A Scientometric Analysis of Publications Related to Predictive Medicine*

Aida Khakimova1[0000-0001-9355-9249], Dongxiao Gu2[0000-0003-3557-009X],
Oleg Zolotarev3[0000-0001-6917-9668], Maria Berberova3[0000-0002-6357-7929],
Michael Charnine4[0000-0003-0450-5156]

[1] Research Center for Physical and Technical Informatics, Nizhny Novgorod, Russia
[2] School of Management, Hefei University of Technology, Hefei, China
[3] Russian New University. Moscow, Russia
[4] FRC CSC of the Russian Academy of Sciences Moscow, Russia, Moscow, Russia
aida_khatif@mail.ru | dongxiaogu@yeah.net |
ol-zolot@yandex.ru | maria.berberova@gmail.com |mc@keywen.com

Abstract. Due to the increasing popularity of new research in medicine this study was conducted to determine recent research trends of predictive, preventive and personalized medicine (PPM). We identified the terms relevant to PPM using own search engine based on neural network processing in PubMed database. We extracted initially about 15000 articles. Then we carried out the statistical analysis for identifying research trends. The article presents the results of solving the problem of evaluating research topics at the level of thematic clusters in a separate subject area. An approach based on the analysis of article titles has been implemented. Identification of terms, connections between them and thematic clustering were carried out using the free software VOSViewer, which allows to extract terms in the form of noun phrases, as well as to cluster them.

Keywords: Predictive Medicine, Preventive Medicine, Personalized Medicine, Biomedicine, Trends, Key Terms, Machine Learning.

1 Introduction

Bibliometric mapping helps transform most of the metadata of publications into maps or visualizations, from which useful information can be obtained through post-processing.

Bibliometry is one of statistical methods to analyze the mass of literature and to reveal historical development [3], as well as a scientific qualitative and quantitative study of publications. Many authors used bibliometric in different areas of medicine, such as ophthalmology [4], rheumatology [5], otolaryngology [6], nephrology [7], geriatrics [8], etc.

* This work is supported by RFBR, grants 18-07-01111, 18-07-00909, 18-07-00225, 19-07-00455 and 20-04-60185.

For example, Zhu & Guan (2013) [9] and Sinkovics (2016) [10] visualized keywords to identify research topics or clusters in specific disciplines. Zhu & Guan [9] looked at keywords and topic categories of publications as actors for mapping a keyword sharing network and a topic sharing network and compared them with corresponding random binary networks. Most of these studies focus on identifying major trends in the form of the most cited studies or the most frequently used terms. While this is an excellent method for identifying major research topics [9], emerging or potentially interesting topics may not be easy to spot.

VOSviewer uses the VOS display technique (visualization of similarities) and is freely available (www.vosviewer.com). Maps are generated from a sharing matrix. The similarity matrix is calculated using the measure of the strength of the association [11]. After transforming the data into visual form, VOSviewer offers two ways to display it: network visualization and density visualization. The network visualization view displays concepts based on their importance. The larger the label and circle, the more important the concept. The color of the circle indicates which cluster the term belongs to. Density visualization shows the importance of areas depending on the number of connected elements.

When creating maps based on a text corpus, the user can choose between binary and full counts. When choosing a binary count, only the presence or absence of the term in the document is considered. In the case of a complete count, all occurrences in the document are considered.

Content analysis is one of the areas of bibliometric analysis and it includes the identification of trends based on words (Huffman et al., 2013 [12]; Menendez-Manjon, Moldenhauer, Wagener, & Barcikowski, 2011 [13]; Sooryamoorthy, 2010 [fourteen]).

Gelman and Unwin [15] warned against overuse of maps or renderings. They recommended supplementing such maps with traditional graphs and tables to provide additional evidence.

2 Subjects and Methods

The general process of visualization of scientific maps and their subsequent analysis consisted of the following stages. As a principle for constructing the matrix, a network based on jointly occurring keywords was chosen

1. Obtaining a semi-structured amount of information from information sources containing document annotations, as well as information about authors, their affiliation,

publication date, keywords, citation information (universal information retrieval system Dimensions; specialized PubMed database).

