MEDS Public Health and Preventive Medicine (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/phpm.2023.030201
ISSN 2616-1915 Vol. 3 Num. 2

Processing of Real World Data in Traditional Chinese Medicine

Yujia Li, Jun Li*

Public Health School, Shaanxi University of Chinese Medicine, Xi'an, Shaanxi, 712046, China
**Corresponding author*

Keywords: Chinese medicine; real-world research; Chinese medicine data characteristics; Chinese medicine data processing

Abstract: Traditional Chinese medicine (TCM) is a unique traditional medicine in China and has been passed down to the present day with its unique theoretical basis and treatment model. The introduction of the concept of evidence-based medicine into traditional Chinese medicine and the scientific evaluation of its clinical efficacy are essential for the better development and transmission of TCM. Randomised controlled trials have always been the "gold standard" of clinical trial evidence due to their strict inclusion and exclusion criteria and strict control of the data collection process, but the specificity of the TCM treatment model, the holistic, ambiguous, diverse and complex nature of the data, and the concomitant events in the research process make it difficult to conduct randomized ontrolled trials in TCM. Real World Study (RWS), which is conducted in actual clinical settings with broad inclusion and exclusion criteria to obtain treatment effects and long-term clinical outcomes as endpoints, can be used for long-term evaluation of treatment measures based on patients' preference. Real world studies allow for the long-term evaluation of treatment measures based on patients' preference, and are able to evaluate the overall effects, adapting to the holistic concept of Chinese medicine and the characteristics of evidence-based treatment. This paper summaries the data characteristics and data processing methods of real world studies in TCM, with a view to providing a reference for real world studies in TCM.

1. Introduction

The clinical practice of Traditional Chinese medicine (TCM) is guided by a Syndrome Differentiation and Treatment a holism, with individualized treatment and comprehensive evaluation being its treatment characteristics. Due to the complexity, non-linearity and ambiguity of TCM evidence, TCM practitioners often use comprehensive and complex interventions in real-world clinical practice. The diversity of individualized interventions and the complexity and long-term nature of clinical outcomes make it more difficult to evaluate the clinical efficacy of TCM. Real-world studies in TCM neighbourhood with treatment acquisition and long-term clinical outcomes as endpoints can provide a scientific evaluation of clinical efficacy while preserving the theoretical characteristics of TCM.

2. Characteristics of TCM data

Chinese medicine uses comprehensive and complex interventions in the process of clinical treatment, with a wide range of data sources and a large volume, which makes the real clinical treatment data of Chinese medicine present: wholeness, ambiguity, diversity and complexity.

2.1 Integrity

Based on the holism, when TCM practitioners carry out clinical practice, they do not only focus on the main symptoms of patients, but also adjust interventions according to the different secondary (mixed) symptoms of patients. According to the content of the “five viscera integrated view” in the holistic concept, the viscera are connected with each other, which is an organic whole. “Huangdi's Inner Meridian” said: “The so-called treatment of disease, see the liver disease, then know that the liver should be transmitted to the spleen, so the first solid temper, no order to receive the evil liver.” Therefore, in the actual clinical diagnosis and treatment process, traditional Chinese medicine collects patient information as comprehensively as possible based on the "holistic view" through “looking, listening, asking and treating”, takes the collected information as a whole, and then carries out data analysis.

2.2 Vagueness

There are a large number of imprecise and non-standard descriptions in Chinese medicine, and the language presents a general, specious character. For example, in his book Liu Ya Xian Medical Discourses and Medical Discourses, the medical practitioner Professor Liu Ya Xian points out that "The Treatise on Typhoid Fever extensively uses vague concepts (i.e. concepts with no clear extension) such as fever, malignant cold, sweating, floating pulse -; why is this? We know that vagueness arises from the fact that the system described is large and complex, with many variables and parameters. It is the result of the combination of multiple variables, not the characteristics of a single variable; it is a manifestation of the system as a whole, not a local or individual characteristic." The diagnosis of disease in TCM does not have clear criteria for indicators as in Western medicine, and the names of diseases are not as specific as in Western medicine, such as the evidence of bulging in TCM, the clinical manifestations of which are: the abdomen swells like a drum, the abdominal skin shows blue tendons and the complexion is pale yellow; similar to the later stages of diseases such as cirrhosis of the liver, tuberculous peritonitis, schistosomiasis, malnutrition and malignant tumours in the abdominal cavity in Western medicine. The words few and several in TCM prescriptions all reflect the ambiguity of TCM data.

