Study on the New Drug Quality Characteristic Evaluation Indictor by Fuzzy Clustering Method

Peng Ruan

Shinawatra University, 99 Moo 10, Bangtoey, Samkhok, Pathum Thani 12160 Thailand
604760047@qq.com

ABSTRACT. *Comprehensive evaluation of the quality characteristics of new drugs is helpful for their marketing. However, the quality evaluation indictors of new drugs are too simple, and too many evaluation indictors may cause respondents to reject and perfunctory, leading to the evaluation results are not reliable enough. In this study, a fuzzy clustering method was proposed to optimize the quality characteristics of new drugs and simplify the evaluation process.*

KEYWORDS: *Quality of new drugs, Quality indicators, Fuzzy clustering method, Quality management*

1. Introduction

After 2015, Chinese government defined a new drug that is not marketed globally. New drugs at this stage are almost always in clinical trials [1]. However, the ecosystem of new drug research and development needs to be further improved, for example, there are too many clinical approvals and hospitals cannot keep up. Zanubrutinib, the first anticancer drug in China, was approved by the U.S food and drug administration (FDA), becoming the first anticancer drug independently developed by a Chinese company and approved for market in FDA [2]. But in China it will take a long time to get approval. Because the review mechanism of new drugs in China restricts the development of new drugs. Before 2015, there were many problems in drug review and approval in China, and the drug review center had a backlog of 23,000 drug approvals [3]. The new drug quality evaluation indictor in Shanghai is relatively simple. According to the code for quality management of drug clinical trials issued by the state food and drug administration, the certification of new drugs mainly includes three indicators: effectiveness, safety and quality controllability [4]. Among them, quality controllability is the indictor of the production process, while effectiveness and safety are only two simple indicators in the hospital drug clinical experiment center. Due to the lack of comprehensive

evaluation of quality characteristics of new drugs, it is difficult for customers to know their true quality, thus hindering their marketing process. This study proposed the quality characteristics of new drugs in multiple dimensions according to the methods of government policies, academic literature reports and consulting customers. However, in the actual investigation and research, many customers of new drugs reported that the survey indicators were too many, which was not conducive to the understanding of customers. Clients with patient identity said that their mental health and ability to think were suppressed due to their own diseases, and there were too many evaluation indicators, which may make some interviewees impatient and refuse to cooperate with the investigation or perfunctory. The results are not real enough. Therefore, it is necessary to explore the reasonable reduction of quality characteristics of new drugs. This study suggests that the best way to reduce quality characteristics is to use artificial intelligence for computational screening. Therefore, this study adopts fuzzy clustering method.

Fuzzy clustering method is a kind of fuzzy mathematical language to things according to certain requirements description and classification of mathematical methods [5], is refers to according to the study of the attributes of the object itself to construct the fuzzy matrix, and on this basis, according to a certain to determine the membership degree of clustering relations, which USES fuzzy mathematics method of the fuzzy relation between the quantitative determination of the sample, so as to objectively and accurately clustering [6]. After the processing of fuzzy clustering method, the indictors are optimized and the evaluation procedure is simplified.

2. Research methods and materials

2.1 Basic contents of fuzzy clustering method

In the sample, clustering analysis is carried out on m corresponding data with P indictors, and P indictors correspond to different attributes and contents. The initial matrix is established with the relevant data of P indictors, as follows:

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{np} \end{bmatrix}$$

Fuzzy clustering analysis is to carry out data mining for fuzzy and uncertain data and its expression connotation and nature, and play the role of simplifying indictor system. The evaluation effectiveness of the simplified part is often consistent with the overall evaluation effect of all the indicators. Simplify a process to highlight core indictor namely, eliminate redundant indictor.

2.2 Calculation method of fuzzy clustering

2.2.1 Data transformation and standard processing

The meaning and importance of each indicator are different, and the judgment experience of experts is also different. Therefore, it is necessary to highlight the main characteristics hidden in the indicator and make the meaning expressed correspond to the subject to be evaluated. Each indictor has different value tendency in different types of evaluation. For example, the significance of a new drug in the evaluation of curative effect and economic benefit is different. In the evaluation of curative effect, the indictor value needs to be converted to better unify its characteristics:

The data processing formula is as follows:

$$X = \frac{X' - \overline{X'}}{C}$$

Type (1), X represents the actual score, ($X \char`\^ \char`\^ '$) is indictor score averages, C as corresponding indictor of variance.

