PhUSE US Connect 2019

Paper AR03

An Open Clinical Data Quality Framework

Vineet Jain, Nimble Clinical Research, Westfield, NJ, USA

ABSTRACT

There has been increasing emphasis on clinical data quality and standardization in recent years in Pharmaceutical Industry. Typical approaches to improve data quality and identify risks are trial specific and costly.

In this paper, a simple yet powerful clinical data quality framework is proposed. It uses formal statistical tests and unsupervised machine learning techniques to find anomalies in the data. The increasing power of computer technology and adoption of CDISC data standards across industry makes it easier to apply such framework. With availability of standardized source data, the approach can be applied on clinical trials with little effort. Thus, data quality checking and risk identification can be made efficient, standardized and broadly applicable across more and more clinical trials.

The supporting software developed by us (Nimble Clinical Research) is freely available to any organization to use. It applies the framework and readily generates actionable list of potential issues for review by cross functional clinical teams.

INTRODUCTION

Ensuring data quality via effective monitoring is critical for sponsors in clinical trials. Accordingly, sponsors spend enormous amount of resources on data review and monitoring. In last two decades, the clinical trials have become increasingly complex in terms of study designs, variability in site experience, infrastructure and geographic spread. In recently published regulatory guidances [4], FDA and ICH have encouraged use of centralized monitoring methods and novel risk-based monitoring approaches over routine visits and 100% source data verification.

Depending on size and complexity of clinical trials, data monitoring can be performed at various levels such as subject level, site level, country level, and cohort level.

At subject level, monitoring and review are extremely tedious. Monitoring at subject level is usually performed thoroughly and therefore this method is extremely expensive. Sponsors usually have matured and thorough processes to monitor subject-level data and ensure data quality. Most Electronic Data Capture (EDC) systems allow creation of exhaustive edit checks for data entry and usually sponsors utilize such checks to flag issues upfront. Monitors are trained and experienced to review subject level data. Subsequently subject level data is thoroughly reviewed by multiple teams. For example, review of patient data reports via patient profiles or narratives is very common. It is not easy for human reviewers to detect cross-subject patterns with manual data review. With manual data review, broader issues are generally discovered randomly and often remain unnoticed until the point where no intervention can be performed.

On the other hand, data monitoring and review at a higher level of data abstraction such as the site level is not as well established and not as thoroughly performed. Monitoring at a higher data abstraction is analogous to centralized monitoring. Key activities as part of centralized monitoring are as follows: identification of key risks and their thresholds; designing a comprehensive monitoring plan and relevant documentation; continuous monitoring of risks or issues; and corrective action planning to manage/mitigate identified risk/issue(s).

Risk based monitoring is key component of centralized monitoring and this approach has been gaining popularity. A review by Hurley et. al. [1] provides a good summary of popular risk-based monitoring tools provided by leading vendors. Usually teams who adopt centralized monitoring, limit its use to creation and review of basic site-level operational data summaries, pre-defined operational/clinical reports and tracking of KRIs and KPIs (key Risk Indicators and Key performance Indicators).

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 [2,3] and from regulatory agencies [5] 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 and have common patterns. 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.

In this article, our novel approach towards CSM is presented. To begin with, objectives of our approach are outlined in the next section.

OBJECTIVES

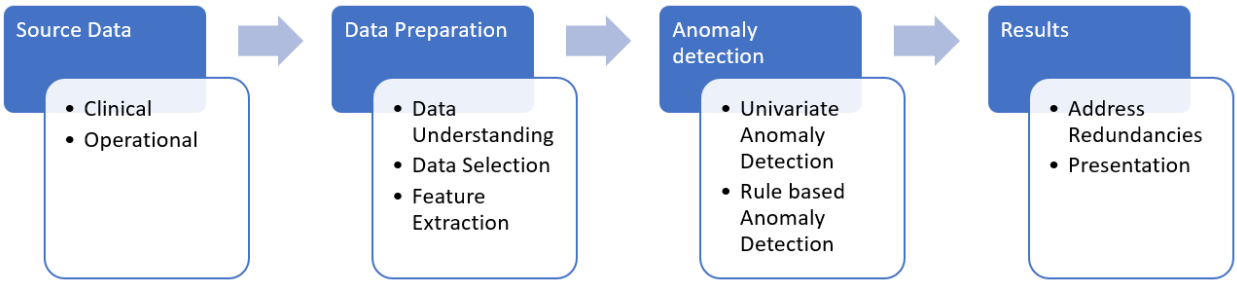
Subjects in clinical data can be pooled into groups in many ways. e.g. by site, country, cohort, age groups, investigator and monitor. Site level grouping is usually most interesting for monitoring. Our approach is designed to find data patterns in such groups deviating from the general behavior observed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

