Paper **AR02**

A Data-driven and cost-effective approach towards Risk Based Monitoring

Vineet Jain, Nimble Clinical Research, New Jersey, USA

ABSTRACT

Pharmaceutical Industry has been increasingly using risk-based approaches to improve data quality and identify risks. Typically, such approaches are trial specific, costly and KRI/KPI driven. However, teams are interested in looking for deviations from study procedures, data entry errors, inaccurate measurements, sloppiness, or fraud in clinical trial data. Detection and resolution of such issues is part of risk based monitoring.

Such issues can be discovered by looking for patterns in data deviating from the general behavior observed in the remaining data. To detect and visualize such patterns (i.e. anomalies) in clinical trial data, a simple yet powerful data driven system is presented. It uses formal statistical tests along with unsupervised machine learning techniques. With availability of source data in CDISC standard, the approach can be applied on a clinical trial readily without any data preparation. The supporting software developed by us is freely available to any organization to use. The findings are presented as a list of data anomalies are supplemented with graphs, allowing end-users to quickly visualize the anomalous patterns.

INTRODUCTION

Ensuring data quality via effective monitoring is critical for sponsors in clinical trials. In recently published regulatory guidance [1], FDA and ICH have encouraged use of centralized monitoring methods and novel risk-based monitoring approaches.

Depending on size, design and complexity of clinical trials, data monitoring can be performed at various levels such as subject level, site level, country level, and cohort level. Most of the effort is usually spent at subject level data review, yet with such review, often cross-subject issues remain unnoticed until a point where no intervention can be performed.

Monitoring at a higher level of data abstraction is analogous to centralized monitoring. Risk based monitoring as a component of centralized monitoring has been gaining popularity. A review by Hurley et. al. [2] provides a good summary of popular risk-based monitoring tools provided by leading vendors.

In contrast to such monitoring methods, Central Statistical monitoring (CSM) is driven based on actual data collection. It has gained more attention in clinical literature [3,4] and from regulatory agencies [1] in the recent years. Although some of the CSM approaches can be quite complex, the concept behind such methods is simple. Data collected within a trial and across centers, countries or cohorts is homogenous. This is since, such data is collected as part of data collection plan outlined in clinical protocol. Any deviations from this general behavior within a site, country or cohort could be due to deviations from study procedures, data entry errors, inaccurate measurements from poorly calibrated medical equipment, sloppiness, or fraud. Hence, it is important to evaluate such anomalies to optimize monitoring activities, detect potential data issues during the early phase of a trial and ensure high level of data quality.

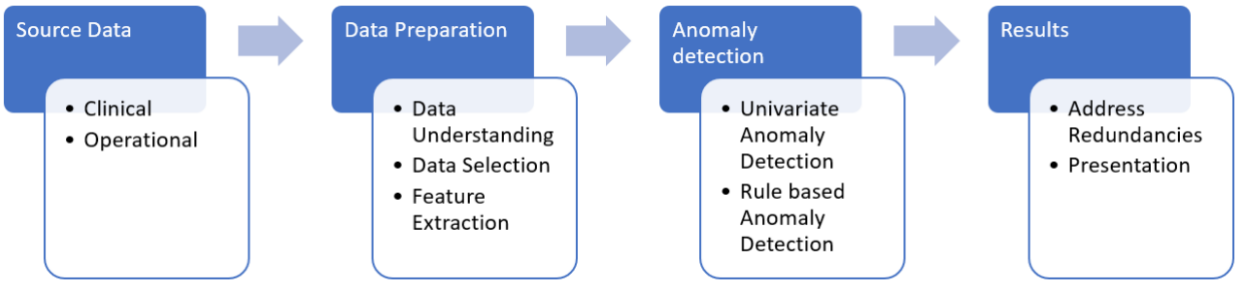
Subjects in clinical data can be pooled into groups in many ways. e.g. by site, country, cohort, age groups, investigator and monitor. Site level grouping is usually most interesting for monitoring. Our approach is designed to find data patterns in such groups deviating from the general behavior observed in rest of the clinical trial data. Such data patterns are referred as anomalies in the paper.

