A Data Anomaly Detection Tool for Site Selection at FDA

Xiaofeng (Tina) Wang, Paul Schuette, Matilde Kam
Center for Drug Evaluation and Research, Office of Translational Sciences, Office of Biostatistics, Immediate Office

Background

- On-site inspections of clinical trial investigators are important to ensure the quality and integrity of the trial data and the reliability of the trial results submitted to the U.S. Food and Drug Administration.
- With the increasing size and complexity of the trials, statistical tools are needed to assist the site selection process and identify potentially problematic sites.

Methods

We describe our experience with a centralized statistical monitoring platform as part of a Cooperative Research and Development Agreement (CRADA) between CluePoints and the FDA.

Statistical Monitoring Applied to Research Trials (SMART):

- Applies battery of statistical tests to different types of variables in multiple domains

All Variables	Continuous Variables	Binary/Categorical Variables	Date Variables
Count Missing	Mean Between-Patient Variability Within-Patient Variability Global Outliers Sequence Outliers Identical Values Propagation Correlation	Proportion Transitions from Transitions to Initial Values	Saturday Sunday

- Compares subject/site level values for a variable to all sites in trial
- Computes a p-value corresponding to the site and the test

$$\mathbf{P} = \begin{array}{c} \\ \text{Site}_1 \\ \\ \text{Site}_i \\ \\ \text{Site}_m \end{array} \begin{array}{c} \begin{array}{cccccc} \text{Test}_1 & & \text{Test}_j & & \text{Test}_n \end{array} \\ \begin{pmatrix} p_{1,1} & \cdots & p_{1,j} & \cdots & p_{1,n} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ p_{i,1} & \cdots & p_{i,j} & \cdots & p_{i,n} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ p_{m,1} & \cdots & p_{m,j} & \cdots & p_{m,n} \end{pmatrix} \end{array}$$

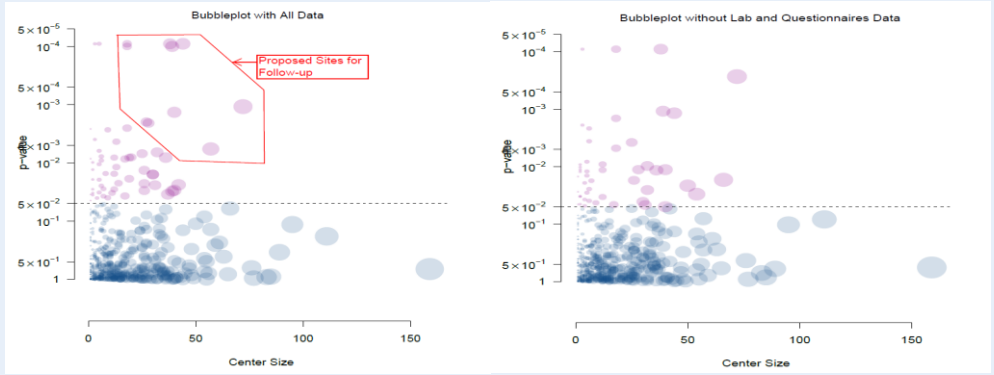
The data inconsistency score for Site_i is then computed as

$$S_i = \frac{\sum_{j=1}^{n} w_j \log(p_{i,j})}{\sum_{j=1}^{n} w_j}$$

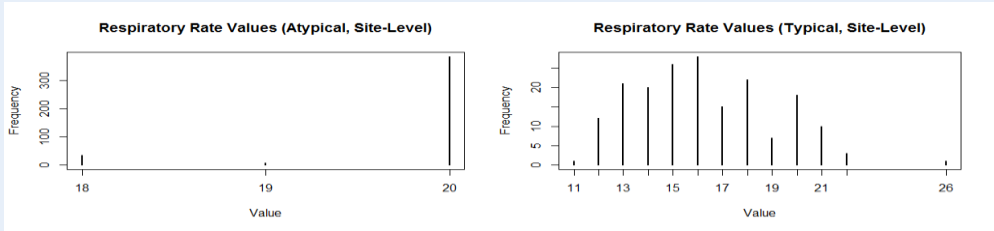
Results

An overall data inconsistency score is calculated from a high-dimensional p-value matrix to assess the inconsistency of the data between one site and the data from all sites. Sites are ranked by the data inconsistency score (-log(p), where p is an aggregated p-value). Operationally, only sites with highest ranks and larger sizes are recommended for inspections.

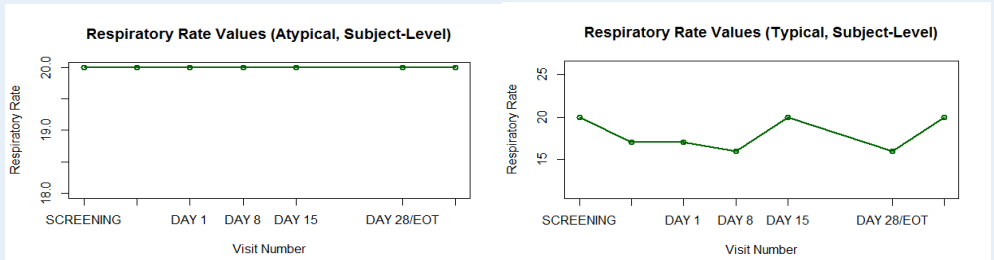
Results from one deidentified application are provided to demonstrate the data anomaly findings through the Statistical Monitoring Applied to Research Trials (SMART) analysis. Sensitivity analysis are performed after excluding laboratory data and questionnaires data.



Graphics from deidentified site-level trial data are provided to illustrate atypical and typical data patterns.



Graphics from deidentified subject-level trial data are provided to illustrate atypical and typical data patterns.



Results

Potential causes of data anomalies may include:

- Errors (technical problems, e.g. mis-calibrated thermometers)
- Sloppiness (incorrect report, e.g. under-reporting of AE)
- Tampering (fabricated, altered, or falsified data, e.g. modification of eligibility of criteria, or propagation of blood pressure)
- Sites that are atypical of the underlying population

Caveats and Limitations

- The software works best if there are at least 10 sites.
- Data preparation can be time consuming, but it is easier with SDTM data.
- If all the data are spurious, then an anomaly will not be detected. The software is designed to deal with problems primarily at the site level.
- The clinical significance of data anomalies may not be obvious.

Conclusions

- A data driven approach can be effective and efficient in selecting sites which exhibit data anomalies.
- Centralized Statistical Monitoring (CSM) can help ensure data quality and data integrity.
- CSM can provide insights to the statistical reviewers for conducting sensitivity analyses, subgroup analyses and site by treatment effect explorations.
- However, challenges exist with messy data and with the lack of conformance to SDTM data standards.