2. Pre-processing of data to improve the quality of the generated map (the formation of groups depending on the period).

3. Construction of scientific maps and their visualization.

4. Analysis of scientific maps, identification of the most intensively developing topics, analysis of temporal, statistical aspects.

In this study, the PubMed database created by the National Center for Biotechnology Information (NCBI) in the United States was identified as the first base for collecting information. The search for standard bibliometric processing was performed based on the terms "predictive" and "personalize (s) e" "preventive" "medicine" with the logical operator "AND". We measured the number of publications in the field of PPM depending on the affiliation of the first author. The output of publications at the country level was estimated using PubMed tools [16].

To isolate key terms using the Word2Wec method, about 15,000 articles (titles and annotations) were selected from the PubMed database, containing the terms "predictive" and "personalize (s) e" in their titles. To extract keywords from the headlines, we used baseline Medline / PubMed database for all years. The annual baseline is released in December of each year [2]. Key terms were extracted from the titles of articles. For statistical processing, a set of programs in the Java language was developed.

Algorithm for finding trends when processing a corpus of natural language texts:

1. Initially, an expert creates a dictionary of key terms, consisting of keywords and phrases in a normalized form. This is a dictionary De or main dictionary.

2. Next, a temporary dictionary-template of new words and phrases is created Dn.

3. Then an array of statistics is created to analyze the neighborhood of keywords.

$$S = \{ D^e_i \{ D^n_k, R_k \} \} \quad (1)$$

D^e_i is an element of key terms vocabulary,

D^n_k – new word (not a stop word or a word from the main dictionary D^e),

R_k – a frequency of occurrence of a new word D^n_k in the vicinity of the key term D^e_i. At first, its value is zero.

4. If a key term is found, then an analysis of its neighborhood is performed. The neighborhood of the word is determined by the size of the sliding window. The words of the analyzed text are normalized. If a word (not a stop word) is found in the vicinity of a keyword it is not included in the main dictionary, it is first checked whether the array S contains a pair of values $\{ D^e_i, \{ D^n_k \}\}$. If there is no such pair of values, then it is added to the dictionary. Then the frequency value (R_k) for a given pair of values, increases by one. Each new element D^n_k can become either a key term or part of it.

5. After processing the text corpus and filling in the S array, the most rated candidates for key terms are selected. Minimum rating threshold (R_{min}) the new term is determined by the expert. нового термина определяет эксперт. All elements of the array S with a rating less than R_{min} are not considered.

6. Next, candidates for a new key terms are generated. If there are several new words for the same keyword with the same R_k, then there is a possibility that the new term could be more than one word. The system generates a possible new multi word term. At the same time, candidates for key terms are generated with already approved elements (from the main dictionary D^e) and possible key terms found in the vicinity of the keyword, but already without it.

7. As a result of the analysis of the values of the array S and approval by the expert new key terms are added into the dictionary. New key terms with the maximum rank determine development trends in this field of medicine.

The “Dimensions” scientific information database was used as the second database. The search was carried out by keywords in the annotations "predictive preventive personalized medicine" for 2008-2020.

VOSviewer version 1.6.15 was used to construct and visualize scientific maps based on the data obtained. The program allows you to carry out scientific mapping based on scientometric analysis (frequency of co-occurrence of keywords).

The construction of a graph of interconnected pairs of objects is based on multidimension scaling (MDS), which is a means for visualizing pairwise connected graph vertices displayed on a plane. The method is used when two or more dimensions need to be examined for data analysis. MDS is used to construct bibliometric maps, which can include either co-citation of sources or co-occurrence of terms (objects, authors, documents, journals, keywords). The number of joint occurrences of elements within a given neighbourhood *i* and *j* is denoted c_{ij}. c_i is a frequency of element *i*.

