www.nature.com/pr

Pediatric RESEARCH

CLINICAL RESEARCH ARTICLE

Pediatric literature trends: high-level analysis using text-mining

Sarina Levy-Mendelovich[1,2,3], Yiftach Barbash[1,4], Ivan Budnik[5], Daniella Levy-Erez[1,6,7,8], Raz Somech[1,9], Shelly Soffer[1,5], Susan Furth[6,7] and Eyal Klang[1,3,10]

BACKGROUND: Pediatric research is a diverse field that is constantly growing. Current machine learning advancements have prompted a technique termed text-mining. In text-mining, information is extracted from texts using algorithms. This technique can be applied to analyze trends and to investigate the dynamics in a research field. We aimed to use text-mining to provide a high-level analysis of pediatric literature over the past two decades.

METHODS: We retrieved all available MEDLINE/PubMed annual data sets until December 31, 2018. Included studies were categorized into topics using text-mining.

RESULTS: Two hundred and twenty-five journals were categorized as Pediatrics, Perinatology, and Child Health based on Scimago ranking for medicine journals. We included 201,141 pediatric papers published between 1999 and 2018. The most frequently cited publications were clinical guidelines and meta-analyses. We found that there is a shift in the trend of topics. Epidemiological studies are gaining more publications while other topics are relatively decreasing.

CONCLUSIONS: The topics in pediatric literature have shifted in the past two decades, reflecting changing trends in the field. Text-mining enables analysis of trends in publications and can serve as a high-level academic tool.

Pediatric Research (2021) 90:212–215; https://doi.org/10.1038/s41390-021-01415-8

IMPACT:

- Text-mining enables analysis of trends in publications and can serve as a high-level academic tool.
- This is the first study using text-mining techniques to analyze pediatric publications.
- Our findings indicate that text-mining techniques enable better understanding of trends in publications and should be implemented when analyzing research.

INTRODUCTION

Pediatrics is a field of large research diversity[1] reflected by the origin of publications, a large number of topics, and a wide range of article types. In order to maintain your expertize as a pediatrician, you constantly need to keep updated regarding new knowledge.[2,3] In 1998, Bergman presented a work analyzing the trends in pediatric publications.[1] This work used manual analysis of pediatric research topics. However, the number of pediatric publications increased exponentially over the past decades making it impossible to manually summarize a topic.[4]

In text-mining, information is extracted from texts using computer algorithms.[5] Text-mining can be applied to identify trends and to investigate the dynamics in a research field.[4,6–8] Examples include anti-epileptic drug research,[9] adolescent substance abuse,[10] and Diatom research.[9] The application of such algorithms has not been frequently used in pediatric research. In this work, we aimed to use text-mining to provide a high-level analysis of trends in pediatric literature over the past two decades.

METHODS

Data set

The U.S. National Library of Medicine produces an annual version of MEDLINE/PubMed data. This data set is freely available to download.[11] We retrieved all available MEDLINE/PubMed annual data sets until December 31, 2018.

We extracted the following data for each entry: PubMed unique article ID (PMID), title, publishing journal, abstract text, keywords (if any), and authors (including the first author affiliation, if available).

We retrieved the number of times each article was cited. For this purpose, we used a National Center for Biotechnology Information application.[11] Data lock and citation retrieval was performed on May 1, 2019.

Data processing

Data processing and result visualization were written on Python (ver. 3.6.5, 64 bits).

[1]Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel; [2]The Israeli National Hemophilia Center and Thrombosis Unit and Amalia Biron Research Institute of Thrombosis and Hemostasis, Chaim Sheba Medical Center, Tel Hashomer, Israel; [3]The Sheba Talpiot Medical Leadership Program, Chaim Sheba Medical Center, Tel Hashomer, Israel; [4]Department of Diagnostic Imaging, Chaim Sheba Medical Center, Tel Hashomer, Israel; [5]Department of Pathophysiology, Sechenov First Moscow State Medical University (Sechenov University), Moscow, Russia; [6]Division of Nephrology, Children's Hospital of Philadelphia, Philadelphia, PA, USA; [7]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA; [8]Pediatric Nephrology unit, Schnieder Pediatric hospital, Petach Tiqva, Israel; [9]Pediatric Department A and the Immunology Service, "Jeffrey Modell Foundation Center Edmond and Lily Safra Children's Hospital", Sheba Medical center, Ramat Gan, Israel and [10]Institute for Healthcare Delivery Science (IHDS), Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, NY, USA
Correspondence: Sarina Levy-Mendelovich (Levysarina@gmail.com)

Received: 22 December 2020 Revised: 13 January 2021 Accepted: 16 January 2021
Published online: 17 March 2021

For text-mining, punctuations and double spaces were removed. First author country was retrieved from the affiliation data of the first author.

