Medicine®

OPEN

A bibliometric analysis and visualization of medical data mining research

Yuanzhang Hu, MD[a], Zeyun Yu, MD[b], Xiaoen Cheng, MD[a,*], Yue Luo, MD[a], Chuanbiao Wen, MD[a,*]

Abstract

Background: Data mining technology used in the field of medicine has been widely studied by scholars all over the world. But there is little research on medical data mining (MDM) from the perspectives of bibliometrics and visualization, and the research topics and development trends in this field are still unclear.

Methods: This paper has applied bibliometric visualization software tools, VOSviewer 1.6.10 and CiteSpace V, to study the citation characteristics, international cooperation, author cooperation, and geographical distribution of the MDM.

Results: A total of 1575 documents are obtained, and the most frequent document type is article (1376). SHAN NH is the most productive author, with the highest number of publications of 12, and the Gillies's article (750 times citation) is the most cited paper. The most productive country and institution in MDM is the USA (559) and US FDA (35), respectively. The Journal of Biomedical Informatics, Expert Systems with Applications and Journal of Medical Systems are the most productive journals, which reflected the nature of the research, and keywords "classification (790)" and "system (576)" have the strongest strength. The hot topics in MDM are drug discovery, medical imaging, vaccine safety, and so on. The 3 frontier topics are reporting system, precision medicine, and inflammation, and would be the foci of future research.

Conclusion: The present study provides a panoramic view of data mining methods applied in medicine by visualization and bibliometrics. Analysis of authors, journals, institutions, and countries could provide reference for researchers who are fresh to the field in different ways. Researchers may also consider the emerging trends when deciding the direction of their study.

Abbreviations: AT = author, EHR = electronic health records, FQ = frequency, KW = keywords, MDM = medical data mining, R = rank, TC = times cited, Y = year.

Keywords: bibliometric analysis, CiteSpace, medical data mining, visualization, VOSviewer

1. Introduction

Data mining, also known as database knowledge discovery, is a powerful method to extract knowledge from data, which is supposed to be able to handle various data types in all formats.[1] Medical data mining (MDM) is defined as an extraction of implicit, potentially useful and novel information from medical data to improve accuracy, decrease time and cost, and construct decision support system with the aim of health promotion. Driven by the rapid development of science and technology, the hospital information construction is becoming more and more perfect, and the medical data storage volume is getting larger. The research on MDM is growing fast, and the application of data mining in medicine is most used by data mining developers and academic researchers.[2] The study of MDM is started from last decades and now it is in a teenage period. How to effectively use the data analysis method to mine the high-value information contained in the massive medical data, and then realize the knowledge discovery, and how to serve the clinical practice and scientific decision-making in hospitals are the great concerns in the field of MDM.

Bibliometrics is the cross-disciplinary science of quantitative analysis of all knowledge carried by mathematical and statistical methods.[3] In light of bibliometric methods, the latest advances, leading topics, current gaps in a certain field of research discipline can be drawn vividly as well as geographically, and it is becoming an import research method for assessing national and international research productivity, international cooperation, citation analysis, research trends, and development of specific fields. At present, many bibliometric analysis methods and tools like CiteSpace and VOSviewer have been developed to help researchers in different field construct knowledge maps, evaluate the collective state of thought about a subject, and identify hotspots in a research field.

Editor: Pricivel Carrera.

YH and ZY contributed equally to this work.

This work was supported by the National Natural Science Foundation of China (No. 81904324) and the National key research and development program of China (2018YFC1707606).

This study did not involve any human or animal experiment. Ethical approval was not required for secondary analysis of anonymous data in this study.

The authors have no conflicts of interest to disclose.

The datasets generated during and/or analyzed during the current study are publicly available.

[a] School of Medical Information Engineering, Chengdu University of Traditional Chinese Medicine, Chengdu, Sichuan, [b] College of Acupuncture and TuiNa, Chengdu University of Traditional Chinese Medicine, Chengdu, China.

[*] Correspondence: Chuanbiao Wen, Xiaoen Cheng, School of Medical Information Engineering, Chengdu University of Traditional Chinese Medicine, No. 37 Shi'er Qiao Road, Jin'niu District, Chengdu 610000, Sichuan Province, China (e-mails: cdutcm_mdm@yeah.net, 1157778219@qq.com).

How to cite this article: Hu Y, Yu Z, Cheng X, Luo Y, Wen C. A bibliometric analysis and visualization of medical data mining research. Medicine 2020;99:22 (e20338).

Received: 23 October 2019 / Received in final form: 24 March 2020 / Accepted: 21 April 2020

http://dx.doi.org/10.1097/MD.0000000000020338

In this paper, we use free available software VOSviewer 1.6.10 to carry out the visualized map, and CiteSpace V to generate diagrams and calculate the betweenness centrality score. The literature on the application of data mining technology in the medical field from the Web of Science database has analyzed to provide a macroscopic overview on the main characteristics of MDM publication. And clear informative pictures presented in this paper demonstrate the research achievements in the domain of the MDM, which could help researchers and practitioners identify the underlying impacts from authors, journals, countries, institution, references, and research topics. Although this work is not structured as an exhaustive analysis of MDM-related literature, it does illustrate the utility of bibliometric techniques for exploring hidden knowledge spaces.[4]

2. Data and methods

2.1. Data collection

The literature data involved in this study are retrieved form the core collection of Web of Science (WOS).[5] The WOS is one of the most comprehensive bibliographic sources available, and provides users an online access port to a number of resources, including massive citation databases, but not all journals or articles are indexed.[6]

For the purpose of this paper, we are interested in exploring the knowledge domain associated with “medical data mining,” and use “medical data mining” as the search term in the WOS database, the literature type is defined as “all types.” For assuring the quality of data, a manual review on search results is adopted in Endnote X9 to remove the unrelated papers, and the CiteSpace function, Alias[7] is used to identify and correct all duplicate values in the databases. Finally, 1575 documents are saved as “Plain Text” with “Full Record and Cited References.” And the timespan is from January 1, 2011 to August 28, 2019, which including information on title, author, keywords, abstract, journal, and year. These records are then exported to CiteSpace and VOSviewer for subsequent analysis, and 5 document types are found (Table 1).

2.2. Analysis methods

Bibliometric analysis offers additional data statistics including author, affiliation, and keywords. In this context, the items of analysis used in the study are detailed like co-authorship, journal analysis, citing, keyword analysis, geo/location collaboration, co-occurrence, and betweenness centrality score.

Betweenness centrality is a way of detecting the amount of influence a node has over the flow of information in a graph, and often used to find nodes that serve as a bridge from one part of a graph to another. For users, betweenness centrality can make it easier to identify pivotal points, which are highlighted in the display with a purple ring in order to stand out in a visualized map.[8] At the document level, the importance of each document in a co-citing map can be partially evaluated by the indicator betweenness centrality.[9]

And we use Price’s law (1) to measure core authors. Price’s law defines that, 50% of the work is done by the square root of the total number of people who participate in the work.[10]

$$MP = 0.749\sqrt{Np\max} \qquad (1)$$

Table 1

Types of retrieved documents.

Type of document	Frequency	Proportion (%)
Article	1376	87.4
Review	122	7.8
Proceeding paper	35	2.2
Editorial material	24	1.5
Meeting abstract	18	1.1
Total	1575	100

3. Findings

3.1. Publication growth trend

The quantity of the publication is an important index that reveals the development trends of scientific research. Figure 1 depicts a chronological view on volume of articles published and cited on MDM.

3.2. Productivity and connectivity

The core authors in the academic community are the important internal strength to promote the development of the discipline,[11] and researchers can identify potential collaborators and understand how their own research fits in MDM research.

There are 6258 authors in the field of MDM, and Table 2 shows the 20 most productive scholars in the MDM research worldwide. These authors have successfully established broad cooperation with researchers in other countries. Among them, the application of data mining methods in the field of medicine can be divided into a few core research teams, and 1 important team is LEWIS P, GANO M, MORO PL, SHIMABUKURO TT, and STEWART B, which focuses on vaccine adverse event reporting, outlier detection and disease risk-factors.[12,13] As for individual researcher, the most productive author is SHAN NH, who based at the Stanford Center for Biomedical Informatics Research (BMIR) and studies pharmacovigilance and text mining,[14] then followed by LI L, who based at the Icahn School of Medicine at Mount Sinai and studies systems biology and machine learning.[15] HU YH, based at the National Chung Cheng University, Taiwan, and studies large scale medical data preprocessing approach.[16] WANG S, based at the University of Queensland, Australia, and studies medical database management. For example, 1 of his studies proposed a framework to effectively assigns the disease labels, medical chart, and note data of a patient are used to extract distinctive features.[17] REINER BI is based at the Department of Radiology, Veterans Affairs Maryland Healthcare System, studies medical imaging data analysis and quantifying analysis of uncertainty in medical reporting.[18–20] LIU HF is based at the Division of Biomedical Statistics and Informatics, Mayo Clinic, studies mining drug–drug interaction adverse events and reporting.[21]

According to Price’s law, there are 163 core authors who have published at least 3 papers, accounting for 1.95% of total scholars (less than 50%), which implies that the application of data mining technology in the field of medicine is in the stage of rapid development.