ANOMALY DETECTION

The brief overview of our anomaly detection framework is discussed in this section. The data preparation, statistical methodology and strategy used in the framework has been presented in detail earlier [5].

Figure 1 below presents the flow of information within the framework.

Figure 1: Anomaly detection Process



Clinical data collected from disparate sources need integration before it can be used for analysis. Additionally, the database design varies from one trial to the other, and one EDC vendor to other. These are obstacles towards use of the source data directly for reporting, data mining and analysis.

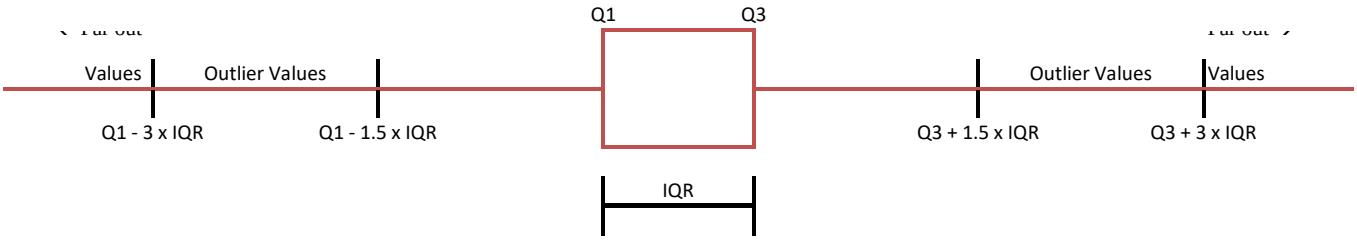
In order to perform unsupervised anomaly detection, we took advantage of the shift of industry towards CDISC standards. In CDISC based data, most data, irrespective of its source and format, is stored in standardized structure, e.g. ISO format for dates. Use of CDISC data allows to minimize upfront effort in data preparation. With CDISC based data, the data preparation can be highly automated. Otherwise, this can be very time-consuming activity.

To find anomalies, the data within a group (e.g. site or country) needs to be compared to remaining data. For this purpose, use of parametric statistical tests is not appropriate, since the distribution of sample data is unknown. Instead, non-parametric Wilcoxon rank sum test is selected to compare distribution, Conover squared ranks test [6] is selected to compare dispersion of numeric variables/features and 2x2 chi-square test is selected to compare binary distribution. Yates correction is applied for chi-square test, if any cell in 2x2 value matrix is less than 5.

Independent statistical comparisons of groups against rest of the data sample result in too many statistically significant yet uninteresting findings. This is because selected non-parametric tests ignore absolute values and do not account for inter-group variability. To address these shortcomings, we supplemented the non-parametric tests with an informal test - Tukey Fences (also known as Tukey Boxplot) [7].

Tukey fences consists of 'Outlier' and 'Far Out' fences beyond which values can be considered as outlier or Far out. These fences are created by using only the Q1, Q3 and IQR (Inter quartile range), i.e. Q3 – Q1. 'Outlier' fences are 1.5 IQR below Q1 and 1.5 IQR above Q3, whereas 'Far out' fences are 3 IQR below Q1 and 3 IQR above Q3 (see Figure 4). These arbitrary fences were proposed by John Tukey [7].

Figure 2: Tukey Fences



For Wilcoxon rank sum test, mean data values for each of the groups is selected as the statistic to perform test for Tukey fences. Similarly, standard deviation is selected to correspond with Conover squared ranks test and ratio of records is selected to correspond with chi-square test.

A pattern is flagged as an anomaly only when it passes through the formal non-parametric test, informal test criteria and minimum data criteria. Selection of thresholds for p-values, Tukey fences, and minimum data are discussed earlier [5].

ANOMALY VISUALIZATION

Visualization of anomalies in form of graphs is extremely useful, as it allows us to present huge amount of data as individual data points, making it possible to identify patterns & trends associated with the anomaly quickly and easily. Here examples of different types of anomalies are presented along with corresponding visualization.