Inclusion criteria

We included articles published in journals categorized as "Pediatrics, Perinatology, and Child Health" based on Scimago ranking for medicine journals.[12] The time frame for article inclusion was the past 20 years (1/1/1999–31/12/2018). As the text-mining technique is dependent on the text in the abstract, we included only entries with abstracts >50 words.

Topic modeling

Included studies were categorized into topics using latent Dirichlet allocation (LDA).[13] This algorithm is well documented and is commonly utilized for topic modeling. LDA splits the corpus into groups by grouping together documents with similar words. Each group is the most significant set of words used by the algorithm for the differentiation. We set the algorithm to output 200 groups of word sets. Manual allocation of topics to each set of words was performed by a domain expert (S.L.-M.) according to Bergman et al.[1] with mild modifications (see Supplement). The final number of topics was 35. Each abstract in the corpus was assigned to one of the 35 topics.

RESULTS

Two hundred and twenty-five journals were categorized as Pediatrics, Perinatology, and Child Health based on Scimago ranking for medicine journals. Out of the total of 29,137,794 entries in PubMed, 610,826 papers were published in 1 of the 225 pediatrics-related journals. Of them, 392,826 had abstracts >50 words. From this corpus, we included papers published between 1999 and 2018, a total of 201,141 papers.

Country of origin

The pediatric articles came from 177 countries and 8 continents. United States had the highest number of publications ($n = 69,263$) followed by United Kingdom ($n = 12,332$), Turkey ($n = 9614$), and Canada ($n = 8676$). When further analyzing the citations/publications ratio (C/P ratio), United States ranked highest (C/P = 6.2), followed by United Kingdom (C/P = 5.9), Sweden (C/P = 5.8), and Canada (C/P = 5.1) (Fig. 1).

Article type

MEDLINE/PubMed article type was specified for the sub-categories: meta-analysis, clinical trial, randomized control trial, multicenter study, editorial, letter, review, and guidelines. As presented in Fig. 2, the majority of these articles were review papers (56%) and clinical trials (13%) (Fig. 2). Clinical guidelines and meta-analyses had the highest C/P ratio (C/P ratio = 12.2 and C/P ratio = 10.6, respectively; Fig. 3).

Trends of topics

The trends of the leading ten topics in the past two decades are presented below. When examining the proportion of each topic along the years, we observed that epidemiological papers (15% in 1999–2004 to 26% in 2014–2018) and psychology (5% in 1999–2004 to 9% in 2014–2018) increased, while neurology (6–3%), infectious diseases (5.4–3.3%), and pulmonology

