Mini Review

Open Access CrossMark

Research data: basis of future research analytics

Abstract

Living in an era of data explosion, there is an urgent need for making high-quality scholarly research data available for technologies like Artificial Intelligence(AI) to make something more out of it, other than publications. AI being portrayed as the sentience of machines, with more computational power and advances analytical methods has the potential to deliver personalised health care to individuals with greater probability. Although various data points such as electronic medical records, insurance company records, financial records, etc. can be used for such purposes research data remains the key building block based on which the deep learning algorithms can generate meaningful results. Hence, we discuss the significance, need and various methods for making high-quality scholarly research data available for the future to identify intricate and potentially unknown patterns hidden in them by harnessing the potential of AI. We also recommend research data deposition to be made a necessary pre-requisite before the publication of the results derived out of them.

Volume 10 Issue 2 - 2021

Sathish Muthu,[1,2] Madhan Jeyaraman,[2,3] Girinivasan Chellamuthu[2]

[1]Department of Orthopaedics, Government Medical College & Hospital, Dindigul, Tamil Nadu, India

[2]Research Associate, Orthopaedic Research Group, Coimbatore, Tamil Nadu, India

[3]Department of Orthopaedics, School of Medical Science & Research, Sharda University, New Delhi, India

Correspondence: Sathish Muthu, Department of Orthopaedics, Government Medical College & Hospital, Dindigul, Tamil Nadu, India, Tel +91 9600856806, Email drsathishmuthu@gmail.com

Received: December 08, 2020 | **Published:** April 08, 2021

Introduction

Any research output that has been collected, observed and analysed to arrive at a result constitutes the research data.[1] Research data goes beyond the entries made in the spreadsheet. Research data includes the raw inputs, processed data, algorithms, protocols, methods, materials, photographs, etc. It is an essential component of research which is need for the reproduction of a given scientific output. Living in an era of data explosion, there is an urgent need for making high-quality scholarly research data available for technologies like Artificial Intelligence(AI) to make something more out of it, other than publications. AI being portrayed as the sentience of machines, with more computational power and advances analytical methods has the potential to deliver personalised health care to individuals with greater probability.[2] Although various data points such as electronic medical records, insurance company records, financial records, etc. can be used for such purposes research data remains the key building block based on which the deep learning algorithms can generate meaningful results.[3,4] Hence, we discuss the significance, need and various methods for making high-quality scholarly research data available for the future to identify intricate and potentially unknown patterns hidden in them by harnessing the potential of AI.

Research data management

If one analyses the lifecycle of research data, it is neither static nor isolated. The lifecycle of data does not end with its creation, processing, analysis, representation, and publication but it also includes its preservation and availability for verification and reuse in the future as shown in Figure 1.[5] Data management is an efficient way of handling data along its lifecycle to ensure that the data is collected in a way that is understandable so that it can be used by other researchers to test its validity or to re-analyse from a different perspective. The most critical part of the data management is to preserve and make the data available to others by providing access to data deposition made in discipline-specific repositories.Publications are no longer considered as the output of research, but data in itself is being considered as the important output of research. This blurs the line between publications and data which leads to an increasing number of data journals like Scientific Data from Nature,[6] GigaScience from Oxford Academic[7] which remain as data banks for future research analysis.

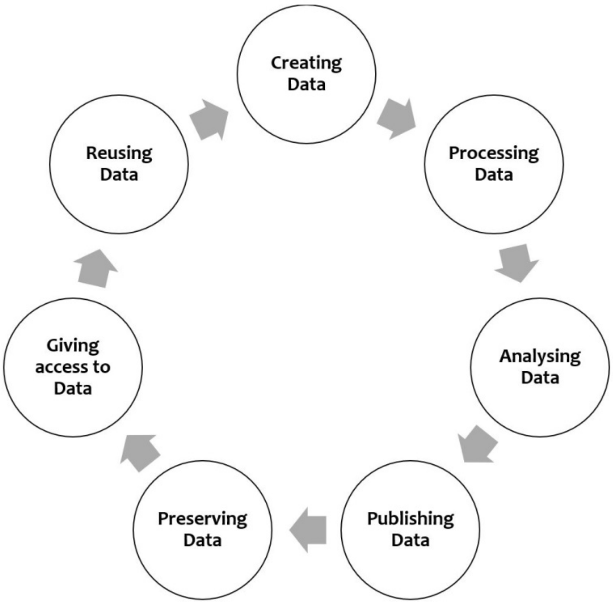

Figure 1 Data life cycle.

