303, USA. E-mail: chaolong@stanford.edu

Zuxun Lu  
School of Public Health, Tongji Medical College, Huazhong University of Science and Technology. No. 13, Hang-Kong Road, Wuhan 430030, China. E-mail: zuxunlu@yahoo.com

Zhiguang Duan  
School of Public Health, Shanxi Medical University, No. 56, Xin-Jian S. Road, Taiyuan 030001, China. E-mail: dzg52827@aliyun.com

Bibliometric analysis based on literature in the Web of Science (WOS) has become an increasingly popular method for visualizing the structure of scientific fields. Keywords Plus and Author Keywords are commonly selected as units of analysis, despite the limited research evidence demonstrating the effectiveness of Keywords Plus. This study was conceived to evaluate the efficacy of Keywords Plus as a parameter for capturing the content and scientific concepts presented in articles. Using scientific papers about patient adherence that were retrieved from WOS, a comparative assessment of Keywords Plus and Author Keywords was performed at the scientific field level and the document level, respectively. Our search yielded more Keywords Plus terms than Author Keywords, and the Keywords Plus terms were more broadly descriptive. Keywords Plus is as effective as Author Keywords in terms of bibliometric analysis investigating the knowledge structure of scientific fields, but it is less comprehensive in representing an article’s content.

Introduction

The bibliometric method is a prominent research tool that can systematically represent the nature of specific scientific disciplines by highlighting research hotspots and detecting research trends. The Web of Science (WOS), maintained by Thomson Reuters, is considered one of the main bibliographic sources of information. WOS records include two types of keywords: Author Keywords, those provided by the original authors, and Keywords Plus, those extracted from the titles of the cited references by Thomson Reuters. Keywords Plus, generated by an automatic computer algorithm, are words or phrases that appear frequently in the titles of an article’s references and not necessarily in the title of the article or as Author Keywords (Garfield, 1990; Garfield & Sher, 1993).

Garfield claimed that Keywords Plus terms are able to capture an article’s content with greater depth and variety.
(Garfield, 1990). Therefore, Keywords Plus terms have been used to identify research trends in a variety of scientific fields, ranging from pentachlorophenol research (Huang, Ao, & Ho, 2008) to gene transfer-related research (Wen, Yu, & Ho, 2009). Other fields, including, but not limited to, world aerosol (Xie, Zhang, & Ho, 2008), solid waste (Fu, Ho, Sui, & Li, 2010), risk assessment (Mao, Wang, & Ho, 2010), global climate change (Li, Wang, & Ho, 2011), and obstructive sleep apnea (Huang, 2009; Huang et al., 2008) have used Keywords Plus in addition to author keyword as part of their bibliometric analyses. However, is Keywords Plus a representative parameter for the content and scientific concepts presented in articles? This question, which emerged as a follow-up question in a study of the knowledge structure of patient adherence (Zhang et al., 2012), has yet to be critically examined to the best knowledge of the investigators of this study.

Author Keywords consist of a list of terms that authors believe best represent the content of their paper (Li, Ding, Feng, Wang, & Ho, 2009). They are often selected prudently. This study presents a comparative assessment of Keywords Plus and Author Keywords to determine whether the former is as reliable an indicator for the scientific concepts presented in articles. The comparison was conducted from two separate perspectives, that of the document and the research field.

Materials and Methods

Articles containing terms related to adherence (e.g., adherence, compliance, non-adherence, persistence) were searched in the WOS (SCI-EXPANDED, SSCI) from January 2000 to May 2011. The search was performed on May 31, 2011 and a total of 3,034 studies were obtained. Investigators examined these studies’ titles and abstracts, selecting only articles that related to patient adherence (adherence was defined by the World Health Organization [WHO] as the extent to which a person’s behavior—taking medication, following a diet, and/or executing lifestyle changes—corresponds with agreed recommendations from a health care provider [WHO, 2003]) and have keywords. Articles that evaluated persistence of specific substances and phenomena (e.g., modified T cells, therapeutic effect, bacteria), reported healthcare providers’ adherence to guidelines/criteria, and/or had no keywords did not meet this criteria and were thus excluded. The remaining 2,397 articles were included in this study.

Document Perspective

Literature containing both Author Keywords and Keywords Plus was selected. The number of Author Keywords and Keywords Plus were counted and compared using a Wilcoxon signed-ranks test with SPSS18.0.