2.3 Diversity and complexity

Chinese medicine obtains information on the clinical symptoms of patients through the four diagnoses, and then generates information on diagnosis and treatment as well as prescriptions. Chinese medicine attaches importance to individualized treatment, so even if the same disease has different interventions due to the pathogenesis and the concurrent symptoms exhibited by the patient, Chinese medicine calls this "different treatment for the same disease" which belongs to the content category of diagnosis and treatment; Chinese medicine There are many classical prescriptions, and the original prescriptions are added or subtracted according to the actual conditions of the patients. The efficacy of the same Chinese medicine varies due to different preparation methods; for example, the efficacy of raw Shou Wu is to detoxify, eliminate sores, laxative and lipid-lowering, while the efficacy of prepared Shou Wu is to nourish the liver and kidney, benefit the essence, darken the hair

and strengthen the muscles and bones.

These characteristics of TCM data make it more difficult to analyse data when conducting real-world research in the field of TCM. How to scientifically process TCM data, obtain high-quality real-world data and improve the quality of real-world research is a key aspect of conducting real-world research.

3. Quality evaluation of TCM clinical data in RWS

The quality of data needs to be controlled and evaluated in RWS, and data that meet the characteristics of the study should be analysed to make it real-world research evidence. The quality evaluation of real-world research data in TCM is mainly carried out in terms of completeness, accuracy, consistency, real-time and convenience.

3.1 Completeness

When collecting clinical information in TCM, it should be as comprehensive as possible, and can be evaluated in terms of the overall completeness and local completeness of the information; overall completeness means that: the data should include complete admission records, first course records, disease course records, and relevant laboratory tests and treatment information; local completeness refers to the local examination, including whether the structure of the paragraph level and word level is complete, whether the inscribed symptoms in the present medical history are structured, and whether the disease and The local completeness refers to local checks, including the completeness of the paragraph-level and word-level structure, the structure of the inscribed symptoms in the present history, and the stratification of the disease and symptom diagnosis.

3.2 Accuracy

The clinical data collected should originate from real clinical practice and conform to the actual clinical situation. The data collected should accurately reflect the patient's clinical situation, and scientific information such as scales and CRF forms should also conform to the patient's actual clinical situation.

3.3 Consistency

The clinical information collected needs to be consistent, especially the information related to clinical efficacy evaluation, and the patient's symptom records before and after admission should be consistent.

3.4 Real-time (timeliness)

Clinicians are required to complete the collection of scientific research data while completing their medical work tasks. It is not through secondary processing, secondary entry and other ways to complete the data of clinical information.

4. Processing methods for RWS TCM data

In view of the holistic, fuzzy, diverse and complex nature of TCM clinical data, the data processing methods used in conducting real-world research in TCM are also special. In the evidence information processing stage, as evidence information is complex, non-linear, fuzzy, dynamic and multi-dimensional, it is mainly done through quantification and dimensionality reduction.

4.1 Quantification and dimensionality reduction

In information processing, quantification, which specifies the research target, is often used in conjunction with dimensionality reduction, which reduces redundant information to improve the accuracy of target identification. The main elements of the discriminatory treatment are "different treatment for the same disease" and "different treatment for different diseases"; for the same disease, depending on the disease mechanism and the patient's manifestation of mixed symptoms, the TCM practitioner will adopt different interventions, i.e. different treatment for the same disease; different treatment for different diseases refers to the development of different diseases in the process of their development If the same disease mechanism emerges, the same interventions are adopted. Chinese medicine is based on the whole person, and the goal of treatment is not only to focus on the regression of the patient's condition, but also to focus on the patient's psychological well-being, and the improvement of quality of life, so the data on evidence in Chinese medicine is often not intuitive and unidimensional. The large variety and volume of TCM data makes it difficult to use conventional data processing methods. The pre-processing of TCM evidence information through quantification and dimensionality reduction can make subsequent data processing more difficult. The main methods of quantification and dimensionality reduction are: scale method, TCM diagnostic instrument parameter evaluation, data envelopment analysis, propensity score method, fuzzy mathematical model, prescription similarity model, etc. to achieve the quantification and dimensionality reduction of TCM evidence information.