More specifically our goals are to:

- Detect anomalies in unsupervised manner as much as possible, with aim to ensure the framework can be efficiently used for a clinical trial and easily reused for subsequent similar clinical trials. Therefore, to achieve the applicability of the system, minimum study specific customization must be needed.
- Perform deeper learning of the data in order to find hidden anomalies beyond just flagging straightforward outliers via univariate statistics
- Use sound, simple and predictable statistical approaches allowing teams with diverse backgrounds to readily understand and analyze the findings.
- Be able to use the system for broader range of clinical trials in terms of its size, rather than limiting its application to large phase III trials with hundreds of sites.
- Keep the framework open for the industry, allowing collaboration and further improvements in the framework.
- Support the framework with stand-alone software that is freely shared with any other organization
- Present the findings as an actionable checklist that are easy to understand for cross-functional teams.
- Achieve higher signal to noise ratio in findings to ensure the end users spend time mostly on high quality and interesting findings. Reviewers would not want to waste time on endless findings that may be highly statistically significant but uninteresting otherwise.
- Visually highlight relative interestingness of the findings to bring attention to most interesting patterns

Figure 1 shows the framework presented in this article to achieve the above goals.

Figure 1: Anomaly detection Process



DATA UNDERSTANDING, SELECTION AND PREPARATION

Most of the clinical data is usually collected via EDC systems while the rest comes from the sources such as lab vendors, IVRS system, and ePRO (electronic patient reported outcomes). Data from these disparate sources usually need reconciliation and integration before it can be used for analysis. Additionally, the data collection design varies from one trial to the other, as typically most trials are custom designed. These are obstacles to use the source data directly for reporting, data mining and analysis.

In last decade, most of the industry has shifted to using CDISC data standard for regulatory submissions. Sponsors are increasingly converting source data into CDISC data further upstream in their processes, prior to database locks.

Due to use of CDISC standards upstream, clinical teams are getting educated with its standards and sponsors are gaining efficiencies with increasing standardization across the clinical data workflow.

In order to perform unsupervised anomaly detection, we took advantage of this 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 us to design framework that can work without the need to custom prepare data for every trial.

Knowledge of CDISC standards can be used to classify the role of variables before performing any analysis. Large number of variables in CDISC data are either redundant or useless for anomaly detection, such as visit identification variables, visit name (VISIT) and Visit number (VISITNUM). To minimize redundancies in the findings and make the anomaly detection process efficient, only one of these two variables should be used in analysis. Similarly, several variables store identifier information, such as record identifier and lab name. Such variables are irrelevant for anomaly detection and result in spurious findings if analyzed. Such redundant and useless variables should be skipped from analysis.

Date and time variables are not directly analyzable as absolute date and time of an event or data collection is not interesting. Instead, relative dates are much more interesting, as usually dates corresponding to scheduled data collection are expected to start/end certain number of days relative to other study dates. For example, visit 2 in a study may be expected to happen 30 days after screening visit 1. A site consistently deviating from the expected timing of second visit may be due to systemic problems at a site. At least one reference date should be picked to make date/time variables useful for analysis. Sometimes teams may want to use multiple reference dates; e.g. using Informed consent date, and treatment start date to find anomalies in events against either of the dates.

Some CDISC variables can be used to split data into subsets, e.g. test code (–TESTCD) variables can be used to perform analysis independently for each test in a vital signs type dataset. By using the understanding of CDISC data structure, variables can be automatically selected for data splitting.

Table 1 shows sample of the proposed rules for variable categorization. Variables that are similar to other existing variable (identified via column ‘Similar to’) or have role as ‘Skip’ are skipped from analysis. Variables, categorized as ‘Split’ variable, are used to split the source datasets to perform independent analysis. Variables not categorized into any of previously mentioned categories are analyzed for anomalies. For certain data, the default proposed variable selection may not be appropriate, and teams should override the default selection in such cases.

Table 1: Sample proposed rules for CDISC Variables

SDTM domain Type	Variable Name	Similar to	Role
All	ARMCD	ARM	
All	ACTARMCD	ACTARM	
Interventions	--DOSTOT	--DOSE	
Findings	--DRVFL		Skip
All	EPOCH		Skip
Findings	--EVALID		Skip
Findings	--LEAD		Skip
Findings	--LLOQ		Skip
Events, Interventions	--MODIFY		Skip
Findings	--NAM		Skip
All	--SEQ		Skip
Events	--SOC	--BODSYS	
All	--STINT		Skip
Findings	--STNRC		Skip
Findings	--TESTCD	--TEST	
Findings	--TEST		Split
DM	RFXSTDTC	RFSTDTC	
DM	RFXENDTC	RFENDTC	
All	USUBJID		Subject ID
All	VISITDY	VISITNUM	

Operational data is not part of CDISCs SDTM and SEND data standardization. If teams want to explore anomalies in operational data along with clinical data, this limitation can be easily overcome by creating operational data in SDTM like structure and including these along with other CDISCs datasets for this analysis. This would allow the system to be more effective by enabling reviewers to explore anomalies in both clinical and operational data together.

