Use of Decentralized-Learning Methods Applied to Healthcare: A Bibliometric Analysis

Carolina Ameijeiras-Rodriguez [1] [0000-0001-8817-7772],
Rita Rb-Silva [1] [0000-0002-1422-0974], Jose Miguel Diniz [1,3] [0000-0002-9950-0579],
Julio Souza [1,2] [0000-0002-8576-1903], Alberto Freitas [1,2] [0000-0003-2113-9653]

[1] MEDCIDS - Department of Community Medicine, Information and Health Decision Science, Faculty of Medicine, University of Porto, Porto, Portugal
[2] CINTESIS@RISE, Faculty of Medicine, University of Porto, Porto, Portugal
[3] PhD Program in Health Data Science Faculty of Medicine of University of Porto. Porto, Portugal

Abstract. The use of health data in research is fundamental to improve health care, health systems and public health policies. However, due to its intrinsic data sensitivity, there are privacy and security concerns that must be addressed to comply with best practices recommendations and the legal framework. Decentralized-learning methods allow the training of algorithms across multiple locations without data sharing. The application of those methods to the medical field holds great potential due to the guarantee of data privacy compliance. In this study, we performed a bibliometric analysis to explore the intellectual structure of this new research field and to assess its publication and collaboration patterns. A total of 3023 unique documents published between 2013 and 2023 were retrieved from Scopus, from which 488 were included in this review. The most frequent publication source was the IEEE Journal of Biomedical and Health Informatics (n=27). China was the country with the highest number of publications, followed by the USA. The top three authors were Dekker A (n=14), Wang X (n=13), Li X (n=12). The most frequent keywords were “Federated learning” (n=218), “Deep learning” (n=62) and “Machine learning” (n=52). This study provides an overall picture of the research literature regarding the application of decentralized-learning in healthcare, possibly setting ground for future collaborations.

Keywords: Federated learning, Decentralized-learning, Privacy-preserving protocols; Machine learning, Smart Healthcare, Medicine, Bibliometric analysis.

1 Introduction

In recent years, data production has increased exponentially in most industries. In the health industry, data is crucial for the continuous improvement of healthcare (HC), health systems and public health policies. On one hand, data generated by this field represents a significant percentage of the global data volume [1]. On the other hand, this type of data is extremely sensitive, so the preservation of the patient's privacy becomes the main objective and the emerging concerns regarding privacy protection led to the creation of demanding regulations and increasing data governance and privacy barriers. Thus, decentralized-learning solutions have gained increased attention

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_25

by allowing to extract the maximum possible value from health data while preserving data privacy. Decentralized-learning methods include collaborative methods, such as split- and federated-learning (FL). FL is probably the most attractive decentralized distributed machine learning (ML) paradigm, with its first publications related to health published in 2013 [2, 3]. In 2017, FL gained great exposure after being presented by Google [4]. Published applications of FL to HC include patient similarity learning [5], cardiology [6], oncology [7, 8], population health [9], among others.

This is the first bibliometric overview of the literature regarding the use of decentralized analysis in HC, fostering new ideas for investigation and potential collaborations.

2 Material and Methods

2.1 Bibliographic Database

Scopus was chosen as a broad and generalized scientific database of peer-reviewed research, comprising all research fields. It allows data extraction and aggregation in several formats, providing detailed information for the analyses.

2.2 Study Design

A bibliometric analysis was performed to evaluate the related research on decentralized-learning in HC, published from January 2013 to the date of the search, January 25, 2023. The search was restricted to documents originally written in English.

2.2.1. **Eligibility Criteria**

For the purpose of this review, decentralized-learning was defined as a ML approach to use data available from multiple parties, without sharing them with a single entity.

Inclusion criteria: Articles had to be peer-reviewed original research or review; and describe an application of decentralized learning techniques to human health.

