frontiers in Reproductive Health

PERSPECTIVE
published: 16 March 2022
doi: 10.3389/frph.2022.800760

Challenges of Utilizing Medical Big Data in Reproductive Health Research

Tianyu Dong[1], *Min Zhu*[2], *Rui Li*[2] *and Xu Wang*[3]*

[1] Tripod (Nanjing) Clinical Research Co., Ltd., Nanjing, China, [2] Department of Health IT Solution, Shanghai Synyi Medical Technology Co., Ltd., Shanghai, China, [3] Department of Endocrinology, Children's Hospital of Nanjing Medical University, Nanjing, China

In the background of the "Three-Child Policy" introduced by the Chinese government, reproductive health has become one of the most important public health issues. With the promotion of digitization management of medical care institutions for women and children in the country, there will be chances to acquire medical big data of obstetrics and pediatrics. Here the authors are presenting their opinions on the challenges of the management and utilization of reproductive big data.

Keywords: reproductive health, clinical research, big data, artificial intelligence, children, women

OPEN ACCESS

Edited by:
Lida Chatzi,
University of Southern California,
United States

Reviewed by:
Sangappa B. Chadchan,
Washington University in St. Louis,
United States

***Correspondence:**
Xu Wang
sepnine@njmu.edu.cn

Specialty section:
This article was submitted to Reproductive Epidemiology, a section of the journal Frontiers in Reproductive Health

Received: *24 October 2021*
Accepted: *04 February 2022*
Published: *16 March 2022*

Citation:
Dong T, Zhu M, Li R and Wang X (2022) Challenges of Utilizing Medical Big Data in Reproductive Health Research. Front. Reprod. Health 4:800760. doi: 10.3389/frph.2022.800760

INTRODUCTION

With the further understanding of Developmental Origins of Health and Disease (DOHaD) theory, early-life is considered as a key period of the whole life health management. Substantial research data have demonstrated the association between early-life and long-time health (1), which emphasizes the particular importance of reproductive health. In the background of the "Three-Child Policy" introduced by the Chinese government, reproductive health has become one of the most important public health issues.

The mode of medical service has changed with the introduction of the policy of accelerating digital development in China. The concept of "Medical Big Data" has come into the scope of the public due to the promising values for improving the quality of medical care and conducting clinical research (2). Health-care professionals, clinical researchers and even the politicians are thinking of how these data should be acquired, managed and used. With the digital development and transformation of the Information System in medical institutions for women and children, medical data of reproductive health need to be fully understood and handled in the context of "Medical Big Data" (3).

CHALLENGES OF THE DATA MANAGEMENT

Usually, clinical information from the medical institutions for women and children's health include medical records, medical orders, nursing, testing, imaging, and pathology, etc. From the perspective of the data manager, the raw information from different functional units should be extracted, cleaned, linked and stored in the cloud of a single data management platform to generate standardized and structured datasets. As the volume, velocity, veracity and variety characters of increasing medical data, the medical engineering and artificial intelligence (AI) techniques are required in large-scale studies (4). There are several challenges during the medical data governance process: data integrity and extraction stage need to improve the accuracy and data precision during; disease data sets define and normalization can improve medical data efficiency; Natural Language

Processing (NLP) technologies help to access and abstract from unstructured data; the machine-learning algorithms to reach to those converted structure data to detect diseases and perform more precision studies (5). NLP is the mainly technology to abstract precision terms from narrative clinical reports and medical big data. The "concepts" need to be defined to recognize the terms. According to different study objectives, the manuscript can be used to the appropriate research field (6). The named entity recognizer (NER) can identify those concepts in clinical free-text notes. Then a relation extraction (RE) method can identify the relations between them. For NER, there has several deep learning-based approaches with embedding various to influence the different results (7). The methods and evaluation will be conducted with mimic data.

Another prominent problem for medical data could be the solution to ensuring the data consistency, integrity and accuracy, as the medical records from a single individual are usually collected from multiple sources especially for children. Unified coding is a prerequisite for matching information from different medical institutions (e.g., records from women's health institutions to children's hospitals) and/or functional units within an institution. Parents' IDs are used as the only identification for babies as they usually do not have their names or IDs at birth. They are also used to track the family history and should be linked with future children's IDs. Identifying information should be uniquely coded, and effective unified coding helps with integration of information according to the flow direction and logical relationship when dealing with data format, repetition, attribute value error, inconsistency and other problems. Confidentiality of private information is of great importance in the modern age (8). Medical records should be anonymized to protect the privacy and information security of patients. On the premise of strict hardware guarantee, the maintenance system and operation standard need to be formulated. The database access authority, administrator authority (7) and docking user authority should be set. Identification information is only assessable to qualified staff and should be stored separately with medical data. It is necessary to monitor the operation of users in order to effectively prevent the human errors leading to unexpected information disclosure.

