# Paper Title

Author1, Author2, Author3, Author4§

1College of XX, XX University, City, Country

2 College of XX, XX University, City, Country

3 College of XX, XX University, City, Country

4 College of XX, XX University, City, Country

§Corresponding author

Email addresses:

Author: xxx@xxx.edu.cn

Author: xxx@xxx.edu.cn

Author: xxx@xxx.edu.cn

Author: xxx@xxx.edu.cn

# Paper Title

# Abstract

**Background**

Social media plays a more and more important role in the research of health and healthcare due to the fast development of internet communication and information exchange. This paper conducts a bibliometric analysis to discover the thematic change and evolution of utilizing social media for healthcare research field.

**Methods**

With the basis of 4,361 publications from both Web of Science and PubMed during the year 2008-2017, the analysis utilizes methods including topic modelling and science mapping analysis.

**Results**

Utilizing social media for healthcare research has attracted increasing attention from scientific communities. *Journal of Medical Internet Research* is the most prolific journal with the USA dominating in the research. Overly, major research themes such as *YouTube analysis* and *Sex event* are revealed. Themes in each time period and how they evolve across time span are also detected.

**Conclusions**

This systematic mapping of the research themes and research areas helps identify research interests and how they evolve across time, as well as providing insight into future research direction.

**Keywords:** Social media, Healthcare research, Topic modelling, Science mapping, Thematic detection, Thematic evolution

# Background

In the past decade, the research field of utilizing social media for healthcare has attracted great interests from scientific communities, which can be observed from the annual increasing of research publications. Internet is becoming a significantly important role as the source of information for public health issues [1]. Health-related information is being actively searched, shared, communicated, and discussed through social media. This kind of online information exchange benefits users in aspects of immediate access to health concern information [2], emotional and psychological support [3], and health-related decision making [4].

# Methods

**Data retrieval and preprocessing**

In this study, bibliometric methodology is applied using data from Web of Science (WoS) and PubMed. WoS is the most authoritative citation database and has been widely applied for bibliometric analysis, while PubMed provides a wide coverage of medical-related publications.

The keywords of social media are developed by domain experts after an extensive literature review.

**Approach for thematic detection analysis**

Proposed by Blei et al. [27], Latent Dirichlet Allocation (LDA) model has been widely applied in topic detection in various domains. It is a Bayesian mixture model for discrete data with an assumption that topics are uncorrelated. Documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

A *document* is represented as a sequence of$ N$ words denoted by$ d=(w\_{1},…,w\_{N})$ , where a *word* is an item from a vocabulary indexed by $\{1,…,V\}$. A *corpus* is a collection of $M$ documents denoted by $D=\{d\_{1},…,d\_{M}\}$. LDA follows the following generation process. 1) The term distribution $β$is as $β\~Dirichlet(δ)$, donating the probability of a word occurring in a given topic; 2) $θ\~Dirichlet(α)$ is the proportions $θ$ of the topic distribution for a document $d$ ; 3) For each word $w\_{i}$ in the document $d$, a topic is chosen by the distribution $z\_{i}\~Multinomial(θ)$, and a word is chosen as $z\_{i}:p(w\_{i}|z\_{i},β)$. The log-likelihood for one document $d\in D$is as Equation (1), and Equation (2) is the likelihood for Gibbs sampling estimation with $k$ topics.

$l\left(α,β\right)=log\left(p\left(α,β\right)\right)=log∫\{\sum\_{z}^{}[\prod\_{i=1}^{N}p(w\_{i}|z\_{i},β)p(z\_{i}|θ)]\}p(θ|α)dθ$ (1)

# Results

**Performance bibliometric analysis**

The statistical result of publication count and citation count from the year 2008 to 2017 are presented in **Figure 1**. It is clear that the research of utilizing social media for healthcare is becoming more and more inﬂuential in scientiﬁc communities evidenced by the significant growth of publications from two databases, i.e. from 18 publications in 2008 to 1,030 publications in 2017.