3.3. The distribution of institutes on MDM study

The analysis of MDM research institutions can clarify the core institutions. There are total 2222 organizations and 34 of them produce more than 9 papers. US FDA possesses the greatest number

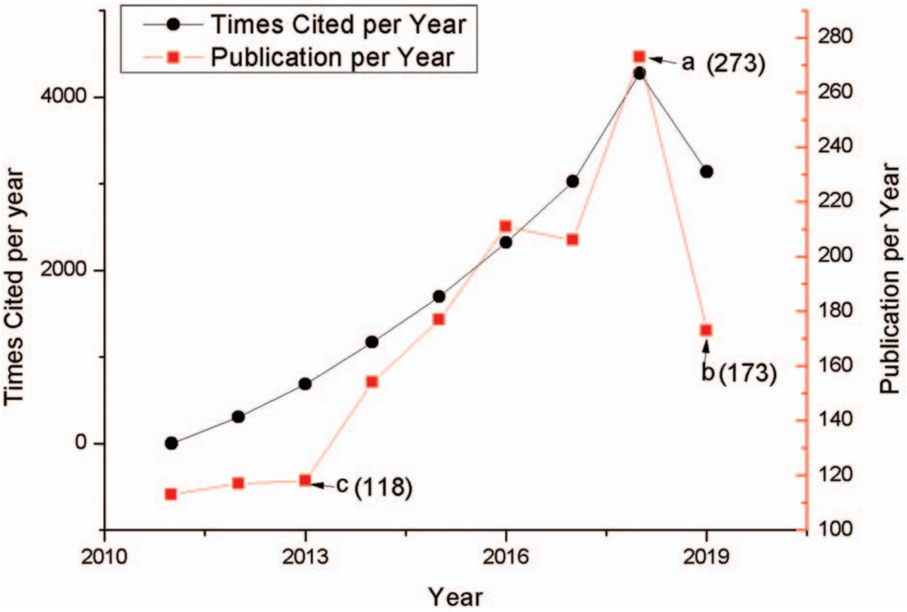


Figure 1. Annual publications and citations of MDM based on WoS. (a) The publication in 2018; (b) the publication between January 1 and August 28, 2019; (c) the publication in 2013.

of publications with a total 35 papers, accounting for 3.6% of all publications in this field. At the second position is the Leland Stanford Junior University with 29 publications, then followed by Mayo Clinic, Columbia University in the City of New York, and Vanderbilt University. The top 10 institutes are listed in Figure 2. Among them, 7 institutes are from American and 3 from China. In addition, the 10 institutes totally cover 204 published papers.

3.4. Countries/regions cooperation analyses

Based on the bibliographic data collected from WOS, the countries co-authorship network visualization map is created by VOSviewer (Fig. 3); the minimum document threshold of a country is 5 and there are 47 countries out of 91 listed are visualization items. Specifically, the USA is identified as the country with the largest amount of studies (559, 1/3th of the total publications), followed by China (238), India (102), Taiwan (76), Australia (72), England (65), and Italy (63), because the leading groups of research and practitioners in MDM are located in the above countries or areas.

Table 2

Most productive scholars in MDM research worldwide.

Rank	Name	Frequency
1	SHAH NH	12
2	LI L	10
3	LEWIS P	9
4	CANO M	8
5	HU YH	8
6	MORO PL	7
7	WANG S	6
8	REINER BI	6
9	SHIMABUKURO TT	6
10	LIU HF	6
11	BALL R	5
12	LI X	4
13	REINER B	4
14	CHUNG K	4
15	TSAI CF	4
16	LIU BY	4
17	STEWART B	4
18	ZHANG Y	4
19	BOTSIS T	4
20	CHEN Y	4

3.5. Journals publishing on MDM

Through the analysis of journals, we can have a better understanding about the structure of core journals in the field of MDM. In total, there are 1575 publications in 650 different journals, but 18. 77% (n=122) journals have published more than 3 papers. A list of the top 10 most productive journals on MDM research is provided in Table 3, and the top 3 most productive journals are *Journal of Biomedical Informatics*, *Expert Systems with Applications*, and *Journal of Medical Systems*.

As shown in Figure 4, the size of the nodes represents the publication amount of a journal, and the color of the nodes demonstrates the subdomains of the MDM research. We use VOSviewer to plot the journal co-citation network and generally the smaller the distance between 2 nodes is, the higher of the citation frequency is. In Figure 4, the largest set of the connected items consists of 119 items and some of the 122 items in the network are not connected to each other. It is manifest that all these journals are divided into 5 clusters; the highly cited journals in the blue cluster are representing biomedical information journals, which starts with *JOURNAL of BIOMEDICAL INFORMATICS*. The red cluster represents Management integration journals, which contains *EXPERT SYSTEMS with APPLICATIONS*. The green cluster contains journals on imaging and genetics, starts with *JOURNAL of DIGITAL IMAGING and FRONTIERS IN GENETICS*. The yellow

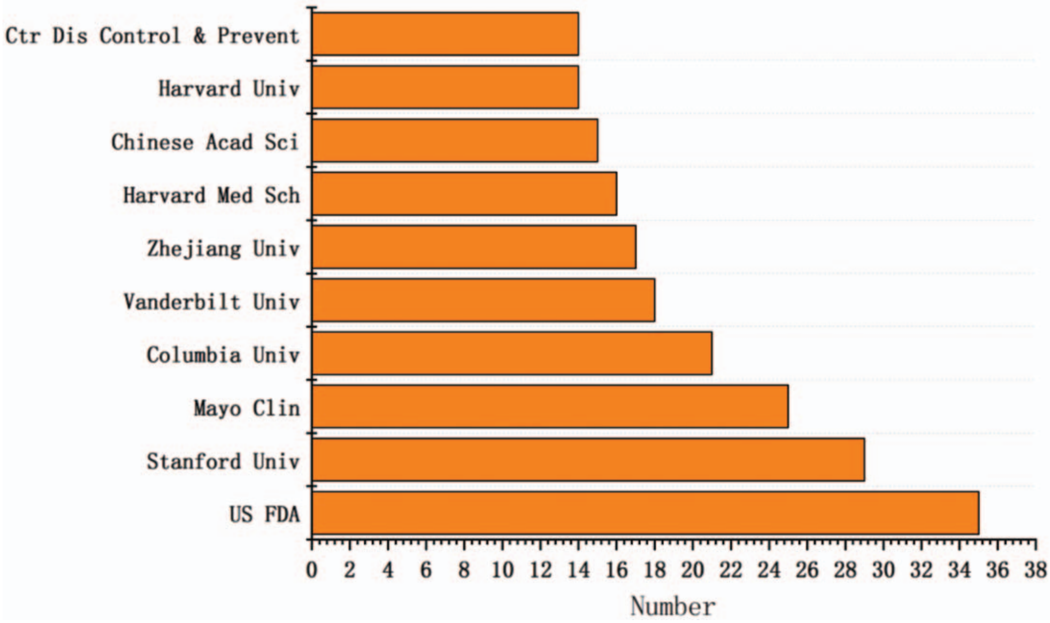


Figure 2. The top 10 institutes with MDM-related publications.

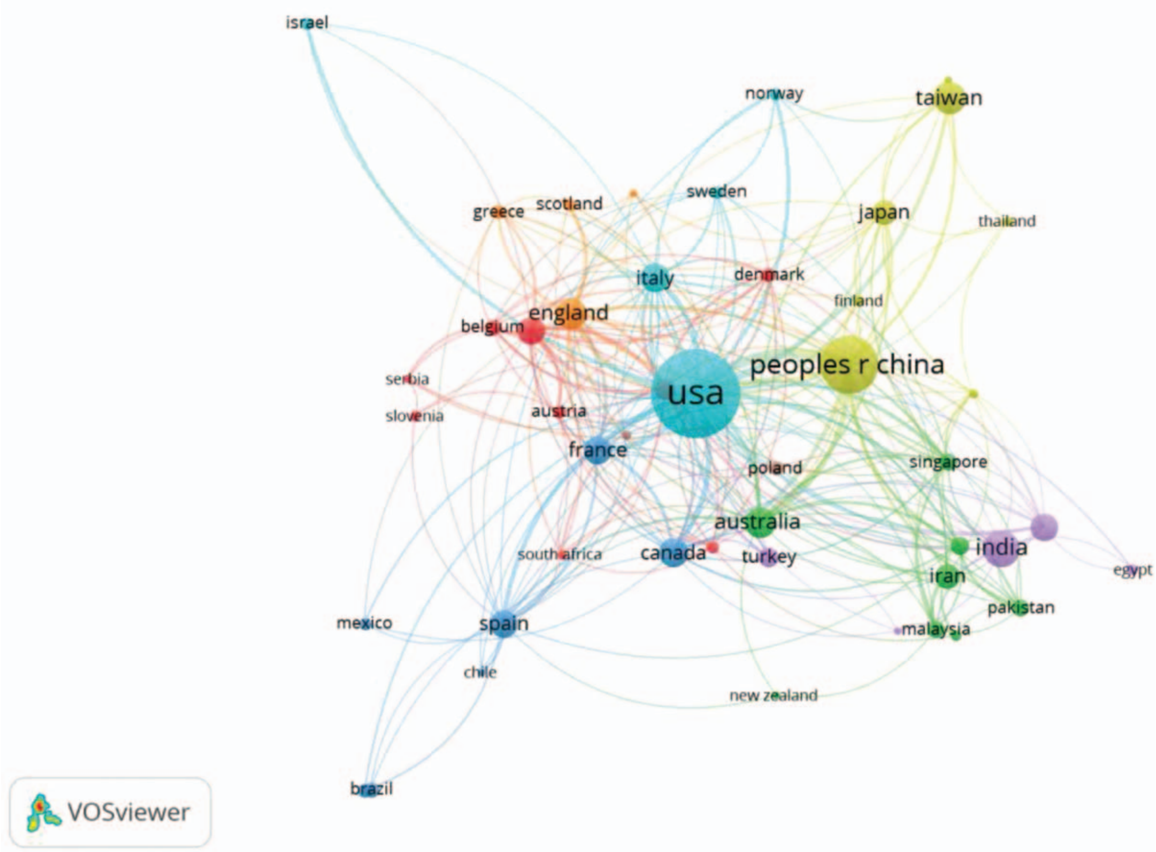


Figure 3. Citation visualization network map of countries/regions based on citation-weights.