Generally, risk-based monitoring approaches emphasize identification of risks for key primary and secondary efficacy and safety variables. Although, this makes monitoring efficient, it restricts the scope of data review as part of central monitoring. In data-driven anomaly detection, it is not as important to focus on just a few targeted variables upfront. Instead it is easier to use most of useful clinical and operational data for anomaly detection. Later, relative interestingness of findings can be used to perform targeted review.

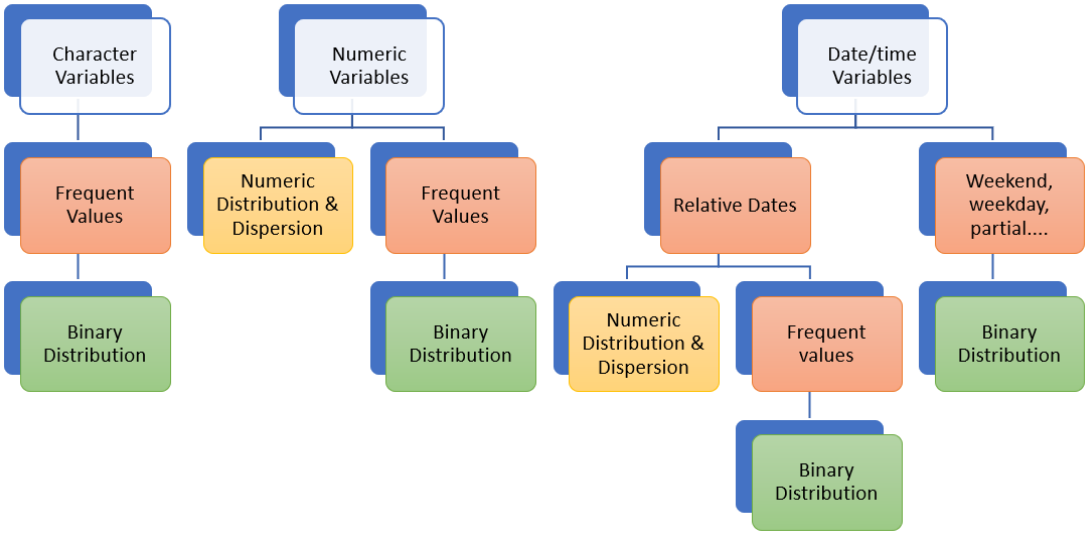
Even if CDISC based data is not available, the framework can be applied to source data by integrating the source data, selecting appropriate data for analysis and formatting the data values into analyzable format.

DATA FEATURES AND DISTRIBUTION

While performing unsupervised anomaly detection, it is not known whether variable is nominal, ordinal, interval or ratio. For numeric variables, without any prior information, it would be a stretch to assume normality or any other distribution. For character variables, it is not known in advance whether a known discrete set of values are expected with a specific distribution. A character variable may contain free text values whereas another character may have interesting pre-specified discrete values. Worse, often in CDISC character variables both pre-specified discrete values and hand-entered text are combined. Applying the prior knowledge of CDISC standard, does not help much in predicting distribution for most of these variables irrespective of whether they are numeric or character.

Keeping the previously mentioned concerns in mind, Figure 2 lists classification of variables, extracted features, and type of analysis applicable on these features. These are carefully selected to ensure anomaly detection is effective as well as efficient. All analyzable source variables are classified as numeric, text or date. Extracted features are either numeric or binary in nature. Numeric distribution and dispersion are compared for numeric features, whereas binary distribution is compared for binary features. Frequent values as a feature is extracted out of most variables. It is a data value that frequently occurs in a dataset (or it split) within in the range: Minimum data threshold (%), 100 - Minimum data threshold (%). The minimum data threshold is defined as percentage such as 5% of data or 10% of data. Missing data value can also be a frequent value. Pre-defined values in source data collection have high chances of being a frequent value. Whereas, hand entered text value in comment fields is much less likely to end up as frequent value. For date/time variables features, such as relative dates, weekend, weekday, partial date, morning, and evening, are calculated.

