Fraudulent data detection in clinical trial using dynamic clustering

Dr. Md. Atiqul Islam, Shahjalal University of Science and Technology, Sylhet, Bangladesh
Md. Rajwanur Rahman, Shafi Consultancy Bangladesh
Shafi Chowdhury, Shafi Consultancy Limited

1. Introduction

Fraudulent data is a severe issue in the clinical trial domain. There are some methods available for identifying data fraudulent and clustering techniques are one of them. However, detecting the homogeneous clusters is always questionable. Therefore, we aim to identify the homogeneous clusters using dynamic clustering technique to detect the fraudulent data. As fraudulent data generally appears at a site level, our technique will focus on detecting sites that may contain fraudulent data.

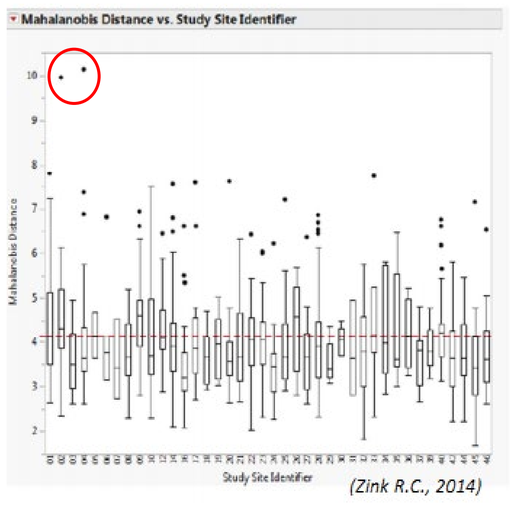
2. Methods and plan

First, hierarchical clustering using Duda and Hart method used for identification of primary numbers of clusters. After identify the actual numbers of clusters per site we checked the one-to-one relationship of each homogenous clusters from different sites to identify inequalities. Based on the findings, we implement additional checking on the identified sites data to determine possible fraudulent data.

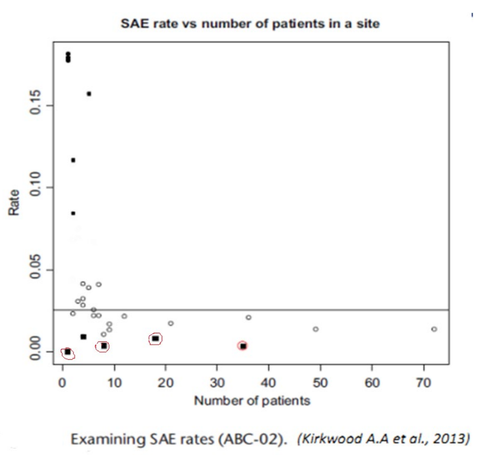
This approach identifies clinical trial data fraudulent with high efficiency compared to the traditional methods. Due to the nature of dynamic modeling, this approach can reduce the use of the extensive resource for identification of fraudulent data.

3. Traditional identification of fraud

Often inliers and outlier data checks provide insight for the fraudulent data. Inliers data are closely related to multivariate mean of some quantitative variables. To identify inliers or outliers, data are interlinked to Mahalanobis Distance, which captures the correlation structure of data.



Plotting visit can be a way to identify fraudulent data. Sometimes, clinical trials may accept visits over the weekend, however, in other cases weekend visits can be considered as an alarm of fabrication. In these cases, assessed visit dates have to be checked to see the actual days when visits took place.



Extreme Serious Adverse Events (SAE) rate can be another measure for fabricated data. All sites with zero SAE or very low SAE rate can be an indicator for fabricated data.

Other than these above-mentioned approach, by targeting usual behavior, rounding to integer and digit preference can also be a good indicator for identifying data fraud. Some basic statistical measures like plotting extreme variance and unusual correlation behavior can also help determining indication of fraudulent data.