Table 3

Top 10 most productive sources.

R	ST	PC	C	TLS
Top 1	JOURNAL OF BIOMEDICAL INFORMATICS	44	748	3063
Top 2	EXPERT SYSTEMS WITH APPLICATIONS	43	515	2712
Top 3	JOURNAL OF MEDICAL SYSTEMS	36	268	1551
Top 4	PLOS ONE	34	422	1504
Top 5	JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION	30	730	2311
Top 6	ARTIFICIAL INTELLIGENCE IN MEDICINE	28	276	1745
Top 7	JOURNAL OF MEDICAL INTERNET RESEARCH	28	797	751
Top 8	BMC MEDICAL INFORMATICS AND DECISION MAKING	26	125	1377
Top 9	IEEE ACCESS	22	76	1125
Top 10	BCM BIOINFORMATICS	21	171	1011

C=citations, PC=publication count, R=rank, ST=source titles, TLS= total link strength.

cluster represents drug and vaccine journals. The last purple cluster represents software engineering journals, which contains *IEEE TRANSACTIONS ON VISUALIZATION and COMPUTER GRAPHICS*. For MDM researchers, IEEE is the world's largest technical organization dedicated to advancing computer technology for the benefit of medicine, and it is important to follow IEEE publications and conferences to keep abreast of their latest research status.

3.6. Keyword analysis with co-occurrence

Keywords are nouns or phrases that reflect the core content of a publication.[22] The co-occurrence network of keywords reflects the research hotpots. In this part, the content is studied by analyzing the distribution of keyword. The bibliometric data shows that there are 6992 keywords involved in this research. To illustrate the research hotspots in MDM, keywords co-occurrence threshold is set as 5 and 413 items are brought into visualization (Fig. 5), which is constructed by the VOSviewer. In the network, visualization items are represented by their label and a circle. The size of the label and the circle of an item are determined by the weight of the item. The higher the weight of an item is, the larger the label and the circle of the item are.[23] The color of an item is determined by the cluster to which the item belongs and lines between items represent links in Table 4. In addition to keyword "data mining," the keywords "classification (790)" and "system (576)" have the strongest strength. The co-keyword network in Figure 5 clearly illustrates 6 distinct clusters and each of them represents a subfield of MDM. As shown in the red cluster (cluster 1, left, 102 items) which contains keywords such as classification, diagnosis, feature selection, artificial neural network, support vector machines, etc, these studies use algorithms to find patterns that make early detection, prediction of the disease, and proper treatment easier.[24–26] The green cluster (cluster 2, upper right, 75items) is associated with text mining, electronic health records, pharmacovigilance, adverse drug reactions, signal-detection, natural language processing, etc, focusing on the main domain of "medical text and language

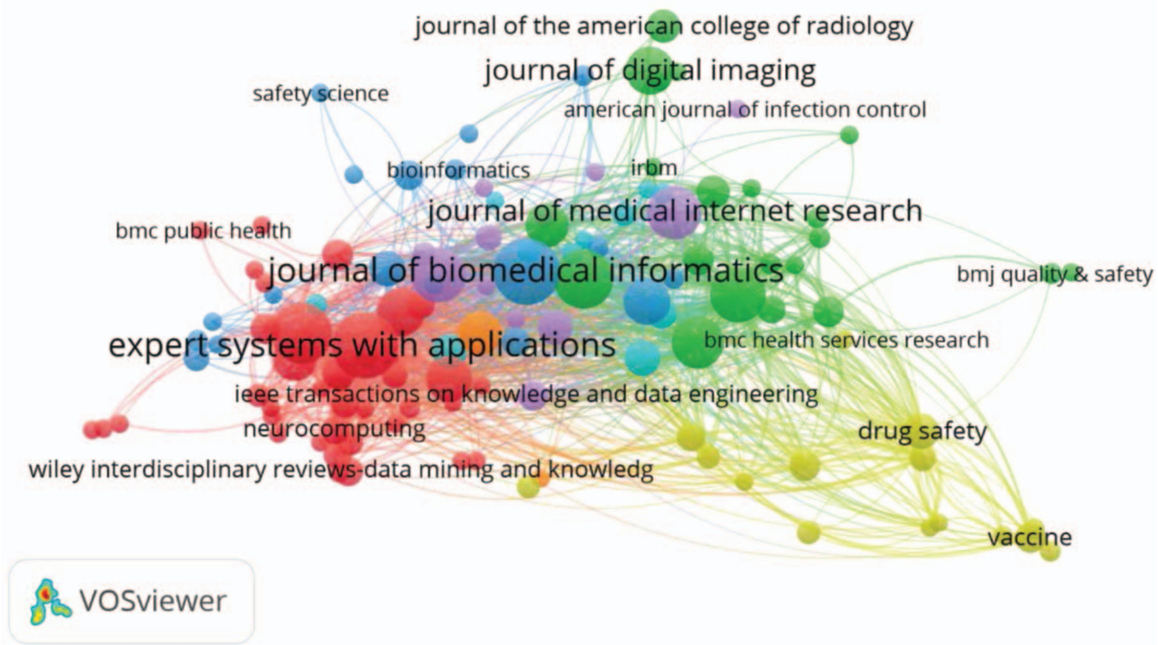


Figure 4. The visualization map of journal publications.

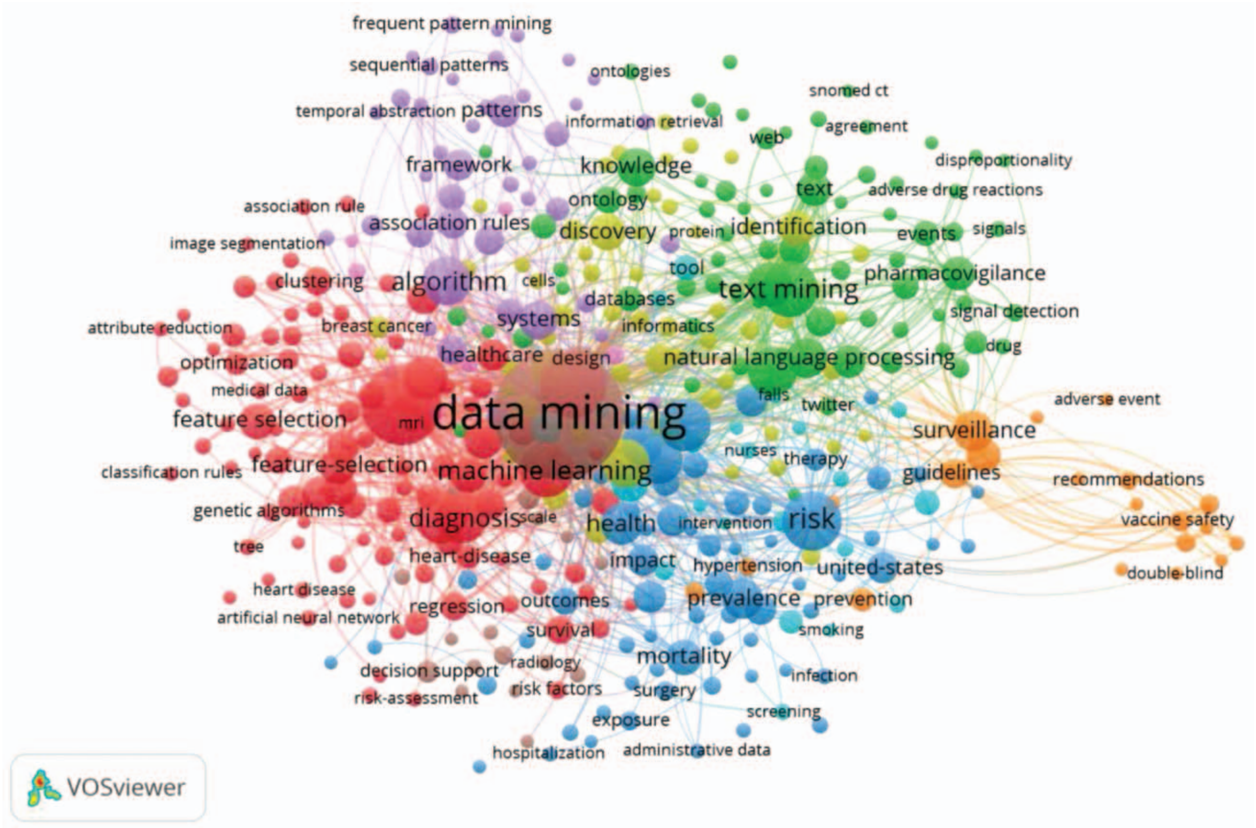


Figure 5. Co-keyword network visualization based on occurrences.

mining." Studies claim that text mining can be applied to extract useful adverse drug event-related information, form multiple textual sources like electronic health records and improve adverse drug event (ADE) detection and assessment.[27–29] Next, in the blue cluster (cluster 3, bottom right, 67 items), keywords like risk, prevalence, hospitalization, mortality, adverse events, disease, infection, etc, are associated with disease topics.[30–33] In this cluster, the machine learning approaches are mainly applied to disease risk model.[34] In the yellow cluster (cluster 4, 58 items), keywords like discovery, identification, genes, protein, informatics, breast cancer, patient, etc, concentrate on the aspect of "drug discovery."[35,36] In the orange cluster (cluster 5, lower right, 20items) comprised keywords like surveillance, vaccine safety, vaccine adverse event reporting system (VAERS), recommendations, etc, are more concerned with "medical safety."[37,38] The last purple cluster (cluster 6, upper left, 44items) gathers keywords like framework, patterns, frequent pattern mining, methodology, decision support system, clinical pathways, etc, mainly concerning "medical mining system."

