Fraud Detection in Clinical Trials: A Graphical Tool

“Data Visualization in Clinical Research”

BIAS

Biometristi dell' Industria ASsociati

Author: Giulia Zardi

Milan, May 29th 2015

cros nt

The Center of Excellence for Clinical Trial Data

Introduction

- A clinical trial database can never be completely free from errors, nevertheless ICH-GCP requires data to be «accurate, complete, and verifiable from source documents».

- Fraud are thought to be rare (estimated to be less than 1%), but its prevalence is underestimated: undetected fraud is unrecorded fraud.

- Thanks to some features of a trial design (i.e. randomization and blinding), some instances of data «cooking» may have little impact on scientific conclusions.

- 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.

Then, why bother looking for fraud?

- People's lives and health at risk : efforts have always been made in verifying the authenticity of clinical trial data to protect the rights and well-being of patients enrolled in a trial.

- To limit serious outcomes: if analysis of data quality may be applied often enough, problems could be identified and addressed early.

- Preserve public perception of research integrity

Types of fraud

- Plagiarism

- **Fabricated data**

  e.g. missing or outlying values replaced by plausible values or fabricating trial participants and all associated data values.

- **Data falsified to reach a desired objective**

  e.g. to make a patient eligible, or to show a treatment effect. These kind of errors can have a devastating effect on the trial credibility, especially if they are aimed at exaggerated treatment effects.

How can we detect fraud?

- Conventional approach: on-site visits to medical centres (Source Data Verification) plus further investigations.

  time-consuming and expensive

«Humans are poor random number generators and are generally forgetful of natural constraints in the data.» (Venet et al, 2012)

Since it is hard to fabricate data convincingly, all kinds of statistical clues are left, which means they can be followed.

⇩

- Alternative method: abnormal data patterns can be detected using specific statistical techniques, which can also be translated into graphical checks to determine the presence of potentially fraudulent data.

Data susceptible to fraud

- Eligibility criteria
  e.g. age, medical history of the patients.
- Repeated measurements ←
  e.g. blood pressure, laboratory data.
- Adverse events reporting ←
- Assessment of medical compliance
- Dates of assessment ←
- Patient’s diaries

Detecting Fraud at a Clinical Site

As fraud basically appears at a site level, we will focus our attention on detecting sites that may contain fraudulent data for individual subjects.

Here, the tools proposed according to literature and our experience:

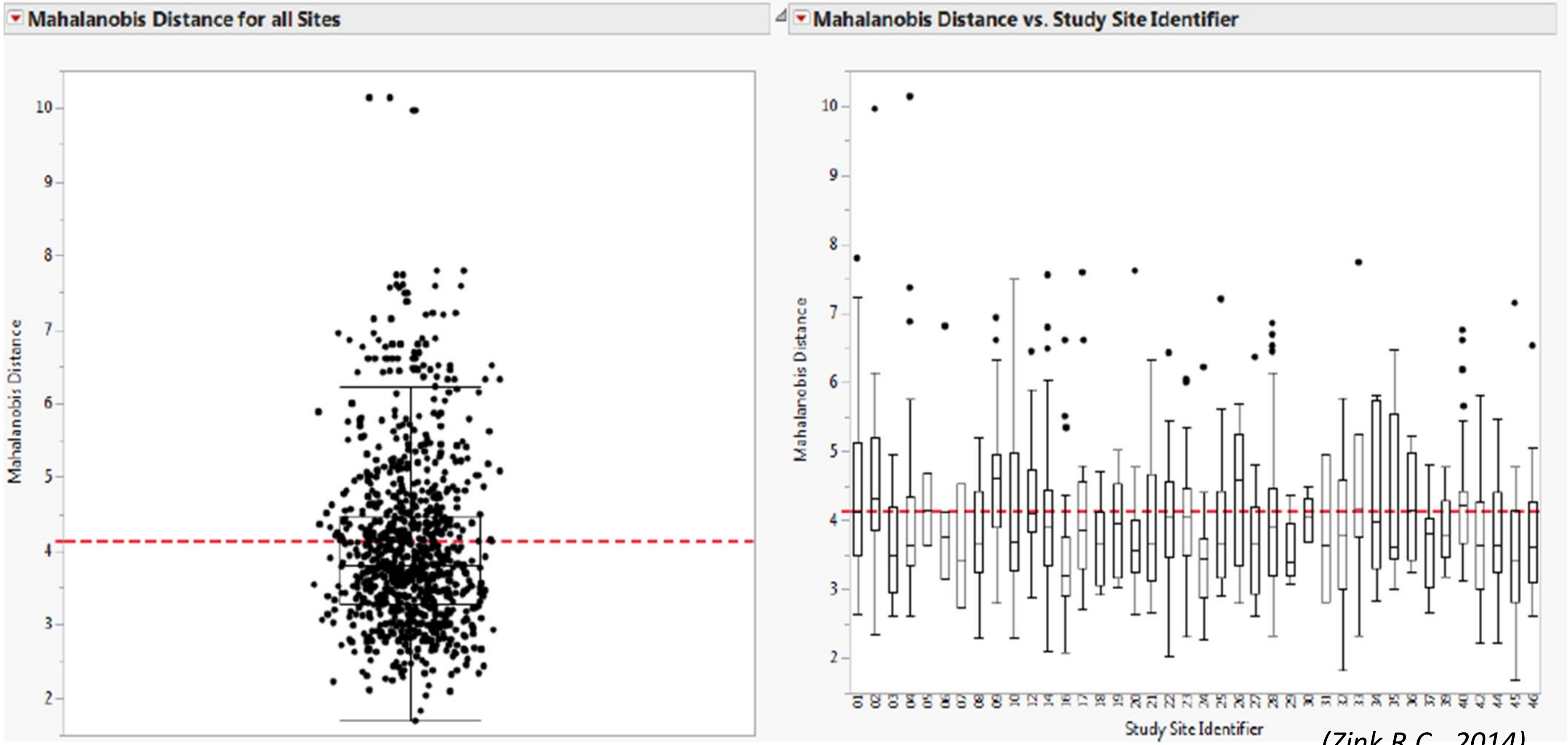
- Inliers: Box plot
- Incorrect dates: Mosaic plot
- Under-reporting of adverse events: Scatter plot
- Rounding to integers: Line plot & Scatter plot
- Digit preference (Last digit): Volcano plot
- Extreme variances: Dot plot & Box plot
- Unusual correlation structure: Heatmap