Figure 2: Classification, feature extraction and analysis type for analyzed variables



ANOMALY DETECTION

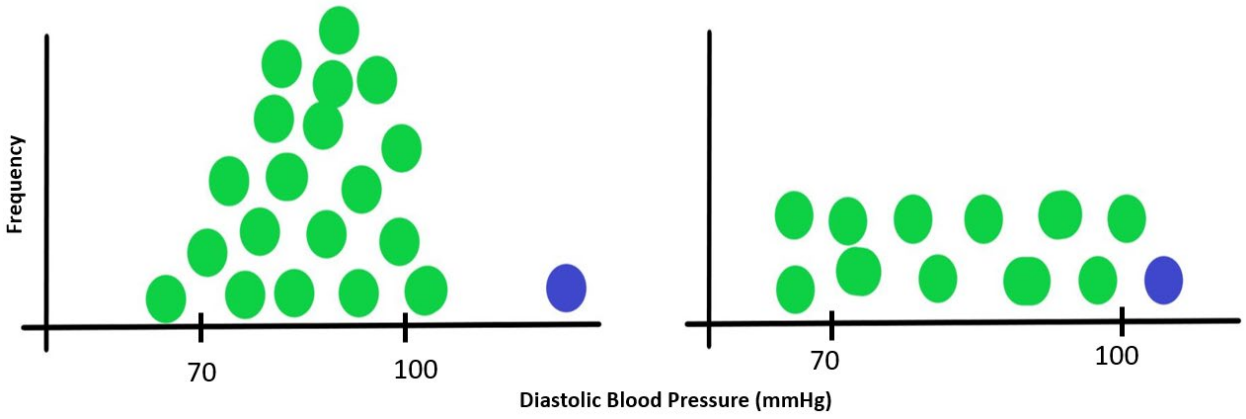
CHOICE OF STATISTICAL MODELS

Since the distribution of sample data and its extracted features is not known, use of parametric statistical tests is not appropriate. Instead, non-parametric Wilcoxon rank sum test is chosen to compare distribution, while Conover squared ranks test [5] is chosen to compare dispersion of numeric variables/features. All binary data are tested via 2x2 chi-square test. Yates correction is applied for chi-square test, if any cell in 2x2 value matrix is less than 5.

Independent statistical comparisons of all the groups against rest of the data sample result in too many statistically significant yet uninteresting findings. When performing over 1,000,000 statistical tests, one can easily end up with 1000s of useless findings with p-values < .00001. In our experiments, use of even smaller p-value threshold did not result in significant better-quality findings. The key reason for this phenomenon is that independent testing of Groups against rest of the data sample does not consider inter-Group variability in the sample data. Additionally, the selected rank-based models ignore the actual data values, which point towards how extreme the data is.

Figure 3 illustrates these issues. Assume two data samples with same size on left and right, where each dot represents an independent data group. For simplicity, all data groups have been assigned same size. With visual examination, we can consider the data group highlighted by blue color as outlier in first sample (left side). Whereas, the blue data group in the second sample (right side) does not seem an outlier. When this blue data group is compared against rest of data, via Wilcoxon Rank Sum test, same statistical significance is achieved in both data samples. This is because Wilcoxon rank sum test uses the relative rank to compare data, while disregarding absolute data values. Similarly, limitations can be demonstrated for Conover Squared Ranks test and Chi-Square test when comparing each data group independently with rest of the data. These limitations need to be addressed to improve the quality of findings.

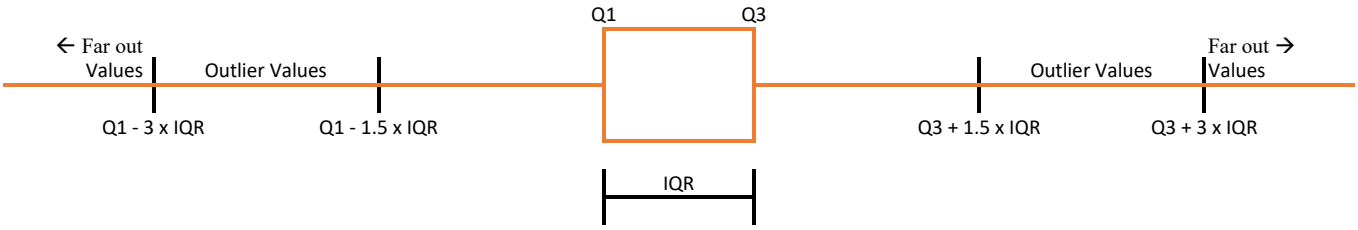
Figure 3: Sample data distribution to illustrate shortcomings in Non-parametric methods for Outlier detection



Anomaly detection, more commonly known as Outlier detection, can be broadly performed via two different ways. Methods based on formal statistics, such as ones selected above, need test statistics for hypothesis testing. Other methods such as boxplots can be classified as informal tests. These generate an interval, beyond which the values can be considered as outlier. Both approaches have its strengths and weaknesses. To account for inter-group data variability and overcome the weaknesses of selected non-parametric statistical tests, we selected an informal test - Tukey Fences (also known as Tukey Boxplot) [6], to complement the non-parametric statistical tests.

