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

Zachary A. Monge, PhD[1], Miao Chen, PhD[1], Viktor Rovskiy, MSc[1], Daniel Kowalski, MS[1], Mohan Jayanna, BE[1], Gordon Thomson, BSc Hons[1], Kristin Stallcup, MS[1], Jeremy D. Scheff, PhD[1], Victor S. Lobanov, PhD[1]

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

Abstract

Recently, there has been increased interest in monitoring clinical trials remotely with the use of statistical methods. Here, we extend current methodologies to detect anomalies in clinical trial data with the use of unsupervised machine learning to increase the reliability of detected anomalies. First, similar to past methods, we conducted a number of statistical tests (e.g., t-test) on various variables (e.g., pulse) comparing each clinical trial site to every other site, which, for each site, generated a p-value for every test and variable. These p-values were converted into site scores (higher scores=more anomalous). To increase the reliability of the site scores, these scores were multiplied by a modulation factor, which was derived from submitting the p-values from each site to an unsupervised machine learning method; the output was the modulation factor (higher scores=more anomalous). With the use of this method, we found on example data that the detected anomalies appeared to be more reliable, where the modulation factor dampened site scores from sites with a small number of participants, which are more likely to yield noisy p-values. We also further extended this method to detect participant-level anomalies. Overall, we believe our work will allow for faster and more accurate detection of clinical trial anomalies.

Research Category: Methodological Research

Background

Recently, there has been increased interest in monitoring clinical trials via centralized monitoring, which refers to monitoring trials remotely with the use of statistical methods (1). This interest stems from the many benefits of centralized monitoring, such as reduced monitoring costs (2), increased efficiency (3), and detection of complicated anomalies (4). In terms of methodology to detect clinical trial anomalies, the extant literature predominately utilizes traditional statistical tests (5,6), which are successful in detecting many types of anomalies, such as those related to fraud and miscalibrated machines (1,6). Similar to others, within our anomaly detection product, Xcellerate Statistical Review (XSR) (7), we utilize statistical tests to detect clinical trial anomalies. Although this approach has been successful, here, we extend this methodology with the use of unsupervised machine learning to (a) increase the reliability of detected anomalies and (b) to detect, in addition to site-level anomalies, participant-level anomalies.

Methods

In the current version of XSR, anomalies are 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. The XSR statistical library contains several tests designed to detect different types of anomalies and these tests are ran on every requested variable. 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. Here, we extend this methodology with two methods.

For the first method, identical to above, first, we calculated site scores. However, differently, the site anomaly scores were modulated by an anomaly modulation factor. The goal of the modulation factor was to (1) maintain high site anomaly scores for sites in which the p-values appear to detect true anomalies and (2) dampen scores from sites which may have yielded noisy p-values (e.g., small sites). 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. This dataset was then submitted to an unsupervised machine learning algorithm – the isolation forest (8). Briefly, the isolation forest assumes that data points that are anomalous can be easily isolated, which is achieved by making random splits within the data until no more splits can be made. This method allows for the detection of patterns of sites' p-values that are less noisy and more likely true anomalies. The output of the isolation forest is an anomaly score, which we

normalized so values ranged from 0 to 1 (higher values=more anomalous); these scores are the modulation factor. Again, the site anomaly scores were multiplied by the modulation factor.

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). For continuous variables, for each variable, participant values were compared to each other by calculating for each participant the robust z-score (9). 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. These scores were converted to robust z-scores.

For methodological development, we used data from a confidential clinical trial study that contained 785 sites and 9,804 participants. For all analyses we examined 54 continuous variables and 115 categorical variables. The results reported below are from this data.

Results

For the first method, to detect site-level anomalies, as can be seen in Figure 1A, 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 anomaly score > 300) would no longer be marked as top anomalous sites. We believe that the new anomaly score is more reliable because of its 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 Figure 1B, for the old anomaly scores (left), there are many sites with a small number of participants that had relatively high anomaly scores, but for the new anomaly scores (right), this issue does not appear to be as prevalent. Also, subjectively, from visual inspection of the top anomalies, the new anomaly scoring system appeared to detect more anomalies that were operationally relevant.