$$cc_{ii} = \sum_{ii \neq ii} cc_{iiii} \tag{2}$$

If we have a set of n vertices, we want to build a graph of connected elements (for example, vertices that indicate the joint appearance of elements, or joint citation of documents), then s_{ij} will mean the strength of the associative connection between vertices *i* and *j* (Van Eck & Waltman, 2009) and represents.

$$ss_{iiii} = \frac{2mmcc_{iiii}}{cc_{ii}cc_{ii}} \tag{3}$$

They call this measure as proximity index. Otherwise, we can say that c_{ij} means the total number of connections of the vertex *i* and *m* means the total number of connections in the network.

$$mm = \frac{1}{2}\sum_{ii} cc_{ii} \tag{4}$$

When constructing a graph, two types of similarity can be used: direct and indirect. When using the indirect type of similarity to build a graph, direct comparison of elements is used, when using the indirect type of similarity, vectors of elements are compared. MDS allows to reduce the dimension of the graph. The method works in a similar way as t-distributed Stochastic Neighbor Embedding (t-SNE).

There are several options for implementing MDS, including Metric multidimensional scaling (mMDS) [28], Non-metric multidimensional scaling (nMDS) [29], Generalized multidimensional scaling (GMD) [30] and so on. The main advantage of this classical approach is the integration into the VOSviewer software product. MDS method implemented in several libraries (for Java, Python, Statistica).

3 Results

3.1 Statistical analysis

We analyzed trends in the field of predictive and personalized medicine (PPM) for the period from 1940 to 2020.

To search in PubMed the keywords “predictive medicine”, “personalized medicine”, “preventive medicine” were used. The collocation “preventive medicine” at first was appeared in 1857. The collocation “predictive medicine” at first was appeared in 1918. The collocation “personalized medicine” at first was appeared in 1952.

From the graph (see Fig. 1) we can see that predictive medicine, personalized medicine and preventive medicine experienced ups and downs in popularity. There is the increase of publications in these areas in last decade, apparently due to the activity of the European Association of Predictive, Preventive and Personalized Medicine (EPMA), which was established in 2008 [17].

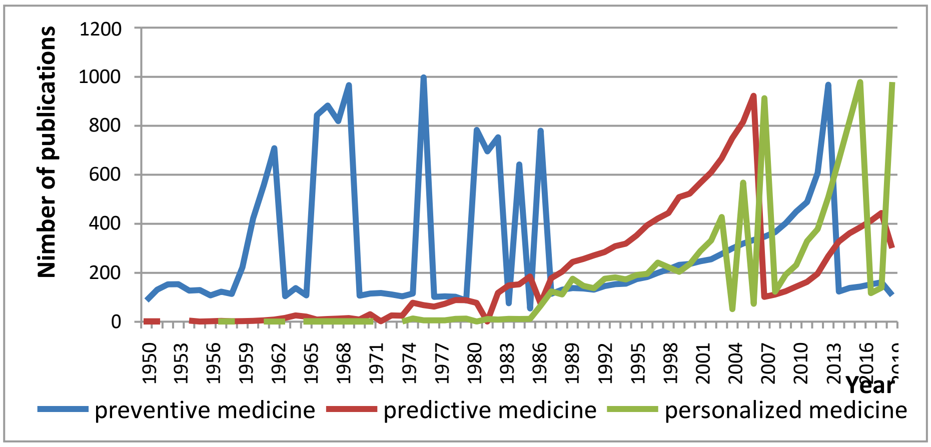

Fig. 1. The number of PubMed publications in the fields of preventive, predictive and personalized medicine from 1950 to 2020.

3.2 Key term mapping with VOSviewer

VOSviewer carries out the extraction of keywords in a matrix, the structure of which forms in the form of a network by linking terms according to the calculated link strength. Bond strength was defined as the total number of occurrences of a term in pairs with other terms.

The first step involved mapping terms extracted from an array of articles indexed in Dimensions for the period 2008–2020. For further visualization, a list of 85 keywords was formed, which included terms from article titles. In VOSviewer, the option "create map from text corpus" was selected. The counting method was set to binary counting. This means that each concept is counted only once per annotation, regardless of how many times it actually appears in a given annotation. VOSviewer's algorithm automatically excludes 40% of least significant terms. In total, 51 terms were used for mapping out of 4979 terms (see Fig. 2).