There are a few variations in boxplots, among which Tukey fences is widely used due to its simplicity for outlier detection. It 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 [6].

Figure 4: Tukey Fences



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

A pattern is flagged as an anomaly only when it passes through both the formal and informal test criteria, i.e. it is statistically significant with p-value lesser than specified threshold and its corresponding summary statistic is outside the Tukey fences. The selection of thresholds for p-values, Tukey fences, and minimum data are discussed in the next section.

THRESHOLDS

An anomaly is a data point that is distant from others. There is no rigid mathematical definition to define a data point as an anomaly. Ultimately, it is a subjective exercise whether an extreme value should be labelled as anomaly. As part of statistical testing, we need to choose thresholds for significance testing to qualify patterns as anomalies. Additionally, we need to pick thresholds for minimum volume of data to qualify and categorize anomalies. All such thresholds and their recommended values are discussed in the following paragraphs:

Significance level for Chi-Sq., Wilcoxon rank sum and Conover squared ranks tests: When the anomaly detection algorithm is used on entire clinical trial data, easily over 10^5 or even over 10^6 statistical comparisons are performed. To account for multiple comparisons, a highly significant cut-off for p-value such as 10^{-5} or 10^{-6} should be used as threshold. In our experience selection of either 10^{-5} or 10^{-6} as p-value cut-off, did not impact the number of findings as much since Tukey Fences criteria further refines the selection of patterns. Our recommended default cut-off for p-values is 10^{-5}, when most of clinical trial data is used in anomaly detection. Otherwise, a custom cut-off based on expected number of statistical comparisons can be picked.

Outlier and Far out Tukey Fences: John Tukey [6] proposed outlier and far out fences as 1.5 IQR and 3 IQR beyond Q1 and Q3. Our recommended values for Tukey fences are the same. Reviewers can later use distance from the IQR to determine relative interestingness of anomalies.

Minimum Data cut for binary data (%): This threshold is used to determine minimum percentage of data that needs to be present in both ends of binary data cut and frequent values. It greatly effects the quality and number of anomalies detected. A value too small (e.g. 2%) would result in too many spurious data patterns appearing as anomalies, whereas a value too high (e.g. 20%) may result in several interesting patterns missed. Our default value is 5% for this threshold.

Minimum No. of Records in a pattern: This threshold determines minimum number of records that must be available for a pattern to be evaluated. This threshold is heavily dependent on the number of subjects available for analysis. A larger study with thousands of subjects should have higher requirement for minimum number of records whereas a study with under hundred subjects should have the threshold set accordingly. Our default value is 80% of number of treated subjects available for analysis. This will ensure data with just one record per subject, such as disposition and demographics data, is used in analysis.

Minimum Data to differentiate a group with few vs ample data records: This threshold can be used to visually differentiate smaller vs larger groups in terms of number of records in the dataset. Distinguishing smaller groups from the larger groups is very useful during review of the findings, since groups with just a few records are more likely to generate spurious findings whereas larger groups with more records are lot more likely to have interesting anomalies.

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.

Distribution and Dispersion test for number of records per subject

These checks compare the distribution and dispersion of number of records per subject within a group against rest of the data sample. The following example supports better understanding of the purpose of these checks. Consider a made-up clinical trial consisting of very sick cancer subjects, resulting in frequent Adverse Events (AE) for most of the subjects. Over the course of treatment period 20-40 AEs are reported per subject by most sites. A site consistently underreporting the number of AEs (e.g. 3-5 AE only per subject) could be of interest. Perhaps the site has not yet entered the data or is not following study procedures per protocol design.

It is useful to evaluate such anomalies early in the study and take corrective action if needed. The test for distribution of number of records per subject would find such deviations. Since number of records per subject in selected data is numeric information, Wilcoxon rank sum test and Conover squared ranks test (along with Tukey fences) are used to test the distribution and dispersion.

Distribution and Dispersion tests for numeric features/variables

All numeric variable and other numeric features are tested for distribution and dispersion via Wilcoxon rank sum test and Conover squared ranks test (along with Tukey fences). A typical example of numeric variable is temperature measurement as part of vital signs data collection. A consistent pattern of higher or lower temperature values for a site compared to values from other sites could signal a data collection issue.