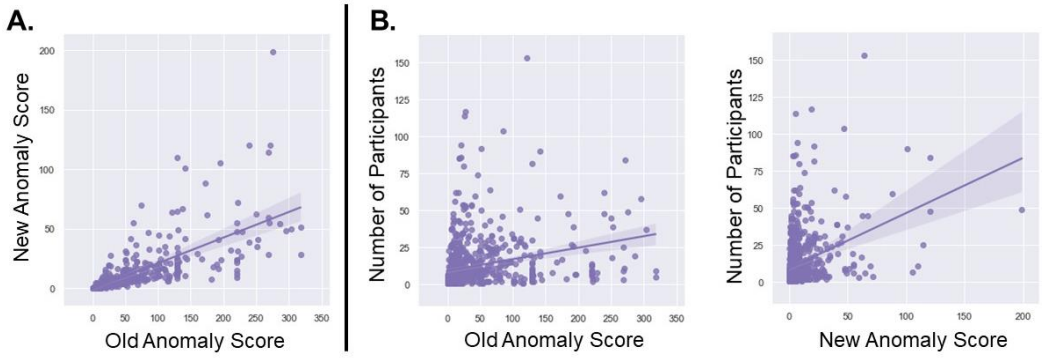

Figure 1. Panel A shows the relation between anomaly scores from the old and new methods. Panel B shows the relation between the anomaly scores (left=old method; right=new method) and number of participants.

For the second method, to detect participant-level anomalies, there was a distribution of anomaly scores skewed toward the right (higher values=more anomalous). Participants with higher anomaly scores would be marked as anomalous and would require further investigation.

Conclusion

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.

References

1. Venet D, Doffagne E, Burzykowski T, Beckers F, Tellier Y, Genevois-Marlin E, et al. A statistical approach to central monitoring of data quality in clinical trials. Clin Trials. 2012;9(6):705–13.
2. Eisenstein EL, Collins R, Cracknell BS, Podesta O, Reid ED, Sandercock P, et al. Sensible approaches for reducing clinical trial costs. Clin Trials. 2008;5(1):75–84.
3. Baigent C, Harrell FE, Buyse M, Emberson JR, Altman DG. Ensuring trial validity by data quality assurance and diversification of monitoring methods. Clin Trials. 2008;5(1):49–55.
4. Buyse M, George SL, Evans S, Geller NL, Ranstam J, Scherrer B, et al. The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. Stat Med. 2005;18(24):3435–51.
5. Knepper D, Fenske C, Nadolny P, Bedding A, Gribkova E, Polzer J, et al. Detecting data quality issues in clinical trials. Ther Innov Regul Sci. 2015;50(1):15–21.
6. Timmermans C, Venet D, Burzykowski T. Data-driven risk identification in phase III clinical trials using central statistical monitoring. Int J Clin Oncol. 2016;21(1):38–45.
7. Dimitris AK, Lobanov VS, Farnum MA, Yang E, Ciervo J, Walega M, et al. Risk-based monitoring of clinical trials : an integrative approach. Clin Ther. 2018;40(7):1204–12.
8. Liu FT, Ting KM. Isolation forest. In: Eighth IEE International Confrence on Data Mining [Internet]. 2008. p. 413–22. Available from: https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf?q=isolation-forest
9. IBM. Modified z score [Internet]. Available from: https://www.ibm.com/support/knowledgecenter/SSEP7J_11.1.0/com.ibm.swg.ba.cognos.ug_ca_dshb.doc/modified_z.html
10. Chakraborty C, Joseph A. Machine learning at central banks. SSRN Electron J. 2017;
11. Kroll B, Schaffranek D, Schriegel S, Niggemann O. System modeling based on machine learning for anomaly detection and predictive maintenance in industrial plants. In: 19th IEEE International Conference on Emerging Technologies and Factory Automation, ETFA 2014. 2014.