UNIVARIATE ANOMALIES

Anomalies based on analysis of just a single variable or a single data feature can be referred to as univariate anomalies. These are easier to detect and visualize. Three different types of univariate anomalies are detected in the framework.

ANOMALIES IN DISTRIBUTION AND DISPERSION FOR NUMBER OF RECORDS PER SUBJECT

Here anomalies in distribution and dispersion of number of records per subject within a group against rest of the data sample are found. The following example provides an example of such anomaly.

EXAMPLE 1:

Consider the made-up clinical trial data where over the course of treatment period average 3-6 AEs are reported per subject by most sites. The site S12 consistently over reporting the number of AEs (e.g. 10-15 AEs per subject) could be of interest.

Figure 3: Visualization of anomalies in distribution of numerical values

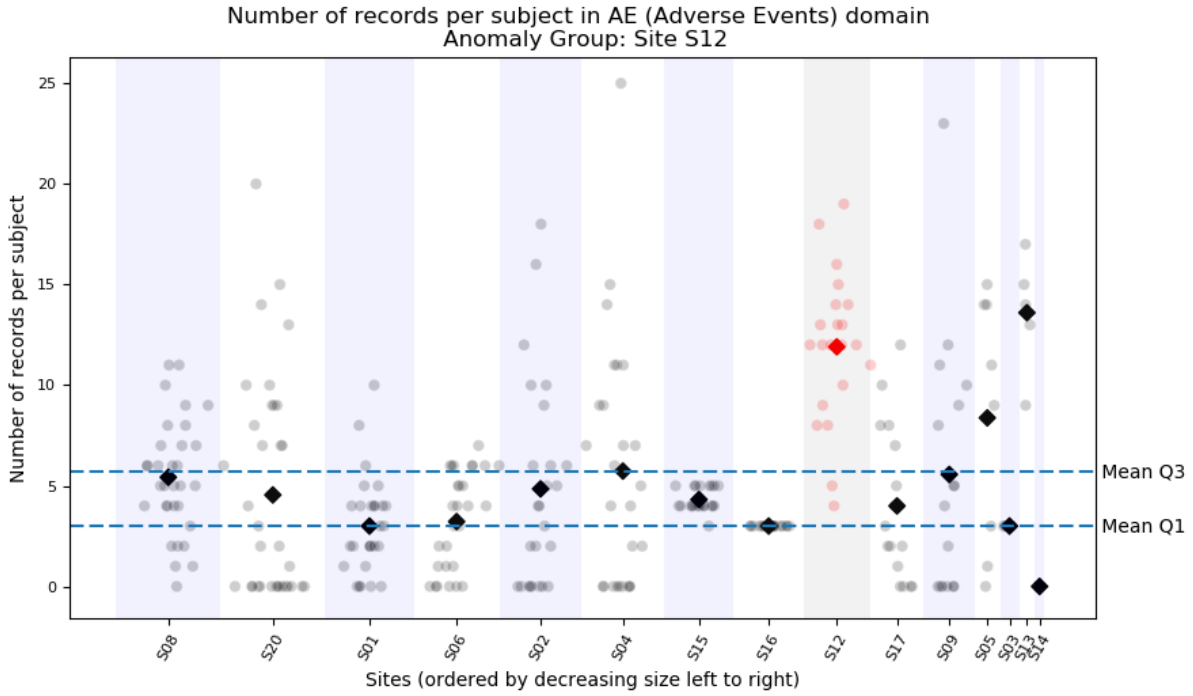


Figure 3 above plots the number of AE records for each subject by site. Each vertical bar represents a site with width of bar proportionate to number of subjects in the site. Sites are ordered by number of subjects within each site, with most subjects on the left and least subjects on the right. The diamonds in the figure represent mean number of records per subject in the site. Site S12 has many more AEs reported per subject (red dots) compared to most other sites. There could be simple and reasonable explanation such as the trial is still live and the site started earlier than

other sites, leading to lot more data per subject. Or it could be a training issue with the collection and reporting of the AEs by the site.