Inliers (and Outliers)

«If it is hard to fabricate convincingly a single variable, it is much harder to make a set of interrelated measurements look genuine.» (Weir et al, 2011)

- Inliers are data too close to the multivariate mean of some quantitative variables taken into consideration.
- Since data are fabricated trying to make them less noticeable and not stand out too much from the rest, inliers detection has been proposed here.
- The procedure to detect inliers (or outliers) is linked to Malhanobis distance which takes into account the correlation structure of the data.

Inliers (and Outliers) - Continued

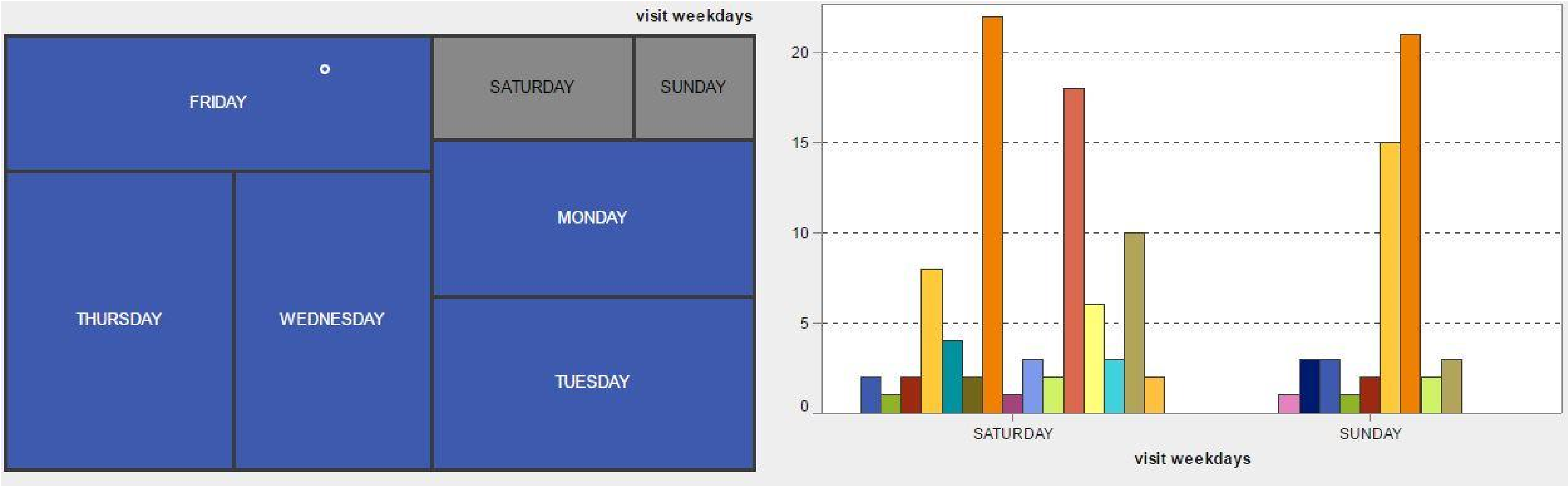


(Zink R.C., 2014)

- **All those observations lying too close to the red dashed line (multivariate mean) may be identified as inliers and sites with too small IQR could be flagged.**

Weekdays

- Depending on the therapeutic areas involved in a specific clinical trial, visits over the weekend may be tolerated. In other cases weekend visits can be an alarm of fabrication.
- All assessed dates of visit have been taken into consideration here.
- The histogram shows site details that have been observed in the mosaic plot on the left.