4.2 Clustering analysis

In addition to quantification and dimensionality reduction, clustering can also be used when processing multidimensional TCM evidence information. Cluster analysis is a tool for processing high-dimensional data by grouping variables of physical or abstract objects into similar categories or distances based on their proximity. The 'group' generated by clustering is a collection of variables that are similar to other variables in the same 'group', but different from the objects in other 'groups'. In Chinese medicine, the "four diagnoses" are used to collect information from the patient through observation, hearing, questioning and cutting, and to make a diagnosis based on the information from the four diagnoses and give corresponding treatment measures. However, the amount of information in the four diagnoses is very complicated. In order to better summaries the information in the four diagnoses, research data can be obtained through epidemiological surveys, and cluster analysis can be applied to explore the classification of Chinese medicine evidence in conditions where the classification of Chinese medicine evidence is still unclear, and use professional knowledge to find out the Chinese medicine evidence patterns that meet the clinical reality. Cluster analysis can make the complex, non-linear, ambiguous, dynamic and multi-dimensional characteristics of TCM evidence information clear and concise, and the conclusions more direct, which is conducive to subsequent data analysis.

Unlike randomized controlled trials, RWS are conducted in real-life clinical settings, such as hospitals, communities or home settings, where multiple data are obtained to evaluate the true impact of a treatment on patients' health [1]. RWS is therefore difficult to randomise to individuals in the same way as a randomized trial, with high data heterogeneity and confounding factors. Confounding may lead to biased study results, and control for confounding is carried out at two main levels: study design and statistical analysis. Diagnosis of the equilibrium of covariates is a prerequisite for statistical analysis. The individualised treatment and comprehensive evaluation of TCM also makes the data volume heterogeneous. To improve the quality of real-world studies, the control and diagnosis of confounding in the data is key when conducting real-world studies in TCM.

5. Indicators of data covariate equilibrium diagnostics

The control of confounding is mainly carried out at two levels: research design and statistical analysis, which presupposes equilibrium diagnosis. In the RWS, equilibrium diagnostic indicators can be used to diagnose the equilibrium of data covariates, and there are currently two main categories of single covariate equilibrium diagnostic indicators and global equilibrium diagnostic indicators.

5.1 Single covariate diagnostic indicators

(1) Absolute difference: It is the absolute value of the difference between the means of the covariates.

$$|D(X)| = |\overline{x_1} - \overline{x_0}| \tag{1}$$

1=exposed group, 0=non-exposed group; the smaller the value the better the equilibrium.

Standard deviation [2-3]: It is the absolute difference divided by the sum of the covariates within the group.

$$SD = |D(X)| / \sqrt{({S_1}^2 + {S_0}^2)/2} \tag{2}$$

S_T^2 is the sample difference of X in the exposed group T. 1 = exposed group, 0 = non-exposed group; the smaller the value the better the equilibrium.

Overlap factor [4-5]: It is the proportion of two covariate probability density functions that overlap. $\int \min_x \left(\hat{f}_1(x), \hat{f}_0(x)\right) dx$ is the density function in the exposed group T estimated using the normal kernel density estimate and Scott's suggested bandwidth; 1 = exposed group, 0 = non-exposed group; the smaller the value the better the equilibrium; usually taken as "$\int \min_x \left(\hat{f}_1(x), \hat{f}_0(x)\right) dx$" to facilitate comparison with other metrics.

K-S distance [6-7]: the maximum vertical distance between two cumulative distribution functions. $max_x \left|\hat{F}_1(x) - \hat{F}_0(x)\right|$ is the empirical cumulative distribution function in the exposed group T; 1 = exposed group, 0 = non-exposed group; range of values: 0-1; the smaller the value the better the equilibrium.

Lévy distance: the maximum perimeter of a square that can be internally tangent between two cumulative distribution functions.