In addition, Site S16 has same number of adverse events for all the subject, which could be interesting pattern for further examination.

It is useful to evaluate such anomalies as early as possible and take corrective action if needed. Such data is flagged as anomaly, after it meets the criteria for Wilcoxon rank sum test (for distribution) or Conover squared ranks test (for dispersion) along with Tukey fences.

ANOMALIES IN DISTRIBUTION AND DISPERSION FOR NUMERIC VARIABLES/FEATURES

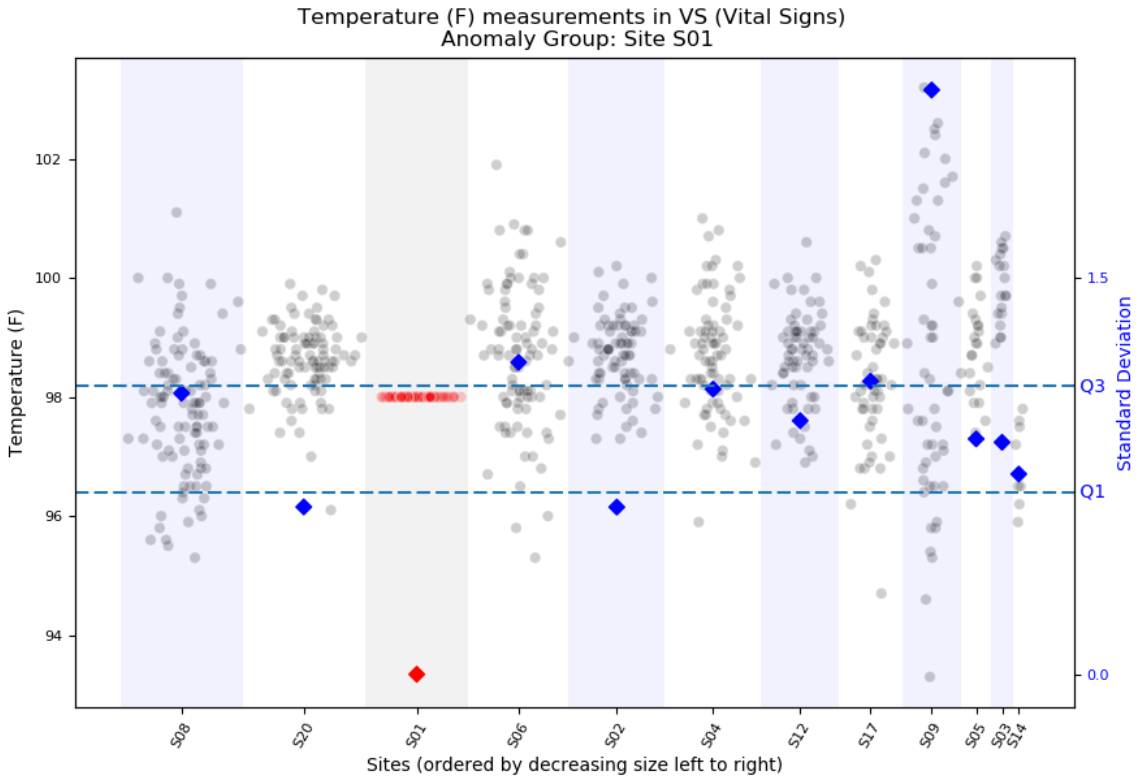
All selected numeric variables and other numeric features in the database are analyzed to look for anomalies in distribution and dispersion via Wilcoxon rank sum test and Conover squared ranks test (along with Tukey fences). Following example shows visualization of an anomaly in typical numerical vital signs data.

EXAMPLE 2:

In the made-up clinical trial data, a site has all the resting body temperature measurements for healthy subjects exactly same, whereas other sites have varying values for temperature for similar subjects.