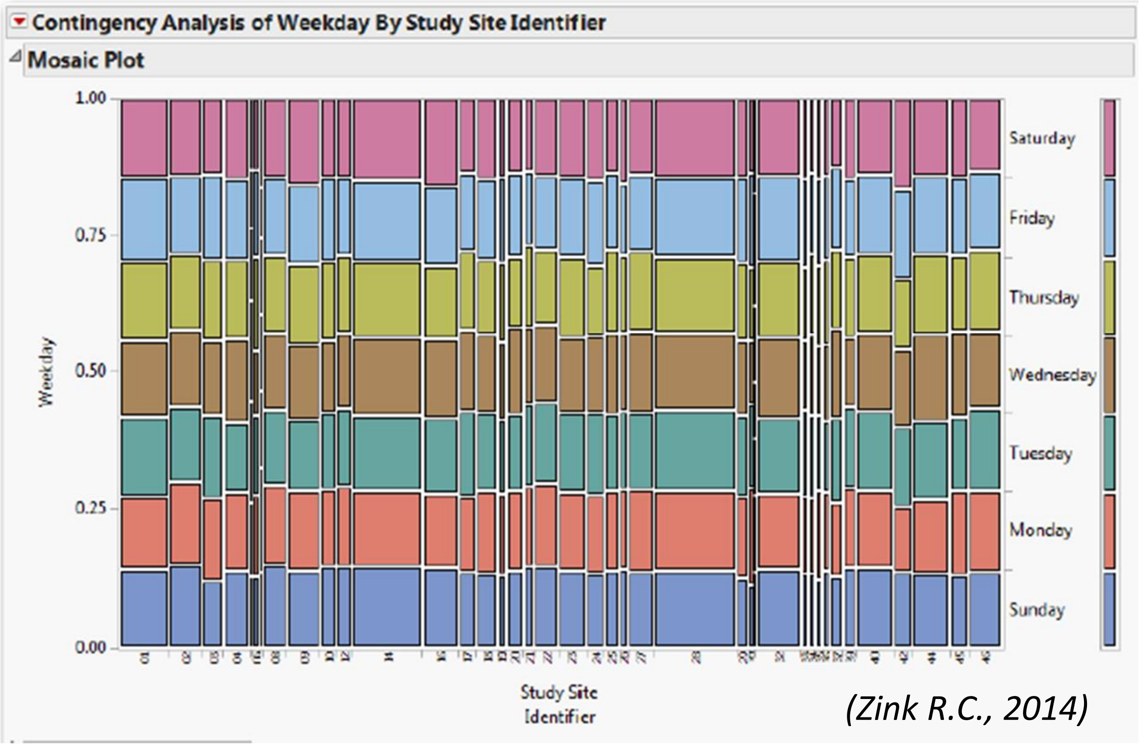


Weekdays

- The following graph is an example of a more complex mosaic plot for a clinical study where, due to the nature of patients pathology, visits on Sundays and Saturdays were allowed.

- Distribution of weekdays seems to be similar across clinical sites for study visits.

Figure 4.3 *Mosaic Plot of Weekday by Study Site Identifier*



Contingency Analysis of Weekday By Study Site Identifier

Mosaic Plot

(Zink R.C., 2014)

Serious adverse events rate

- SAE have been considered, but this graph can be adapted to any type of AE (or event in general).
- Rate is calculated as the number of subjects with a SAE in a site divided by the number of total subjects in that site.
- The horizontal line reported represents the average of SAE rate.
- Action can be taken for sites furthest toward the bottom right.