Overlap (O) and Redundancy (R) were calculated for the literature containing both types of keywords (Mikki, 2010).

Overlap (O) is defined by

\[ O = \frac{|K(v_i) \cap A(v_i)|}{|K(v_i) \cup A(v_i)|} \]

where \( K(v_i) \) is the set of Keywords Plus and \( A(v_i) \) is the set of Author Keywords. The value of \( O \) is 1 if \( K(v_i) \) is equal to \( A(v_i) \), and 0 when the two sets of keywords do not intersect at all.

Redundancy (R), the reappearance of Author Keywords in Keywords Plus, is defined by

\[ R = \frac{|K(v_i) \cap A(v_i)|}{|A(v_i)|} \]

where the value of R ranges from 0 for 0% coverage to 1 for 100% coverage.

Research Field Perspective

Author Keywords were extracted from each paper, and those with similar meanings were aggregated into a single primary term (e.g., “adherence,” “nonadherence,” “compliance,” and “noncompliance” were merged into “adherence”). In such instances, a term may occur two or more times in the same paper but only be counted once. For example, if “nonadherence,” “compliance,” and “chronic hepatitis C” were the three listed keywords, only “adherence” (representing both “nonadherence” and “compliance”) and “chronic hepatitis C” would be used for further study.

The number of occurrences of an author keyword is defined as the frequency of a keyword, whereas co-occurrence frequency is defined as the frequency of a pair of keywords occurring simultaneously. High frequency words were selected to form the co-word networks that were then visualized in two-dimensional knowledge maps by the network visualization software NetDraw2.084. The same applied for Keyword Plus terms.

Results and Discussion

From January 2000 to May 2011, there were a total of 2,397 research papers in the WOS regarding patient adherence. Of those papers, 1,672 contained both types of terms, totaling 8,085 Author Keywords and 11,893 Keyword Plus terms. Statistical analysis (Wilcoxon signed-ranks test) showed that there were significantly more Keyword Plus terms than Author Keywords (\( Z = -23.052, p < 0.001 \)).

| TABLE 1. The percentiles of overlap and redundancy. |
|----------------|------|------|------|------|
|               | min  | 25   | 50   | 75   | max  |
| Overlap       | 0.050| 0.071| 0.083| 0.125| 0.333|
| Redundancy    | 0.083| 0.167| 0.200| 0.250| 0.600|
Of the 1,672 papers with both types of terms, 324 (19.38%) had overlap, containing one to three terms shared by both term sets. Of those, most (n = 281; 86.73%) contained one shared term, 40 (12.35%) papers contained two shared terms, and three (0.93%) papers contained three shared terms. Overlap between the two sets ranged from 0.050 to 0.333 with a median of 0.083, and the median for redundancy of Author Keywords by Keyword Plus terms was 0.200 (Table 1).

Significant differences in the content and focus of studies were found between the Keywords Plus and Author Keywords analyses. Garfield drew similar conclusions in a study from the perspective of retrieval (Garfield & Sher, 1993). A literature search using source title words alone retrieved 100 relevant articles, whereas using a combination of Keywords Plus and Author Keywords produced 63 additional articles. Thirteen of the 63 items were both Keywords Plus and Author Keywords, while 45 supplementary items were uniquely found by only Keywords Plus and five by only Author Keywords (Garfield & Sher, 1993). This suggests that Keywords Plus terms are more comprehensive and general than Author Keywords.

Research Field Perspective

In total, 2,397 articles were used to analyze the knowledge structure of patient adherence research from January 2000 to May 2011. After processing, there were 3,012 Author Keywords and 3,498 Keyword Plus terms that could be used for analysis.

Table 2 compares the number and frequency of the two types of terms. The number of Keywords Plus with high frequency is significantly larger than for Author Keywords, and the frequencies of Keywords Plus are similarly higher than that of Author Keywords. Furthermore, the proportion of Keywords Plus terms occurring at least 100 times among all Keywords Plus terms is greater in both number and frequency than that of Author Keywords occurring at least 50 times among all Author Keywords. The percentage of Keywords Plus terms with frequencies greater than 35 and the percentage of Author Keywords with frequencies greater than 10 was equivalent in both number and frequency.

The knowledge structure of patient adherence research is mapped with 81 Author Keywords with a frequency of greater than 10 and 84 Keywords Plus terms with a frequency of greater than 35. Among these, 35 terms were shared by both.

