Core Issues and Solution Strategies for Data-Driven Research on Patterns of Syndrome Differentiation and Treatment in Traditional Chinese Medicine: Postprint

Authors: Zhen Qian, Zhu Rong, Wang Zhongrui, Cui Weifeng, Yan Shuxun, Shao Mingyi, Yu Haibin, Fu Yu, Fu Yu

Date: 2023-07-18T00:00:00+00:00

Abstract

Pattern differentiation and treatment determination constitutes the core diagnostic and therapeutic thinking in Traditional Chinese Medicine (TCM) and represents the key determinant of clinical efficacy. At present, research based on clinical data serves as the principal methodology for exploring TCM syndrome-treatment patterns; however, such approaches have failed to genuinely and comprehensively dissect the intrinsic relationships among the critical factors of "disease - pattern - formula - herb - efficacy," consequently yielding research outcomes of limited clinical value. Therefore, we systematically delineate core challenges including the suboptimal alignment between electronic medical records and clinical research requirements, the influence of data governance on data accuracy, and the inadequacy of existing data analysis methodologies in excavating TCM syndrome-treatment patterns. Within a data-driven framework, we propose the establishment of a big data platform for TCM clinical research and the development of artificial intelligence-centric data governance and analytical technologies, thereby realizing the integration of clinical practice and scientific research, furnishing novel concepts and methodologies for investigating TCM syndrome-treatment patterns, and advancing the development of Traditional Chinese Medicine.

Full Text

Core Problems and Solving Strategies of the Research on the Law of TCM Syndrome and Treatment Based on Data Driven

ZHEN Qian[1], ZHU Rong[1], WANG Zhongrui[1], CUI Weifeng[2], YAN Shuxun[3], SHAO Mingyi[3], YU Haibin[3], FU Yu[3]*

[1]Henan University of Traditional Chinese Medicine, Zhengzhou 450046, China
[2]Henan Integrated Hospital of Traditional Chinese and Western Medicine, Zhengzhou 450003, China
[3]The First Affiliated Hospital of Henan University of Chinese Medicine, Zhengzhou 450000, China
*Corresponding author: FU Yu, Associate chief physician; E-mail: kybfuyu@126.com

Abstract

Treatment based on syndrome differentiation constitutes the core diagnostic and therapeutic thinking of Traditional Chinese Medicine (TCM) and is critical to determining clinical efficacy. Currently, research based on clinical data represents the primary method for exploring the law of TCM syndrome and treatment. However, these studies have not truly and comprehensively analyzed the intrinsic relationships among the key factors of "disease-syndrome-formula-medicine-effect," resulting in research findings with low clinical value. Therefore, we systematically examined the core problems of poor matching between electronic medical records (EMR) and clinical research requirements, the impact of data governance on data accuracy, and the difficulty of current data analysis methods to uncover TCM syndrome and treatment patterns. In the context of data-driven research, we propose establishing a big data platform for TCM clinical research and developing AI-centered data governance and analysis technologies to achieve the integration of clinical practice and research. This approach provides new ideas and methods for studying the law of TCM syndrome and treatment, thereby promoting the development of TCM.

Keywords: Traditional Chinese medicine therapy; Law of syndrome and treatment; Data driven; Data mining; Electronic medical record; Core problems; Solving strategies

---

1. EMR as an Important Carrier for Research on the Law of TCM Syndrome and Treatment

Treatment based on syndrome differentiation is considered the core diagnostic thinking of TCM, representing the fundamental principle guiding clinical practice and the key to determining therapeutic efficacy. However, due to different

syndrome differentiation systems, non-standardized terminology, complex clinical situations, and the lack of holistic assessment of disease occurrence and development—coupled with susceptibility to false appearances and subjective factors—syndrome differentiation often varies among practitioners. This creates a situation of "different treatments for different individuals," limiting the inheritance and innovation of TCM. TCM follows a cyclical, spiraling development pattern of "practice-theory-guiding practice-perfecting theory." Research on the law of TCM syndrome and treatment based on clinical data can uncover implicit knowledge of syndrome differentiation and treatment patterns, summarizing diagnostic and therapeutic principles in response to holistic disease changes. This helps enhance the initiative and predictability of clinical syndrome differentiation and improve therapeutic effectiveness.