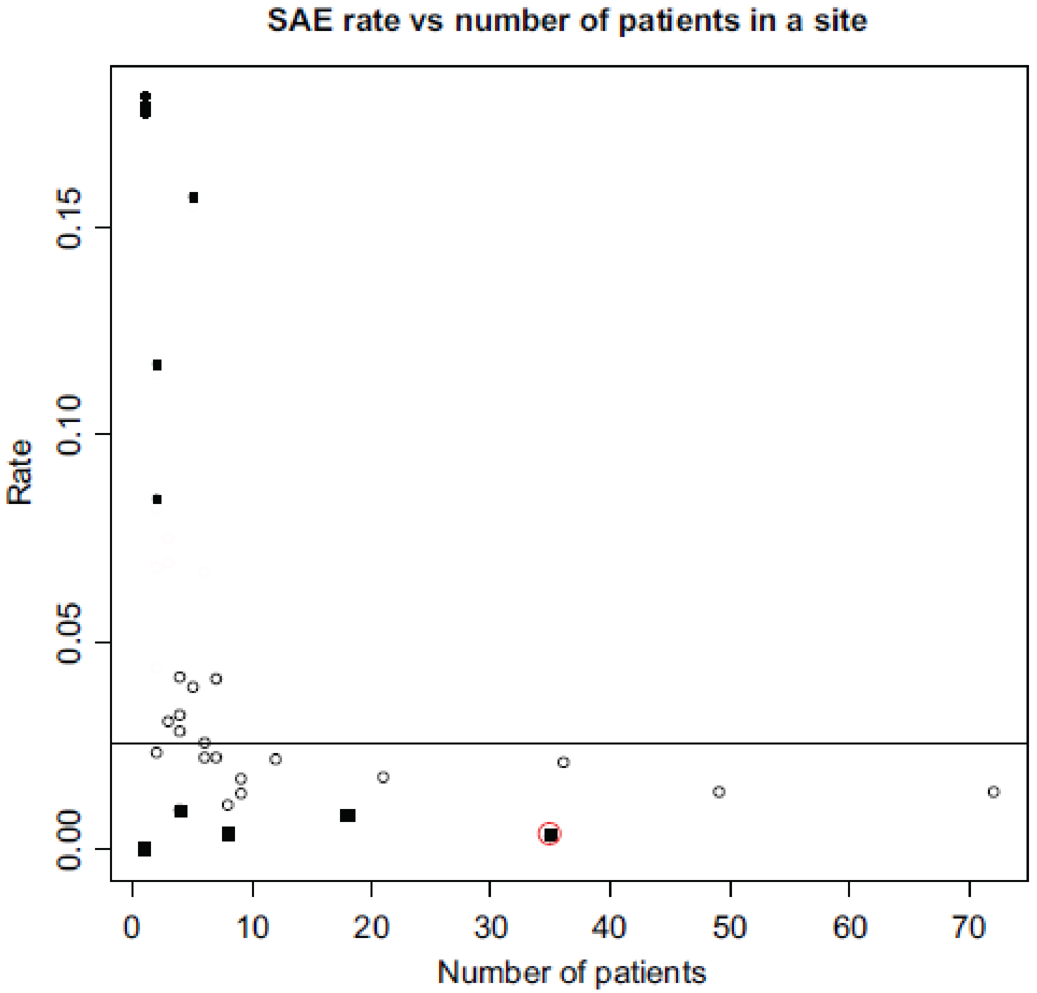
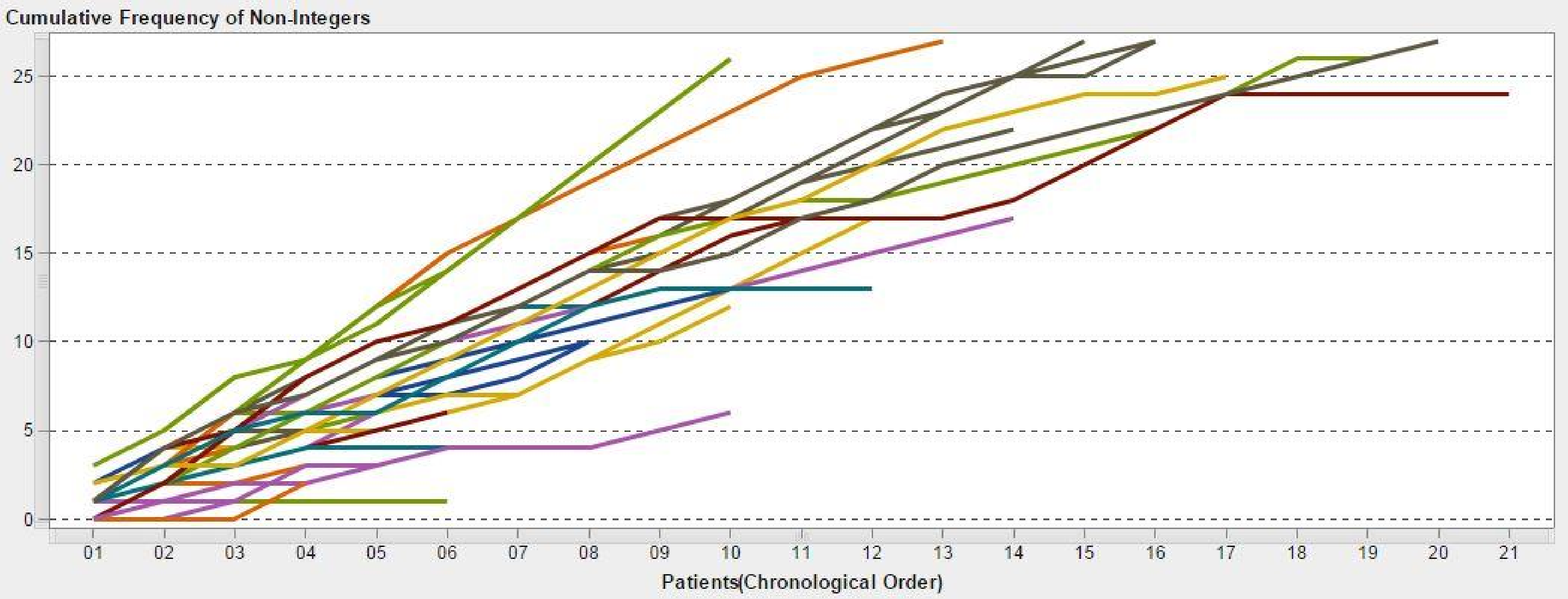