Binary Distribution test for frequent values and other extracted features for date/time variables

Since no distribution is assumed for any of the analyzed variables, frequent values are searched among all the analyzed variables either directly (numeric and character variables) or indirectly (date/time variables, where frequent values are obtained for calculated relative days). Anomalies in binary distribution of frequent values and other extracted features for date/time variables are analyzed via 2x2 Chi-square test (along with Tukey fences). In the chi-square test, counts of records with data values absent vs present are compared within a group against the rest of the data sample.

A typical example of frequent value would be severity of Adverse Event as 'Severe'. Much higher or lower ratio of Adverse Event observations as 'Severe', for a site compared to other sites, could point to site training issue.

It is not usual to expect frequent values in continuous numerical variables such as vital signs measurements. Nevertheless, we often find interesting anomalies in such data, similar to 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, typical sites in the trial had only 1-4% of values as 70 mmHg. Such site will be flagged as anomaly (if it meets all necessary thresholds).

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 of a clinical trial, severity as 'Severe' vs mid/moderate and seriousness of an AE are strongly correlated such that typical sites have just 0-3% records with severity as 'Severe' and seriousness as 'Not serious'. A site having 25% records of its AE records with severity as 'Severe' and seriousness as 'Not serious', can point to poor training of the site and may lead to corrective action.

There are many popular techniques for anomaly detection such as density-based techniques, and neural networks. These enable discovery of complex patterns in data, but such patterns can be hard to examine, visualize and utilize. Instead, we chose to go with intuitive rule-based anomalies which has its origins from association rules, a very popular data mining technique.

Rule discovery can be approached in many ways [7, 8]. We approach anomalous rule discovery first by detecting strongly correlated cross-variable data values/features (referred to as rules or simply as associations) and then by detecting the groups that deviate from these associations. Although rules with any number of variables can be created in rule-based discovery, we limit the discovery of rules to bivariate rules (i.e. based on combination of values/features from two independent variables). This is due to multiple reasons: rules with more variables tend to be redundant to simpler rules with lesser variables; It is harder to work with more complicated rules; and it gets computationally expensive to detect anomalies involving three or more variables.

Rule-based discovery is effective for categorical data but does not apply to continuous numerical data. To include numerical variables in rule-based discovery, numerical variables are transformed into three separate binary values – value < Q1 vs ≥ Q1, value > Q3 vs ≤ Q3, value within IQR (interquartile range Q3-Q1) vs outside IQR. Although the ranges are arbitrary, these are helpful to split data into familiar categories. The values obtained from binary division of data, along with frequent values extracted during data preparation, are referred as frequent values for the purpose of rule-based anomaly discovery.

As a first step of anomalous rule detection, strongly correlated associations are found by comparing binary distribution of frequent values via 2x2 chi-squared test. The frequent values in such association originate from two independent variables. The sum of two of the four smallest counts in the 2x2 contingency table need to meet the

minimum data criteria, otherwise the association is dropped. Such associations are dropped as almost perfectly correlated associations are not much useful to find anomalies. To find groups deviating from an association, the association is split into four components, analogous to four data cells in 2x2 contingency table. The binary distribution of each of these components is separately tested, comparing each group against rest of the data sample via 2x2 chi-square test.

To visualize the algorithm, consider following made-up clinical trial data to compare bivariate distribution of AE severity (severe=Yes/No) and AE seriousness for the site 101 against rest of the data. The following tables present counts of AEs for Serious AEs (Yes/No) vs. Severe AEs (Yes/No) for overall AE data and site 101 AE data. By quick visual check, it is evident that Seriousness and Severity are correlated in overall AE data. Secondly, the data for site 101 does not follow the same pattern as in overall data, mainly due to unusually high AEs counted as not serious but severe.

Overall AE Data			
		Severe	
		Yes	No
Serious	Yes	30	10
	No	10	30

Site 101 AE Data			
		Severe	
		Yes	No
Serious	Yes	5	2
	No	8	5

In the overall AE data, Seriousness and Severity are correlated with a statistically significant p-value of 0.000022. Assuming this bivariate association meets the necessary thresholds, each of its 4 components are then compared for the site 101 against rest of the data. Note, 'rest of the data' is obtained after removing Site 101 data from overall data.