Table 4

The top 20 keywords of the MDM publications.

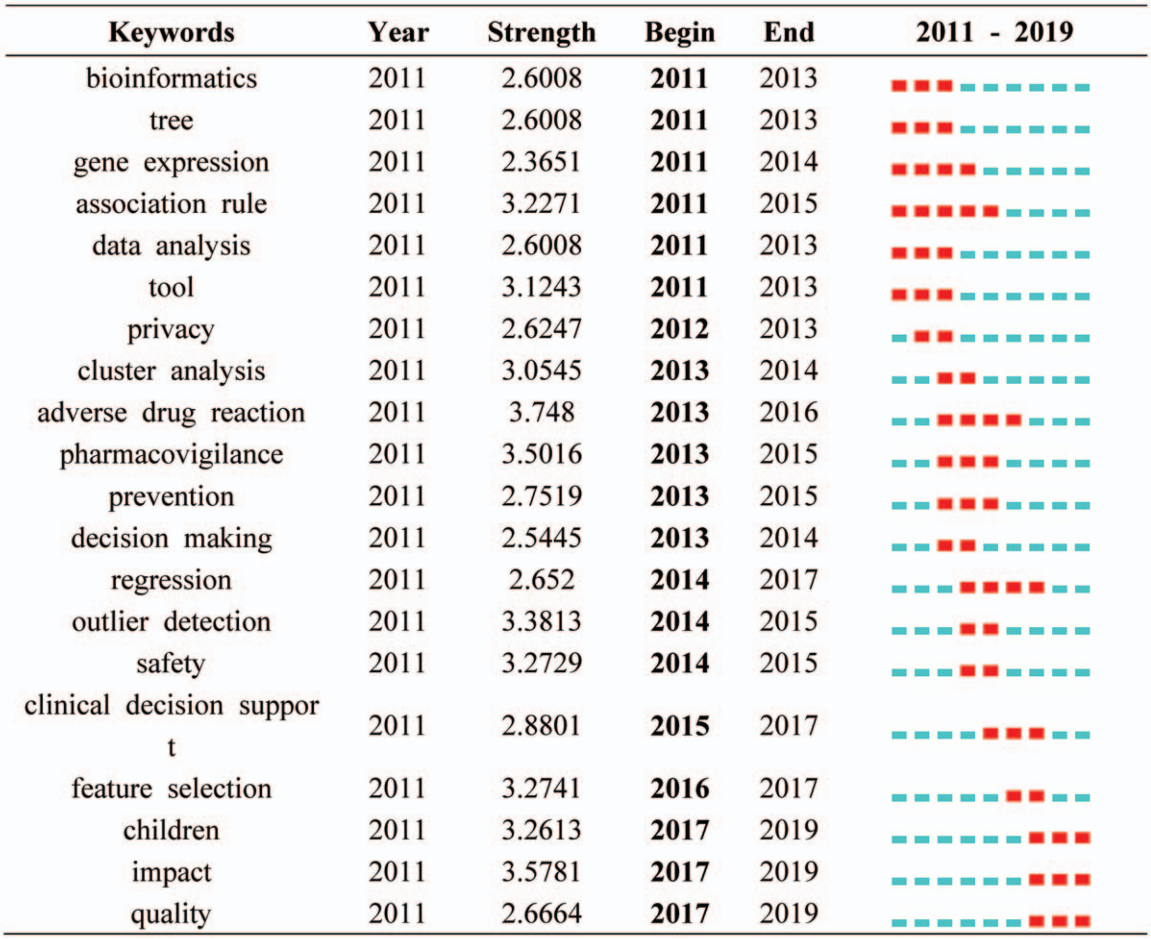
R	KW	FQ	TLS
1	classification	168	790
2	system	116	576
3	risk	94	474
4	text mining	82	402
5	Electronic health records	62	382
6	care	72	379
7	diagnosis	73	363
8	prediction	66	352
9	model	53	249
10	Natural language processing	44	238
11	surveillance	39	209
12	algorithms	39	195
13	pharmacovigilance	32	191
14	cancer	36	158
15	models	34	154
16	risk-factors	28	125
17	association	25	116
18	Logistic-regression	16	108
19	Neural-networks	19	99
20	Support vector machines	20	87

We use CiteSpace to generate cluster labels, usually the LLR (log-likelihood tests) algorithm gives the best result in terms of the uniqueness and coverage. There we have got 11 clusters based on LLR: #0 extraction system, #1 deep learning, #2 knowledge-based systems, #3 decision tree, #4 drug discovery, #5 medical imaging, #6 gene ontology, #7 instance selection, #8 hepatitis, #9 vaccine safety, and #10 clinical decision support.

Table 5 shows which keywords have the strongest bursts and which period of time the strongest bursts takes place (settings: years per slice: 1, node types: keyword, top N per slice: 50). The red part represents the period when the citation burst has happened. "Bioinformatics" is the first keyword proposed in MDM research. "Association rule" has the longest period of burst from 2011 till 2015. The keyword "clinical decision support," "feature selection," "children," "impact," and

Table 5

Top 20 keywords with the strongest citation bursts.



Keywords	Year	Strength	Begin	End	2011 - 2019
bioinformatics	2011	2.6008	**2011**	2013	🟥🟥🟥🟦🟦🟦🟦🟦🟦
tree	2011	2.6008	**2011**	2013	🟥🟥🟥🟦🟦🟦🟦🟦🟦
gene expression	2011	2.3651	**2011**	2014	🟥🟥🟥🟥🟦🟦🟦🟦🟦
association rule	2011	3.2271	**2011**	2015	🟥🟥🟥🟥🟥🟦🟦🟦🟦
data analysis	2011	2.6008	**2011**	2013	🟥🟥🟥🟦🟦🟦🟦🟦🟦
tool	2011	3.1243	**2011**	2013	🟥🟥🟥🟦🟦🟦🟦🟦🟦
privacy	2011	2.6247	**2012**	2013	🟦🟥🟥🟦🟦🟦🟦🟦🟦
cluster analysis	2011	3.0545	**2013**	2014	🟦🟦🟥🟥🟦🟦🟦🟦🟦
adverse drug reaction	2011	3.748	**2013**	2016	🟦🟦🟥🟥🟥🟥🟦🟦🟦
pharmacovigilance	2011	3.5016	**2013**	2015	🟦🟦🟥🟥🟥🟦🟦🟦🟦
prevention	2011	2.7519	**2013**	2015	🟦🟦🟥🟥🟥🟦🟦🟦🟦
decision making	2011	2.5445	**2013**	2014	🟦🟦🟥🟥🟦🟦🟦🟦🟦
regression	2011	2.652	**2014**	2017	🟦🟦🟦🟥🟥🟥🟥🟦🟦
outlier detection	2011	3.3813	**2014**	2015	🟦🟦🟦🟥🟥🟦🟦🟦🟦
safety	2011	3.2729	**2014**	2015	🟦🟦🟦🟥🟥🟦🟦🟦🟦
clinical decision suppor t	2011	2.8801	**2015**	2017	🟦🟦🟦🟦🟥🟥🟥🟦🟦
feature selection	2011	3.2741	**2016**	2017	🟦🟦🟦🟦🟦🟥🟥🟦🟦
children	2011	3.2613	**2017**	2019	🟦🟦🟦🟦🟦🟦🟥🟥🟥
impact	2011	3.5781	**2017**	2019	🟦🟦🟦🟦🟦🟦🟥🟥🟥
quality	2011	2.6664	**2017**	2019	🟦🟦🟦🟦🟦🟦🟥🟥🟥

"quality" are the nearest hot-spot keyword in the burst. Keyword burst also shows that the theme of the study changes quickly as time goes on, and many branches of MDM research are synchronously thriving (Table 6), like "reporting system," "precision medicine," and "inflammation."

3.7. Most cited papers

Citation analysis is one of the parameters for assessing the quality of research. Table 7 lists the total citations, titles, authors, and publication years of the top 20 most cited papers of MDM. Among these 20 papers, 12 papers receive more than 120 citations and 4 papers receive more than 180 citations. Lambin describes the process of radiomics, and provides a guidance for investigating the standardized evaluation of both the scientific integrity and the clinical relevance of the numerous publishes.[39] Gillies put forward the opinion that converting radiomics image to higher-dimensional data and mining of these data could improve clinical decision support.[40] Parmar proposes a semiautomatic region that can grow volumetric segmentation algorithm, which investigates in terms of its robustness for quantitative imaging feature extraction and uses 14 feature selection methods and 12 classification methods to examine in terms of their performance and stability for predicting overall survival. The variability analysis indicates that the choice of classification method is the most dominant source of performance variation.[41,42] Besides that, Jensen proposes that integrating electronic health records (EHR) data with genetic data will give a finer understanding of genotype–phenotype relationship.[29] In sum, these articles mentioned above showed the part application about data mining methods in MDM from different aspects.