Figure 8. Examining SAE rates (ABC-02). *(Kirkwood A.A et al., 2013)*

cros nt The Center of Excellence for Clinical Trial Data

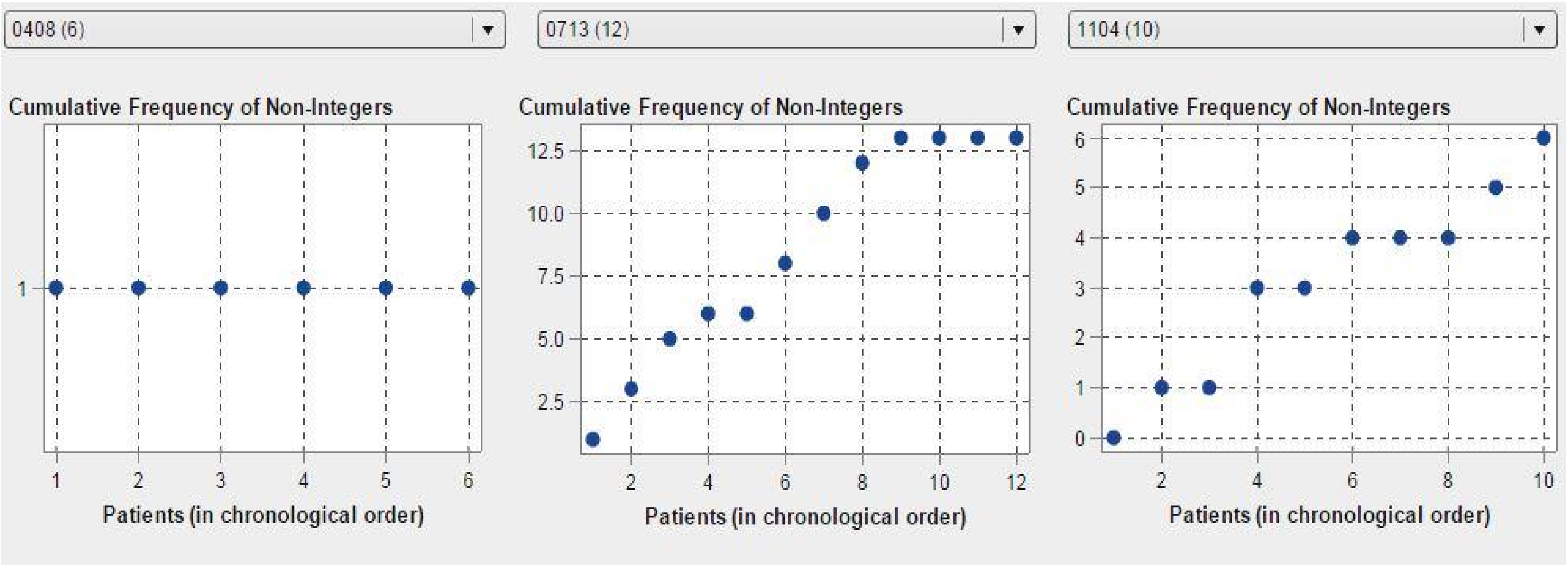
Rounding to integers

- Checks are going to be performed on whether too many continuous data are rounded into whole integers.
- Taylor *(Taylor et al., 2012)* proposed a graphical solution examining the pattern of integers recorded over time.
- The first step would be observing the general trend of the cumulative frequency of non-integers on a certain repeated variable.



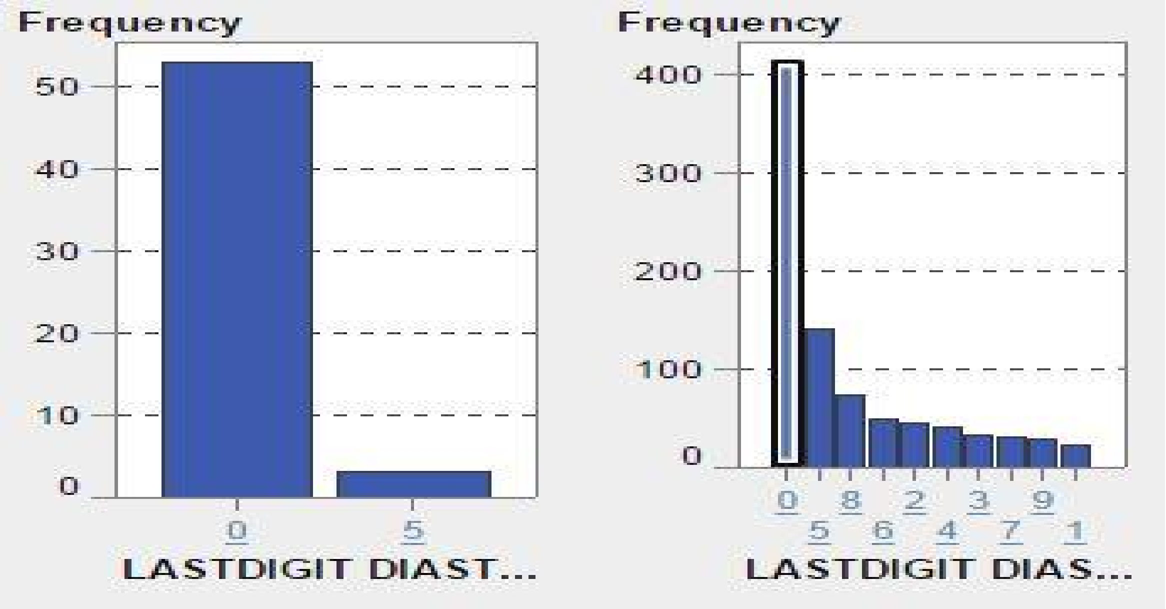
Rounding to integers

- Then a plot for single selected centers can be printed.
- A monotonic increase will be seen if all numbers are non integers, otherwise a horizontal line will show too much rounding (as for the first site considered).