Exclusion criteria: All the results comprising books, book chapters, conference proceedings, editorials, perspectives, study protocols, commentaries, pre-prints and other unpublished studies; studies with quality issues; publications made before 2013; and full text publications not written in English were excluded. This was intended to reduce the potential bias caused by different publications from the same study (duplication) and restrict the analysis to high-quality original studies.

2.2.2. **Search strategy**

A composite search strategy was adopted to include terms related to “decentralized-learning” and “healthcare”. The final search expression was accomplished by two queries, using TITLE-ABS-KEY filter as presented in the Supplementary Material (Sup Mat).

2.3 Software and Data Analysis

After a review by two independent reviewers and a third reviewer for disagreements, we selected the articles that met the inclusion criteria. To facilitate this selection, we used the Rayyan tool [10]. Bibliometrix R package [11] and the Biblioshiny platform were used to perform the analysis of the selected articles. The Keyword Plus engine from the Web of Science was used to analyze the most used keywords.

2.4 Ethical Considerations

This study has been granted exemption from requiring ethics approval as it did not involve animal or human participants and only publicly available data was used.

3 Results

3.1 General information

A total of 3050 documents were retrieved from Scopus, corresponding to 3023 unique documents published between January 1, 2013, to the day of the search, January 25, 2023. After the eliminating duplicated records, title and abstract screening, and full-text review, 488 documents were included (16.1% of total retrieved documents) - Figure 1 of Sup Mat. The list of included articles is available in the Sup Mat.

The annual growth rate was 16.49%. However, from 2013 to 2019, the number of published articles was relatively stable, with a mean of 7.4 articles published per year. Since 2020, there has been exponential growth, with 41 articles published in 2020, 101 articles in 2021, 271 articles in 2022, and 23 in the first 25 days of 2023. The average of citations per document was 12.3.

3.2 Countries

A total of 45 countries presented a corresponding author with at least one publication. Nevertheless, 7 countries alone concentrated approximately half of the world's production: China (71 publications), the USA (68), Germany (23), South Korea and Netherlands (22), India (21), and Canada (15). International co-authorship comprised 49.8% of the total publications. In terms of total citations, publications from the USA stood out in the first position (1895 citations), followed by the Netherlands (702), Germany (432), Canada (373) and China (338). However, analyzing the average number of citations per article, the Netherlands took first place (31.9 citations/article), followed by the USA (27.9), Canada (24.9), Italy (23.2) and Germany (18.8)

A country-collaboration bibliometric network was constructed and is presented in Figure 1. Researchers from 50 countries reported international collaborations. Some collaboration clusters can be seen, namely one dominated by the USA (red cluster), another dominated by China (blue cluster) and another one mostly constituted by European countries, with a prominent position for Germany and the UK (green cluster).

DOI: 10.1007/978-3-031-36024-4_25

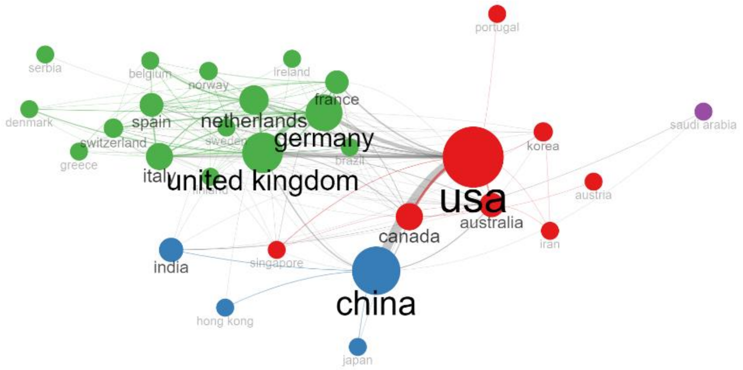

Fig. 1. Countries collaboration network regarding research in decentralized analysis in HC.