Currently, with the increasing digitization of healthcare information, TCM EMRs not only reflect the comprehensive application of principles, methods, formulas, and medicines but also surpass other data sources in terms of authenticity and reliability. Therefore, research on the law of TCM syndrome and treatment based on EMRs can discover hidden therapeutic knowledge. However, existing studies have not truly dissected the intrinsic relationships among the key factors of "disease-syndrome-formula-medicine-effect," yielding low clinical value. This deficiency stems from a series of underlying problems.

2. Core Problems in Data-Driven Research

2.1 Poor Matching Between EMR and Clinical Research Requirements EMRs were originally designed to facilitate clinical workflows and hospital management, inevitably creating mismatches with clinical research requirements. First, the effectiveness of clinical treatment forms the foundation for data-driven TCM research. Currently, there is no suitable method for evaluating clinical efficacy in TCM, making it difficult to scientifically and objectively interpret the actual therapeutic effects documented in EMRs. Consequently, the applicability of research findings based on EMRs for guiding clinical practice remains questionable. Second, although there has been increasing emphasis on establishing integrated clinical-research platforms in recent years, several objective factors persist regarding clinical data: (1) Low quality: Due to factors such as lack of standardized TCM clinical terminology, incomplete documentation by healthcare staff, inadequate quality control by medical institutions, and varying patient conditions, the obtained data information is often not detailed or accurate. (2) Complex structure: Various hospital information system databases—including Laboratory Information System (LIS) data, Picture Archiving and Communication Systems (PACS) data, and Hospital Information System (HIS) data—have not achieved true integration, resulting in complex structures with predominantly unstructured text information that hinders data processing and application. (3) Information "islanding": Data from different regions, hospital levels, and categories exhibit variations and relative isolation. Combined with the inherent characteristics of non-standardized data that limit sharing, as well

as immature collaboration mechanisms, uneven distribution of interests, and incomplete legal frameworks, data remains fragmented.

2.2 Data Governance Affecting Data Traceability, Completeness, and Accuracy

2.2.1 Data Extraction The process of extracting data from storage modules presents numerous challenges. Variable information across different modules varies in format, values, completeness, and accuracy. Moreover, extraction methods—including manual entry, system export, and electronic data capture (EDC) technology tools—affect data accuracy and security. Simultaneously, the lack of unified, standardized collection guidelines complicates and prolongs the extraction process.

2.2.2 Data Cleaning Currently, there remains no complete, standardized, and effective system of TCM clinical medical terminology standards. Although researchers can refer to textbooks, pharmacopeias, and syndrome guidelines to standardize disease names, syndromes, and herbal medicines, this still leads to data inaccuracies. For example, categorizing "Fa Banxia" (processed pinellia) or "Jiang Banxia" (ginger pinellia) simply as "Banxia" (pinellia), or classifying "vertigo," "headache," and "dementia" under chronic cerebral ischemia inevitably introduces biases into research results. When structuring text data, natural language processing (NLP) technology is primarily employed, but it suffers from labor-intensive processes, sensitivity to data conditions, and poor rule transparency. Additionally, various methods for handling outliers, anomalies, and missing values—such as mean imputation for missing data and smoothing for outliers—each have their own advantages and disadvantages and fail to adequately resolve these issues.

2.3 Difficulty in Analyzing Complex Relationships

2.3.1 Limitations of Data Analysis Methods Themselves Based on constructed databases, data analysis methods are employed to uncover implicit syndrome and treatment patterns. However, current approaches are constrained by the inherent limitations of various algorithms, lack technological innovation, and have not integrated TCM academic thinking. provides a brief summary of the defects in data analysis methods.

2.3.2 Difficulty in Analyzing the Complex System of "Disease-Syndrome-Formula-Medicine-Effect" Complex, non-uniform, asymmetric, and non-additive causal relationships exist between syndromes and their related factors. Meanwhile, treatment methods follow syndrome establishment and change, while formulas derive from methods and are composed of medicines. This flexibility and principled nature of therapeutic methods and formulas reflects the complexity of TCM's dynamic thinking. However,

various data analysis algorithms explore correlations with syndromes based on extracted high-frequency syndrome elements and laboratory indicators, ignoring the fact that these relationships are not simple linear associations and separating the intrinsic connection between treatment methods and syndromes. Additionally, when analyzing core formulas and medicines, it is important to note that high-frequency drugs are not necessarily the core efficacy components in formulas, and the reasonableness of relationships between drugs and therapeutic effects remains questionable. Furthermore, the lack of interdisciplinary talent combining TCM, computer science, and mathematical statistics prevents effective communication and integration across professional domains, leading to fragmentation. Consequently, current data mining technologies remain incapable of analyzing the implicit knowledge patterns of TCM.