Digit preference

- People seem unable to generate random long sequences of numbers, hence this can be easily revealed by analyzing the digit preference.
- The distribution of the last digit can be evaluated comparing the frequencies of digits for a variable in a site (plot on the left) against all the others (plot on the right).
- Some instances of rounding may be identified as well.



Digit preference: Introduction to volcano plot

- Volcano plots are space-saving scatter plots that can be used to emphasize important differences.
- A suitable way to present last digit preference analysis would be through a volcano plot.
- A p-value for each site is calculated using a CMH row mean score statistic applied on a site vs all other sites table for digit from 0 to 9.
- On Y-axis: -$\log_{10}$ transformation of the p-value, which means the smaller the p-value, the larger the number on the y-axis.
- On X-axis: maximum difference between a suspect site vs all the other sites across all observed digits.
- Adjustments for multiple testing problems may also be included in a manner that is straightforward to interpret (FDR line).

Digit preference

- Last digit preference is here globally represented with a volcano plot for study measurements analysis.

- All sites p-value combinations are shown and some instances may be identified as significant differences in digit distribution.

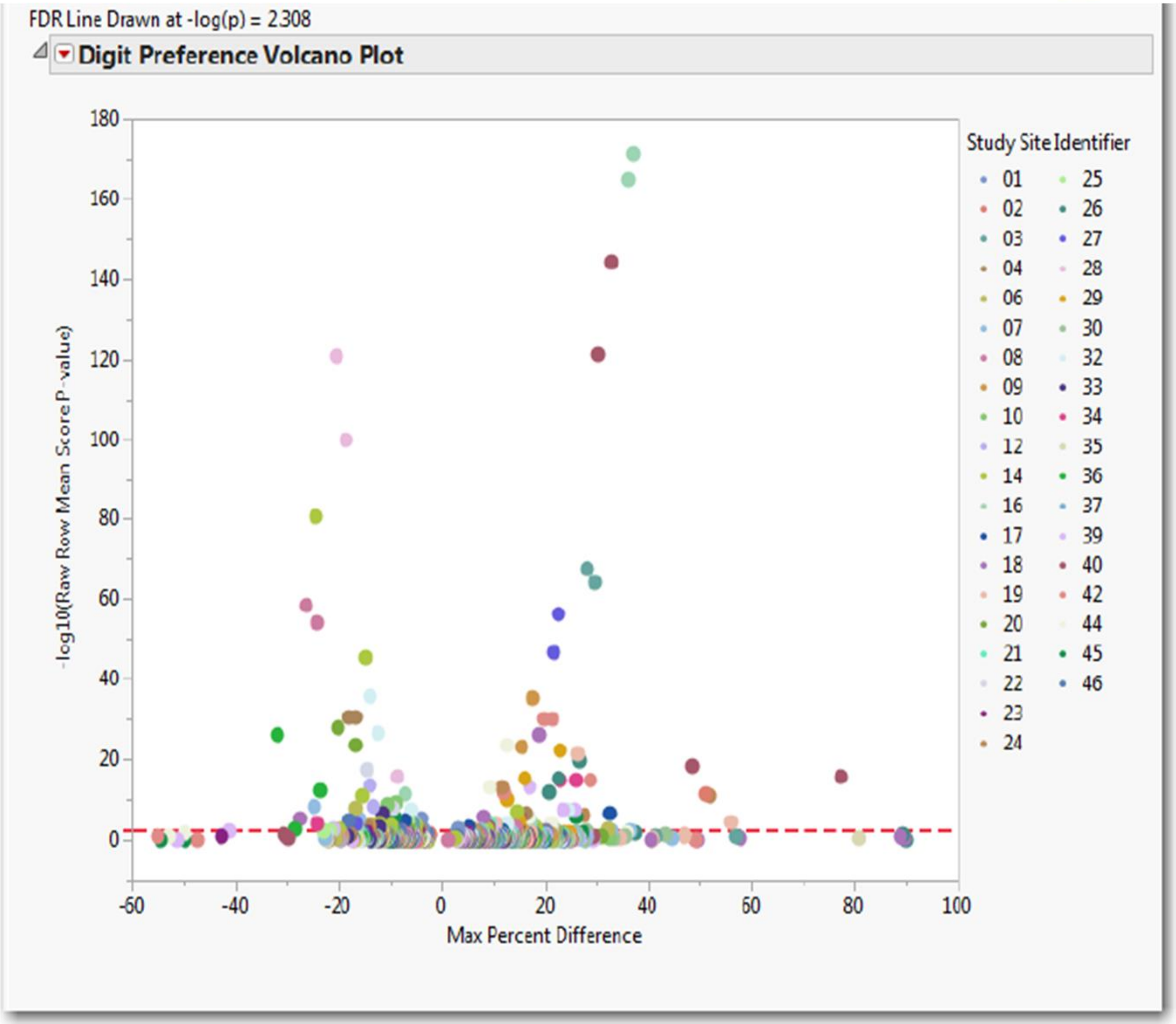


Figure 13. Volcano Plot of Digit Preference.