$\min_\epsilon\{\epsilon > 0, \hat{F}_0(x-\epsilon) - \epsilon \le \hat{F}_1(x) \le \hat{F}_0(x+\epsilon) + \epsilon \text{ for all } x\}$, 1=exposed group, 0 = non-exposed group, range of values: 0-1; the smaller the value the better the equilibrium.

The above indicators only provide equilibrium diagnostics for a single covariate. There may be some minor differences between covariates and the accumulation of these differences may also cause bias in the study results; therefore, some scholars have proposed global equilibrium diagnostic indicators.

5.2 Global equilibrium diagnostic indicators

(1) Marginal equilibrium: It is a matrix of the variance and covariance of two sets of covariates.

L1 metric: a multidimensional covariate box is created by a specific stratification law, and the proportion of two sets of data falling into the box is observed.

$0.5 \sum_{H \in \mathcal{H}} |(H) - \hat{f}_0(H)|$, $\widehat{f_T}(H)$ is the proportion of exposed group T falling into the covariate box, 1=exposed group, 0=non-exposed group; takes values in the range 0-1; 0 indicates the most perfect equilibrium.

General weighted difference: It is a weighted sum of standardised differences calculated for all covariates, covariates squared and interaction terms.

C-statistic after propensity score matching: given by the area under the ROC curve of the matched sample propensity score model.

$C = \frac{1}{m}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1}\mathbb{I}\{p_i - p_j\}$ i = 1, ... ; 1 is a dummy variable, 1 if pi < pj, 0 otherwise; 1 = exposed group, 0 = non-exposed group; the range of values is 0.5-1; the minimum value indicates that the propensity score model cannot distinguish between exposed and non-exposed groups after matching, i.e. perfect equilibrium; for comparison with other indicators, it is usually taken as: "$0.5 - C = \frac{1}{m}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1}\mathbb{I}\{p_i - p_j\}$".

6. Conclusion

Chinese medicine is a traditional medicine with Chinese characteristics, and a scientific and accurate efficacy evaluation of the clinical efficacy of Chinese medicine is conducive to guiding Chinese medicine to conform to the development of the times and is an effective inheritance of Chinese medicine. Chinese medicine is a clinical practice guided by the theory of dialectical treatment and holistic concept, and the clinical data present diversity, complexity and dynamics, which are not suitable for the concept of traditional randomized controlled trials, so more real-world studies are conducted in the neighbourhood of Chinese medicine clinical efficacy research. Real-world studies, which are patient-centred, conducted in real clinical settings, and pursue long-term comprehensive efficacy evaluation, coincide with the holistic concept of Chinese medicine and evidence-based treatment. However, real-world research data come from a wide range of sources, with high heterogeneity and confounding, which may cause bias in research results and require proper statistical analysis, for which control of confounding is a prerequisite. This paper innovatively proposes to introduce balanced diagnostic indicators into real-world studies of TCM, with a view to providing methodological guidance for the balanced treatment of research data in real-world studies of TCM. However, there are various types of TCM data and the control of data quality needs to be explored and innovated in daily research.

References

[1] Sherman R E, Anderson S A, Dal Pan G J, et al. Real-World Evidence - What Is It and What Can It Tell Us? [J]. N Engl J Med, 2016, 375(23):2293-2297.

[2] Austin PC. Assessing balance in measured baseline covariates when using many-to-one matching on the propensity-score. Pharmacoepidemiology and Drug Safety 2008; 17:1218–1225.

[3] Austin PC. Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score. Pharmacoepidemiology and Drug Safety 2008; 17:1202–1217.

[4] Bradley E. Overlapping coefficient. Encyclopedia ofStatistical Sciences 1985; 6:546–547.

[5] Inman HF, Bradley EL. The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. Communications in Statistics-Theory and Methods 1989; 18: 3851–3874.

[6] Stephens MA. Use of the Kolmogorov-Smirnov, Cramer-Von Mises and related statistics without extensive tables. Journal ofthe Royal Statistical Society, Series B 1970; 32:115–122.

[7] Pestman WR. Mathematical Statistics: An Introduction. Walter De Gruyter Inc: Berlin, 1998.