3. Solving Strategies

3.1 Establishing a Big Data Platform for TCM Clinical Research In recent years, the state has attached great importance to the analysis and utilization of medical big data, issuing numerous policies to encourage hospital exploration and emphasizing that medical data should originate from patients and serve patients. This "data-driven" concept makes the construction of an integrated clinical-research platform imperative. In implementation, it is necessary to establish a national-level major disease-specific research data management platform, upgrade and optimize clinical systems, integrate HIS, LIS, PACS, and other data, and collect clinical data information in a large-scale, structured, and dynamic manner. Policies should be improved to explore disease-specific data sharing mechanisms and integrate intra- and extra-hospital data sources to achieve interconnectivity. The database should feature high quality with macro-micro and pathophysiological integration, covering multi-level content including disease-symptom-syndrome-formula-medicine-related diagnostic information, clinical biological samples, and multi-omics data. Additionally, to ensure high data quality, a clinical-research integrated EMR quality control system and framework must be established. Based on healthcare workers' emphasis on medical record quality and accurate documentation of patient conditions, standardized electronic medical records should be created and quality-controlled through corresponding clinical-research integration standards. Simultaneously, to standardize TCM terminology, principles of simplification, systematicity, and consensus should be followed, combining the characteristics of TCM terminology to establish conceptual and terminological systems for TCM-related "etiology and pathogenesis, diagnosis, diseases, treatment principles and methods, and formulas," standardizing TCM terms and definitions. These methods should be continuously revised according to social conditions and disciplinary development. Therefore, fundamental data problems can be solved, achieving national-level, high-quality, structured clinical research data that promotes clinical-research integration and facilitates research on the law of TCM syndrome and treatment.

3.2 Developing AI-Centered Data Governance and Analysis Technologies

3.2.1 Strengthening Data Governance Methods For data-driven clinical research scenarios, the core technology of clinical-research integrated platforms is the clinical data governance engine. Every step—from data governance to analysis—requires technological support. In terms of data governance, future trends point toward electronic and intelligent data collection and management. A standardized data collection and entry guide must be specified, strictly followed, and suitable EDC systems or electronic data management tools for TCM should be actively created. Knowledge graphs, with their powerful semantic processing and data organization capabilities, combined with advanced NLP and other AI technologies, will enable automatic transformation of data into structured, standardized, and normalized formats. This fundamentally addresses data quality and structural issues.

3.2.2 Organically Integrating TCM Thinking with AI Existing data analysis methods still have inherent defects. Researchers must correctly understand the performance characteristics and implications of different methods, selecting analysis approaches accurately and appropriately from various research perspectives, or even combining multiple methods to leverage complementary advantages through repeated mining, thereby enhancing result completeness and systematicity. For example, analyzing TCM syndromes from symptom and patient population perspectives using different methods and angles can yield better research results. However, future data analysis methods oriented toward simulating TCM syndrome and treatment thinking require continuous development and innovation as crucial measures for achieving intelligent syndrome differentiation and treatment. Integrating knowledge graphs with TCM characteristics and deep learning in AI, combined with algorithmic modeling of logical rules and their introduction into machine neural networks, enables more orderly and progressively deeper machine knowledge learning and understanding under rule guidance. This promotes the construction and improvement of deep neural network models, advancing AI-centered technology development. Researchers have also proposed dual data-and-knowledge-driven methods, constructing domain-specific knowledge graphs and embedding them into neural networks for training. Examples include embedding semantic network knowledge graphs into Graph Neural Networks (GNN) for fusion research, and implementing multi-relational reasoning for disease diagnosis based on laboratory knowledge graphs and Logistic Regression (LR) algorithms. These approaches provide ideas and references for developing targeted TCM syndrome differentiation and treatment knowledge graphs and big data-fused AI algorithm models, achieving data analysis technology that simulates the human brain and fully integrates TCM thinking for intelligent syndrome differentiation and treatment.