3.3 Affiliation and Institutions

Apart from Harvard Medical School, no institutions that have concentrated production in the field of decentralized learning in HC. There were 1250 different institutions worldwide having published at least one article in this area, averaging 2.6 institutions per article. Harvard Medical School (133 publications), Technical University of Munich (45), University of Michigan (42), University of Pennsylvania (39) and University of California (38) ranked in the top-5 most productive institutions.

3.4 Sources and publications

The 488 articles were published in 245 different international scientific journals with an impact factor. No single source concentrated a significant percentage of the overall production. The most relevant sources, with at least 10 published articles were IEEE Journal of Biomedical and Health Informatics (27), IEEE Access (17), IEEE Transactions On Medical Imaging (13), IEEE Internet Of Things Journal (12), Journal Of Biomedical Informatics (12), Journal Of The American Medical Informatics Association (12), IEEE Transactions On Industrial Informatics (10), JMIR Medical Informatics (10). Table 1 presents the most cited articles and their respective Normalized Total Citations (NTCs). Brismiti (2018) presents the highest number of global citations, whereas Xu (2021) presents the highest NTC, despite being the fourth most cited article.

Table 1. Top-5 most globally cited articles regarding research in decentralized analysis in HC.

Paper	DOI	Citations	NTC
Brisimi TS, 2018, Int J Med Inform	10.1016/j.ijmedinf.2018.01.07	297	4.18
Sheller MJ, 2020, Sci Rep	10.1038/s41598-020-69250-1	216	7.47
Chen Y, 2020, IEEE Intell Syst	10.1109/MIS.2020.2988604	201	6.95
Xu J, 2021, J Healthc Inform Res	10.1007/s41666-020-00082-4	175	9.81
Chang K, 2018, J Am Med Infor As	10.1093/jamia/ocy017	158	2.23

3.5 Authors

A total of 4038 researchers participated in the 488 articles, corresponding to a mean of 8.3 authors per article. Only 6 publications (1.2 % of the total) were single-authored articles. Regarding the most relevant authors in this field, the top-10 authors with the highest number of publications are Dekker A (14 publications), Wang X (13), Li X (12), Chen Y (11), Lambin P (11), Van Soest J (10), Wang J (10), Wang Z (9), Jiang X (8) and Liu J (8). Figure 2 further presents the production of these authors over time, where the size of the circle is proportional to the number of published articles and the intensity of the color is proportional to the number of citations.

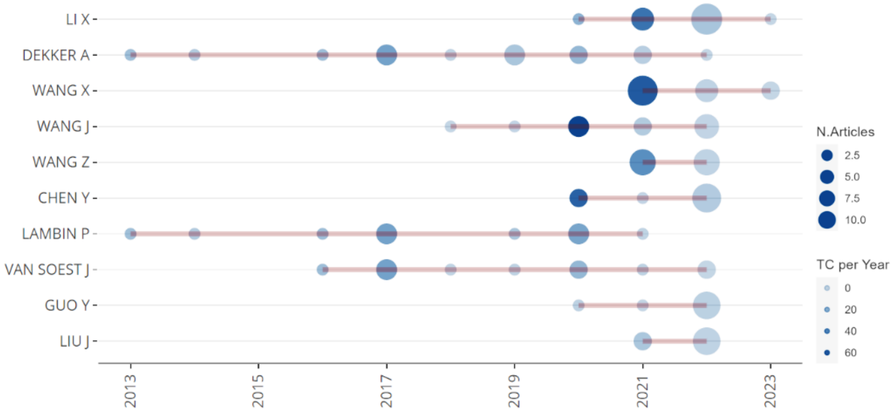

Fig. 2. Production of the top-10 most relevant authors over time.

3.6 Keywords and concepts

In the last 3 years, the most common keywords, considering the Keyword Plus engine from Web of Science, were “federated learning” (217 counts), “deep learning” (62), “machine learning” (52), “distributed learning” (29), “medical imaging” (9), “electronic health records” (8), “big data” (7), “pharmacoepidemiology” (7) and “distributed computing” (5). Other common keywords reveal some commonly targeted research fields, namely “COVID-19”, “hospital care”, “medical imaging”, “electronic health records” and “internet of things” (IoT).

