A Hybrid Statistical-Machine Learning Approach to Anomaly Detection in Clinical Trial Data

Zachary A. Monge[1], Miao Chen[1], Viktor Rovskiy[1], Daniel Kowalski[1], Mohan Jayanna[1], Gordon Thomson[1], Kristin Stallcup[1], Jeremy D. Scheff[1], & Victor S. Lobanov[1]

[1]Covance, Informatics, Princeton, New Jersey, United States

Introduction

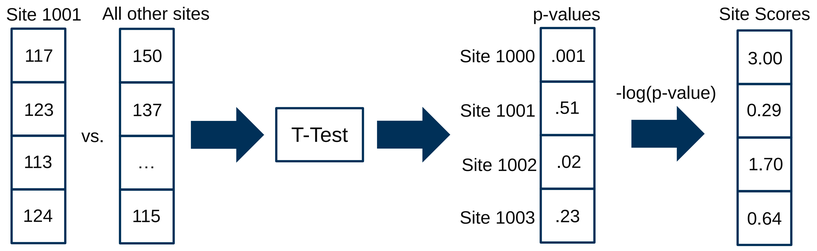
- There are many benefits to monitoring clinical trials via *centralized monitoring*, which refers to monitoring clinical trials remotely with the use of statistical methods, such as reduced monitoring costs, increased efficiency, and detection of complex anomalies.
- In terms of methodology to detect anomalies, the extant literature predominately utilizes traditional statistical tests, which are successful in detecting many types of anomalies, such as those related to miscalibrated machines or fraud.
- Within our anomaly detection product, Xcellerate Statistical Review, we utilize a battery of statistical tests to detect anomalies.
- Here, we extend our methodology with the use of unsupervised machine learning to (1) increase the reliability of detected anomalies and (2) to detect, in addition to site-level anomalies, participant-level anomalies.

Methods: Site-Level Anomalies

Former Approach to Anomaly Detection

In the former version of XSR, anomalies were detected by running a series of statistical tests (e.g., t-test, chi-square test) that compare variable data (e.g., diastolic blood pressure, adverse event severity counts) from each site to every other site, which generated p-values. To assign an anomaly score to each site, for each site, we calculated the negative log of the p-values (corrected; higher values = more anomalous), and the highest value from a site, which corresponds to a specific test and variable, is assigned as the site's anomaly score (i.e., site score). Below is an example with systolic blood pressure values:

Systolic Blood Pressure Example

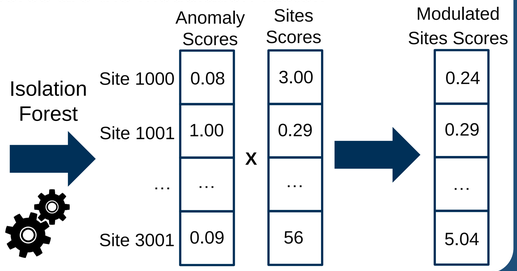


New Approach to Anomaly Detection

First, identical to the former method, first we calculated site scores. However, differently, the site anomaly scores were modulated by an anomaly modulation factor. This modulation factor was achieved by curating a dataset of p-values (uncorrected), where each row is a site and each column is a variable test. The dataset was submitted to an unsupervised machine learning algorithm, the isolation forest, which outputs an anomaly score, which were normalized so values arranged from 0 to 1 (higher values = more anomalous); these scores are the modulation factor.

P-Values from Each Variable-Test

Site ID	Blood Pressure T-test	…	AE Severity Type Chi-Square	Number of Subjects
1000	0.025	…	0.45	30
1001	0.1	…	0.03	15
…	…	…	…	…
3001	0.0004	…	0.1	2