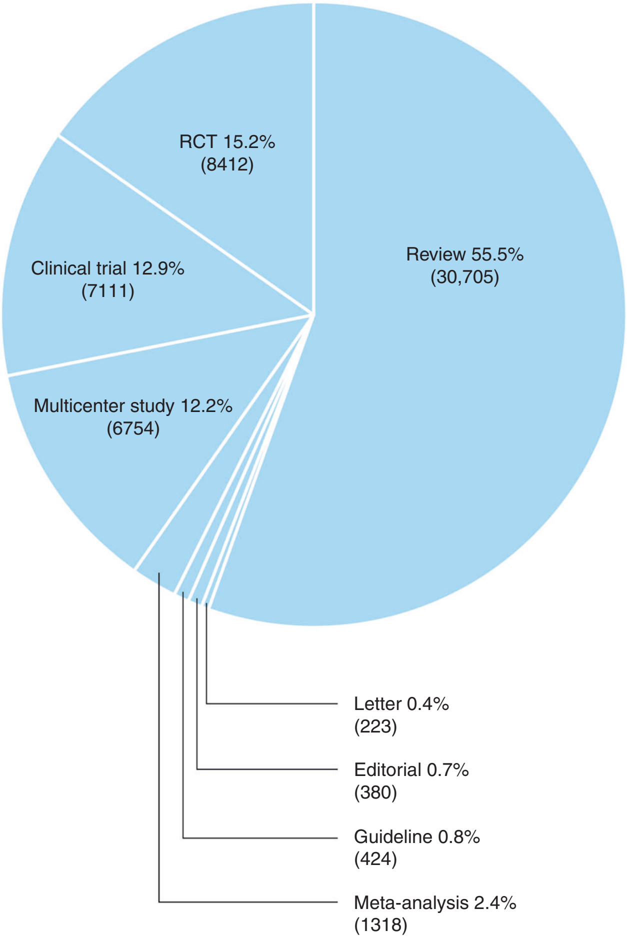

Fig. 2 Number of articles per type of article. Total number of articles per type of article, Each article type has a total number of articles over the past 20 years as well as a percentage out of the total.

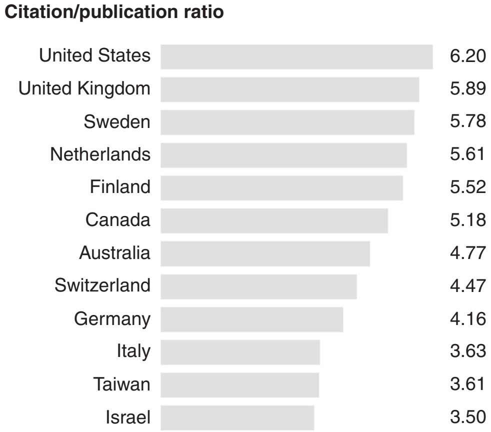

Fig. 1 Citations/publications ratio per country. Citations/publications ratio per country over the past 20 years.

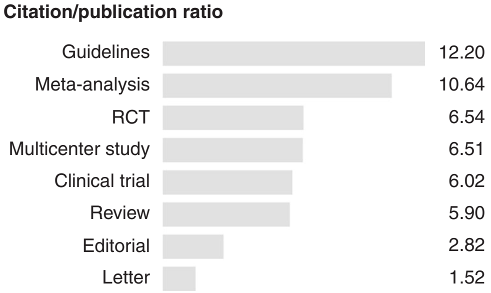

Fig. 3 Citations/publications ratio per article type. Citations/publications ratio per article type over the past 20 years.