4 Discussion

This is the first bibliometric review focused on the application of decentralized-learning to the HC field. This research domain has emerged in recent years as a response to the privacy and security barriers to the use of health data from multiple sources. The number of included articles is relatively small due to the novelty of this topic.

The exponential production after 2019 may be explained by several factors. First, the use of FL by Google, a company with worldwide projection, may have attracted

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_25

research groups attention to these methods. Second, the General Data Protection Regulation (GDPR) became enforceable beginning 25 May 2018 [12], which forced collaborative research network to find solutions to overcome some privacy barriers. In addition, the SARS-COV-2 worldwide pandemic promoted an intense development and dissemination of privacy-preserving technologies in the HC field.

China and the USA stood out for their research volume, considering the corresponding author's affiliation, with the latter being by far the country with the highest number of citations, presenting a dominance of the intellectual structure in this field. Moreover, four of the top-5 most productive institutions are based in the USA, highlighting the USA' investments in this field. In terms of country collaboration, China and the USA present a substantial network with countries geographically widespread, whereas there are robust collaboration networks within European countries, who must jointly comply with the GRDP [12].

The most relevant sources are journals on health informatics, rather than computer science or engineering, which might be explained by the specificity of the HC field. In terms of methods, FL is the most popular one, followed by blockchain, which has introduced a wide range of applications in HC, such as protection of HC data and management of genomics databases, electronic health records (EHRs) and drug supply chain [13].

Common keywords related to application areas include EHRs, imaging analysis, COVID-19 and IoT. The increasing adoption of EHRs facilitates multi-institutional collaboration, although concerns regarding infrastructure, data privacy and standardization introduces several constraints. In this sense, decentralized analysis, in particular FL, allows multiple medical and research sites to jointly train a global predictive model without data sharing, fostering collaborative medical research [14]. Moreover, a decentralized approach is useful for avoiding the issue of having a single point of failure, which is common in current EHR systems, which are usually centralized [15]. Imaging analysis, paired with advanced techniques such as deep learning, has become an effective diagnostic method, e.g., detection of several types of cancer [16]. For COVID-19, imaging analysis and decentralized learning have been employed for processing chest X-ray images to predict clinical outcomes [17]. This disease brought unprecedented challenges for health systems worldwide, including mandatory testing and geographic tracking and mapping of outbreaks, with FL being mainly proposed for COVID detection using large amounts of multisite data [18]. Emerging technologies such as IoT devices forced Industry and Academia to join efforts for building medical applications [19]. Research on decentralized analysis in this field addresses the several security threats resulting from the use of these types of devices. In particular, the use of blockchain and its decentralized storage system facilitates secure data storing and sharing through networks composed of distributed nodes, as noted with IoT-based systems [20].

We recognize several limitations of this study, including: 1) the use of a multiple bibliographic databases would be preferable, however adding results from searches in other databases would not be feasible due to time restrictions; 2) the existence of errors in the Scopus database, e.g. duplicated publications; 3) considering the continuous update of Scopus and the interval between the day of search and the report of the review results, there is a lag between these results and the actual research progress, and 4)

DOI: 10.1007/978-3-031-36024-4_25

limitations of the bibliometric methodology itself [21]. However, this study provides a useful and comprehensive overview of the intellectual structure in this research domain.

5 Conclusion

This bibliometric review analyzed 488 published articles to identify the intellectual structure and trends of research on decentralized-learning applied to the HC field. This review may be a useful resource for gaining insights into this emergent research domain and fostering novel ideas for investigation.

Acknowledgments: This research work was developed under the project "Secur-e-Health" (ITEA 20050), co-financed by the North Regional Operational Program (NORTE 2020) under the Portugal 2020 and European Regional Development Fund (ERDF), with the reference NORTE-01-0247-FEDER-181418.