3.8. Co-cited papers in the field of MDM

The co-citation analysis assesses if articles are cited together and their corresponding frequencies and scales. If 2 articles are both cited as references in another article, then those 2 papers have a co-citation relationship. In this essay, VOSviewer is used to build a co-citation paper network for MDM research, the network of articles represents the intellectual basis of the field (Fig. 6). The MDM papers identified here cite collectively 55,175 unique publications, among them, 139 papers which have been co-cited more than 10 times are analyzed. As the visualization illustrated, each cluster has a color that indicates the group to which the cluster is assigned. We can see that all these papers are divided into 4 clusters. The red cluster, in terms of citations received, is led by L Breiman's article (*Breiman, 2001, Machine Learning: Random Forests*),[43] followed by C Cortes (*1995, Support-*

Vector Networks),[44] Esfandiari (*2014, Knowledge discovery in medicine: Current issue and future trend*),[1] LeCun Y (*2015, Deep learning*),[45] and Alex Krizhevsky (*2017, ImageNet Classification with Deep Convolutional Neural Networks*).[46] These studies propose some deep learning algorithms like deep convolutional neural network, random forests, and support-vector network. The blue cluster has Rojas E (*2016, Process mining in healthcare: A literature review*)[47] and Rakesh Agrawal (*1993, Mining association rules between sets of items in large databases; 1994, Fast Algorithms for Mining Association Rules and 1995, Mining sequential patterns*),[48–50] which studies association rules and processes mining for healthcare processes. The yellow and green are tightly connected to each other, indicating shared relevant literatures compared to the rest of the network. The yellow cluster contains Savova GK (*2010, Mayo clinical Text Analysis and Knowledge Extraction System*),[51] Nat Genet (*2000, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium*),[52] and Olivier Bodenreider (*The Unified Medical Language System (UMLS): integrating biomedical terminology*)[53] studying the medical mining system. George Hripcsak (*2013, Next-generation phenotyping of electronic health records*) and Jensen Peter B (*2012, Mining electronic health records: towards better research applications and clinical care*) apply text mining in electronic health records for providing assistance to the physician.[29,54] The green cluster has DuMouchel W (*1999, Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system: Discussion*),[55] Aronson AR (*2010. An overview of MetaMap: historical perspective and recent advances*),[56] P LePendu (*2013, Pharmacovigilance Using Clinical Notes*),[28] Bate A (*2009, Quantitative signal detection using spontaneous ADR reporting*),[57] and David C. Classen (*2011, 'Global Trigger Tool' Shows That Adverse Events In Hospitals May Be Ten Times Greater Than Previously Measured*) studying bayes model, adverse medical events, and biomedical information.[58]

Table 6

Analysis of the keywords and centrality of MDM.

Year	Keyword	*C	†C
2011	Decision support system	16	0.10
	Electronic health record	69	0.07
	Artificial neural network	31	0.06
	Text mining	17	0.05
	Computer aided diagnosis	6	0.05
	Support vector machine	37	0.03
2012	association	26	0.08
	Particle swarm optimization	5	0.04
	Adverse event	17	0.03
	Image mining	3	0.02
	Risk-factor	36	0.02
2013	framework	18	0.7
	challenge	13	0.04
	Health care	23	0.03
	Adverse drug reaction‘	14	0.02
	Time series	2	0.01
2014	Big data	56	0.02
	Primary care	7	0.01
	epidemiology	18	0.01
2015	Clinical decision support	9	0.04
	Reporting system	3	0.01
	protein	3	0.01
2016	validation	17	0.00
	Clinical trial	4	0.00
2017	Precision medicine	4	0.01
	Emergency department	4	0.01
	Frequent pattern mining	4	0.00
2018	depression	6	0.00
	Deep learning	11	0.00
2019	stroke	3	0.00
	Random forest	4	0.00
	Inflammation	3	0.00

* C=count.

† C=centrality.

Table 7

The 20 most cited documents in MDM according to WOS.

R	Title	Y	TC	AT
1	Radiomics: images are more than pictures, they are data	2016	750	Gillies, RJ
2	Mining electronic health records: towards better research applications and clinical care	2012	468	Jensen, PB
3	The long non-coding RNAs, a new cancer diagnostic and therapeutic gold mine	2013	271	Qi, P
4	Machine learning methods for quantitative radiomic biomarkers	2015	186	Parmar, C
5	Radiomics: the bridge between medical imaging and personalized medicine	2017	169	Lambin, P
6	Robust radiomics feature quantification using semiautomatic volumetric segmentation	2014	163	Parmar, C
7	Business process analysis in healthcare environments: a methodology based on process mining	2012	149	Rebuge, A
8	GEMINI: integrative exploration of genetic variation and genome annotations	2013	148	Paila, U
9	Health effects and toxicity mechanisms of rare earth elements–knowledge gaps and research prospects	2015	142	Pagano, G
10	Healthcare information systems: data mining methods in the creation of a clinical recommender system	2011	129	Duan, L
11	Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels	2011	120	Tatonetti, N
12	Data mining technique for automated diagnosis of glaucoma using higher order spectra and wavelet energy features	2012	119	Mookiah, MRK
13	An ontology-based measure to compute semantic similarity in biomedicine	2011	119	Batet, M
14	Constrictive bronchiolitis in soldiers returning from Iraq and Afghanistan	2011	118	King, MS
15	Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features	2015	117	Nikfarjam, A
16	A review of approaches to identifying patient phenotype cohorts using electronic health records	2014	114	Shivade, C
17	Getting more out of biomedical documents with GATE's full lifecycle open source text analytics	2013	108	Cunningham, H
18	Evolution of reporting *P* values in the biomedical literature, 1990–2015	2016	104	Chavalarias, D
19	Performance of pharmacovigilance signal-detection algorithms for the FDA adverse event reporting system	2013	102	Harpaz, R
20	A survey on deep learning for big data	2018	99	Zhang, QC

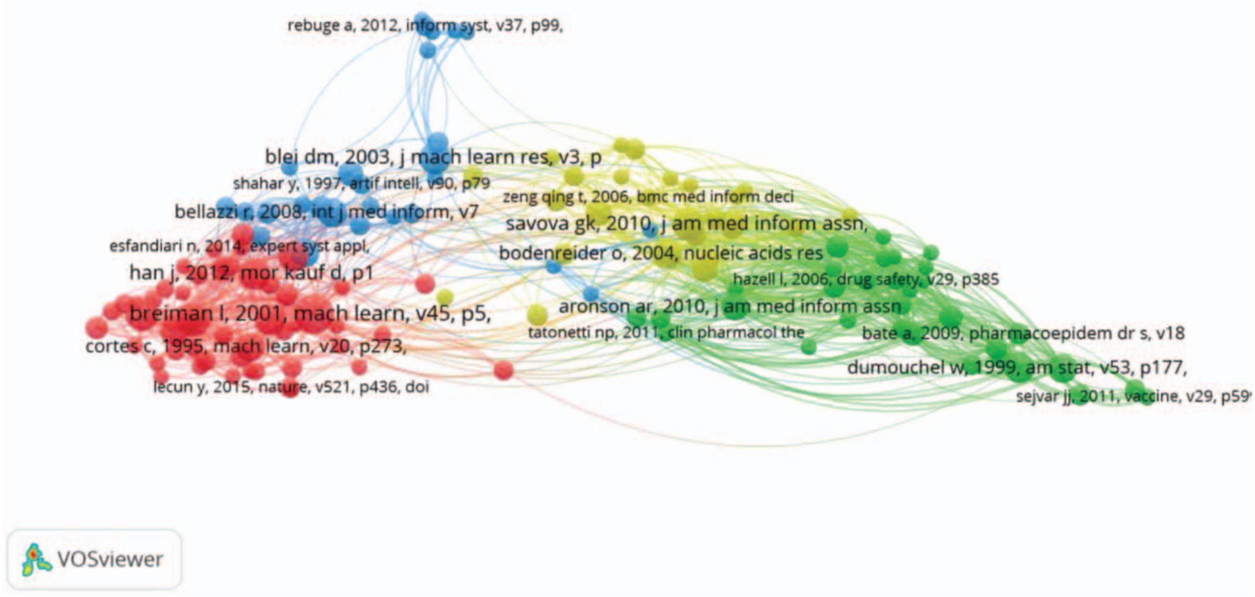

rebuge a, 2012, inform syst, v37, p99,
blei dm, 2003, j mach learn res, v3, p
shahar y, 1997, artif intell, v90, p79
bellazzi r, 2008, int j med inform, v7
zeng qing t, 2006, bmc med inform deci
savova gk, 2010, j am med inform assn,
esfandiari n, 2014, expert syst appl,
bodenreider o, 2004, nucleic acids res
han j, 2012, mor kauf d, p1
hazell l, 2006, drug safety, v29, p385
breiman l, 2001, mach learn, v45, p5,
aronson ar, 2010, j am med inform assn
tatonetti np, 2011, clin pharmacol the
bate a, 2009, pharmacoepidem dr s, v18
cortes c, 1995, mach learn, v20, p273,
dumouchel w, 1999, am stat, v53, p177,
lecun y, 2015, nature, v521, p436, doi
sejvar jj, 2011, vaccine, v29, p59
VOSviewer

Figure 6. Cited references network of MDM.