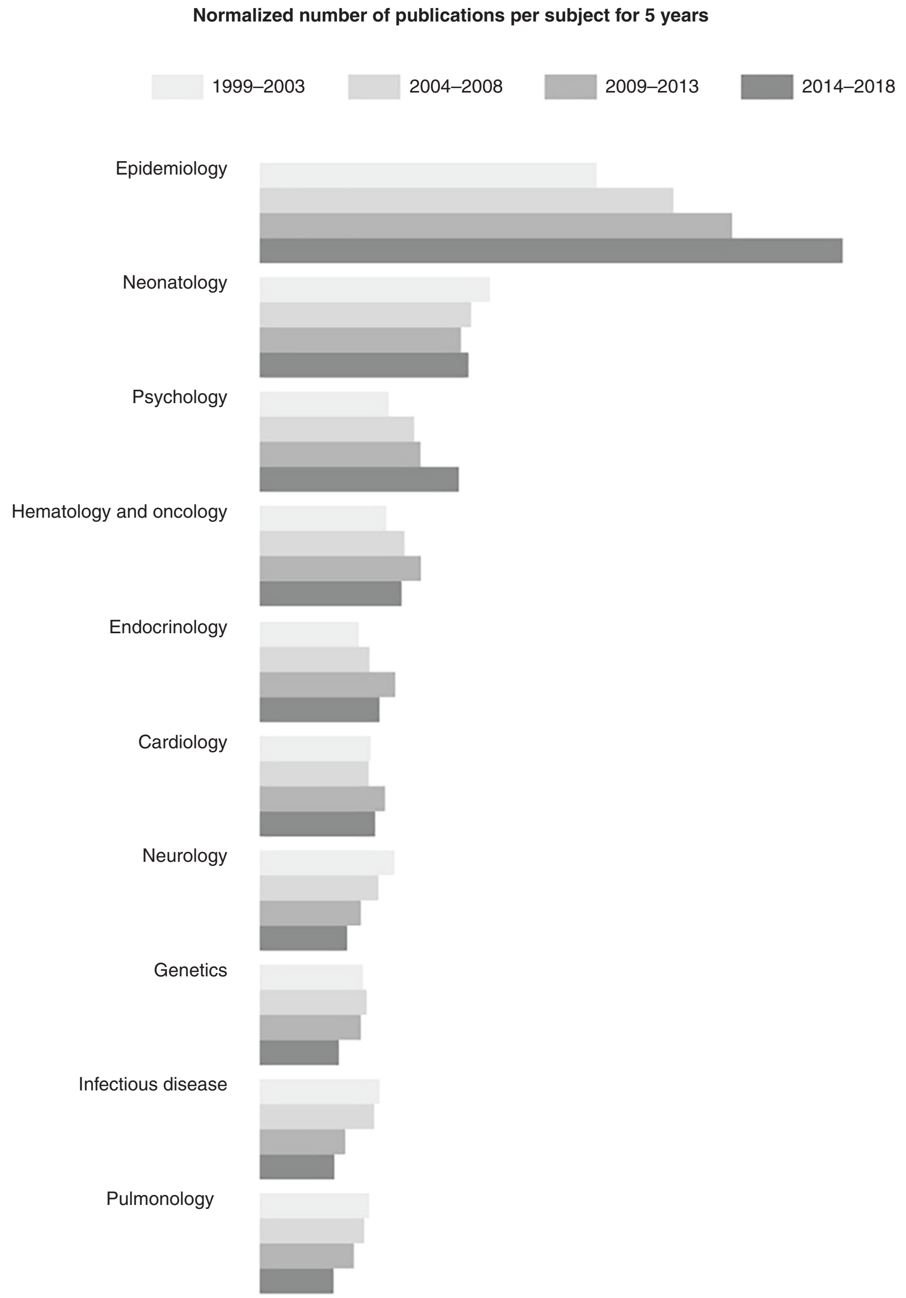

Fig. 4 Number of publications per topic over the 2 decades. Number of publications per topic every 5 years over the past 2 decades. Each topic has 4 bars each representing a 5 year period.

(4.9–3.3%) relatively decreased (Fig. 4). Psychology and psychiatry had the highest C/P ratio (6.8) followed by neurology, epidemiology, and endocrinology (Fig. 5).

DISCUSSION

This study employed text-mining to provide a high-level view of pediatric research in the past two decades. The number of articles published in the past 20 years has grown, which reflects the growing interest in research in the pediatric world.

United States leads in the number of pediatric publications, which reflects the prominent role of the United States pediatric community in global pediatric research. Although low-income countries have higher morbidity, they have a disproportionally lower number of publications, as was previously reported by Keating et al.[14] This is most probably attributed to the lack of resources directed to research and academic centers in those countries.

When analyzing article types, we demonstrated that the most frequently cited were clinical guidelines and meta-analyses, both provide analyses and summary of a large amount of data that can influence the clinical setting.

The topic change of focus to epidemiology papers may be a result of a number of factors. The understanding of the epidemiology of diseases can help in disease prevention and better resource planning to improve quality and longevity in an era in which economy has an important impact on delivering medical treatment. The field of psychology and psychiatry has also grown, supporting the importance of mind in medicine and should be addressed in pediatric research as well.

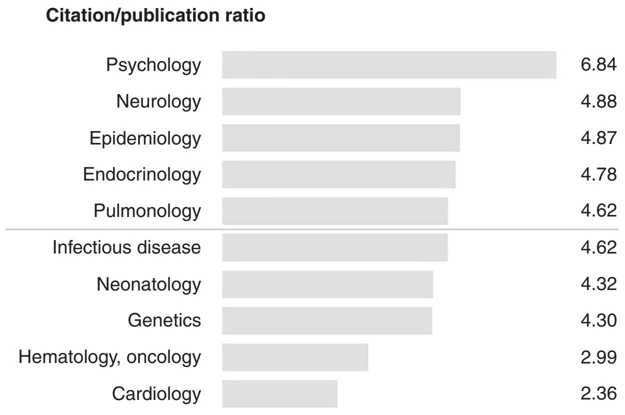

Fig. 5 Citations/publications ratio per topic. Citations/publications ratio per topic over the last 20 years.