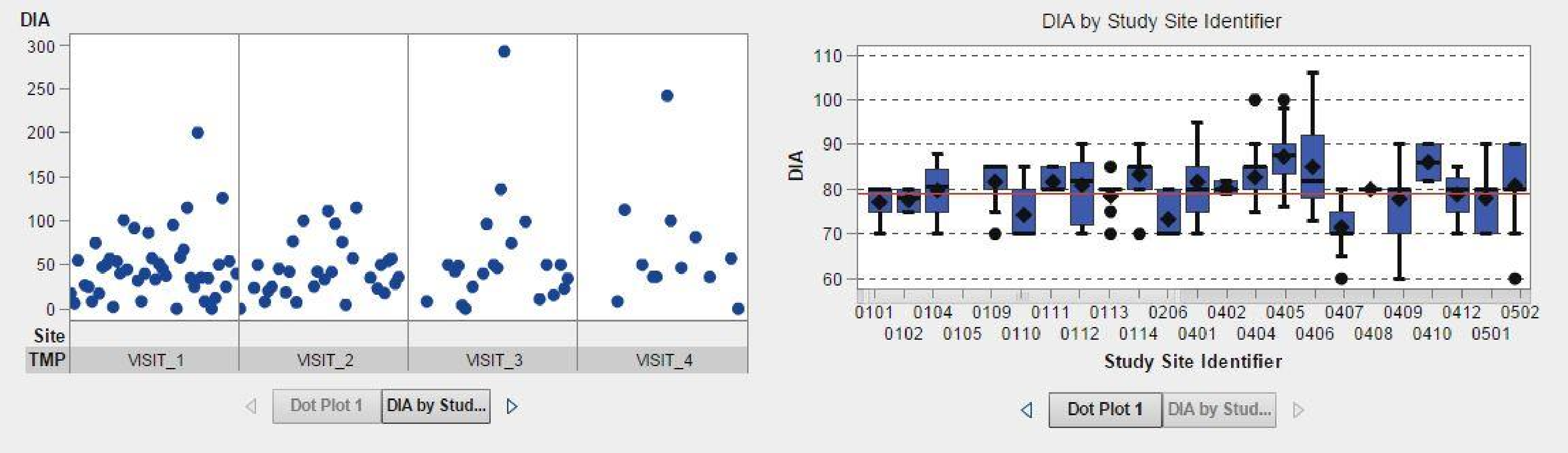
(Zink R.C., 2014)

Extreme variances

- Appropriate mean values may be easily invented, but creating appropriate levels of variability isn’t that straightforward.
- In fact, fabricated data usually shows an abnormally small variability (confirmed through simulations).
- Changes in the variance of a measurement over time may then be used to highlight the start (or end) of a period of data fabrication.

Extreme variances

- The first figure below helps identifying particular site-visit combinations for diastolic blood pressure measurements where variability (variances on Y-axis) appears to be too small compared to those of other sites.
- The boxplot summarizes diastolic blood pressures across all visits by site highlighting any departures in mean, median or variability.