Stage 1 involved rendering a text corpus of 15,000 titles identified during the preparation phase. These names were extracted from the PubMed database (see section 2).

The purpose of this step is to create an overview of concepts within a given body of literature.

Stage 2 consisted of two parts. Part 1 was aimed at extracting key terms using special software tools.

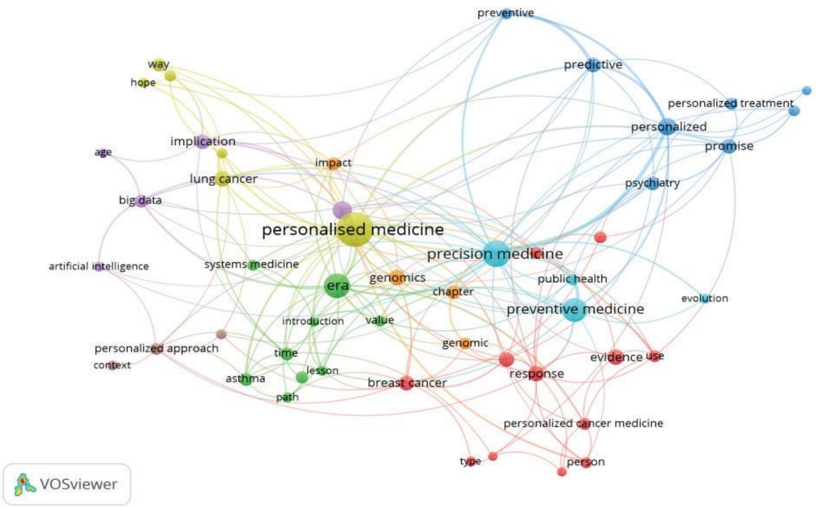

Fig. 2. The visualization of 51 concepts.

The method is based on the sequential linear division of the sentence into various phrases. The statistics of various options for parsing phrases in the corpus were calculated using a floating "window" - a numeric hyperparameter that shows how many words to the left and right of the central one is its environment. In this way, more than 9200 derivatives and phrases were extracted.

Part 2 focused on mapping key terms with overlapping windows (see Fig. 3).

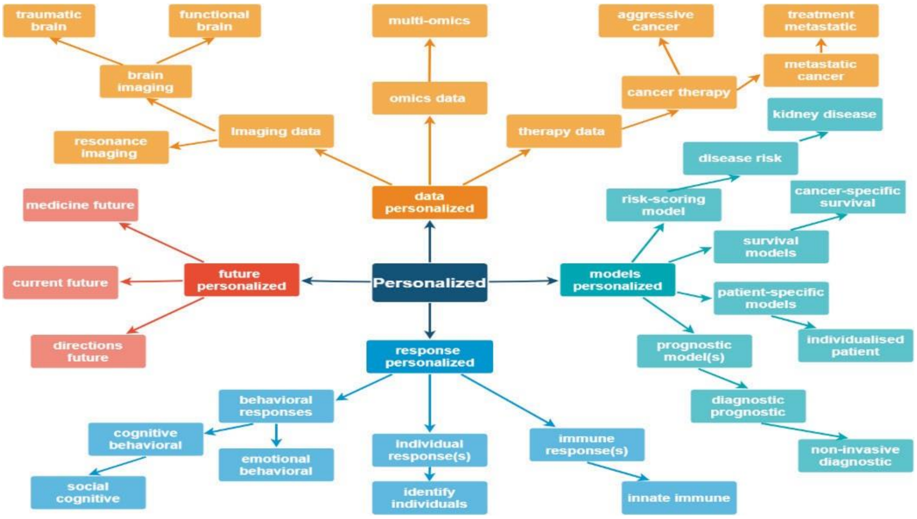

Fig. 3. Contextual mapping of terms obtained by statistical analysis of the body of texts based on «personalized» term.