UTILIZATION OF MEDICAL BIG DATA

The utilization of Medical Big Data is still in its early development period in China. It is recommended that clinical researchers cooperate with clinicians to generate good scientific research questions. Clinicians are familiar with the problems in the area of reproductive health that need to be solved, and clinical researchers, including epidemiologists and biological statisticians, can provide valuable suggestions on the feasibility and performance of the study. The database should be user friendly for searching for information and summarizing subsets of the whole dataset. Reproductive health data are collected from parents-child pairs, which makes the datasets very complex and hard to explore the causality. Traditional statistical methods such as logistic regression and survival analysis are still very useful, but sometimes cannot meet the needs of mining complex data. The values of machine learning in data mining have been demonstrated in plenty of previous epidemiological studies (9–11). Novel tools are warranted to model the multivariate information. Taking the advantages of machine learning and big data, researchers are able to obtain robust results and simulated extrapolation in more common populations.

DISCUSSION

Reproductive health is always a concern in modern societies due to the aging population. Great attention is paid to women and children's health in China. Several large birth cohorts have been constructed in the country in recent years. However, the utilization of unstructured medical data are still under development. With the promotion of digitization management of medical care institutions for women and children, there will be chances to acquire medical big data of reproductive health. These data could be well used for researches in the context of good data standardization. AI techniques will help with both the data structuring and data analysis/interpretation. The accurate identification of babes and their parents is an important issue of ensuring the data consistency and integrity. Confidentiality of private information should be protected seriously especially for children. In addition to the collection of medical records, biobanks are promising for further development in the area. Biobanks for samples collected from parent-child pairs enables the employment of advanced systems biology techniques and special study design for discovery of intergeneration effects at the molecular levels. These precious resources will contribute to future reproductive researches, and greatly promote the health care of women and children.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

XW designed the work and made critical revisions on the manuscript. TD drafted the manuscript. MZ and RL made critical revisions on the manuscript. All authors contributed to the article and approved the submitted version.

REFERENCES

1. Gluckman PD, Hanson MA, Cooper C, Thornburg KL. Effect of *in utero* and early-life conditions on adult health and disease. *N Engl J Med*. (2008) 359:61–73. doi: 10.1056/NEJMra0708473
2. Zhang L, Wang H, Li Q, Zhao MH, Zhan QM. Big data and medical research in China. *BMJ*. (2018) 360:j5910. doi: 10.1136/bmj.j5910
3. Jaddoe VWV, Felix JF, Andersen AN, Charles MA, Chatzi L, Corpeleijn E, et al. The LifeCycle Project-EU Child Cohort Network: a federated analysis infrastructure and harmonized data of more than 250,000 children and parents. *Eur J Epidemiol*. (2020) 35:709–24. doi: 10.1007/s10654-020-00662-z
4. Wang L, Alexander CA. Big data analytics in medical engineering and healthcare: methods, advances and challenges. *J Med Eng Technol*. (2020) 44:267–83. doi: 10.1080/03091902.2020.1769758
5. Hofer IS, Halperin E, Cannesson M. Opening the black box: understanding the science behind big data and predictive analytics. *Anesth Analg*. (2018) 127:1139–43. doi: 10.1213/ANE.0000000000003463
6. Barco TL, Kuchenbuch M, Garcelon N, Neuraz A, Nabbout R. Improving early diagnosis of rare diseases using Natural Language Processing in unstructured medical records: an illustration from Dravet syndrome. *Orphanet J Rare Dis*. (2021) 16:309. doi: 10.1186/s13023-021-01936-9
7. Alfattni G, Belousov M, Peek N, Nenadic G. Extracting drug names and associated attributes from discharge summaries: text mining study. *JMIR Med Inform*. (2021) 9:e24678. doi: 10.2196/24678
8. McMahon AW, Cooper WO, Brown JS, Carleton B, Doshi-Velez F, Kohane I, Goldman JL, Hoffman MA, Kamaleswaran R, Sakiyama M, Sekine S, Sturkenboom M, Turner MA, Califf RM. Big Data in the Assessment of Pediatric Medication Safety. Pediatrics (2020) 145: doi: 10.1542/peds.2019-0562
9. McIntosh C, Conroy L, Tjong MC, Craig T, Bayley A, Catton C, et al. Clinical integration of machine learning for curative-intent radiation treatment of patients with prostate cancer. *Nat Med*. (2021) 27:999–1005. doi: 10.1038/s41591-021-01359-w
10. Mbatchou J, Barnard L, Backman J, Marcketta A, Kosmicki JA, Ziyatdinov A, et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet*. (2021) 53:1097–103. doi: 10.1038/s41588-021-00870-7
11. D'Ascenzo F, De Filippo O, Gallone G, Mittone G, Deriu MA, Iannaccone M, et al. Machine learning-based prediction of adverse events following an acute coronary syndrome (PRAISE): a modelling study of pooled datasets. *Lancet*. (2021) 397:199–207. doi: 10.1016/S0140-6736(20)32519-8

Conflict of Interest: TD was employed by Tripod (Nanjing) Clinical Research Co., Ltd. MZ and RL were employed by Shanghai Synyi Medical Technology Co., Ltd.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.