Overall, future efforts will create a clinical-research integrated platform with high-quality, multi-dimensional data and intelligent research analysis, requiring

collaborative efforts from interdisciplinary, applied, and cross-disciplinary talents in TCM, clinical research management, bioinformatics, computer science, and statistics. This will emphasize comprehensive and coordinated development of medical care and scientific research, forming a professional talent team.

4. Thoughts and Prospects

In today's big data era, as data resources have become a national strategic resource, China continuously advocates improving data utilization efficiency, promoting data empowerment, and enhancing data value. However, some studies merely "speak with data" through simplistic data acquisition and analysis for decision-making and action—processes that are insufficiently automated, intelligent, or valuable. In contrast, data-driven approaches provide guiding decisions and actions through more automated, intelligent, and scientific data analysis and processing, continuously generating positive feedback loops and promoting decision optimization. Centered on lean analysis and data closed-loop concepts, this ultimately forms a decision-making and action system based on data. Based on this, researchers use continuously emerging clinical data to explore the law of TCM syndrome and treatment, representing both the inheritance of TCM theoretical thinking and its continuous exploration and innovation. However, due to issues such as low-quality TCM clinical data, complex structures, fragmentation, and limitations in data governance and analysis technologies, results cannot comprehensively and accurately cover the complex relationships among disease, symptom, formula, and medicine, yielding low value for guiding clinical syndrome differentiation and treatment.

This paper systematically examined existing core problems and identified key difficulties in using clinical data for scientific research. Therefore, following the principle of clinical-research integration, we propose establishing a big data platform for clinical research and developing AI-centered data governance and analysis technologies, providing ideas and directions for building a national research platform and technological innovation. However, most medical institutions currently lack the platform conditions and professional research teams needed to conduct standardized clinical research, inevitably facing challenges in data resource integration, sharing, management, and analysis. Therefore, achieving clinical-research integration is a long-term and arduous process involving the evolution and optimization of many core aspects of clinical practice, medical management, and technology. This requires national government attention to the comprehensive development of medical care and scientific research, continuous efforts from medical institutions and interdisciplinary talents, coordinated cooperation, and the gradual establishment of mechanisms and platforms conducive to translating research findings into clinical application, ultimately realizing TCM inheritance and innovation.

Author Contributions: Zhen Qian was responsible for conceptualization and writing; Zhu Rong and Wang Zhongrui conducted literature retrieval; Cui Weifeng, Yan Shuxun, Shao Mingyi, and Yu Haibin guided the research think-

ing; Fu Yu was responsible for the overall article, proposed the research ideas, revised the manuscript, and performed quality control and proofreading.

Conflict of Interest: The authors declare no conflicts of interest.

References

[1] Liu YF, Sun MY, Yao HZ, et al. Current status and reflections on the application of big data technology in the field of Traditional Chinese Medicine [J]. Chinese Journal of Evidence-Based Medicine, 2018, 18(11): 1180-1185. DOI:10.7507/1672-2531.201804072.

[2] Xu JY, Lou ZH, Deng ZYT, et al. Analysis of the connotation and system construction of TCM treatment theory [J]. China Journal of Traditional Chinese Medicine and Pharmacy, 2023, 38(1): 63-66.

[3] Leng YL, Gao H, Fu XX, et al. Research progress on clinical research methods for TCM syndromes [J]. China Journal of Traditional Chinese Medicine and Pharmacy, 2021, 36(10): 6002-6005.

[4] Fu Y, Shao MY, Zhao RX, et al. Discussion on evaluation methods for clinical efficacy of Traditional Chinese Medicine based on TCM evidence [J]. Journal of Traditional Chinese Medicine, 2020, 61(13): 1124-1129. DOI:10.13288/j.11-2166/r.2020.13.004.

[5] Sheng H. Research on key issues of TCM electronic medical records in the context of big data [D]. Jinan: Shandong University of Traditional Chinese Medicine, 2017.