2.2.2 Correlation matrix calculation

The score data is often large, or small, not conducive to judgment; The numerical value can be limited to the range of [0,1], then the calculation formula is:

$$X = \frac{X' - X'_{min}}{X'_{max} - X'_{min}}$$

When $X' = X'_{max}$, X=1; when $X' = X'_{min}$, X=0

After calculation, a matrix element r_{ij}, can be obtained to form a similarity coefficient matrix R between indicators (this step can also be completed by using the correlation analysis function of SPSS software) :

$$R = \begin{bmatrix} r_{11} r_{12} x r_{13} ... r_{1n} \\ r_{21} r_{22} \ r_{23} \ \cdots r_{2n} \\ \vdots \ \ \vdots \ \ \ \vdots \ \ \ \vdots \ \ \ \vdots \\ r_{n1} r_{n2} \ r_{n3} \ \cdots r_{nn} \end{bmatrix}$$

Matrices are often fuzzy relationships, which can be further treated by equivalent computation.

(3) Clustering

R must be a fuzzy equivalence relation in order to cluster. The R obtained from step (2) cannot be satisfied in general conditions, so it shall be modified.

2.3 Fuzzy equivalence relations

2.3.1 Definition of fuzzy equivalence relation

Fuzzy equivalence relation mainly expresses the correlation between indictors. Usually, the correlation between multiple indictors is not strong or even weak, which is called fuzzy matrix relation. Fuzzy matrices need to be self - reflective, symmetric and transitive.

If it satisfies: ①reflexivity:r_{ij} =1; ②symmetry:$r_{ij}=r_{ji}$; ③transitivity:$R\ ^0R⊆R$. said $R = (r_{ij})_{n×n}$ is a fuzzy equivalence relation.

When the matrix does not satisfy the transitivity, it needs to be transformed to obtain the fuzzy equivalent matrix.

2.3.2 Fuzzy equivalent matrix transfer package method

A matrix A=$(a_{ij})_{m×s}$,B=$(b_{ij})_{s*n}$, Is the fuzzy matrix, in this case, matrix A=B, then:

$A^OB=(C_{ij})_{m×n}$

$C_{ij}=\max\{(a_{ij}∧ b_{kj}|1≤k≤s)\}$

"^{O}" is the Zadeh operator,∧ it means the minimum of the two, ∨ it's the maximum of the two. The calculation method for the two identical matrices is as follows: the first row of matrix A is compared with the first column of matrix B; the minimum value is obtained by p2wise comparison; the maximum value is obtained; and so on, the new matrix C is formed.

The classification algorithm adopts intercept method and sets intercept value λ=1.0, 0.8, 0.6, 0.5, 0.4, 0.2, The specific intercept value can be divided according to the characteristics of the whole matrix elements, and the operation can be carried out from high to low in principle.Ifλ=0.8, then≥0.8 turn to 1.0, <0.8turn to 0. After transformation, observe the distribution of 1 and 0 in each row of the new matrix. And so on, until you can no longer classify.

This algorithm can also get the dynamic clustering graph by the "classification-system clustering" function of SPSS software, which is more convenient and intuitive for classification processing.

3. Practical application

3.1 Basic Situation of Examples

Source of indicators: according to literature, referring to relevant policies and regulations of the ministry of health of China, consulting customers' opinions, etc.

Invite 31 customers to participate in the evaluation. Customers include medical personnel, patients, etc. Are highly educated. Contains indicators: the cure rate, effective rate, anxiety, fear, physical pain, discomfort, adverse reaction, recurrence rate, sequela, drug dependence, such as 10 indicators, revised finally identified six effective indicators: the cure rate (X1), effective speed (X2), psychological comfort (X3), body comfort (X4), adverse reactions (X5), recurrence rate (X6). Likert five-level scoring method was used for each indictor. The customer scores of the six indicators are shown in table 1. The Alpha of the questionnaire is 0.715, indicating that it is acceptable.

Table 1 Score of 6 indicators

No.	X1	X2	X3	X4	X5	X6
01	5	5	4	4	5	5
02	5	4	5	5	4	5
03	5	4	4	5	5	4
04	5	5	4	4	5	4
05	4	4	3	4	4	5
06	5	4	5	4	4	5
07	5	5	5	5	5	5
08	5	4	4	3	4	4
09	5	5	5	5	5	5
10	5	5	4	5	5	5
11	5	5	5	5	5	5
12	5	5	4	4	4	4
13	5	4	4	4	5	5
14	4	4	4	3	5	4
15	4	3	3	4	4	3
16	5	5	4	3	5	4
17	5	4	5	4	5	4
18	5	4	3	4	4	3
19	5	5	5	5	5	5
20	5	5	5	4	4	4
21	5	5	5	5	3	3
22	5	5	4	5	5	4
23	4	4	4	4	4	4
24	5	5	5	5	5	4
25	5	5	5	4	5	4
26	5	5	4	4	4	5
27	5	5	5	5	5	5
28	5	5	4	4	5	5
29	5	5	4	4	5	5
30	5	5	5	4	4	4
31	5	5	5	4	4	4

3.2. Data normalization

There were 31 customer samples, involving 10 indicators for screening, and finally 6 key indicators were determined, and the data of the cure rate of the indicators were transformed and processed, and the rest indicators were analogized.