Why deposit research data?

Orthopaedic surgery has evolved with various techniques and technologies developed in recent decades but high-quality evidence to support their usage in everyday practise is lacking due to various ethical and cost concerns. This gives the necessary ground for solutions derived from deep learning approaches of AI. Although multiple clinical data registries maintain high-quality health care data, essential data on the current research that is being published remains a critical element of analysis through AI. Hence, Research data sharing remains the way forward for scientific progress. Research data sharing allows for the validation, replication, re-analysis, re-interpretation, new analysis or inclusion into meta-analysis. It increases the reproducibility and credibility of the research.[8] It increases the value of the investment made in funding scientific research. It also reduces the burden of the authors in managing data access requests. By linking the research data in the associated publication, it increases the visibility and ensures greater recognition.

Principles of human research data deposition

Appropriate ethical committee approval along with patient consent following all applicable local laws must be sought before sharing patient-related data in public domains.[9] Data sharing should never compromise participant privacy. Data that result in the identification of the participant such as name, physical address, birth dates, contact information, etc. should not be included in research data deposition. Even data that does not directly identify the participant may also be inappropriate when they are used in combination such as data from a small group of vulnerable populations or private groups. Steps necessary to de-identify the research data towards the participant identification is always recommended. Various guidelines have been put forth on these grounds by national and international agencies on research data deposition.[10–16]

Methods of data deposition

Data repository

All the research data and the related metadata for the reported findings are better managed by deposition in a data repository. It can be deposited in a specialty-specific repository that accepts specific structured data types or cross-disciplinary repositories that accepts various data types. However, generalisation from cross-disciplinary repositories remains challenging making specialty-specific repositories as the ideal mode of data deposition.[17]

Supporting files

Although repositories are the preferred method of research data availability, authors can also provide the research data as a supporting file linked to the research publication. Authors should use formats that are standard to their discipline to allow wide dissemination.

Choosing a data repository

For the management of research data, data repositories remain the most preferred method of data deposition. FAIR data principles provide the necessary guidelines in the selection of an ideal data repository which is a critical step to achieve the goals of data deposition.[18]

Findable

To make sure others can find our data, we must ensure that it is hosted by a stable recognised repository which assigns a globally unique persistent identifier such as DOI to your research data so that it is findable for future human and machine use. To ensure the findability of our research data, all the necessary fields that contribute to the metadata records must be filled.

Accessible

Granting access to medical research data has its ethical concerns and hence open sharing may not be possible all the time. However, specific research data supporting the publication can be made available with an appropriate level of security.

Interoperable

For an integrative analysis by humans and machines, data deposition must be made in an open file format using standard vocabulary. Specific file formats and vocabularies are dictated by disciple-specific repositories to maintain the interoperability of the research data.

Reusable

Research data that is made findable, accessible and interoperable is always fit for reuse. Sometimes additional documentation may be required alongside to make the data understandable and thus reusable to anyone who is not familiar with the data that is being provided. The sample list of discipline-specific and inter-disciplinary repositories available for data deposition for research involving the orthopaedic spine surgery is shown in Table 1. There are various registries available like FAIRsharing[19] and re3data[20] which give information on the data repositories available based on the discipline of choice along with the list of journals supporting their use.

Table 1 Sample list of discipline specific and inter-disciplinary repositories available for data deposition for research involving spine

Data Repository	How article and data are linked	Dataset Size Limits	Repository URL
Discipline Specific Repository - Spine:			
ClinicalTrials.gov (NCT)	Authors should specify NCT accession numbers	1 GB per dataset	http://clinicaltrials.gov/
Neuroimaging Informatics Tools and Resources Collaboratory (NITRC)	Authors should specify NITRC accession numbers.	Image Repository	http://www.nitrc.org/
Neuroscience Information Framework (NIF)	Authors should mention Research Resource IDentifier (RRID)	Image/Dataset Repository	http://www.neuinfo.org/
OpenNeuro	Authors should specify OpenNeuro accession numbers	Image Repository	http://www.openneuro.org
Inter-disciplinary Repository:			
Mendeley Data	Mendeley Data banners will be shown on ScienceDirect when the repository has data for the article	10 GB per dataset	https://data.mendeley.com/
Harvard Dataverse	Some journals have a dedicated Dataverse repository set up for authors to upload their data that belongs with the article. Authors should include the dataset DOI in the article.	10 GB per dataset	https://dataverse.harvard.edu/

Table Continued...