A citation burst has 2 attributes: the intensity of the burst and how long the burst status lasts. Table 8 shows references with the strongest citation bursts across the entire dataset during the period of 2011 to 2019. The first burst article is Marylyn D. Ritchie (2012–2013, 2. 6377), who studies EMR-linked DNA bio-repository to detect known genotype–phenotype associations; the result demonstrates that phenotypes representing clinical diagnoses can be extracted from EMR systems.[59] The strongest strength is Mohammed Saeed (2016–2017, 7. 2014), who develops an intensive care unit research database and applies automated techniques to aggregate high-resolution diagnostic and therapeutic data from a large, diverse population of adult intensive care unit patients.[60] Followed by SEJVAR JJ and MARTIN D (both 2015–2016, 6. 4047), SEJVAR JJ provides the case definitions and guidelines for the standardized collection and assessment of information about Guillain Barré syndrome (GBS) and Fisher syndrome (FS).[61] MARTIN D focuses on using data mining methods to find a novel safety signal for vaccine safety monitoring.[62] The nearest burst reference is HRIPCSAK G (2016–2019, 3. 2683) and Wei CH (2017–2019, 3. 8182). HRIPCSAK G focuses on next-generation phenotyping of EHR for the complex, inaccurate, and frequently missing of the EHR data; and Wei CH describes PubTator, an automated text mining tool for curating knowledge from biomedical literature into structured databases.[54,63]

4. Discussions and conclusions

4.1. Findings and discussions

The knowledge map of MDM was visualized by information visualization software Citespace and VOSviewer based on the literature retrieved from WOS for 2011 to 2019 years. Through the author analysis, journal analysis, country analysis, institution analysis, co-cited references network analysis, co-occurrence keywords network analysis and burst keywords analysis, the research achievements, and potential impacts of MDM have been identified in a multipurpose and comprehensive way. Some interesting results concerning the MDM-related publications can be summarized as follows:

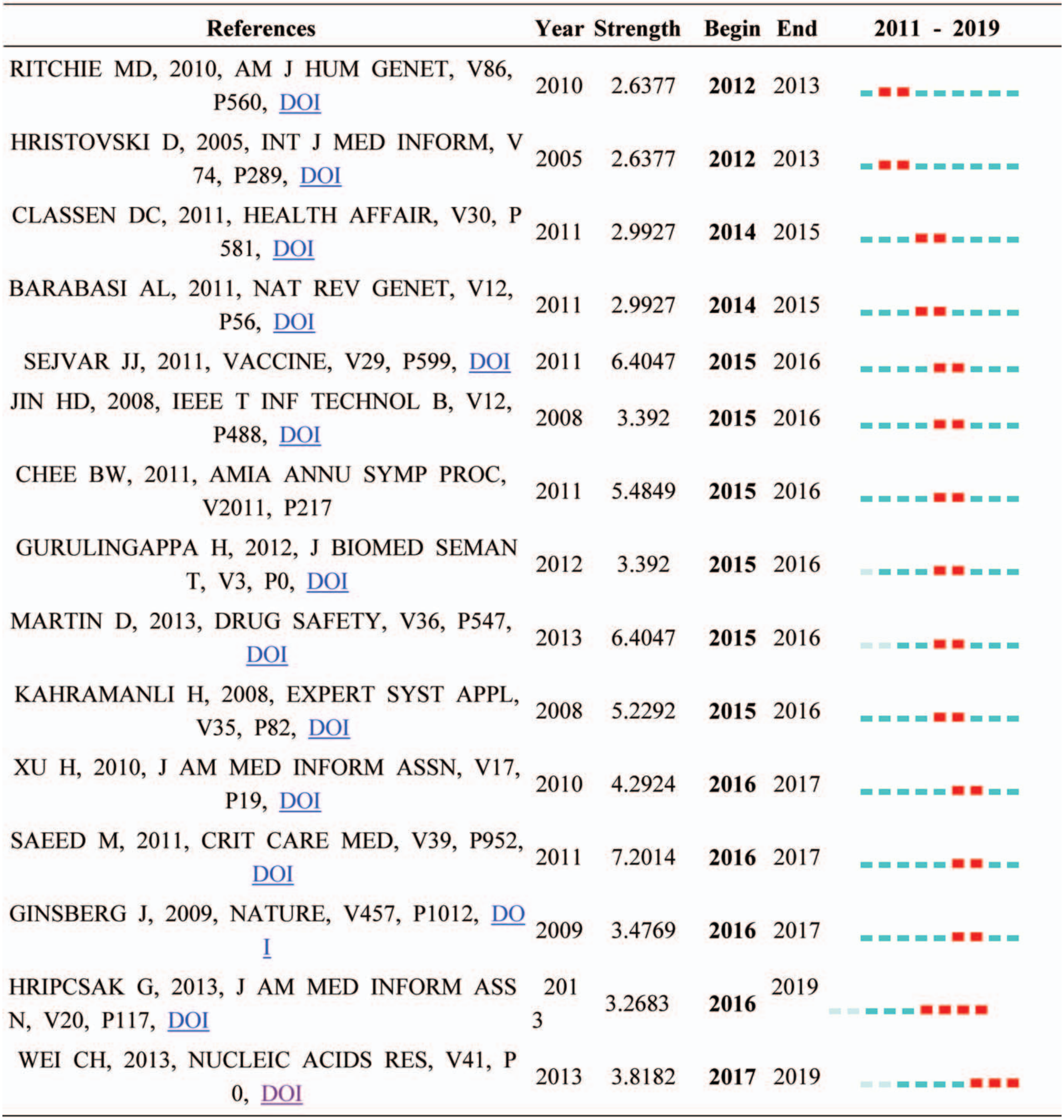
First, from the research analyze of publications, the annual number of published and cited papers have gradually increased during the last decades based on the data from WOS. Most notably the publication output on MDM cities has increased exponentially since 2013, which represents the kick-off of MDM due to the success of application of data mining technology in other fields. There is a growing interest in the researches related to the MDM, which corresponds to the urgent need for discovering medicine knowledge, assisting physicians, improving public health, and supporting patients.[64–66]

Second, in terms of institutes, the US FDA has the highest number of publications. The USA has 7 institutes among the top 10 institutes with regard to the number of MDM-related publications, which implies that the USA is the bellwether in this field and *JOURNAL OF BIOMEDICAL INFORMATICS* ranks first among the journals.

Third, keywords burst is an indicator of a most active area of research, which refers to these keywords increase particular attention from the related scientific communities in a certain period of time.[67] Based on the co-keyword network and burst analysis, we have found that there are some new study trends like extracting information from the text of electronic medical records (EMR) and mining genetic data. Also, the patient-centered model is an inevitable trend in future medical development, and there are great discussions about precision medicine.[68–70] The big-data revolution will vastly improve the granularity and timeliness of available epidemiological information with hybrid systems augmenting.[71] And data mining analysis is the key to precision medicine treatment.[72] Denny also mentions that natural language processing methods to process narrative text data may be needed.[73] Wagner provides a tool (DGIdb, www.dgidb.org) and Pinero developed a platform DisGeNET (www.disgenet.org) for mining the druggable genome for precision medicine hypothesis generation.[74,75] In addition to the genome data,

Table 8

A total 15 references with the strongest citation bursts over the period between 2011 and 2019 are shown.



References	Year	Strength	Begin	End	2011 - 2019
RITCHIE MD, 2010, AM J HUM GENET, V86, P560, DOI	2010	2.6377	**2012**	2013	🟦🟥🟥🟦🟦🟦🟦🟦🟦
HRISTOVSKI D, 2005, INT J MED INFORM, V 74, P289, DOI	2005	2.6377	**2012**	2013	🟦🟥🟥🟦🟦🟦🟦🟦🟦
CLASSEN DC, 2011, HEALTH AFFAIR, V30, P 581, DOI	2011	2.9927	**2014**	2015	🟦🟦🟦🟥🟥🟦🟦🟦🟦
BARABASI AL, 2011, NAT REV GENET, V12, P56, DOI	2011	2.9927	**2014**	2015	🟦🟦🟦🟥🟥🟦🟦🟦🟦
SEJVAR JJ, 2011, VACCINE, V29, P599, DOI	2011	6.4047	**2015**	2016	🟦🟦🟦🟦🟥🟥🟦🟦🟦
JIN HD, 2008, IEEE T INF TECHNOL B, V12, P488, DOI	2008	3.392	**2015**	2016	🟦🟦🟦🟦🟥🟥🟦🟦🟦
CHEE BW, 2011, AMIA ANNU SYMP PROC, V2011, P217	2011	5.4849	**2015**	2016	🟦🟦🟦🟦🟥🟥🟦🟦🟦
GURULINGAPPA H, 2012, J BIOMED SEMAN T, V3, P0, DOI	2012	3.392	**2015**	2016	⬜🟦🟦🟦🟥🟥🟦🟦🟦
MARTIN D, 2013, DRUG SAFETY, V36, P547, DOI	2013	6.4047	**2015**	2016	⬜⬜🟦🟦🟥🟥🟦🟦🟦
KAHRAMANLI H, 2008, EXPERT SYST APPL, V35, P82, DOI	2008	5.2292	**2015**	2016	🟦🟦🟦🟦🟥🟥🟦🟦🟦
XU H, 2010, J AM MED INFORM ASSN, V17, P19, DOI	2010	4.2924	**2016**	2017	🟦🟦🟦🟦🟦🟥🟥🟦🟦
SAEED M, 2011, CRIT CARE MED, V39, P952, DOI	2011	7.2014	**2016**	2017	🟦🟦🟦🟦🟦🟥🟥🟦🟦
GINSBERG J, 2009, NATURE, V457, P1012, DO I	2009	3.4769	**2016**	2017	🟦🟦🟦🟦🟦🟥🟥🟦🟦
HRIPCSAK G, 2013, J AM MED INFORM ASS N, V20, P117, DOI	201 3	3.2683	**2016**	2019	⬜⬜🟦🟦🟦🟥🟥🟥🟥
WEI CH, 2013, NUCLEIC ACIDS RES, V41, P 0, DOI	2013	3.8182	**2017**	2019	⬜⬜🟦🟦🟦🟦🟥🟥🟥

medical images are also the important data sources of MDM. Radiomics has been defined as the conversion of images to minable data, which benefit to yield quantitative image-based phenotypes for data mining with other-omics for discovery (i.e., imaging genomics) or yield predictive image-based phenotypes of disease for precision medicine.[76] In the future, since the continuous development of computer software and hardware, the application of data mining technology in radiology may allow radiologists to further integrate their knowledge with their clinical colleagues in other medical specialties, and promote the development of precision medicine. Also, many medical mining systems could help physicians in daily clinical practice, like improving diagnosis accuracy,[77] reducing diagnosis time,[78] precisely providing quantified temporal order information of critical medical behaviors in clinical pathways, and reducing errors in medicine.[79,80]