Finding the Mean Value of Indicators: $X_{11}{}' =(4+5+…+5) /31=4.55$

Variance: $C_1=\{[(4-4.55)^2+(5-4.55)^2…+(5-4.55)^2]/31\}^{1/2} \}=0.568$

Standard values are obtained from formula (1): $X_{11}{}'' =(4-4.55)/0.568=-0.968$

$X_{21}{}'' =(5-4.55)/0.568=0.792$

…

Thus get: $X_{max1}{}'' = 0.792$; $X_{min1}{}'' =-0.968$

According to formula (2):

X11=(-0.968+0.968)/(0.792+0.968)=0

X21=(0.792+0.968)/(0.792+0.968)=1

…

By analogy, data normalization of six indictor attributes is calculated and processed.

3.3 Calibration

Calibration is to establish a fuzzy similarity relationship. In this case, the correlation coefficient method is used.

$$x = \frac{\sum_{k=1}^{m}(X_{ik} - \bar{X}_\iota)\left(X_{jk} - \bar{X}_J\right)}{\sqrt{\sum_{k=1}^{m}(X_{ik} - \bar{X}_\iota)^2}\sqrt{\sum_{k=1}^{m}\left(X_{jk} - \bar{X}_J\right)^2}}$$

$$\bar{X}_\iota = \frac{1}{m}\sum_{k=1}^{m} X_{ik}$$

$$\bar{X}_J = \frac{1}{m}\sum_{k=1}^{m} X_{jk}$$

From this, the fuzzy similarity matrix can be calculated.

$$R = \begin{bmatrix} 1 & 0.26 & 0.271 & 0.086 & 0.150 & 0.441 \\ 0.206 & 1 & 0.605 & 0.506 & 0.315 & 0.210 \\ 0.271 & 0.605 & 1 & 0.471 & 0.293 & 0.294 \\ 0.086 & 0.506 & 0.475 & 1 & 0.413 & 0.236 \\ 0.15 & 0.315 & 0.293 & 0.413 & 1 & 0.253 \\ 0.441 & 0.21 & 0.294 & 0.236 & 0.253 & 1 \end{bmatrix}$$

3.4 Constructing Fuzzy Equivalent Matrix

It can be seen from the fuzzy similarity matrix that: R has reflexivity and symmetry, but it lacks transmissibility, and there may be large errors in direct classification. Therefore, the transfer package method was used to transform the matrix and make it transitive. The specific algorithm was the matrix A0B, and 0 was the zadok operator, namely "inverted" was taken into the minimum value and "inverted" into the maximum value. The first row of the matrix A, multiplied by the first column of the matrix B, is minimized on both sides of the multiplication, and the maximum value is retained as the first element of the new matrix C. And so on, you get your new matrix C, which in this case is A is equal to B. The new matrix is as follows:

$$R = \begin{bmatrix} 1 & 0.293 & 0.294 & 0.294 & 0.294 & 0.441 \\ 0.506 & 1 & 0.413 & 0.413 & 0.413 & 0.271 \\ 0.471 & 0.605 & 1 & 0.506 & 0.506 & 0.271 \\ 0.271 & 0.506 & 0.506 & 1 & 0.605 & 0.471 \\ 0.271 & 0.413 & 0.413 & 0.413 & 1 & 0.506 \\ 0.441 & 0.294 & 0.294 & 0.294 & 0.293 & 1 \end{bmatrix}$$

3.5 Clustering Method

When λ=0.1 or 0.8, the truncation relation is transformed into:

$$R = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

At this time, six indicators are divided into six categories: [X1], [X2], [X3], [X4], [X5], [X6], obviously, the classification work has not been completed, so it is necessary to continue clustering.

When λ=0.6 the truncation relation is transformed into:

$$R = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

At this time, six indicators are divided into six categories: [X1], X2], [X3], [X4], [X5], [X6], obviously, the classification work has not been completed, so it is necessary to continue clustering.

When λ=0.5 the truncation relation is transformed into:

$$R = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

At this time, six indicators are divided into five categories: [X1], [X2], [X3, X4], [X5], [X6], However, the classification work has not been completed, so it is necessary to continue clustering.

When λ=0.64 the truncation relation is transformed into:

$$R = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

At this point, the six indicators are still divided into three categories: [X1, X6], [X2, X3], [X4, X5].