**Figure 1** Publication count and citation count.

Researches in the field have been published in a wide range of nearly one thousand publication sources. Some of these publication sources are highly relevant to the ﬁeld, while others are partially related. **Table 1** lists the top 20 publication sources ranked by publication count in the research field. According to both publication percentage and H-index, *Journal of Medical Internet Research*, *PLoS One*, and *Cyberpsychology, Behavior and Social Networking* are the most inﬂuential journals in the ﬁeld.

**Table 1** Prolific publication sources

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Rank | Publication sources | C | PC | P% | H | IF (2017) |
| 1 | *Journal of Medical Internet Research* | Canada | 308 | 7.06 | 59 | 4.671 |
| 2 | *PLoS One* | The USA | 253 | 5.80 | 51 | 2.766 |
| 3 | *Cyberpsychology, Behavior and Social Networking* | The USA | 229 | 5.25 | 58 | 2.689 |
| 4 | *Studies in Health Technology and Informatics* | The USA | 92 | 2.11 | 11 | NA |
| 5 | *Health Communication* | The USA | 43 | 0.99 | 18 | 1.71 |
| 6 | *Journal of Adolescent Health* | The USA | 37 | 0.85 | 21 | 4.098 |
| 7 | *Journal of Health Communication* | The USA | 36 | 0.83 | 18 | 1.648 |
| 8 | *BMC Public Health* | England | 34 | 0.78 | 17 | 2.42 |
| 9 | *Tobacco Control* | England | 26 | 0.60 | 16 | 4.151 |
| 10 | *American Journal of Pharmaceutical Education* | The USA | 23 | 0.53 | 15 | 1.495 |
| 11 | *Journal of Biomedical Informatics* | The USA | 23 | 0.53 | 16 | 2.882 |
| 12 | *Medical Teacher* | England | 23 | 0.53 | 15 | 2.45 |
| 13 | *American Journal of Public Health* | The USA | 22 | 0.50 | 14 | 4.38 |
| 14 | *Vaccine* | England | 22 | 0.50 | 14 | 3.285 |
| 15 | *BMC Medical Informatics and Decision Making* | England | 20 | 0.46 | 9 | 2.134 |
| 16 | *Journal of the American Medical Informatics Association* | England | 20 | 0.46 | 11 | 4.27 |
| 17 | *Nurse Education Today* | Scotland | 20 | 0.46 | 10 | 2.067 |
| 18 | *Pediatrics* | The USA | 20 | 0.46 | 8 | 5.515 |
| 19 | *Journal of Cancer Education* | The USA | 19 | 0.44 | 11 | 1.547 |
| 20 | *Proceedings of the National Academy of Sciences of the United States of America* | The USA | 19 | 0.44 | 13 | 9.504 |

Note: C: countries or regions; PC: publication count; %P: percentage of publications among all the 4,361 publications; H: H-index score; IF (2017): impact factor (2017).

Among the 4,361 publications, there are 3,311 affiliations and 14,154 authors from 88 countries/regions. 18.18% of the countries/regions, 65.06% of the affiliations, or 84.41% of the authors contribute only one publication. **Table 2** lists top 20 most prolific countries/regions, affiliations, and authors.

From the country/region perspective, the USA dominates in the ﬁeld with 2,394 publications, accounting for 54.90% of the total publications. The USA also has the highest H-index as 125, indicating the high quality of its publications. Other prolific countries/regions with more than 100 publications include England, Australia, Canada, China, Germany, and Spain.

**Thematic detection analysis**

With the optimal topic number as 20 and the initialized *α* as 0.028204, LDA model using Gibbs sampling is conducted for overall thematic detection. The 20 topics with their top 15 representative terms is shown in **Table 3**, along with their possible themes, e.g., *YouTube analysis*, *Sex event*, *Web-based medical education*, *Students’ use of Facebook*, and *Twitter use*.