		Serious = Yes & Severe = Yes		Serious = Yes & Severe = No		Serious = No & Severe = Yes		Serious = No & Severe = No	
		Yes	No	Yes	No	Yes	No	Yes	No
Sites	101	5	15	2	18	8	12	5	15
	Rest	25	35	8	52	2	58	25	35
P-values:		0.28		1		0.000095		0.28	

In the four separate comparisons presented above, only p-value in the third set is significant. This confirms our earlier visual analysis. Assuming this finding meets all the necessary thresholds, the third set will be reported as anomaly.

A component of an association is considered anomaly for a data group, when all three criteria are met: the chi-square p-value threshold, Tukey fences criteria, and minimum data criteria.

The rule-base detection method described in this framework, enables thorough search of anomalous bivariate patterns. Next section addresses redundancy and presentation of detected anomalies.

RESULTS

ADDRESSING REDUNDANCY

In our approach, we analyze same data in many possible ways by dissecting it, extracting its multiple features, and looking for cross-variable relationships. This often results in too many similar findings, ultimately pointing towards same underlying data issue. For example, all the following anomalies are similar: most of AEs selected as 'Severe', no AEs selected as 'Mild', and most AEs selected as 'Severe' & 'Serious' together. Independent review of such findings is a waste of effort. When any of these is carefully reviewed, checking the others is not necessary. With this premise, we consolidate the detected anomalies using the following approach (in the order specified):

- Within a group, when multiple anomalies originating from the same source variable are found, pick the simplest anomaly as representative anomaly, and discard rest of the patterns. In our approach we pick univariate anomalies over bivariate anomalies, numeric distribution related anomaly over numeric dispersion related anomaly, and numeric distribution/dispersion related anomaly over binary distribution of frequent values.
- If multiple anomalies with same level of simplicity are found, pick the most dominant (highest distance from IQR in terms of IQR) pattern.
- If exactly same anomaly exists across multiple data splits, then pick the anomaly from the first data split in the dataset. This consolidation is especially needed for normalized CDISC based datasets. Sometimes an anomaly exists in common data that repeats across multiple splits in a dataset. For example, same date and

time information of the vital sign's measurement is usually present across multiple vital measurement parameters. If there were an anomaly related to date/time of measurement, it would repeat across all related data splits and should be consolidated.

The previous consolidation steps bring decrease the overall number of anomalies to a more manageable number for reviewers.

PRESENTATION OF ANOMALIES

After consolidating the anomalies, we are left with targeted anomalies. Table 2 shows how these can be listed for review. One example for each type of anomaly is presented. Most of the information here is self-explanatory. The 'Records' column has number of records in the group out of total records in the dataset (or its data split) for numeric tests. Whereas for character tests, it has number of records meeting the data selection criteria out of both the group and rest of the data in the dataset (or data split). Percentage (%) column has the percentage of records in the group relative to total records in the dataset (or data split). Q1 and Q3 for Tukey fences are created using the entire data sample, whereas statistic value is calculated for the data group specified in first column. Using these values, distance from IQR is calculated in terms of IQR.

$$Distance\ from\ IQR = \frac{Absolute\ (Statistic - Qx)}{Q3 - Q1}, \quad where\ Qx\ is\ either\ Q1\ or\ Q3, whichver\ is\ closer\ to\ Statistic$$

For example, in the first finding Distance from IQR is calculated as (15.6 – 4.5)/(4.5 – 0) = 2.5.

Table 2: Sample findings

	Data					Records		Non-Parametric Statistics		Tukey Fences		
Group	Data-set	Data Split	Variables	Data Selection	Finding Type	Numbers	%	Model	p-value	Stat-istic	Q1, Q3	Distance from IQR
SITEID = '103'	CE	NA	NA		No. of records per subject	78 out of 345	22.6	Wilcoxon rank-sum	<.00001	Mean: 15.6	0, 4.5	2.5
SITEID = '101'	CM	Category for Medication (CMCAT) = 'CONCOM-ITANT MEDS'	Start Date (CMSTDTC) w.r.t to Treatment Start Date		Numeric Distribution	20 out of 523	3.8	Wilcoxon rank-sum	0.00005	Mean: 100.6	25, 50	2.0
SITEID = '103'	VS	Vital Signs Test Name (VSTEST) = 'Temperature'	Numeric Result/Finding in Standard Units (VSSTRESN)		Numeric Dispersion	210 out of 2212	9.5	Conover squared ranks	<.00001	SD: 1.13	.11, 0.27	5.4
SITEID = '105'	AE	NA	Serious Event (AESER)	AESER= 'Mild'	Univariate Binary Distribution	5 out of 34 vs. 352 out of 383	8.9	2x2 Chi-Square	<.00001	Ratio: .15	.75, 1	2.4
SITEID = '107'	CM	Category for Medication (CMCAT) = 'Rescue Medication'	Reported Name (CMTRT) & Dose per Administration (CMDOSE)	CMTRT = 'Dummy' & CMDOSE > 5	Bivariate Binary Distribution	15 out of 100 vs. 5 out of 353	22.1	2x2 Chi-Square	<.00001	Ratio: .15	0, .03	4.0

Knowing relative interestingness of the anomalies is useful for end users to perform targeted review. Columns '%' and 'Distance from IQR' can be used for this purpose.