Figure 4 plots all the temperature measurements across all subjects by site in the same way as in Figure 3. In addition, standard deviation is plotted with a different scale on the right Y axis. Most sites have standard deviation for temperature between .7 & 1.1. Site S01 has all the temperature measurements at 98 F. It would be worth exploring to find out the reason for such consistency in measurements. It could be due to data entry error or incorrect calibration of measuring instrument.

Figure 4: Visualization of anomalies in dispersion of numerical values



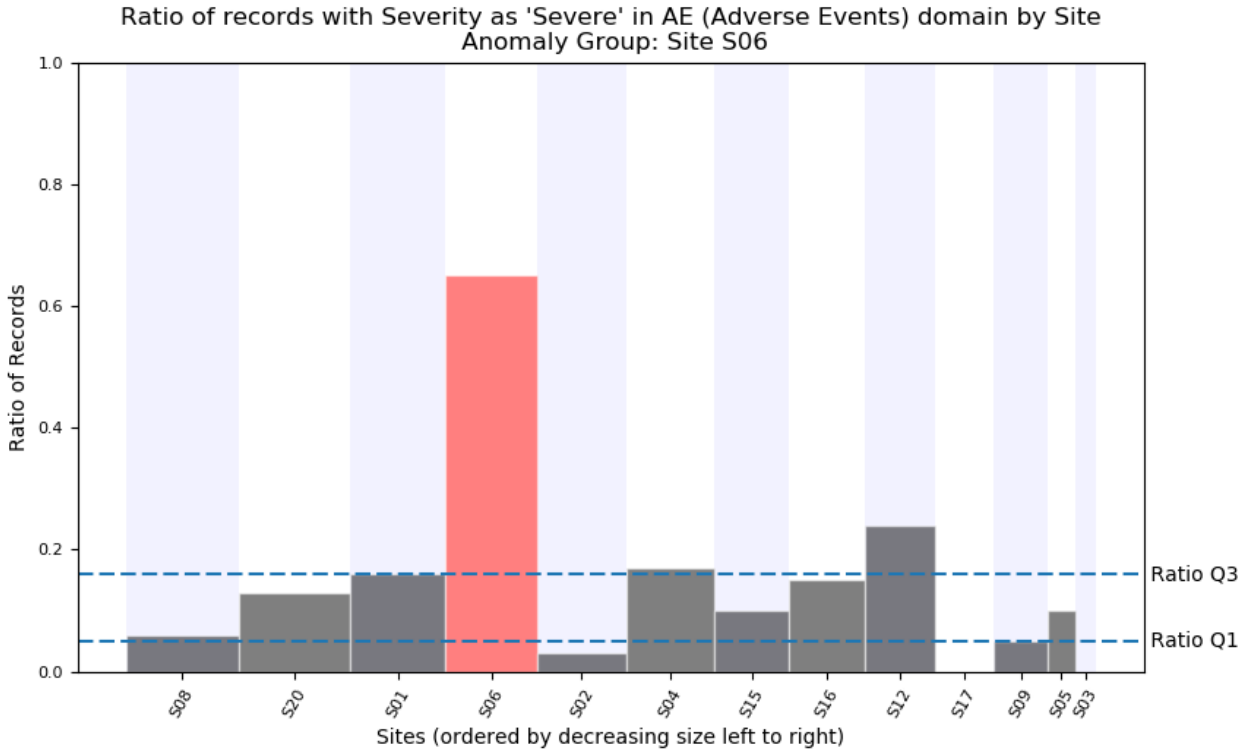
ANOMALIES IN BINARY DISTRIBUTION

Anomalies in binary distribution of frequent values and other binary features are identified via 2x2 Chi-square test (along with Tukey fences). In the chi-square test, counts of records with a data value absent vs present within a group is compared against the rest of the data sample. Following example shows visualization of an anomaly in such data using incidence of severity as 'Severe' in Adverse Event.