Data Repository	How article and data are linked	Dataset Size Limits	Repository URL
Figshare	Authors should include the dataset DOI in the article.	1 TB per dataset	http://figshare.com/
Dryad Digital Repository	Authors should include the dataset DOI in the article.	300 GB per dataset	https://datadryad.org/stash
Open Science Framework	Authors should include the dataset DOI in the article.	5 GB per dataset	https://osf.io/
Zenodo	Authors should include the dataset DOI in the article.	50 GB per dataset	https://zenodo.org/

URL, Uniform resource locator; GB, Giga Byte; TB, Tera Byte

Concerns in data deposition

If the data is restricted for public deposition due to ethical or security reasons, only restricted access can be given to the researchers and reviewers under specific conditions. Research data deposition has its data protection issues which need to be given adequate attention. Sharing of data over the internet may be a concern when it is too large to be feasibly hosted by a repository which needs to be sorted on a case to case basis.[21] In case of data obtained from a third party, further restrictions apply to the availability of the research data.

Challenges in orthopaedic surgery

Even if data deposition is made mandatory for research publications in orthopaedics, there are certain challenges ahead for making them useful for AI-based analysis. First, uniform appropriately labelled dataset templates have to be established for universal use in orthopaedic surgery research. [22] Second, research in the orthopaedics frequently involves image-based analysis which needs manual labeling of the data for classification for machine learning to occur.[23–25] Although unsupervised algorithms were developed to allow the ML models to analyse and classify such image-based data, with poor quality and quantity of the training datasets for the ML algorithms, there are chances of erroneous decisions thereby reducing the validity of their decisions.[26] Finally, many AI-based algorithms are trained and validated for use within an institution and hence its transferability into for universal application so that it undergoes continuous learning and evolution from the new datasets available remains a challenge.[27]

Directions for the future

With the advancement in the field of Artificial Intelligence(AI), with appropriate research data availability, computer-based algorithms can perform intricate and extremely complex analysis to detect potential previously unknown patterns in them. Machine learning(ML) is one such advancement of AI which is based on artificial neural networks that involve the construction and application of statistical algorithms that make observations from the existing data and continuously learn to create a predictive model based on the data. There are various ML-based models developed to assist surgeons in decision making[28–30] and predicting outcomes of treatment offered and estimating their probability of failure[31–34] on an individual basis. The potential and the probability of the generated conclusions are increased with the availability of baseline high-quality research data.[35] Hence, the deposition of research data must be considered as an essential step in every research publication to extend the scope of the research beyond its limits.

Conclusion

With the continuous evolution in the computational capacity of AI, high-quality scholarly data remain the essential prerequisite for increasing the validity of their predictive outcomes. While this technology is still in its infancy, preventing its full-fledged integration and implementation into the health care system, making the necessary baseline dataset by research data deposition would help to harness their potential towards patient care in near future. Hence, we recommend research data deposition to be made a necessary pre-requisite before the publication of the results derived out of them.

Acknowledgments

The authors esteem and appreciate Miss Helen Nmesomachi Wokem for having pain stakingly done the word processing of the manuscript and other formatting necessary for the finishing.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Funding

None.

References

1. Surkis A, Read K. Research data management. *J Med Libr Assoc.* 2015;103(3):154–6.
2. Han X-G, Tian W. Artificial intelligence in orthopedic surgery: current state and future perspective. *Chinese Medical Journal.* 2019;132(21):2521–2523.
3. Ristevski B, Chen M. Big data analytics in medicine and healthcare. *J Integr Bioinform.* 2018;15:20170030.
4. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deeplearning for electronic health records. *NPJ Digit Med.* 2018;1:18.
5. Griffin PC, Khadake J, LeMay KS, et al. Best practice data life cycle approaches for the life sciences. *F1000Res.* 2017;6:1618.
6. Scientific Data. [Internet] Nature. Accessed on April10, 2020.
7. GigaScience Oxford Academic [Internet]. OUP Academic. Accessed on April10, 2020.
8. Alsheikh-Ali AA, Qureshi W, Al-Mallah MH, et al. Public availability of published research data in high-impact journals. *PLoS One.* 2011;6(9):e24357.
9. Savage CJ, Vickers AJ. Empirical study of data sharing by authors publishing in PLoS journals. *PLoS One.* 2009;4:e7078.
10. NIH. National institutes of health plan for increasing access to scientific publications and digital scientific data from NIH funded scientific research. 2015. Accessed on April10, 2020.
11. Welcome Policy on data, software and materials management and sharing 2017. Accessed on April10, 2020.
12. Medical Research Council UK Data Sharing Policy. Accessed on April10, 2020.