The top frequent keywords used for AP clustering analysis include *social media* (3,484), *human* (2,109), *internet* (1,323), *female* (886), *male* (817), *adolescent* (694), *adult* (624), *young adult* (522), *Facebook* (473), and *social networking* (463). **Figure 2** shows that the 139 keywords are classified into 28 clusters with exemplars, e.g., *self concept*, *male*, *middle aged*, *internet*, *cancer*, *Youtube*, and *weight loss*.

# Discussion

Based on the 4,361 research publications from Web of Science and PubMed during the year 2008-2017, a bibliometric analysis of utilizing social media for healthcare research is conducted, aiming at exploring the thematic detection and evolution of the research ﬁeld.

The first finding worth noting is that the research field has attracted more and more attention from scientific communities throughout the last ten years. Most prolific publication sources are *Journal of Medical Internet Research*, *PLoS One*, and *Cyberpsychology, Behavior and Social Networking*. The USA dominates in the research with a comparatively higher publication count. Its dominant role can also be observed from the top prolific authors and affiliations, most of which belong to the USA.

# Conclusions

Aiming at understanding the thematic change and evolution of utilizing social media for healthcare research during the last decade, this paper presents a quantitative analysis of publications from Web of Science and PubMed. Topic modelling analysis is used to identify major areas from an overall perspective. An approach of science mapping combining performance analysis is applied to quantify and visualize the thematic evolution. This systematic mapping of the research themes and research areas helps identify research interests and how they evolve across time, as well as providing insight into future research direction.

## Abbreviations

**USA**: United States; **WoS**: Web of Science; **MeSH**: Medical Subject Headings; **TF-IDF**: Term Frequency-Inverse Document Frequencies

# Declarations

## Authors’ contributions

Author leaded the method application, experiment conduction and the result analysis. Author participated in the data extraction and preprocessing. Author participated in the manuscript revision. Author provided theoretical guidance and the revision of this paper.

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Availability of data and materials

The datasets used and analyzed during the current study are available from the first author upon reasonable requests. / The dataset is publically available via XXX.

## Competing interests

The authors declare that they have no competing interests.

## Funding

The work is supported by grants from National Natural Science Foundation of China (No. XXX).

## References

1. Baker L, Wagner TH, Singer S, Bundorf MK. Use of the Internet and E-mail for Health Care Information: Results from a National Survey. *JAMA*. 2003; 289(18): 2400-2406.
2. Oh KM, Jun JM, Zhao XQ, Kreps GL, Lee EE. Cancer Information Seeking Behaviors of Korean American Women: A Mixed-Methods Study Using Surveys and Focus Group Interviews. *Journal of Health Communication*. 2015; 20(10): 1143-1154.
3. Lee SY, Hawkins R. Why Do Patients Seek an Alternative Channel? The Effects of Unmet Needs on Patients' Health-Related Internet Use. *Journal of Health Communication*. 2010; 15(2): 152-166.
4. Fox S, Purcell K. Chronic Disease and the Internet. Pew Internet & American Life Project Washington, DC, 2010.
5. Lavrač N, Keravnou ET, Zupan B. Intelligent Data Analysis in Medicine and Pharmacology: An Overview. In: Lavrač N, Keravnou ET and Zupan B, (eds.). Intelligent Data Analysis in Medicine and Pharmacology. Boston, MA: Springer US, 1997, pp. 1-13.
6. Sinnenberg L, DiSilvestro CL, Mancheno C, Dailey K, Tufts C, Buttenheim AM, et al. Twitter as a Potential Data Source for Cardiovascular Disease Research. *JAMA Cardiology*. 2016; 1(9): 1032-1036.
7. Ireland ME, Schwartz HA, Chen QJ, Ungar LH, Albarracn D. Future-Oriented Tweets Predict Lower County-Level HIV Prevalence in the United States. *Health Psychology*. 2015; 34(S): 1252-1260.