EXAMPLE 3:

In the made-up clinical trial data, a site has 65 % of AEs flagged as 'Severe', whereas other sites have 5-15% of records flagged as 'Severe' for AEs. Figure 5 plots ratio of AE records flagged as severe grouped by site as a bar plot. The layout of sites and IQR (inter quartile range) via Q1 & Q3 is same as in in previous figures. There could be reasonable explanation for the pattern, such as that patients are sicker at this particular site.

Figure 5: Visualization of anomalies in binary values



One would not expect frequent values in continuous numerical variables such as vital signs measurements. Nevertheless, we often find interesting anomalies in such data, as shown in following example. Consider diastolic blood pressure data collection in a trial. It is observed that 70 mmHg as diastolic blood pressure is a frequent value in about 6% data values. One of the sites had about 50 diastolic blood pressure measurements with 40% of all its values as 70 mmHg. Whereas, remaining sites in the trial had only 1-4% of values as 70 mmHg. Such site will be flagged as an anomaly via tests of binary distribution.

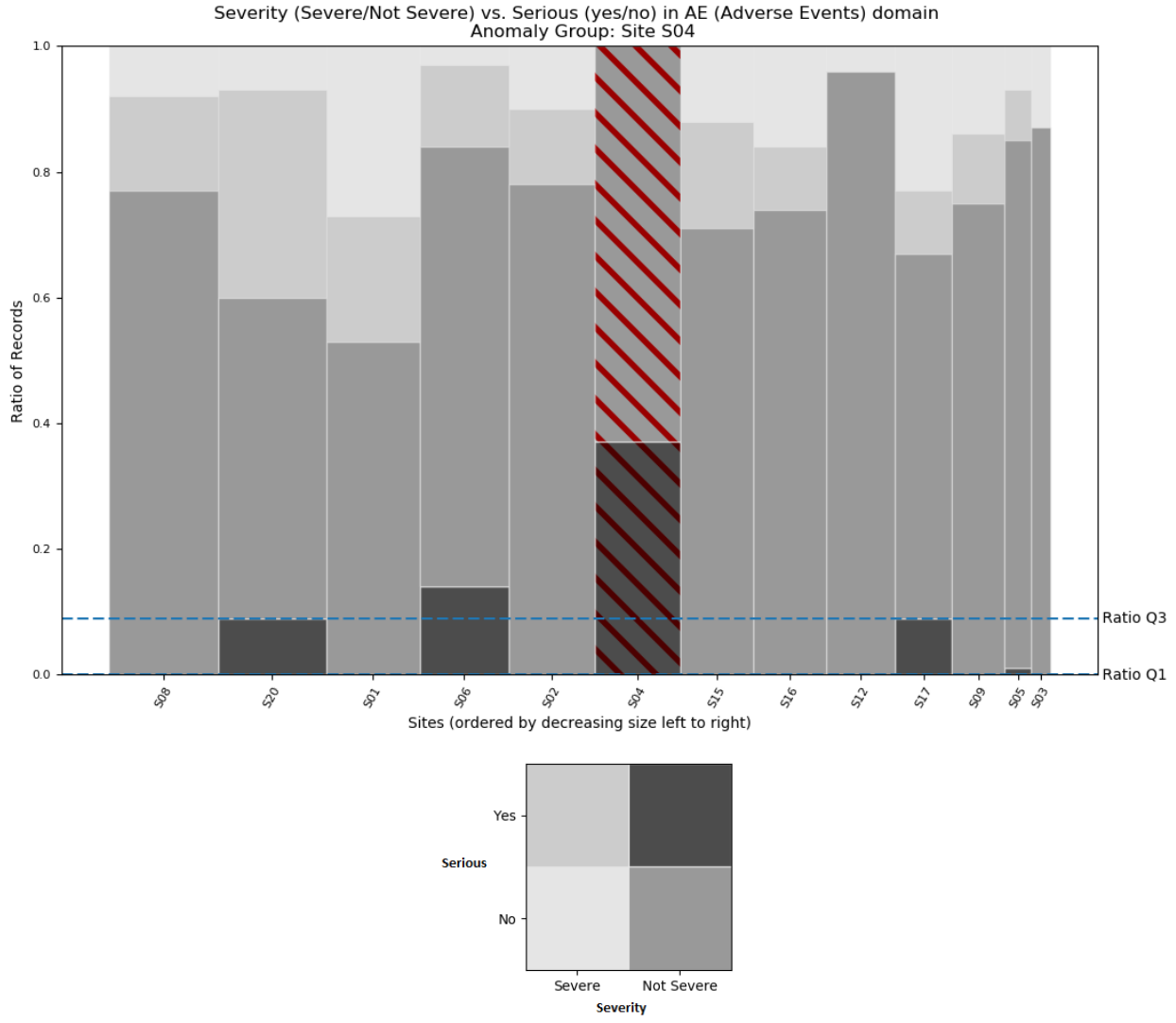
BIVARIATE RULE-BASED ANOMALIES

One of the goals of this framework is to perform deeper learning of the data to find hidden anomalous relationships across variables beyond just flagging outliers in univariate data. For example, in adverse event data, severity and seriousness of an AE are usually strongly correlated. Data for a site, deviating from such pattern can be flagged as anomaly. Our approach for discovering such bivariate rule-based anomalies is explained in detail in our previous paper [5]. Following example illustrates such anomalies visually.

EXAMPLE 4:
Consider made up clinical trial data, where typical sites have just 0-9% AE records flagged as ‘Severe’ and ‘Not serious’. A site with 37% records of its AE records flagged as ‘Severe’ and ‘Not serious’, can point to poor training at the site and may need corrective action.