Part 3 focused on answering a research question about how software tools can help identify interesting ideas.

An additional visualization was performed in which the threshold for the minimum number of occurrences was set to 5 and the number of concepts included in the visualization was set to 123 (see Fig. 4). This step was done with a binary count. Non-specific terms were excluded (Asia, October, USA, New York, 21st century, etc.).

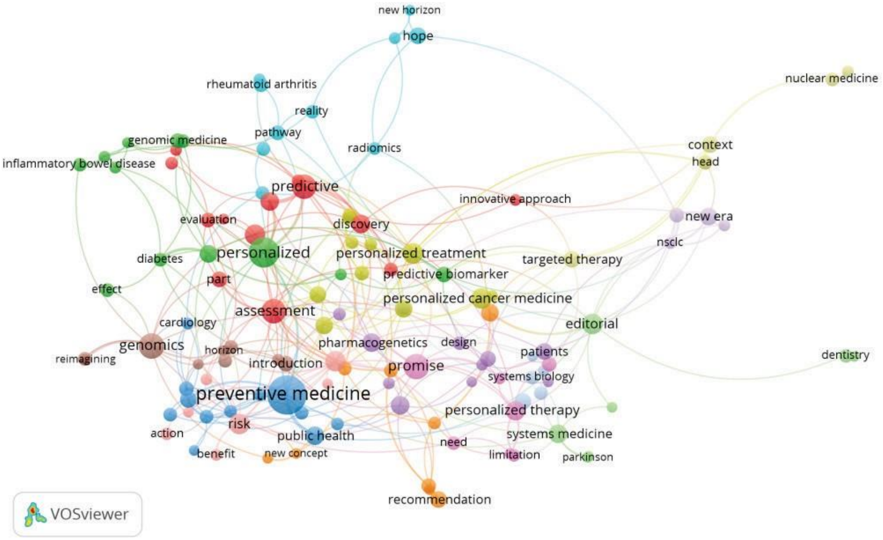

Fig. 4. Visualization of 123 concepts (14 clusters)

The list of terms / concepts defined in VOSviewer can be used as an initial contextual mapping template. Contextual analysis can also help to identify potential relationships between two or more concepts.

For example, let us select the first cluster. It includes the terms assessment, clinical, discovery, evaluation, innovative approach, personalized medicine approach, predictive, preventive, rare disease, translation, translational.

We found these terms in our list of 9200 terms and phrases. The term «predictive» is central to the cluster in terms of the number of links. We built a context map of terms for this set of terms, considering the results of VOSviewer and our statistical processing of the corpus of articles.

4 Results

As discussed in the previous section, 123 concepts were included in the visualization in Phase 1 of the analysis. The VOSviewer map gave 14 conceptual clusters. After exploring the concepts in each cluster, we selected 1 cluster to build the contextual conceptual mapping.

Table 1. Topic titles and cluster terms.

Label	**Cluster**	**Topic**	**Weight<Total link strength>**	**Weight<Occurrences>**	**Score<Avg. pub. year>**
innovative approach	1	Innovative approaches in predictive, preventive, personalized medicine	5	6	2018.3
personalized medicine approach	1		5	19	2017.4
predictive	1		26	27	2016.6
preventive	1		17	16	2017.1
diabetes	2	Personalized management and application of biomarkers in the treatment of a range of diseases (diabetes, stomach cancer)	11	8	2017.6
diabetes mellitus	2		7	6	2017.7
gastric cancer	2		4	5	2018.2
personalized	2		21	39	2017.1
personalized management	2		6	9	2016.9
predictive biomarker	2		6	11	2017.4
prognostic biomarker	2		7	6	2018.7
cardiology	3	Preventive medicine and cardiology perspectives	5	6	2015.5
future perspective	3		6	7	2015.3
preventive medicine	3		22	67	2016.3
public health	3		11	15	2015.9
molecular diagnostic	4	Personalized diagnostics and medicine in oncology	6	7	2015.1
ovarian cancer	4		8	14	2017.6
personalized cancer medicine	4		5	19	2016.3
personalized treatment	4		12	18	2015.4
pharmacogenomic	4		9	10	2015.9
paradigm shift	5	Paradigm and strategy of personalized medicine	4	7	2017.3
personalized medicine strategy	5		3	6	2015.8
pharmacogenetics	5		7	15	2016.5

