Fraudulent Data Checks and How to Develop them

Shafi Chowdhury, Shafi Consultancy Limited
Aminul Islam, Md.Jabrul Islam, Shafi Consultancy Bangladesh

1. Introduction

Fraudulent data checks are now required by the FDA (US Food and Drug Administration) as part of routine data checks. However, there can be many different methods used on many different types of data. This poster looks at some of the statistical techniques used to try and identify fraudulent data, the different types of data available, and which methods can be used on each type of data.

The results can also be presented in many different formats, from charts and plots to complex analysis tables. This paper will also review the different methods to see if one method gives clearer results and interpretation than others.

As fraudulent data check is something that must be done for each trial, the paper will offer valuable suggestions as to how these checks can be performed and standardised to minimise efforts.

2. Fraudulent Data

Fraudulent data means the deliberate fabrication or falsification of data which is a serious scientific misconduct in clinical trials. The occurrence or prevalence of data fraud in clinical trials is generally assumed to be quite low but, by its nature, this occurrence is difficult to estimate.

Identifying and documenting fraud can be a time-consuming and expensive process that once started can damage the perception and reputation of a research institution and may also lead to the unsuccessful proving of misdoing.

3. Why Look at Fraudulent Data?

- Because people's lives and health will be at risk.
- To protect the rights and well-being of patients enrolled in a trial by verifying the authenticity of clinical trial data.
- To identify and address problems early to limit the serious implications by analyzing data quality.
- Maintain the research integrity in the public eye.

4. Types of Fraud Data

- Plagiarism.
- Piracy.
- Fabricated data.
- Altered or ignored inclusion / exclusion criteria.
- Data falsified to reach a desired outcome.

5. Data Susceptible to Fraud

- Eligibility criteria e.g. age, medical history of the patients.
- Patient diary data.
- Repeated measurements e.g. blood pressure and laboratory data.
- Adverse events reporting.
- Assessment of medical compliance.
- Dates of assessment.
- Efficacy results.

6. Detection of Fraud at Clinical sites

At clinical site level it is possible to acquire data by fraud, and it is vital that we detect the sites that are affected by fraudulent data.

To detect fraudulent data the **conventional approach** is to perform on-site visits to the medical centres (i.e. Source Data Verification) plus further investigations. It can also be done by **monitoring for patterns and trends in documentation, checking the dates** and **examining returned clinical trial materials and drug usage.** However, it is best practice to use different statistical techniques by utilizing theoretical knowledge and experiences. For that purpose, we can use the following:

- Check Inliers and Outliers
- Incorrect dates: Mosaic Plots
- Under-reporting of Adverse Events: Scatter Plots, 2D and 3D.
- Rounding to integers: Line and Scatter Plots
- Last Digit Preference: Volcano Plots
- Check for duplicate patients by comparing key data, including age, height, weight, vital signs and visit schedules. Box plots can be used for these checks.

7. Check Inliers and Outliers

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.

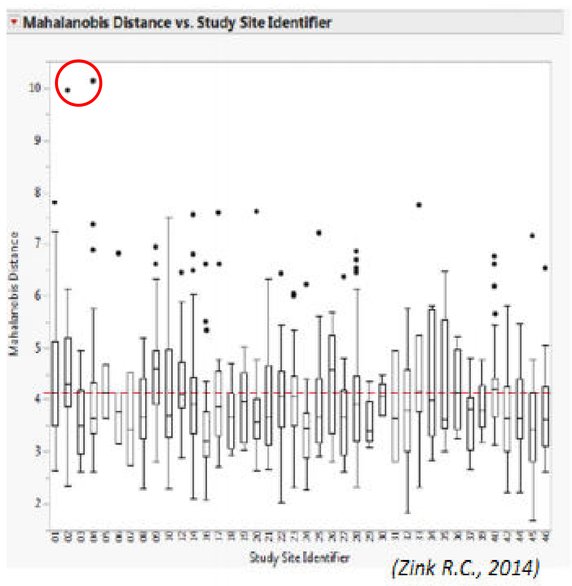


Figure-1

The red line in Figure-1 represents the multivariate mean of the data and all the observations lying too close to the line are indentified as inliers. The sites with too small IQR (Inter Quartile Ranges) could also be flagged.

8. Data Collection in Weekdays and National Holidays

Clinical trials in some therapeutic areas 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.

The site details may be represented using a bar chart analysis in the following way:

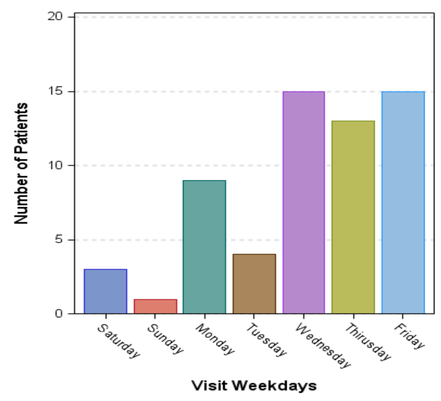
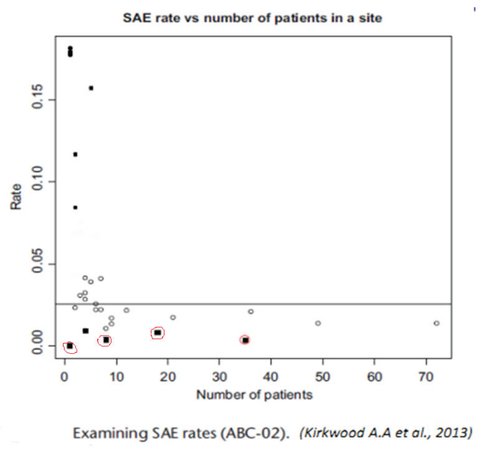