Fourth, the research areas can be divided into clinical application (including screening, diagnosis, treatment, prognosis, monitoring, and management) and data mining approaches (classification, regression, clustering, prediction, association rule mining, and hybrid).[81–85] The clinical support system can be used to in several conditions such as emergency situation, shortage of physicians, and to decrease human errors. The algorithms that are applied in medicine such as logistic-regression, decision tree, neural-networks, support vector machines (SVM), and association rule. Time series and random forest algorithms are the most popular. But each data mining algorithm has its advantages and disadvantages. There are some common strength of the data mining techniques like suitable computational accuracy and ability to handle complex relationship among different features.[86] Besides, each of them has its own strength like the simplicity and comprehensibility of decision tree, popularity, and ability of neural-networks in general model extraction, association rule is suitable for describing frequent patterns among dataset; k-means is easy to implement and understand, and SVM is efficient.[87–89] Despite the advantage, the limitation also should be considered like time consuming and inability to support for large dataset. And each data mining algorithm has its own limitation, as an example, in SVM the generated models are a black box and it is designed essentially for binary classification; in random forest, it cannot estimate values of the variable outside the range of the training data; and in K-Means, it do not explain why and how these samples are grouped into a cluster. Data mining algorithms are capable to obtain valuable knowledge form raw dataset, but models are too complex to understand and interpret by human experts especially in black-box phenomenon. The final goal of modeling in medicine is providing understandable knowledge for physicians to conduct care strategies, so for overcoming this interpretability problem, extraction of rules and visualization could be applied.

Fifth, the number of times an article cited as a reference in another article reflects its scientific impact. And the citation can determine the distribution of the most influential literature in the field of MDM. Among the top 20 citation publications, 5 articles are published in 2011, 10 articles are published from 2012 to 2015, and 5 articles are published after 2016. Because high degree of cooperation with other countries and regions, USA is the most active (documents: 11) and influential (citations: 2566) country and ranks first, which indicates that the United States is the central region of MDM research, then followed by Canada (documents: 4, citations: 918) and Netherlands (documents: 3, citations: 742). China ranks fourth, which had 2 publications and had a citation of 456, shows that the degree of international academic cooperation is not as close as that of the USA. China should pay attention to the scientific of papers published and strengthen cooperation with other countries and regions in the future MDM study.

4.2. Limitations and future outlook

Although we have review the papers of MDM and demonstrate the research achievements and the potential impacts according to the authors, journals, countries, institution, references, and research topics, the limitation of this study should be addressed:

(1) The substantial works have published in MDM field, but it is impossible to discuss all of them in a single work. This type of research depends on bibliometric datasets and the datum collection is limited the WOS. In order to have a better understanding for MDM relevant research, better and bigger datasets are needed.
(2) In this study, we only focus on publications between 2011 and 2019 that mentioned MDM in the database which are publicly available on the website, the authors admit the possibility of missing some import research publications on MDM, and some relevant records could be missing, if the query phrases used for topic searches did not match some records. We are hoping that future studies covering longer time period will shed more light on the field, researchers, and publications. We hope this paper will make it possible to explore previous works by visualization and bibliometrics to provide guideline for researchers who are new to the field.

However, there are also other challenges such as structural EMR data, signal, radiomics image, integration with the hospital workflow, and lack of data mining package for medical domain. Noise and missing value also are common challenges in MDM. For process mining in healthcare, there are no portable solutions for all different hospital environments, and lack of a visualization tool of the process models as well as a great reliance on experts.[90] Medical data mining framework could be described as 5 steps:

(1) medical problem understanding;
(2) medical data preview;
(3) discovering relations between data;
(4) extracting relations to be a model;
(5) verification of extracted model by background knowledge.

For the mining framework, although there are some standards such as ICD-10 for disease information integration (WHO),[87] appropriate collection and transmission standards are still absent.

Author contributions

Conceptualization: Yue Luo, Chuanbiao Wen.
Data curation: Yuanzhang Hu.
Formal analysis: Yuanzhang Hu, Zeyun Yu.
Methodology: Yuanzhang Hu, Xiaoen Cheng.
Project administration: Chuanbiao Wen.
Supervision: Yue Luo, Xiaoen Cheng.
Writing – original draft: Yuanzhang Hu, Zeyun Yu.
Writing – review & editing: Xiaoen Cheng, Yue Luo, Chuanbiao Wen.

Correction

Dr. Xiaoen Cheng’s name was misspelled in the original publication as Xiaoen Chen. This has since been corrected.

References

[1] Esfandiari N, Babavalian MR, Moghadam AME, et al. Knowledge discovery in medicine. Curr Issue Future Trend 2014;41:4434–63.
[2] Cios KJ, Moore GW. Uniqueness of medical data mining. Artif Intell Med 2002;26:1–24.
[3] Borgman CL, Furner J. Scholarly communication and bibliometrics. Ann Rev Inform Sci Technol 2002;36:2–72.
[4] Wei F, Grubesic TH, Bishop BWJPG. Exploring the GIS knowledge domain using CiteSpace. Prof Geogr 2015;67:374–84.
[5] Bakkalbasi N, Bauer K, Glover J, et al. Three options for citation tracking: Google Scholar, Scopus and Web of Science. Biomed Digit Libr 2006;3:1–8.

[6] Chadegani AA, Salehi H, Yunus MM, et al. A comparison between two main academic literature collections: Web of Science and Scopus Databases. Asian Soc Sci 2013;9:18–26.

[7] Hu Z, Chen C, Liu Z. Salah AA, Tonta Y, Akdag Salah AA, Sugimoto C, Al U. The recurrence of citations within a scientific article. Proceedings of ISSI 2015 Istanbul: 15th International Society of Scientometrics and Informetrics Conference, Istanbul, Turkey, 29 June to 3 July, 2015, Bogaziçi University Printhouse 2015.

[8] Chen C. CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature. J Assoc Inform Sci Technol 2006;57:359–77.

[9] Li M, Porter AL, Wang ZL. Evolutionary trend analysis of nanogenerator research based on a novel perspective of phased bibliographic coupling. Nano Energy 2017;34:93–102.

[10] Liu H, Zhu Y, Guo Y, et al. Visualization analysis of subject, region, author, and citation on crop growth model by CiteSpace II software. Adv Intell Syst Comput 2014;278:243–52.

[11] Ortega JL, Aguillo IFJIP. Visualization of the Nordic academic web: link analysis using social network tools. Inform Process Manage 2008;44:1624–33.

[12] Haber P, Moro PL, Cano M, et al. Post-licensure surveillance of quadrivalent live attenuated influenza vaccine United States, Vaccine Adverse Event Reporting System (VAERS), July 2013–June 2014. Vaccine 2015;33:1987–92.

[13] Moro PL, Woo EJ, Paul W, et al. Post-marketing surveillance of human rabies diploid cell vaccine (Imovax) in the vaccine adverse event reporting system (VAERS) in the United States, 1990–2015. PLoS Neglect Trop Dis 2016;10:

[14] Leeper NJ, Bauer-Mehren A, Iyer SV, et al. Practice-based evidence: profiling the safety of cilostazol by text-mining of clinical notes. PloS One 2013;8:e63499.

[15] Ye M, Zhang HZ, Li L. Research on data mining application of orthopedic rehabilitation information for smart medical. IEEE Access 2019;7:177137–47.

[16] Hu YH, Lin WC, Tsai CF, et al. An efficient data preprocessing approach for large scale medical data mining. Technol Health Care 2015;23: 153–60.

[17] Wang S, Chang XJ, Li X, et al. Diagnosis code assignment using sparsity-based disease correlation embedding. IEEE Trans Knowl Data Eng 2016;28:3191–202.

[18] Reiner BI. Medical imaging data reconciliation, Part 3: Reconciliation of historical and current radiology report data. J Am Coll Radiol 2011;8:768–71.

[19] Reiner BI. Quantifying analysis of uncertainty in medical reporting: creation of user and context-specific uncertainty profiles. J Digit Imaging 2018;31:379–82.

[20] Reiner BI. Quantitative analysis of uncertainty in medical reporting: creating a standardized and objective methodology. J Digit Imaging 2018;31:145–9.

[21] Jiang GQ, Liu HF, Solbrig HR, et al. Mining severe drug-drug interaction adverse events using Semantic Web technologies: a case study. Biodata Min 2015;8.

[22] Chen YY, Li CM, Liang JC, et al. Health information obtained from the internet and changes in medical decision making: questionnaire development and cross-sectional survey. J Med Internet Res 2018;20: e47.

[23] Jing Y, Cheng C, Shi S, et al. Comparison of complex network analysis software: Citespace, SCI 2 and Gephi. IEEE International Conference on Big Data Analysis 2017.

[24] Azadeh A, Saberi M, Kazem A, et al. A flexible algorithm for fault diagnosis in a centrifugal pump with corrupted data and noise based on ANN and support vector machine with hyper-parameters optimization. Appl Soft Comput 2013;13:1478–85.

[25] Chen LF, Su CT, Chen KH, et al. Particle swarm optimization for feature selection with application in obstructive sleep apnea diagnosis. Neural Comput Appl 2012;21:2087–96.

[26] Mookiah MRK, Acharya UR, Lim CM, et al. Data mining technique for automated diagnosis of glaucoma using higher order spectra and wavelet energy features. Knowl Based Syst 2012;33:73–82.