4. Dynamic clustering based identification

Hierarchical clustering (HC) is a classical clustering technique to produce a nested sequence of clusters. It can produce an informative ordering of the objects and produce small clusters which may be helpful for discovery, for example fraudulent detection in clinical trial data.

But clustering techniques provide little information about the clusters structure in the clinical trial data. There is no unified approach on what essentially constitutes a cluster and no conclusive answer for choosing the number of clusters.

Duda and Hart (DH) is one of the indices to choose the right number of clusters using certain characteristics from per site of clinical trial data.

Briefly, DH is the ratio of the two within sum of squares to decide whether a cluster can be divided into two clusters.

For large number of observations in cluster *c*, SSW_K is approximately normally distributed with mean $n_c p\sigma^2$ and variance $2n_c p\sigma^4$ and SSW_{K+1} follows approximately a normal distribution with mean $n_c p(1-2/\pi p)\sigma^2$ and variance $2n_c p\{1-8/(\pi^2 p)\}\sigma^4$. The variance σ^2 represents the variance of the complete population. From these observations the following criterion was used to sub-divide whenever the following holds

$$\frac{SSW_{K+1}}{SSW_K} < \left[1 - \frac{2}{\pi p} - z\sqrt{\frac{2\{1-8/(\pi^2 p)\}}{n_c p}}\right], K \geq 1$$

with *z*, a standard normal quantile with value 3.20.

Somers' delta is generally used for measuring the strength and direction of association between an dependent and independent variable. One to one relationship of each pair of homogenous clusters will be checked using Somers' delta statistic to rank the clusters and possibly indicate problematic sites.

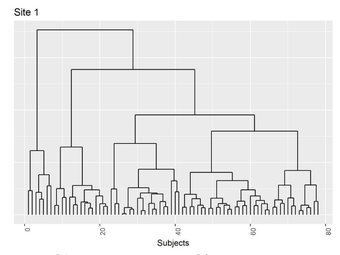
5. Analysis and Findings

A dummy trial with following attributes was used to show the proposed method in action.

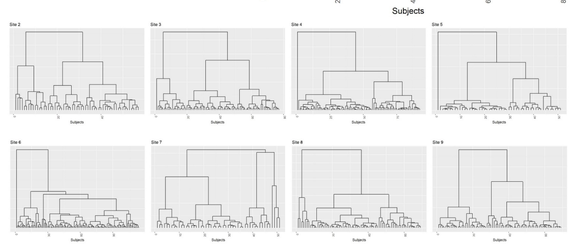
- Trial consists of 676 subjects, 9 sites
- Age, Gender and BMI used from demographic variables
- Baseline laboratory data for CRE, SGPT, SGOT and ALKP were used as a concerned variable for possible fraudulent data detection.

Hierarchical clustering results with Duda & Hart index values

Site	1	2	3	4	5	6	7	8	9
No. of Cluster(s)	1	2	3	1	1	1	1	2	4
DH Index	0.741	0.588	0.620	0.794	0.663	0.863	0.696	0.670	0.602



Based on clustering information and Somers' delta statistics, site 9 has indication of fraudulent data as it belongs to less homogenous group.



6. Conclusion

Compare to the traditional method, model-based clustering should produce more accurate identification, because of considering large set of information, robust method, and use of dynamic approach. The evidence found on traditional approach for identifying fraudulent data can be compared with the model-based findings to ensure a higher rate of correctly identification of problematic sites and thus can be taken appropriate measure to prevent further.

Thus, dynamic clustering technique can be effectively used in identifying nearly real-time identification of fraudulent data in clinical trial domain.

7. References

- "Central Statistical Monitoring in clinical trial, IMMPACT-XVIII 2015" by Amy A Kirkwood.
- "Risk-Based Monitoring and Fraud Detection in Clinical Trials Using JMP® and SAS®" by Richard C. Zink

Please feel free to contact us for further information.

www.shaficonsultancy.com

Shafi Consultancy Ltd.