13. BMC's policy on Open Data. Accessed on April10, 2020.

14. BMJ Data sharing. Accessed on April10, 2020.

15. PLoS Collection Open Data. Accessed on April10, 2020.

16. Taichman DB, Sahni P, Pinborg A, et al. Data sharing statements for clinical trials: a requirement of the International Committee of Medical Journal Editors. *PLoS Med.* 2017;14(6):e1002315.

17. Banzi R, Canham S, Kuchinke W, et al. Evaluation of repositories for sharing individual-participant data from clinical studies. *Trials*. 2019;20(1):169.

18. Wilkinson MD, Dumontier M, Aalbersberg J, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*. 2016;3(1):1–9.

19. Sansone S-A, McQuilton P, Rocca-Serra P, et al. FAIRsharing as a community approach to standards, repositories and policies. *Nat Biotechnol.* 2019;37(4):358–67.

20. Pampel H, Vierkant P, Scholze F, et al. Making research data repositories visible: the re3data.org Registry. *PLoS One.* 2013;8(11):e78080.

21. Ienca M, Ferretti A, Hurst S, et al. Considerations for ethics review of big data health research: A scoping review. *PLoS One*. 2018;13(10):e0204937.

22. Austin CC, Brown S, Fong N, et al. Research data repositories: review of current features, gap analysis, and recommendations for minimum requirements. IASSIST quarterly: winter 2015; International Association for Social Science, Information Services, and Technology.

23. Antani SK, Long LR, Thoma GR. Content-based image retrieval for large biomedical image archives. *Stud Health Technol Inform.* 2004;107(Pt 2):829–833.

24. Liu R, Wang Y, Baba T, et al. SVM-based active feedback in image retrieval using clustering and unlabeled data. *Pattern Recognition*. 2008;41:2645–2655.

25. Rahman MM, Bhattacharya P, Desai BC. A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback. *IEEE Trans Inf Technol Biomed.* 2007;11:58–69.

26. Shah A, Conjeti S, Navab N, et al. Deeply learnt hashing forests for content based image retrieval in prostate MR images. Medical Imaging 2016: Image Processing. Bellingham, WA*: The International Society for Optics and Photonics (SPIE)*. 2016;9784.

27. van Hooff ML, van Loon J, van Limbeek J, et al. The Nijmegen decision tool for chronic low back pain. Development of a clinical decision tool for secondary or tertiary spine care specialists. *PLoS One.* 2014;9:e104226.

28. Tan WK, Hassanpour S, Heagerty PJ, et al. Comparison of natural language processing rules-based and machine-learning systems to identify lumbar spine imaging findings related to low back pain. *Acad Radiol.* 2018;25:1422–1432.

29. Hopkins BS, Weber KA 2nd, Kesavabhotla K, et al. Machine learning for the prediction of cervical spondylotic myelopathy: a post hoc pilot study of 28 participants. World Neurosurg. 2019;127:e436-e442.

30. Huber FA, Stutz S, de Martini IV, et al. Qualitative versus quantitative lumbar spinal stenosis grading by machine learning supported texture analysis—experience from the LSOS study cohort. *Eur J Radiol.* 2019;114:45-50.

31. Kim JS, Arvind V, Oermann EK, et al. Predicting surgical complications in patients undergoing elective adult spinal deformity procedures using machine learning. *Spine Deform.* 2018;6: 762–770.

32. Merali ZG, Witiw CD, Badhiwala JH, et al. Using a machine learning approach to predict outcome after surgery for degenerative cervical myelopathy. *PLoS One.* 2019;14: e0215133.

33. Shah AA, Karhade AV, Bono CM, Harris MB, Nelson SB, Schwab JH. Development of a machine learning algorithm for prediction of failure of nonoperative management in spinal epidural abscess. *Spine J.* 2019;19(10):1657–1665.

34. Karhade AV, Shah AA, Bono CM, et al. Development of machine learning algorithms for prediction of mortality in spinal epidural abscess. *Spine J.* 2019;19(12):1950–1959.

35. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. *N Engl J Med.* 2018;378:981–983.