Table 3 shows the 20 Keywords Plus terms and Author Keywords that appeared most frequently, with their rankings and frequencies. There are broad terms in the Keywords Plus set, such as “therapy,” “care,” and “disease.” Keyword Plus terms were significantly more likely to describe methods and techniques, whereas Author Keywords were more likely to describe specific adherence issues in association with a particular disease. Both represent the study population and risk factors of adherence equally (Table 4).

Figure 1 shows the knowledge map of patient adherence research constructed from the Author Keywords of articles from January 2000 to May 2011. To highlight the most significant relationships, weak linkages with a
| Group          | Common terms                                                                                     | Author Keywords                                                                                     | Keywords Plus                                                                                     |
|---------------|-------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------|
| diseases      | HIV, asthma, schizophrenia, diabetes, depression, hypertension, psychosis                      | chronic disease, osteoporosis, renal transplant, heart failure, cardiovascular disease, breast cancer, COPD, glaucoma, TB, multiple sclerosis, bipolar disorder, rheumatoid arthritis, inflammatory bowel disease, obstructive sleep apnea | disease, myocardial infarction, coronary heart disease                                             |
| population    | adolescent, child, patient, elderly, women                                                    | African American, pediatric prevention, primary care, therapy, intervention, drug, HAART, protease inhibitor, antiretroviral | population, HIV patient, adult blood pressure control, glucose control, health-care               |
| conditions    |                                                                                                 |                                                                                                    |                                                                                                  |
| methods       | RCT, validity, questionnaire, self report, clinical trial                                      | qualitative research, electronic monitor                                                            | trial, meta analysis, follow up, randomized trial, cohort, pharmacy record, efficacy, validation, reliability, controlled-trial, double-blind |
| others        | adherence, persistence, drug adherence, outcome, quality of life                               | quality of care, Pakistan, guideline adherence, satisfaction, screening                             | care, management, impact, risk, health, mortality, prevalence, blood pressure, infection, survival, guidelines, disorder, pattern, program, self reported adherence, quality, model, USA, efficacy, viral load |

**Note.** COPD = Chronic Obstructive Pulmonary Disease; CPAP = Continuous Positive Airway Pressure; HAART = Highly Active Antiretroviral Treatment; DOTS = Directly Observed Treatment Strategy; RCT = Randomized Controlled Trial.

![FIG. 1. Knowledge structure of research on patient adherence formed with Author Keywords (frequency ≥10 times). The size of the nodes indicates the keywords’ frequency, and the thickness of the lines indicates the cooccurrence frequency of keyword pairs.](image-url)
co-occurrence frequency of less than 25 are not shown in the map (Figure 2). It is immediately apparent that the map is dominated by the keyword “adherence,” and that most other terms are linked directly to it. It might be concluded that the research focus was concentrated on drug adherence, adherence of special populations, adherence of patients with specific diseases and interventions for adherence. Interestingly, the term “HIV” is a satellite center forming a subnetwork, which indicates that the treatment adherence of people living with HIV/AIDS (PLWHA) was a major theme during this period.

The main structure of the knowledge map formed by Keywords Plus terms, is slightly more complex than that of Author Keywords (Figure 3). The word “adherence,” which dominates the whole network, is linked to subnetworks by words such as “drug,” “drug adherence,” “therapy,” and “HIV.” The subnetwork indicates that adherence to therapy regimen, drug adherence, and adherence of PLWHA are the research hotspots in this period. This finding is consistent with that of the Author Keywords map. Certain terms such as “trial” and “meta-analysis” appeared in the Keywords Plus map, suggesting that trial and meta-analysis are important research methods and techniques.

Using both Keywords Plus and Author Keywords terms when analyzing knowledge structure of science may be more comprehensive. However, the following arguments can be made against doing so, as they demonstrate that the combination of both is not necessarily a better choice than using Keywords Plus alone.

First, Keywords Plus covers the majority of Author Keywords; in the high frequency words set, more than half of the Author Keywords were mirrored in the Keywords Plus sets. Furthermore, if the threshold for Keywords Plus is reduced to ≥10 times, the coverage of Author Keywords increases to more than 75%.

Second, the merger will inflate the total number of words and the frequencies of the words shared between the two sets, which were all high frequency originally. Correspondingly, the frequencies of other words unique to one set remain unchanged, and they have little chance of being selected as high frequency words used to map the knowledge structure.