Figure-2

9. Serious Adverse Event (SAE) Rate

Serious Adverse Events (SAE) rate is calculated by the following formula:
SAE Rate = (Number of subjects with a SAE in a site) / (Total subjects present in that site)

All sites with zero SAEs, plus the lowest 10% of rates are shown as black squares.



Examining SAE rates (ABC-02). *(Kirkwood A.A et al., 2013)*

Figure-3

10. Check Patient Profile Report

Participant level checks can be performed to find recording and data entry errors. The date checks may also detect fraud if falsified data has been created carelessly. Procedural errors may be picked up by looking at the order of the that dates occurred, for example patients treated before randomisation, consent date after 1st AE or follow-up visit after death etc are shown below:

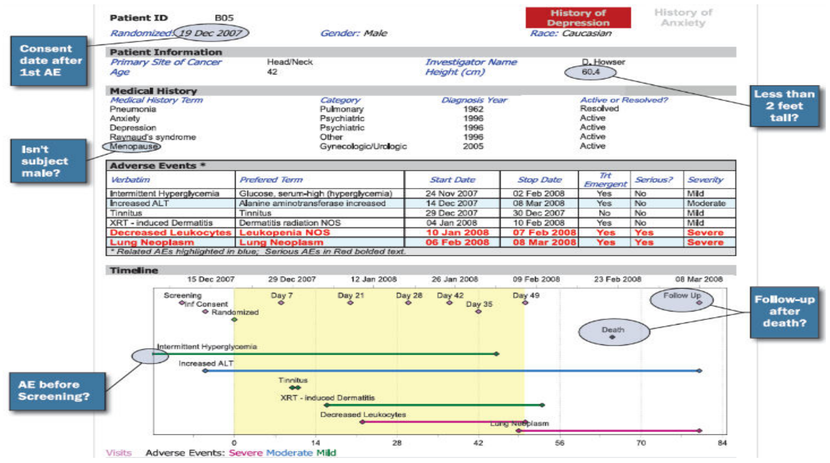


Figure-4

11. Digit Preference and Rounding

The distribution of the leading digits (1-9) in each site can be compared with the distribution of the leading digits in all the other sites are put together.

Digit	Site frequency	Site percent	All other frequency	All other percentage
1	344	33.05	8419	32.29
2	159	15.27	4048	15.53
3	181	17.39	3500	13.42
4	100	9.61	3201	12.28
5	73	7.01	1881	7.21
6	60	5.76	1430	5.49
7	53	5.09	1242	4.76
8	35	3.36	1134	4.35
9	36	3.46	1216	4.66
p-value:	0.00277738			

Figure-5

Also rounding can be checked in a similar way using the last digit rather than the first, or graphically.

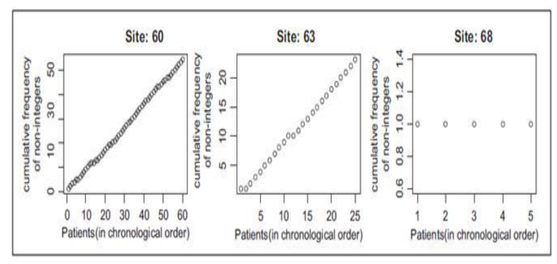


Figure-6

12. Other Methods

Apart from the methods described above we can detect the fraudulent data by using alternative methods such as the following:

- Check the difference between planned and actual visit dates.
- Use scatter plot of consent dates.
- Check incidence of CTs, baseline conditions and discontinuation.
- Use scatter plots of discontinuation dates.
- Check the change from baseline of vital statistics.

13. Conclusion

Generally, the main responsibility for detecting fraudulent data lies with the clinical monitors and quality assurance group. However, statistical techniques also have an important role in the detection of fraudulent data. It can be seen here that different techniques are required for different types of data, but a combination of plots to visually review the trends backed up by statistical analysis to detect significance produces the best results.

Company standard programs can be developed to implement these statistical techniques based on standard data structure. This will further reduce the efforts required to perform a comprehensive set of checks to detect fraudulent data.

14. References

- "Central Statistical Monitoring in clinical trial, IMMPACT-XVIII 2015" by Amy A Kirkwood.
- "Fraud Detection in Clinical Trials: A Graphical Tool" by Giulia Zardi.
- "Fraud and Misconduct in Clinical Trial" by Frank Wells.

Please feel free to contact us for further information.

www.shaficonsultancy.com

Shafi Consultancy Ltd.