[27] Rave H, Alison C, Suzanne T, et al. Text mining for adverse drug events: the promise, challenges, and state of the art. Druf Saf 2014;37:777–90.

[28] Lependu P, Iyer SV, Bauer-Mehren A, et al. Pharmacovigilance using clinical text. AMIA Jt Summits Transl Sci Proc 2013;109.

[29] Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012;13:395–405.

[30] Silvera SAN, Mayne ST, Gammon MD, et al. Diet and lifestyle factors and risk of subtypes of esophageal and gastric cancers: classification tree analysis. Ann Epidemiol 2014;24:50–7.

[31] Collins SA, Kenrick C, David A, et al. Relationship between nursing documentation and patients' mortality. Am J Crit Care 2013;22:306–13.

[32] Richardson AM, Lidbury BA. Infection status outcome, machine learning method and virus type;interact to affect the optimised prediction of hepatitis virus;immunoassay results from routine pathology laboratory assays in unbalanced data. BMC Bioinform 2013;14:1–8.

[33] Collins SA, Cato K, Albers D, et al. Relationship between nursing documentation and patients mortality. Am J Crit Care 22:306–13.

[34] Bandyopadhyay S, Wolfson J, Vock DM, et al. Data mining for censored time-to-event data: a Bayesian network model for predicting cardiovascular risk from electronic health record data. Data Min Knowl Disc 2015;29:1033–69.

[35] Yuan F, Zhang YH, Wan S, et al. Mining for candidate genes related to pancreatic cancer using protein–protein interactions and a shortest path approach. Biomed Res Int 2015;2015:623121.

[36] Papanikolaou N, Pavlopoulos GA, Theodosiou T, et al. DrugQuest –a text mining workflow for drug association discovery. BMC Bioinform 2016;17(Suppl 5):333–41.

[37] David M, David M, Marthe BG, et al. Data mining for prospective early detection of safety signals in the Vaccine Adverse Event Reporting System (VAERS): a case study of febrile seizures after a 2010–2011 seasonal influenza virus vaccine. Drug Saf 2013;36:547–56.

[38] Moro PL, Theresa H, Tom S, et al. Adverse events after Fluzone? Intradermal vaccine reported to the Vaccine Adverse Event Reporting System (VAERS), 2011–2013. Vaccine 2013;31:4984–7.

[39] Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 2017;14:749.

[40] Gillies RJ, Kinahan PE, Hricak HJR. Radiomics: images are more than pictures, they are data. Radiology 2015;278:151169.

[41] Parmar C, Grossmann P, Bussink J, et al. Machine learning methods for quantitative radiomic biomarkers. Sci Rep 2015;5:13087.

[42] Parmar C, Rios VE, Leijenaar R, et al. Robust Radiomics feature quantification using semiautomatic volumetric segmentation. PLoS One 2014;9:e102107.

[43] Breiman L. Random forests. Mach Learn 2001;45:5–32.

[44] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273–97.

[45] Lecun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436.

[46] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Paper presented at: International Conference on Neural Information Processing Systems 2012.

[47] Rojas E, Munozgama J, Sepúlveda M, et al. Process mining in healthcare: a literature review. Methodol Rev 2016;61:224–36.

[48] Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. Paper presented at: ACM sigmod record 1993.

[49] Srikant R, Naughton JF. Fast Algorithms for Mining Association Rules and Sequential Patterns. 1996.

[50] Agrawal R, Srikant R. Mining Sequential Patterns. 1995.

[51] Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17:507.

[52] Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. Gene Ontol Consort 2000;25:25–9.

[53] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32(database issue):267–70.

[54] Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc 2013;20:117–21.

[55] Dumouchel W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. Am Stat 1999; 53:177–90.

[56] Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2015;17:229–36.

[57] Fujimoto M, Higuchi T, Hosomi K, et al. Association between statin use and cancer: data mining of a spontaneous reporting database and a claims database. Int J Med Sci 2015;12:223–33.

[58] Classen DC, Roger R, Frances G, et al. 'Global trigger tool' shows that adverse events in hospitals may be ten times greater than previously measured. Health Aff 2011;30:581–9.

[59] Ritchie MD, Denny JC, Crawford DC, et al. Robust replication of genotype–phenotype associations across multiple diseases in an electronic medical record. Am J Hum Genet 2010;86:560–72.

[60] Mohammed S, Mauricio V, Reisner AT, et al. Multiparameter Intelligent Monitoring in Intensive Care II: a public-access intensive care unit database. Crit Care Med 2011;39:952–60.

[61] Sejvar JJ, Kohl KS, Gidudu J, et al. Guillain–Barré syndrome and Fisher syndrome: case definitions and guidelines for collection, analysis, and presentation of immunization safety data. Vaccine 2011;29:599–612.

[62] Martin D, Menschik D, Bryant-Genevier M, et al. Data mining for prospective early detection of safety signals in the Vaccine Adverse Event Reporting System (VAERS): a case study of febrile seizures after a 2010–2011 seasonal influenza virus vaccine. Drug Saf 2013;36:547–56.

[63] Wei CH, Kao HY, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Res 2013;41:W518–22.

[64] Larose DT, Larose CD. Discovering knowledge in data (an introduction to data mining). Univariate Statistical Analysis. vol. 9 2014;91–108. 10.1002/9781118874059.

[65] Loglisci C, Malerba D. A Temporal Data Mining Approach for Discovering Knowledge on the Changes of the Patient's Physiology. 2009;26–35.

[66] Brossette SE, Sprague AP, Hardin JM, et al. Association rules and data mining in hospital infection control and public health surveillance. J Am Med Inform Assoc 1998;5:373–81.

[67] Li XX, Shen J. Visualization analysis on key technologies of technical evolution – in the field of 3G mobile communication. Adv Mater Res 2013;694–697:2394–9.

[68] Hodson R. Precision medicine. Nature 2016;537:S49.

[69] Ashley EA. The precision medicine initiative: a new national effort. JAMA 2015;313:2119–20.

[70] Roy Choudhury A, Cheng T, Phan L, et al. Supporting precision medicine by data mining across multi-disciplines: an integrative approach for generating comprehensive linkages between single nucleotide variants (SNVs) and drug-binding sites. Bioinformatics 2017;33:1621–9.

[71] Bansal S, Chowell G, Simonsen L, et al. Big data for infectious disease surveillance and modeling. J Infect Dis 2016;214(Suppl 4):S375–9.

[72] Aronson SJ, Rehm HL. Building the foundation for genomics in precision medicine. Nature 2015;526:336–42.

[73] Denny JC. Chapter 13: mining electronic health records in the genomics era. PLoS Comput Biol 2012;8:e1002823.

[74] Wagner AH, Coffman AC, Ainscough BJ, et al. DGIdb 2.0: mining clinically relevant drug–gene interactions. Nucleic Acids Res 2015;44: D1036–44.

[75] Pinero J, Bravo A, Queralt-Rosinach N, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res 2017;45:D833–9.

[76] Giger ML. Machine learning in medical imaging. J Am Coll Radiol 2018;15:512–20.

[77] Bashir S, Qamar U, Khan FH, et al. HMV: a medical decision support framework using multi-layer classifiers for disease prediction. J Comput Sci 2016;13:10–25.

[78] Liu X, Lu R, Ma J, et al. Privacy-preserving patient-centric clinical decision support system on naïve bayesian classification. IEEE J Biomed Health Inform 2017;20:655–68.

[79] Huang Z, Lu X, Duan H. On mining clinical pathway patterns from medical behaviors. Artif Intell Med 2012;56:35–50.

[80] Fujihara H, Yamada C, Furumaki H, et al. Evaluation of the in-hospital hemovigilance by introduction of the information technology-based system. Transfusion 2015;55:2898–904.

[81] Zierk J, Ganslandt T, Rauh M, et al. Data mining of reference intervals for coagulation screening tests in adult patients. Clin Chim Acta 2019;499:108–14.

[82] Jia F, Lei YG, Lin J, et al. Deep neural networks: a promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. Mech Syst Signal Proc 2016;72–73:303–15.

[83] Chekroud AM, Zotti RJ, Shehzad Z, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry 2016;3:243–50.

[84] Xiang Y, Zhang CQ, Huang K. Predicting glioblastoma prognosis networks using weighted gene co-expression network analysis on TCGA data. BMC Bioinform 2012;13:8.

[85] Xu BY, Xu LD, Cai HM, et al. The design of an m-Health monitoring system based on a cloud computing platform. Enterp Inf Syst 2017; 11:17–36.

[86] Shaikhina T, Lowe D, Daga S, et al. Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. Biomed Signal Process Control 2002;37:1025–42.

[87] Zavodni AEH, Wasserman BA, Mcclelland RL, et al. Carotid artery plaque morphology and composition in relation to incident cardiovascular events: the Multi-Ethnic Study of Atherosclerosis (MESA). Radiology 2014;271:381–9.

[88] Liu Y, Liu Q, Zheng D, et al. Application and improvement discussion about Apriori algorithm of association rules mining in cases mining of influenza treated by contemporary famous old Chinese medicine. Paper presented at: IEEE International Conference on Bioinformatics & Biomedicine Workshops 2012.

[89] Rojas E, Munoz-Gama J, Sepúlveda M, et al. Process mining in healthcare: a literature review. J Biomed Inform 2016;61:224–36.

[90] Thygesen SK, Christiansen CF, Christensen S, et al. The predictive value of ICD-10 diagnostic coding used to assess Charlson comorbidity index conditions in the population-based Danish National Registry of Patients. BMC Med Res Methodol 2011;11:83.