Additionally, the knowledge structures of patient adherence resulting from Author Keywords and Keywords Plus analysis are vastly similar. This result is consistent with findings in the research fields of world aerosol (Xie et al., 2008), solid waste (Fu et al., 2010), risk assessment (Mao et al., 2010), and obstructive sleep apnea (Huang, 2009).

Finally, it would take more time and effort to use both sets of terms.

**Conclusion**

On the whole, Keywords Plus and Author Keywords revealed similar research trends. The two types of terms characterized the study population and risk factors of non-adherence equally. However, Keywords Plus terms emphasized research methods and techniques, whereas Author Keywords tended to hone in on specific diseases and conditions.

Bibliometric analyses of the structure of scientific fields should use Keywords Plus, because the large quantity of terms and their broad meanings confer several advantages. Thus, relative to Author Keywords, Keywords Plus terms are less specific descriptors of the content of articles. It should be noted that Keywords Plus was a pioneering compromise.
by Garfield to make extra use of cited references in bibliographic records by indexing the words in the cited references in the pandisciplinary collection of Science Citation Index (SCI), Social Sciences Citation Index (SSCI), and Arts & Humanities Citation Index (A&HCI) in lack of a pandisciplinary thesaurus or controled vocabulary.

The aim of this preliminary study was to examine the efficacy of Keywords Plus terms in the field of patient adherence research, and as such the conclusions drawn do not necessarily extend to other research fields. More studies are still needed to explore the efficacy of Keywords Plus in other research arenas.

Acknowledgments

The research reported in this paper was done as part of the project “Scientific and Technological Collaboration in the Field of Biomedicine—Using Co-authorship and Co-inventorship Analysis (No. 71240006)”, supported by National Natural Science Foundation of China.

References

Fu, H.Z., Ho, Y.S., Sui, Y.M., & Li, Z.S. (2010). A bibliometric analysis of solid waste research during the period 1993–2008. Waste Management, 30(12), 2410–2417.
Garfield, E. (1990). Keywords Plus®: ISI’s breakthrough retrieval method. Part I. Expanding your searching power on Current Contents on Diskette. Current Contents®, 1(32), 5–9.
Garfield, E., & Sher, I.H. (1993). Keywords Plus™ algorithmic derivative indexing. Journal of the American Society for Information Science, 44(5), 298–299.
Huang, C.P. (2009). Bibliometric analysis of obstructive sleep apnea research trends. Journal of the Chinese Medical Association, 72(3), 117–123.
Huang, Y., Ao, X.L., & Ho, Y.S. (2008). Use of citation per publication as an indicator to evaluate pentachlorophenol research. Scientometrics, 75(1), 67–80. doi: 10.1007/s11192-007-1849-y
Li, L.L., Ding, G.H., Feng, N., Wang, M.H., & Ho, Y.S. (2009). Global stem cell research trend: Bibliometric analysis as a tool for mapping of trends from 1991 to 2006. Scientometrics, 80(1), 39–58.
Li, J.F., Wang, M.H., & Ho, Y.S. (2011). Trends in research on global climate change: A Science Citation Index Expanded-based analysis. Global and Planetary Change, 77(1–2), 13–20.
Mao, N., Wang, M.H., & Ho, Y.S. (2010). A bibliometric study of the trend in articles related to risk assessment published in Science Citation Index. Human and Ecological Risk Assessment: HERA, 16(4), 801–824.
Mikki, S. (2010). Comparing Google scholar and ISI web of science for Earth sciences. Scientometrics, 82(2), 321–331.
Wen, D.H., Yu, T.C., & Ho, Y.S. (2009). Bibliometric tools applied to analytical articles: The example of gene transfer-related research. OCLC Systems & Services: International Digital Library Perspectives, 25(3), 186–199.
WHO. (2003). Adherence to long-term therapies: Evidence for action. Geneva: Swiss.
Xie, S.D., Zhang, J., & Ho, Y.S. (2008). Assessment of world aerosol research trends by bibliometric analysis. Scientometrics, 77(1), 113–130.
Zhang, J., Xie, J., Hou, W.L., Tu, X.C., Xu, J., Song, F.J., & Lu, Z.X. (2012). Mapping the knowledge structure of research on patient adherence: Knowledge domain visualization based co-word analysis and social network analysis. PLoS ONE, 7(4), e34497. doi: ARTN.