Methods: Site-Level Anomalies

Test Data

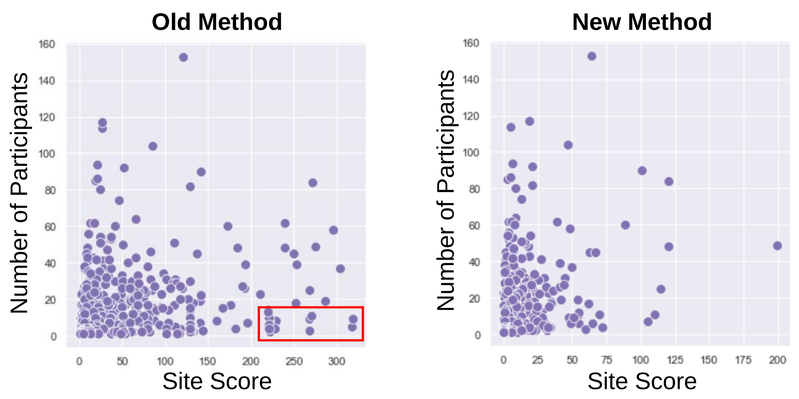
For methodological development, we used data from a confidential clinical study that contained 785 sites and 9,804 participants. For all analyses we examined 54 continuous variables and 115 categorical variables. The variables covered the following domains: adverse events, clinical events, demographics, exposure as collected, inclusion/exclusion criteria, laboratory results, medical history, pharmacokinetics, physical exam, procedures, questionnaires, reproductive system findings, subject characteristics, subject status, substance use, subject visits, trial arms, and, vital signs.

Results: Site-Level Anomalies

Comparing Old and New Anomaly Detection Scores

The old and new site anomaly scores are related to each other ($r(783) = 0.75$, $p = 2.08e\text{-}143$), but the sites that would have been marked as the top anomalous sites (old site score > 300) would no longer be marked as top anomalous sites.

Relation Between Site Scores and Number of Participants



We believe the new site scores are more reliable because of their relation to the number of participants, where sites with a small number of participants will likely yield noisy p-values. As can be seen in the figures above, for the old site scores (left), there are many sites with a small number of participants that had relatively high site scores, but for the new sites scores (right), this issue does not appear to be as prevalent. Also, subjectively, from inspection of the top anomalies, the new anomaly scoring system appeared to detect more anomalies that were operationally relevant.

Methods: Participant-Level Anomalies

For the second method, we detected participant-level anomalies. We developed an approach for continuous variables (e.g., diastolic blood pressure) and categorical variables (e.g., AE severity count).

Continuous Variables

For continuous variables, for each variable, participant values were compared to each other by calculating for each participant the robust z-score:

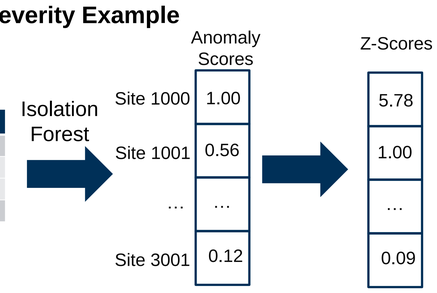
$$z = \frac{x_i - Median(x)}{1.486 * Median\ Absolute\ Deviation}$$

Categorical Variables

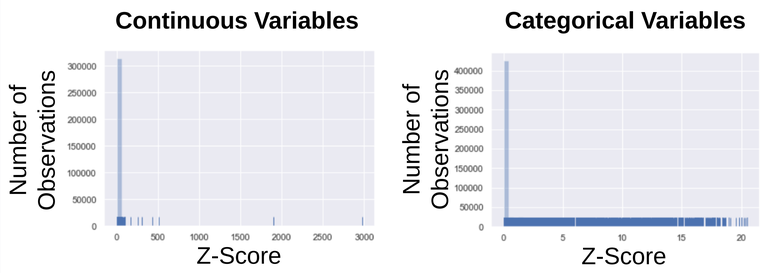
For categorical variables, for each variable and participant, the occurrence of each category was counted (e.g., number of mild, moderate, and severe adverse events) and then divided by the total count of occurrences, which yielded the relative proportion of each category. These values were submitted to an isolation forest, which yielded an anomaly score for each participant. To keep the scale of the anomaly scores similar to the continuous variables, these anomaly scores were converted to robust z-scores. Below is an example:

AE Severity Example

Site ID	Mild AE	Moderate AE	Severe AE
1000	0.25	0.50	0.25
1001	0.00	0.00	0.00
…	…	…	…
3001	0.90	0.10	0.00



Results: Participant-Level Anomalies



Above are histograms of the z-scores for each participant-variable combination, where higher scores indicate a more anomalous value.

Conclusions

In sum, we believe that our hybrid statistical-machine learning approach identifies more reliable anomalies and remains highly interpretable. In addition, the capability to detect participant-level anomalies allows for the detection of anomalies that may not be visible at the site-level view. In the future, we plan to extend this methodology to detect multivariate anomalies.

zachary.monge@covance.com