Text-mining enabled us to summarize the past 20 years of pediatric literature and get a better understanding regarding trends in this field. Text-mining can enable the medical academia to deal with the ever growing number of publications that otherwise would not be possible.

Our research has several limitations. First, this is a comprehensive study that includes 20 years of research. As such, it can only provide a high-level view of pediatric research. Second, it should be noted that papers that were published in the last year of our analysis have not had a long time to be cited, therefore this may influence their C/P ratio. Third, there was some topic overlap, therefore some papers had more than one topic. This may have affected the results. Fourth, LDA used words taken out of abstracts and not full text, that said, abstracts are a reflection of the essence of the paper. Fifth, this study analyzed pediatric journals. Pediatric papers published in non-pediatric journals were not included.

In conclusion, the topics in pediatric literature have shifted in the past two decades, reflecting changing trends in the field. Text-mining enables analysis of trends in publications and can serve as a high-level academic tool emphasizing where there is a need for additional medical education as well as research.

ACKNOWLEDGEMENTS

We would like to thank MS. Amy Zhong at Icahn School of Medicine at Mount Sinai, New York for her expert assistance with the design of the tables and figure. No external funding was used for this research.

AUTHOR CONTRIBUTIONS

S.L.-M., E.K., and Y.B. contributed to conception and design and acquisition of data. Y.B., I.B., S.S., S.F., R.S., and D.L.-E. contributed to interpretation of data. S.L.-M., E.K., and D.L.-E. drafted the article, revised it critically for important intellectual content, and approved the version to be published. I.B., S.F., and R.S. revised the article critically for important intellectual content.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41390-021-01415-8.

Competing interests: The authors declare no competing interests.

Consent: No patient consent was needed in this study.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Bergman, A. B. 50 years of pediatrics: 1948-1998. the journal in 1947 and 1997: a dramatic change. *Pediatrics* **102**, 186–190 (1998).
2. Ozuah, P. O. Residency research requirement as a predictor of future publication productivity. *J. Pediatr.* **155**, 1–2 (2009).
3. Alvira, C. M. et al. Enhancing the development and retention of physician-scientists in academic pediatrics: strategies for success. *J. Pediatr.* **200**, 277–284 (2018).
4. Singh S. P., Swagata, K., Sudhir, S. M. & Singh V. P. The application of text mining algorithms in summarizing trends in anti-epileptic drug research. *Int. J. Stat. Probability* https://doi.org/10.5539/ijsp.v7n4p11 (2018).
5. Thuraisingham, B. M. *Data Mining: Technologies, Techniques, Tools, and Trends* (CRC Press, 1999).
6. Alfalqi, K. & Alghamdi, R. A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl*. https://doi.org/10.14569/IJACSA.2015.060121 (2015).
7. Hao, T. A bibliometric analysis of text mining in medical research. *Soft Comput.* **22**, 7875–7892 (2018).
8. Song, M. Detecting the knowledge structure of bioinformatics by mining full-text collections. *Scientometrics* **96**, 183–201 (2013).
9. Zhang, Y. et al. Trends in diatom research since 1991 based on topic modeling. *Microorganisms* https://doi.org/10.3390/microorganisms7080213 (2019).
10. Wang, S. H. et al. Text mining for identifying topics in the literatures about adolescent substance use and depression. *BMC Public Health* **16**, 279–016 (2016).
11. N.I.H. of U.S. National Library of Medicine, download MEDLINE/PubMed data. www.nlm.nih.gov/databases/download/pubmed_medline.html (2020).
12. SCImago, (n.d.). SJR — SCImago journal & country rank. http://www.scimagojr.com (2020).
13. Blei, D. M. Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).
14. Keating, E. M. et al. Global disparities between pediatric publications and disease burden from 2006 to 2015. *Glob. Pediatr. Health* https://doi.org/10.1177/2333794X19831298 (2019)., corrected publication 2021