Supplementary material: The Sup Mat can be found at: https://bit.ly/41rCjDL.

References

1. Callaway A. The healthcare data explosion [Available from: https://www.rbccm.com/en/gib/healthcare/episode/the_healthcare_data_explosion.
2. Doiron D, Burton P, Marcon Y, Gaye A, Wolffenbuttel BH, Perola M, et al. Data harmonization and federated analysis of population-based studies: the BioSHaRE project. Emerging themes in epidemiology. 2013;10(1):1-8.
3. El Emam K, Samet S, Arbuckle L, Tamblyn R, Earle C, Kantarcioglu M. A secure distributed logistic regression protocol for the detection of rare adverse drug events. Journal of the American Medical Informatics Association. 2013;20(3):453-61.
4. McMahan B, Moore E, Ramage D, Hampson S, y Arcas BA, editors. Communication-efficient learning of deep networks from decentralized data. Artificial intelligence and statistics; 2017: PMLR.
5. Lee J, Sun J, Wang F, Wang S, Jun C-H, Jiang X. Privacy-preserving patient similarity learning in a federated environment: development and analysis. JMIR medical informatics. 2018;6(2):e7744.
6. Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W. Federated learning of predictive models from federated electronic health records. International journal of medical informatics. 2018;112:59-67.
7. Naeem A, Anees T, Naqvi RA, Loh W-K. A comprehensive analysis of recent deep and federated-learning-based methodologies for brain tumor diagnosis. Journal of Personalized Medicine. 2022;12(2):275.
8. Pati S, Baid U, Edwards B, Sheller M, Wang S-H, Reina GA, et al. Federated learning enables big data for rare cancer boundary detection. Nature communications. 2022;13(1):7346.
9. Huang L, Shea AL, Qian H, Masurkar A, Deng H, Liu D. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay

time using distributed electronic medical records. Journal of biomedical informatics. 2019;99:103291.

10. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Systematic reviews. 2016;5:1-10.

11. Ahmi A. Bibliometric Analysis using R for Non-Coders: A practical handbook in conducting bibliometric analysis studies using Biblioshiny for Bibliometrix R package2022.

12. General Data Protection Regulation (GDPR) Regulation (EU) 2016/679, (2016).

13. Haleem A, Javaid M, Singh RP, Suman R, Rab S. Blockchain technology applications in healthcare: An overview. International Journal of Intelligent Networks. 2021;2:130-9.

14. Dang TK, Lan X, Weng J, Feng M. Federated learning for electronic health records. ACM Transactions on Intelligent Systems and Technology (TIST). 2022;13(5):1-17.

15. Kimovski D, Ristov S, Prodan R. Decentralized Machine Learning for Intelligent Health Care Systems on the Computing Continuum. arXiv preprint arXiv:220714584. 2022.

16. Subramanian M, Rajasekar V, VE S, Shanmugavadivel K, Nandhini P. Effectiveness of Decentralized Federated Learning Algorithms in Healthcare: A Case Study on Cancer Classification. Electronics. 2022;11(24):4117.

17. Dayan I, Roth HR, Zhong A, Harouni A, Gentili A, Abidin AZ, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. Nature medicine. 2021;27(10):1735-43.

18. Naz S, Phan KT, Chen YPP. A comprehensive review of federated learning for COVID-19 detection. International Journal of Intelligent Systems. 2022;37(3):2371-92.

19. Al-Kahtani MS, Khan F, Taekeun W. Application of Internet of Things and Sensors in Healthcare. Sensors. 2022;22(15):5738.

20. Sivasankari B, Varalakshmi P. Blockchain and IoT Technology in Healthcare: A Review. Challenges of Trustable AI and Added-Value on Health: Proceedings of MIE 2022. 2022;294:277.

21. Wallin JA. Bibliometric methods: pitfalls and possibilities. Basic & clinical pharmacology & toxicology. 2005;97(5):261-75.