Usually, groups with little data are more likely to generate spurious findings whereas larger groups with more data are more likely to have an interesting anomaly. In order to immediately bring attention to groups with more data, findings corresponding to groups with adequate data can be highlighted in '%' column as shown in the Table 2. There the minimum Data cut to differentiate a group with few vs ample data records is assumed to be default value of 5%.

Farther the data for a group is from typical data, more interesting the anomaly is. Distance from IQR provides exactly this information. Hence, the findings with distance from the IQR >= 3 (3 is default Tukey fence for 'far out' values) are highlighted to bring attention to more extreme data.

The findings with both '%' of records and distance from IQR exceeding the thresholds are likely to be most interesting.

CHALLENGES

This framework has been successfully tested on studies with the range of approximately 50 subjects to over 1000 subjects and proved to be effective for most of the clinical and operational data. When the framework is directly used on all available data (without any data selection), one may face computational challenges or end up with too many irrelevant findings.

By applying the knowledge of CDISC standards, default data selection is effective in selecting the representative variables for the known CDISC variables. However, clinical data has many more custom variables added to supplemental datasets (SUPP--). If all these variables are analyzed without careful data selection, many spurious findings may be generated. If an EDC dataset with many identifier and coded variables is directly used without any data selection, again many spurious findings would be generated.

Large normalized CDISC datasets such as labs and questionnaires can pose computational challenges. This is because the statistical comparisons are performed for each data split independently. A lab dataset with 100 lab tests may use 100 times (or more) the computational resources compared to typical adverse event dataset which does not need any data splits. Such problem is exacerbated for large phase III trial.

Another challenge is discovering bivariate patterns in datasets with too many variables. For instance, a dataset consisting of only two analysis variables, with five features each, will result in 25 statistical comparisons per group to discover bivariate patterns. Similarly, ten analysis variables, with five features each, will require 1125 statistical comparisons per group, which is still manageable. As the number of variables within a dataset keep increasing, such exhaustive cross-variable pattern discovery can go out of hand. This is not an issue with typical CDISC based datasets that usually have less than 20 interesting variables (including any variables in supplemental datasets). However, this can be an issue for raw horizontal datasets that in some cases run into hundreds of variables.

Both computational needs and quality of findings, can be controlled by targeted data selection and appropriate choice of thresholds. Especially for the troublesome data sources, it is recommended to select interesting variables/ tests/data and leave the rest, select representative data among interesting yet similar data, remove repeating data, and use de-normalized data if needed. Lenient values for the thresholds can lead to too many data comparisons and poor-quality findings.

CONCLUSION

In this article, a novel and robust approach to central statistical monitoring is introduced. It takes advantages of data standardization in industry and use it to detect data anomalies autonomously. Simple and sound statistical techniques have been combined with rule-based discovery to find anomalous patterns in data. Ease of use for end users is kept in mind to structure the findings such that these are easier to understand and utilize in further data examination.

While running large number of statistical comparisons (usually more than 10^5 and often over 10^6), many anomalous patterns are bound to emerge. Many of these of anomalies would be due to random odd patterns or due to demographic differences. On the other hand, many of the findings may point towards deviations from study procedures, data entry errors, inaccurate measurements from poorly calibrated medical equipment, sloppiness, or fraud. This framework can help in detecting such data issues early in the trial and help achieve better overall quality of clinical data.

REFERENCES

1. 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.
2. 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.
3. 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.
4. 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.
5. Conover WJ. Practical Nonparametric Statistics. 3rd ed. Hoboken, NJ: Wiley; 1999. pp. 300-303.
6. Tukey JW. Exploratory Data Analysis. Reading (PA): Addison-Wesley;1977.

7. Zhao Q, Bhowmick SS (Nanyang Technological University, Singapore). Association rule mining: A survey. .Technical Report, CAIS: Singapore; 2003. Report No.: No. 2003116..
8. Novak PK, Lavrac N, Webb GI. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging patterns and subgroup mining. Journal of Machine Learning Research. 2009; 10: 337-403.

CONTACT INFORMATION

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

Vineet Jain
Nimble Clinical Research
Westfield, NJ 07090
Email:vineet.jain@nimble-cr.com
Web:nimble-cr.com