[6] Shao MY, Liu BY, Xie Q, et al. Discussion on the development status and trends of clinical research data in Traditional Chinese Medicine [J]. Modernization of Traditional Chinese Medicine and Materia Medica-World Science and Technology, 2015, 17(8): 1743-1747. DOI:10.11842/wst.2015.08.027.

[7] Wang W, Liu YM, Tan J, et al. Concepts, planning, and research database construction of retrospective database studies [J]. Chinese Journal of Evidence-Based Medicine, 2018, 18(2): 155-160.

[8] Zhao YX, Shao MY, Chen XQ, et al. Discussion on the authenticity of real-world data and its influencing factors [J]. Journal of Traditional Chinese Medicine, 2021, 62(4): 303-306, 311. DOI:10.13288/j.11-2166/r.2021.04.007.

[9] Fu Q, Mao C. Epidemiological research driven by health medical big data: opportunities and challenges [J]. Chinese Journal of Disease Control & Prevention, 2023, 27(2): 125-126, 237. DOI:10.16462/j.cnki.zhjbkz.2023.02.001.

[10] Wang W, Gao P, Wu J, et al. Technical specifications for constructing research databases based on existing health medical data [J]. Chinese Journal of Evidence-Based Medicine, 2019, 19(7): 779-786.

[11] Hu KR, Gong HZ, Wan X. Research on standard operating procedures for real-world clinical data management: taking syphilis clinical data as an

example [J]. Chinese Journal of Hospital Administration, 2021, 37(9): 761-765. DOI:10.3760/cma.j.cn111325-20210316-00230.

[12] Ren HL, Guo JJ, Sun HX, et al. Reflections on standardization research of medical terminology [J]. Journal of Medical Informatics, 2018, 39(5): 2-7. DOI:10.3969/j.issn.1673-6036.2018.05.001.

[13] Wang YQ, Wang T, Zhang TS, et al. Analysis of syndrome and treatment patterns of Traditional Chinese Medicine in treating post-PCI angina in coronary heart disease based on data mining [J]. Chinese Journal of Gerontology, 2023, 43(4): 776-780. DOI:10.3969/j.issn.1005-9202.2023.04.003.

[14] Hou XF, Wang YL, Xu ZP. Study on the distribution pattern of TCM syndromes in chronic cerebral ischemia [J]. Chinese Journal of Basic Medicine in Traditional Chinese Medicine, 2023, 29(1): 116-119, 168. DOI:10.19945/j.cnki.issn.1006-3250.2023.01.042.

[15] Bao XY, Huang WJ, Zhang K, et al. A customized method for information extraction from unstructured electronic medical records [J]. Journal of Peking University (Health Sciences), 2018, 50(2): 256-263. DOI:10.3969/j.issn.1671-167X.2018.02.010.

[16] Wu C, Wang ZY, Xu L, et al. Quality control of electronic medical record data based on artificial intelligence [J]. Hospital Administration Journal of Chinese People's Liberation Army, 2021, 28(2): 134-135, 168. DOI:10.16770/J.cnki.1008-9985.2021.02.010.

[17] Wu ZY, Bai KL, Yang LR, et al. A survey on electronic medical record text mining [J]. Journal of Computer Research and Development, 2021, 58(3): 513-527. DOI:10.7544/issn1000-1239.2021.20200402.

[18] Chen ZK, Song X, Gao J, et al. Research progress on TCM diagnosis and treatment based on data mining [J]. Chinese Archives of Traditional Chinese Medicine, 2020, 38(12): 1-9. DOI:10.13193/j.issn.1673-7717.2020.12.001.

[19] Ma MY, Shen L, Wen TC, et al. Application of data mining technology in analysis of TCM diagnosis and treatment data [J]. Chinese Journal of Information on Traditional Chinese Medicine, 2016, 23(7): 132-136. DOI:10.3969/j.issn.1005-5304.2016.07.037.

[20] Xu YL, Sheng MY, Wang Z, et al. Comparative study of several data mining methods for analysis of TCM syndromes [J]. Chinese Journal of Information on Traditional Chinese Medicine, 2019, 26(12): 97-102. DOI:10.3969/j.issn.1005-5304.2019.12.020.