Cluster 1 focuses on topics related to innovative approaches in predictive preventive personalized medicine (shown in table 1). Cluster 2 is devoted to topics related to per- sonalized management and application of biomarkers. Cluster 3 can be referred to as

Preventive Medicine and Cardiology Perspectives. Cluster 4 focuses on personalized diagnosis and medicine in oncology. Cluster 5 brings together topics related to the paradigm and strategy of personalized medicine. Cluster 6 deals with Omics in the treatment of rheumatoid arthritis. Cluster 7 focuses on topics related to personalized preventive medicine in cancer treatment. Cluster 8 includes issues of genetics and genomics in the PPM. Clusters 9 to 12 cover topics related to PMP in the management of chronic non-cancer diseases (multiple sclerosis, diabetes, Parkinson's disease COPD). Clusters 13 and 14 focus on innovative PPM in oncology, including Non-small-cell lung carcinoma (NSCLC). Data from 8 cluster have not included in the table 1.

The above list of research areas may already be of interest from a scientific landscape perspective. At this point, you can go back to the beginning of the process and develop a search strategy based on one cluster.

The starting point for contextual conceptual mapping (see Fig. 5) was a list of concepts of the first cluster.

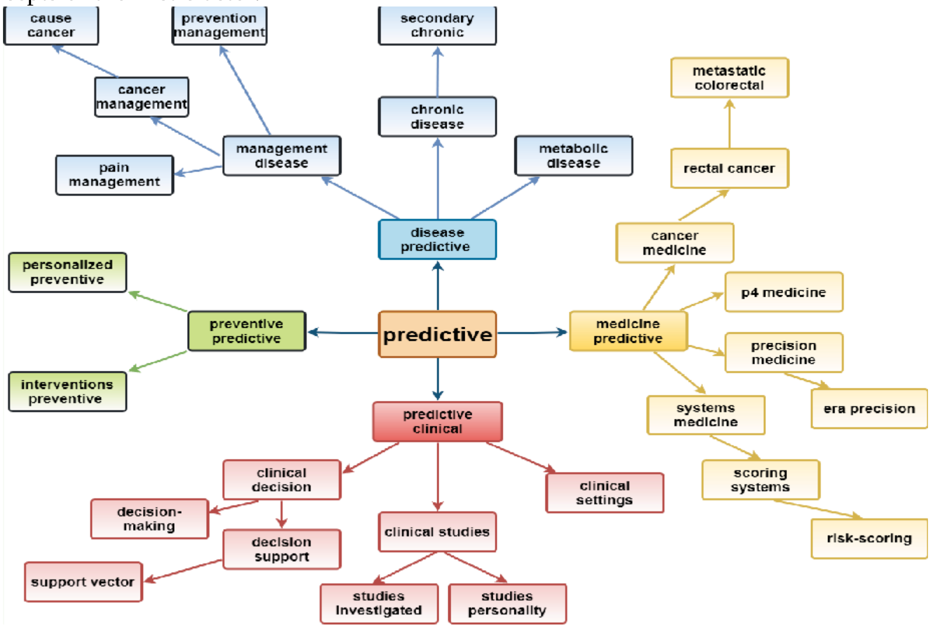

Fig. 5. Contextual mapping of terms obtained by statistical analysis of the text corpus based on the predictive term.

The list was supplemented with additional concepts that proved to be important. The question of whether a term or concept was considered potentially significant was determined in terms of the frequency of its occurrence in titles and annotations (as determined by frequency analysis).

Figure 5 shows the results of the extended contextual analysis. This map can serve as a starting point for a deeper review of the literature.