When λ=0.2, It can only be divided into one class, so the clustering ends.

3.6 Making Dynamic Clustering Diagram

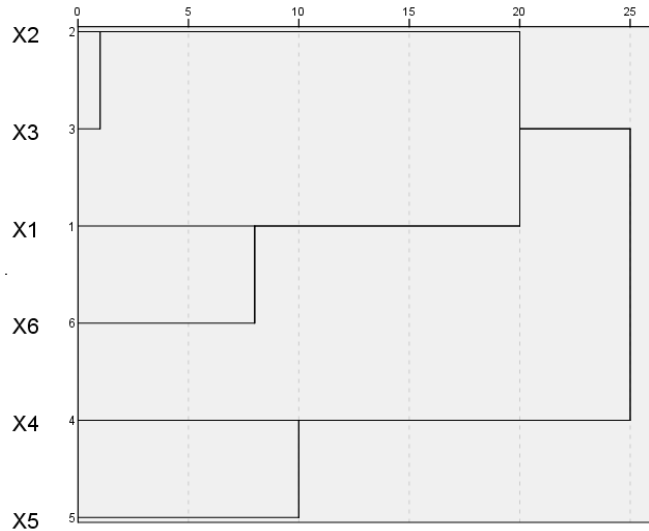The dynamic clustering map is made with the classification function of SPSS software as follows:

Figure. 1 1 = cure rate; 2 = effective speed; 3 = mental comfort; 4 = physical comfort; 5 = adverse reactions; 6 = recurrence rate

First Classification: [X4], [X2, X1]

Second Classification: [X4, X5], [X2, X1, X6]

…

As can be seen from the dynamic clustering diagram, psychological comfort (X4) is the first class, and cure rate (X1) and response rate (X2) are the first class. The significance is that the evaluation of X4, X1 or X2can represent the overall level, and it is not necessary to evaluate each indictor.

3.7 Verification effect

The overall average value of the six indicators was calculated as the total group, and then the average value of X4+X1 and X4+X2 was calculated as the X4+X1 group and X4+X2 group, and the three groups were tested by non-parametric test. The results showed that: compared with the total group, there was no difference in $P > 0.05$, indicating that the combination of X4, X1 and X2 could represent the overall level. However, there was a significant difference between the two groups of X4+X1 and X4+X2, $P < 0.05$, indicating that the two groups of data could not be added together to represent the overall level. See table 2 for details

Table 2 Results of nonparametric tests

Group	χ^2	Df	P
Total vs X4+X1	0.222	1	0.637
Total vs X4+X2	1.470	1	0.225
X4+X1 vs X4+X2	4.571	1	0.033

When screening the quality characteristics of new drugs, only psychological comfort needs to be evaluated, and the evaluation of cure rate or effectiveness speed can reflect the overall quality of new drugs, instead of evaluating all indicators. In other words, to evaluate the quality of a new drug, we need only extract and calculate the data of two indicators. The result is an effective reduction of the indicators, which greatly reduces the workload of manual evaluation.

4. Conclusion

In the face of complex evaluation indictors and data, this study use fuzzy clustering method to further classify the indictor system, so as to obtain key laws hidden in the data, effectively simplifying the complexity of new drug quality evaluation.

However, there are also shortcomings in this study: because of the variety of new drugs, there are also indictor differences between different drugs. Although the quality characteristics of new drugs have been simplified, more practice is needed to realize the comprehensive evaluation of quality.

References

[1] Deng yong. Suggestions on improving the drug registration administration measures (draft for soliciting opinions) [N]. Democracy and legal times, 2019-10-31 (006).

[2] Jue xiao. Chinese new anticancer drug was approved for the first time in the us. People's Daily overseas edition. November 29, 2019

[3] Tong xiao, Chen yuwen. Analysis on the examination and approval of registration and application of five types of new traditional Chinese medicine drugs in China from 2009 to 2015 [J]. Chinese journal of pharmaceutical industry, 2008, 49 (09): 1327-1330.

[4] Lai shuqing, Wang qiwei, Liu jianhui. Analysis of new drug approval and registration application in China in 2017 [J]. Chinese journal of pharmaceutical industry, 2008, 49 (03): 392-396.

[5] Indrajit Saha, Jnanendra Prasad Sarkar, Ujjwal Maulik. Integrated Rough Fuzzy Clustering for Categorical data AnalysisFuzzy Sets and Systems, 2018, 23 (2): 7-8

[6] Sin-Jin Lin. Hybrid Kernelized Fuzzy Clustering and Multiple Attributes Decision Analysis for Corporate Risk Management. International Journal of Fuzzy Systems, 2017, 19 (3): 659-670