[21] Xu WF, Liu GP, Wang YQ. Application and prospect of multivariate statistical methods in classification and identification of TCM syndromes [J]. Chinese Journal of Information on Traditional Chinese Medicine, 2015, 22(8): 124-128. DOI:10.3969/j.issn.1005-5304.2015.08.039.

[22] Zhao M, Li JS. Evaluation research on TCM service capacity based on principal component analysis and factor analysis [J]. Journal of Jiangxi University of Traditional Chinese Medicine, 2020, 32(6): 103-107.

[23] Zhang LT. Study on TCM syndromes in early onset of coronary heart disease based on factor analysis [D]. Shenyang: Liaoning University of Traditional Chinese Medicine, 2013.

[24] Gong YB, Ni Q, Wang YY. Review of modern methodology for TCM syndrome research (I): Data mining technology for TCM syndromes [J]. Journal of Beijing University of Traditional Chinese Medicine, 2006, 29(12): 797-801. DOI:10.3321/j.issn:1006-2157.2006.12.001.

[25] Tao ZL, Chen HJ. Application progress of data mining in TCM syndrome research [J]. Shanghai Journal of Traditional Chinese Medicine, 2021, 55(6): 91-95. DOI:10.16305/j.1007-1334.2021.1910162.

[26] Lv QL. Fusion of data mining and complex networks and its application in Traditional Chinese Medicine [J]. Chinese Traditional and Herbal Drugs, 2016, 47(8): 1430-1436. DOI:10.7501/j.issn.0253-2670.2016.08.031.

[27] Bai JX, Hu XJ, Xu JT. Research progress on TCM diagnosis and treatment patterns based on complex network technology [J]. Lishizhen Medicine and Materia Medica Research, 2020, 31(9): 2207-2209. DOI:10.3969/j.issn.1008-0805.2020.09.051.

[28] Guo L, Wang YY. Discussion on complex phenomena in TCM syndromes and corresponding research approaches [J]. Chinese Journal of Basic Medicine in Traditional Chinese Medicine, 2004, 10(2): 3-5. DOI:10.3969/j.issn.1006-3250.2004.02.002.

[29] Liu ML, Zhang XY, Ding L, et al. Application and progress of data mining methods in research on TCM compatibility patterns [J]. China Journal of Chinese Materia Medica, 2021, 46(20): 5233-5239. DOI:10.19540/j.cnki.cjcmm.20210303.501.

[30] White T, Blok E, Calhoun VD. Data sharing and privacy issues in neuroimaging research: opportunities, obstacles, challenges, and monsters under the bed [J]. Hum Brain Mapp, 2022, 43(1): 278-291. DOI:10.1002/hbm.25120.

[31] Wu HK, Li XD, Yang F, et al. New exploration of clinical research integrated electronic medical record quality control system [J]. Chinese Journal of Integrative Medicine on Liver Diseases, 2013, 23(3): 181-182. DOI:10.3969/j.issn.1005-0264.2013.03.023.

[32] Zhang SN, Chen LY, Yan SY. Discussion on standardization of TCM terminology [J]. Shanghai Journal of Traditional Chinese Medicine, 2019, 53(6): 7-10.

[33] Deng YZ, Yu J, Liu C, et al. Current status and future trends of clinical data management in the electronic era [J]. Chinese Journal of New Drugs, 2014,

23(8): 879-884.

[34] Chen X, Li YF. Discussion on data mining of famous veteran TCM doctors' experience based on medical record deconstruction [J]. China Journal of Traditional Chinese Medicine and Pharmacy, 2019, 34(6): 2608-2611.

[35] Wu Z, Pan S, Chen F, et al. A comprehensive survey on graph neural networks [J]. IEEE Trans Neural Netw Learn Syst, 2021, 32(1): 4-24. DOI:10.1109/TNNLS.2020.2978386.

[36] Jing J. Construction of a knowledge- and data-driven laboratory artificial intelligence disease diagnosis system [D]. Shanghai: Naval Medical University of Chinese People's Liberation Army, 2021.

[37] He LY, Li XL, Liu Y, et al. Ideas and methods for constructing knowledge graphs of TCM syndrome differentiation and treatment [J]. Journal of Traditional Chinese Medicine, 2017, 58(19): 1650-1653. DOI:10.13288/j.11-2166/r.2017.19.008.

Note: Figure translations are in progress. See original paper for figures.