5 Conclusion

We have presented a method for analyzing publications and identifying trends in predictive medicine. The proposed method allows to search for new terms in preventive medicine based on the neighborhood approach.

Experts took part in identifying new trends. The use of software to automatically highlight trends significantly reduces the time it takes to generate new terms. To process the corpus of selected articles, one expert would need several months, while using the proposed software and considering the work of experts, this work was completed in 2 weeks.

The modified word2vec method was used by the authors to highlight key terms and to build forecasts. The diagrams are built using the MDS method. The search and analysis of information was carried out in several electronic libraries using their analytical mechanisms and visualization tools. A limitation of the current study is that the analysis relied on information provided by the authors in the titles. Future research may include full text analysis. Two contextual conceptual term maps were created in the prototype. These maps were created to provide an overview of research topics and to identify promising directions.

Acknowledgments

This work is supported by RFBR, grants 18-07-01111, 18-07-00909, 18-07-00225, 19-07-00455 and 20-04-60185.

References:

1. Hicks D, Wouters P, Waltman L, de Rijcke S, Rafols I. The Leiden Manifesto for research metrics. Nature 2015;520:429-31.
2. PubMed Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2005-. PubMed Help. [Updated 2019 Jul 25]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK3827/
3. Young H: Glossary of Library and Information Science. 1983, Chicago: American Library Association Google Scholar. Indian J Ophthalmol. 2015 Jan;63(1):54-8. doi: 10.4103/0301-4738.151471.
4. Mansour AM, Mollayess GE, Habib R, Arabi A, Medawar WA. Bibliometric trends in ophthalmology 1997-2009. Semin Arthritis Rheum. 2017 Jun;46(6):828-833. doi: 10.1016/j.semarthrit.2016.12.002.
5. Redondo M, Leon L, Povedano FJ, Abasolo L, Perez-Nieto MA, López-Muñoz F.A bibliometric study of the scientific publications on patient-reported outcomes in rheumatology. Clin Otolaryngol. 2017 Dec;42(6):1338-1342. doi: 10.1111/coa.12910.
6. Saunders TFC, Rymer BC, McNamara KJ. A global bibliometric analysis of otolaryngology: Head and neck surgery literature. G Ital Nefrol. 2016 Nov-Dec;33(6). pii: gin/33.6.10.
7. Torrisi AM, Granata A. Bibliometric indicators of nephrology journals: strengths and weaknesses. [Article in Italian] Geriatr Gerontol Int. 2017 Feb;17(2):357-360. doi: 10.1111/ggi.12880.