Figure 6 visualizes this data by plotting ratio of records based on severity (Severe/Not Severe) and Seriousness (Yes/No) for each site in a stacked bar plot. The stacked vertical bars have 4 possible shades of grey for each site representing 4 possible values in 2 x 2 cross-combination of severity and seriousness. The layout of sites and IQR (inter quartile range) via Q1 & Q3 is same as in in previous figures.

Figure 6: Visualization of anomalies in bivariate associations



CONCLUSION

Analyzing anomalous findings from any risk-based monitoring system can be a difficult & tedious activity requiring support from cross functional team in digging and visualizing data. This anomaly detection framework is designed to help with this challenge. The anomalous findings are presented such that these easier to understand. All observed anomalies in the framework are visually depicted to get a deeper understanding of anomalous data w.r.t other data

groups. This makes it easier to quickly examine and select anomalies as potential issues before spending further resources on closer review of the data.

REFERENCES

1. U.S. Department of Health and Human Services, Food and Drug Administration. Guidance for Industry: Oversight of Clinical Investigations — A Risk-Based Approach to Monitoring. Available from: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM269919.pdf.

2. Hurley C, Shiely F, Power J, Clarke M, Eustace JA, Flanagan E, Kearney PM. Risk based monitoring (RBM) tools for clinical trials: A systematic review. Contemp Clin Trials. 2016; 51:15-27. doi: 10.1016/j.cct.2016.09.003.

3. Kirkwood AA, Cox T, Hackshaw A., Application of methods for central statistical monitoring in clinical trials. Clin Trials. 2013; 10(5):783-806. doi: 10.1177/1740774513494504.

4. François T. Practical Implementation of Central Statistical Monitoring. PhUSE. Mint-Saint-Guibrt (Belgium): CluePoints; 2013. Available from: https://www.lexjansen.com/phuse/2013/sp/SP10.pdf.

5. Jain V. An Open Clinical Data Quality Framework. PhUSE US Connect 2019.

6. Conover WJ. Practical Nonparametric Statistics. 3rd ed. Hoboken, NJ: Wiley; 1999. pp. 300-303.

7. Tukey JW. Exploratory Data Analysis. Reading (PA): Addison-Wesley;1977.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Vineet Jain
Nimble Clinical Research
New Jersey, USA
Email:vineet.jain@nimble-cr.com
Web:nimble-cr.com