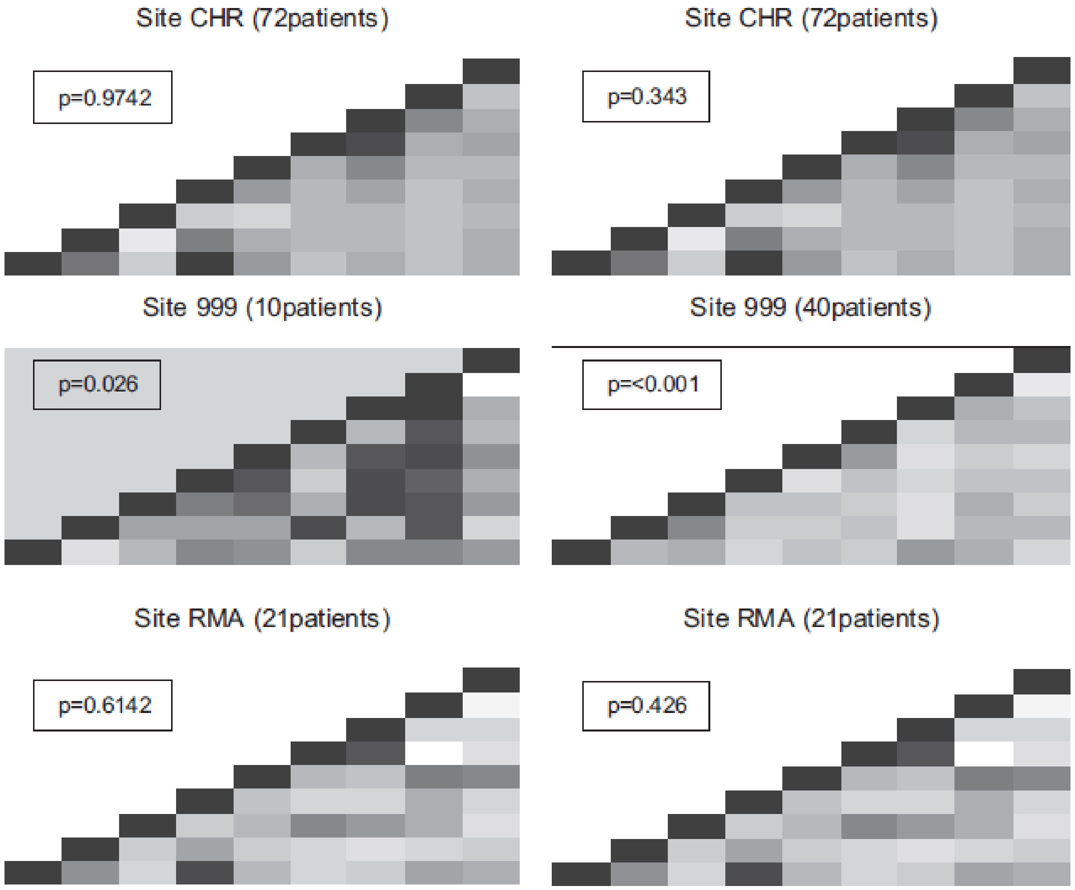


Anomalous correlation structure

- While it has been shown in many articles that the variability for fabricated data is lower than the one observed on real data, it has been demonstrated as well that high correlation coefficients may be a sign of data fabrication.
- A heatmap (when values contained in a matrix are presented) has been proposed here.
- Correlation checks may be evaluated for continuous variables and each square in the heatmap represents the correlation between a pair of variables.
- Each square color is related to the grade of correlation:

■ Highly positive □ Highly negative

Anomalous correlation structure - Continued



(Kirkwood A.A et al., 2013)

- **Correlation checks for 9 variables are presented here for 3 (2 real and 1 fabricated) sites divided into 2 panels (which differs for the method of fabrication used for the site in the middle).**

- **It can be seen that the correlation structures for fabricated sites seem quite unusual compared to those of real sites (simulated p-values confirmed).**

Error or Fraud?

- The boundary between data recording errors and fraud is often fuzzy, although the latter is characterized by a deliberate attempt to deceive.

- If fraud is confirmed, it may not be sufficient to correct or discard problematic data. The validity of the whole trial may come into question and further actions will eventually be taken.

What to do next?

Further investigation is always needed:

- **Monitors have to be notified of the nature and magnitude of problems and the potentially problematic nature of certain sites could be demonstrated by an on-site audit.**
- **Anyway, one of the goals would be making monitoring visits more targeted and informed so that the resulting process could be more efficient: which leads to on-site monitoring in «suspected sites» (key concept for CSM).**
- **Eradicating fraud may be impossible, but its occurrence can be minimised by driving this proposed analysis together with other statistical techniques as support.**

Summary

cros nt

The Center of Excellence for Clinical Trial Data

Summary

- Many types of analysis can be performed. In general, graphs have always been easiest to interpret.
- Interactive softwares would be more appropriate to perform this kind of analysis (i.e. JMP, SAS Visual Analytics).
- In spite of this, other softwares as SAS or R, may be used as well, even if the process could be more time-consuming.
- All these checks need to be run during and after the course of the trial.
- Detecting unusual patterns on the data is useful to guide study monitors actvities.
- The presented tool is flexible: different quantitative variables (repeated measurements or not) may be analyzed at the same time and different types of graphs may be added in addition to those proposed here and many other types of analysis might be explored (such as Holidays distribution, visit occurrence or duplicated sample analysis, etc..)

Thank you for your attention!

References

- Weir C. & Murray G. (2011). Fraud in clinical trials: detecting it and preventing it. *Significance* 8: 164-168.
- Buyse M. et al. (1999). The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. *Statistics in Medicine* 18:3435-3451.
- Zink R. C. (2014). Risk-based monitoring and Fraud Detection in Clinical Trials using JMP® Clinical.
- Zink R. C. (2014). Risk-based monitoring of Clinical Trials using JMP® and SAS ®.
- Zink R. C., Wolfinger R. D. & Mann G. (2013). Summarizing the incidence of adverse events using volcano plots and time windows. *Clinical Trials* 10: 398-406.
- Venet D., Doffagne E., Beckers F., et al. (2012). A statistical approach to central monitoring of data quality in clinical trials. *Clinical Trials* 9: 705-713.
- Kirkwood A., Cox T., Hackshaw A. (2013). Application of methods for central statistical monitoring in clinical trials. *Clinical Trials* 10: 783-806.
- International Conference of Harmonization. (1996). E6: Guideline for Good Clinical Practice. Available at: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6/E6_R1_Guideline.pdf
- Wu X., Carlsson M. (2010). Detecting data fabrication in clinical trials from cluster analysis perspective. *Pharm Statistics* 10: 3435-3451.
- Taylor R. N., McEntegart D., Stillman E. (2002). Statistical techniques to detect fraud and other data irregularities in clinical questionnaire data. *Drug Information Journal* 36: 115-125.
- Akhtar-Danesh N. & Dehghan-Kooshkghazi M. (2003). How does correlation structure differ between real and fabricated data-sets? *BMC Medical Reasearch Methodology* 3(18): 1-9.