8. Ang HM, Kwan YH. Bibliometric analysis of journals in the field of geriatrics and gerontology. J Neuropsychiatry Clin Neurosci. 2015 Fall;27(4):354-61. doi: 10.1176/appi.neuropsych.15010024.
9. Zhu, W., & Guan, J. (2013). A bibliometric study of service innovation research: based on complex network analysis. Scientometrics, 94(3), 1195-1216. Retrieved from https://EconPapers.repec.org/RePEc:spr:scient:v:94:y:2013:i:3:d:10.1007_s11192-012-0888- 1.
10. Sinkovics, N. (2016). Enhancing the foundations for theorising through bibliometric mapping. International Marketing Review, 33(3), 327-350. https://doi.org/10.1108/IMR-10-2014-0341
11. van Eck, N.J. and Waltman, L. (2007), "Bibliometric mapping of the computational intelligence field", International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, Vol. 15 No. 5, pp. 625-645
12. Huffman, M. D., Baldridge, A., Bloomfield, G. S., Colantonio, L. D., Prabhakaran, P., Ajay, V. S., Prabhakaran, D. (2013). Global Cardiovascular Research Output, Citations, and Collaborations: A Time-Trend, Bibliometric Analysis (1999-2008). Plos One, 8(12), 7. doi: 10.1371/journal.pone.0083440
13. Menendez-Manjon, A., Moldenhauer, K., Wagener, P., & Barcikowski, S. (2011). Nanoenergy research trends: bibliometrical analysis of nanotechnology research in the energy sector. Journal of Nanoparticle Research, 13(9), 3911-3922. doi: 10.1007/s11051-011-0344-9
14. Sooryamoorthy, R. (2010). Medical research in South Africa: a scientometric analysis of trends, patterns, produc-tivity and partnership. Scientometrics, 84(3), 863-885. doi: 10.1007/s11192-010-0169-9
15. Gelman, A., & Unwin, A. (2013). Infovis and Statistical Graphics: Different Goals, Different Looks. Journal of Computational and Graphical Statistics, 22(1), 2-28. https://doi.org/10.1080/10618600.2012.761137
16. Yang H, Lee HJ. Research Trend Visualization by MeSH Terms from PubMed. Int J Environ Res Public Health 2018;15:1113-27.
17. Jazayeri SB, Alavi A, Rahimi-Movaghar V. Situation of medical sciences in 50 top countries from 1996 to 2010 - based on quality and quantity of publications. Acta Med Iran 2012;50:273–8.
18. Garfield E. Keywords plus – ISI's breakthrough retrieval method. 1. Expanding Your Searching Power on Current Contents on Diskette. Current Contents 1990;32:5-9.
19. Xu Q, Boggio A, Ballabeni A. Countries' Biomedical Publications and Attraction Scores. A PubMed-based assessment [version 2; peer review: 2 approved]. F1000Research 2015;3:292-8.
20. Klimenko S., Khakimova A., Charnine M., Zolotarev O., Merkureva N. Semantic approach to visualization of research front of scientific papers using web-based 3D graphic. В сборнике Proceedings of the 2018 International Conference Web 3D. The 23rd International Proceedings - Web3D 2018: 23rd International ACM Conference on 3D Web Technology 23, 3D Everywhere. 2018. С. a20.
21. Golubnitschaja, Olga & Kinkorova, Judita & Costigliola, Vincenzo. (2014). Predictive, Preventive and Personalised Medicine as the hardcore of 'Horizon 2020': EPMA position paper. The EPMA journal. 5. 6. 10.1186/1878-5085-5-6.
22. Geographic directory «About countries». http://ostranah.ru/_lists/population.ph.
23. Wikipedia. Lists of countries by GDP. https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal).

24. Chen C, Wang Z, Li W, Sun X. Modeling Scientific Influence for Research Trending Topic Prediction. Thirty-Second AAAI Conference on Artificial Intelligence, 2018. https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16721.
25. A.Kh. Khakimova, O.V. Zolotarev, M.A. Berberova. Visualization of bibliometric networks of scientific publications on the study of the human factor in the operation of nuclear power plants based on the bibliographic database Dimensions. Scientific Visualization, 2020, volume 12, number 2, pages 127 - 138, DOI: 10.26583/sv.12.2.10, E-ISSN:2079-3537.
26. Zolotarev, O.; Solomentsev, Y.; Khakimova, A.; Charnine M. Identification of semantic patterns in full-text documents using neural network methods. In Proceedings of the 29th International Conference on Computer Graphics and Vision. Graphicon-2019. 2019. Available online: http://ceur-ws.org/Vol-2485/paper64.pdf.
27. Klimenko, S.; Charnine, M.; Zolotarev, O.; Merkureva, N.; Khakimova, A. Semantic Approach to Visualization of Research Front of Scientific Papers Using Web-Based 3d Graphic. In Proceedings of the 23rd International ACM Conference on 3D Web Technology. 2018, 1-6.
28. Kruskal, J. B. (1964), "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis", Psychometrika, 29 (1): 1–27, doi:10.1007/BF02289565.
29. Bagdonavicius, V., Kruopis, J., Nikulin, M.S. (2011). "Non-parametric tests for complete data", ISTE & WILEY: London & Hoboken. ISBN 978-1-84821-269-5.
30. Bronstein AM, Bronstein MM, Kimmel R (January 2006). "Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching". Proc. Natl. Acad. Sci. U.S.A. 103 (5): 1168–72. Bibcode:2006PNAS.103.1168B. doi:10.1073/pnas.0508601103.