Paper IC-10

Collaborative Development of a Clinical Trial Data Anomaly Detection Tool within an Industry Consortium

Pekka Tiikkainen, Bayer Pharma and IMPALA Consortium, Espoo, Finland

ABSTRACT

Identifying anomalies early in on-going clinical trials data is crucial for ensuring the integrity and quality of data for submissions. However, this can be a time-consuming and challenging task, often requiring extensive manual review. To improve this, the IMPALA consortium, a collaboration of world-leading bio-pharmaceutical companies, is working to cooperatively develop an anomaly detection tool originally developed at Bayer for flagging timeline outliers at study sites and subjects. Our vision is to make the tool a widely accepted best practice, used by both the industry and the regulatory authorities alike. In an effort to promote transparency and collaboration, the tool will be made publicly available as an open-source R package in early 2024. As the IMPALA members continue to enhance the tool, FDA and other authorities are encouraged to join the project. Together, we could enhance clinical trial outcomes, and ultimately, improve patient care.

INTRODUCTION

Clinical trials play a pivotal role in advancing medical research, and the integrity of data generated from these trials is of utmost importance. However, anomalies in clinical trial data, which can compromise the validity of the trials, are not uncommon. These anomalies can arise due to a variety of reasons.

For instance, procedural errors during the trial, such as incorrect application of the study protocol, can lead to inconsistencies in the data. Similarly, data entry errors, where information is incorrectly recorded or input into the database, can also result in anomalies.

In some cases, anomalies can be a result of more serious issues like fraudulent activities at the study site, including data fabrication or falsification. Other potential causes include equipment malfunction, which can generate erroneous readings, and biological variability among study subjects, which can lead to unexpected outliers in the data.

The IMPALA consortium [1], a collaborative effort of leading bio-pharmaceutical companies, is actively working on addressing these challenges. IMPALA are developing an anomaly detection tool known as Clinical Trial Anomaly Spotter (CTAS), which is designed to identify anomalies efficiently and accurately in clinical trial data.

As of December 2023, CTAS is freely available an open-source R package for any interested party to use [2]. The consortium invites regulatory bodies like the FDA to join this endeavor, with the shared goal of enhancing the quality and integrity of clinical trial data, and ultimately, improving patient care and outcomes.

IMPALA

The Inter coMPany quALity Analytics (IMPALA) Consortium is a collaborative initiative that brings together leading biopharmaceutical companies. The consortium's primary objective is to revolutionize quality assurance methods in clinical trials by leveraging advanced analytics and best practices.

IMPALA provides a platform for its member companies to work collectively, sharing knowledge and resources to develop innovative solutions that address the challenges faced in clinical trials. The consortium's work extends beyond anomaly detection, focusing on a broad range of aspects related to clinical trial data quality and integrity.

In addition to CTAS, IMPALA has been working on a variety of other products and initiatives. These initiatives include the development of a tool for detection of adverse event underreporting [3], and workshops on quality data analytics [4]. IMPALA also offers a wealth of resources to its members, including access to work products, webinars, and events. The consortium regularly publishes updates on its initiatives, providing the public with the latest information on the progress of their projects.

The consortium actively engages with health authorities, promoting a collaborative approach to enhance the quality and integrity of clinical trial data. This engagement is crucial in fostering a regulatory environment that supports the adoption of advanced analytics and best practices in clinical trials.

In summary, the IMPALA consortium is a pioneering initiative in the biopharmaceutical industry, dedicated to revolutionizing quality assurance in clinical trials. Through its collaborative approach and innovative projects, IMPALA is making significant strides in enhancing the quality and integrity of clinical trial data, ultimately improving patient care and outcomes. For more information, visit the IMPALA website at https://impala-consortium.org/.

CENTRAL STATISTICAL MONITORING

Central Statistical Monitoring (CSM) is an increasingly utilized approach in clinical trials that leverages statistical algorithms to review and manage data. This method is designed to ensure the quality and integrity of data collected during clinical trials, which is crucial for the approval of new therapies and interventions.

CSM works by identifying anomalies and outliers in clinical trial data that may indicate issues such as procedural errors, data entry mistakes, or even fraud. The statistical algorithms used in CSM are capable of analyzing large volumes of data, providing a more efficient and practical means of monitoring compared to traditional methods.

One of the traditional methods of data verification in clinical trials is Source Data Verification (SDV). SDV involves a 100% review of all data collected in a clinical trial against the original source of the data. While this method is thorough, it is also time-consuming, labor-intensive, and costly [5, 6]. Moreover, SDV does not necessarily guarantee the detection of more complex patterns of anomalies, particularly those that may arise from systemic issues or fraud [7].

In contrast, CSM offers several advantages over 100% SDV. Firstly, CSM is more efficient, as it can process and analyze large volumes of data quickly, freeing up resources that can be used elsewhere in the trial. Secondly, CSM is more cost-effective, as it reduces the need for extensive manual review and on-site monitoring. Thirdly, CSM can identify systemic issues that may be missed by SDV, as it analyzes data at an aggregate level and can detect patterns that may indicate underlying problems.

Furthermore, CSM promotes a risk-based approach to monitoring, as it allows for the prioritization of sites and data points based on the risk of anomalies. This approach is more strategic and targeted compared to the blanket approach of 100% SDV and is in line with the guidance provided by regulatory bodies such as the FDA [8].

CLINICAL TRIAL ANOMALY SPOTTER (CTAS)

The Clinical Trial Anomaly Spotter (CTAS) is a powerful tool developed and published by the IMPALA Consortium for Central Statistical Monitoring. The package is designed with a high degree of user flexibility in mind, with allowing the user to steer the algorithms with various input parameters.

The underlying algorithm of the CTAS operates by defining one or more time series for each parameter. The algorithm summarizes time series into a set of features such as Mean, Standard deviation, Unique value count, and Autocorrelation (Figure 1). These features help in identifying individual subjects with suspicious data.

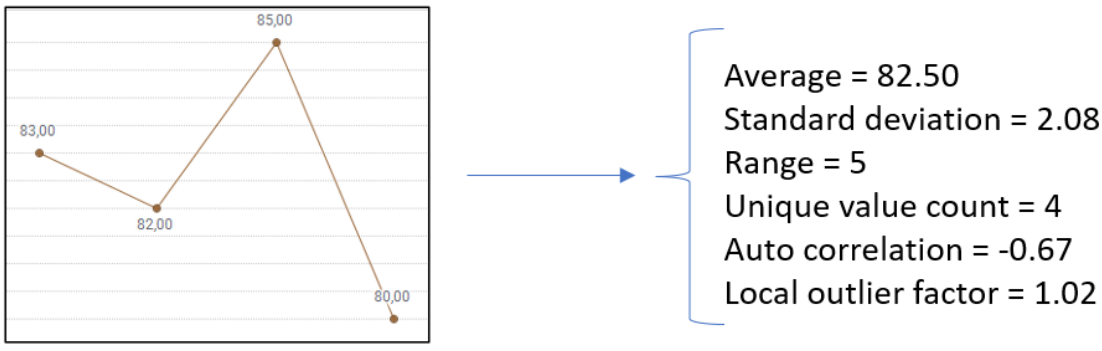

Figure 1. Example of features which summarize a time series.

Furthermore, trial sites with systematic bias in their data are flagged. For example, all albumin measurements at a site might have a growing trend while there is no trend at the study level. The algorithm uses the Kolmogorov-Smirnov test to compare sites' time series feature distributions and quantify bias. The raw p-value obtained from the test is corrected for multiple testing, and the negative logarithm of the corrected p-value is taken to come up with the final score for the site (Figure 2).

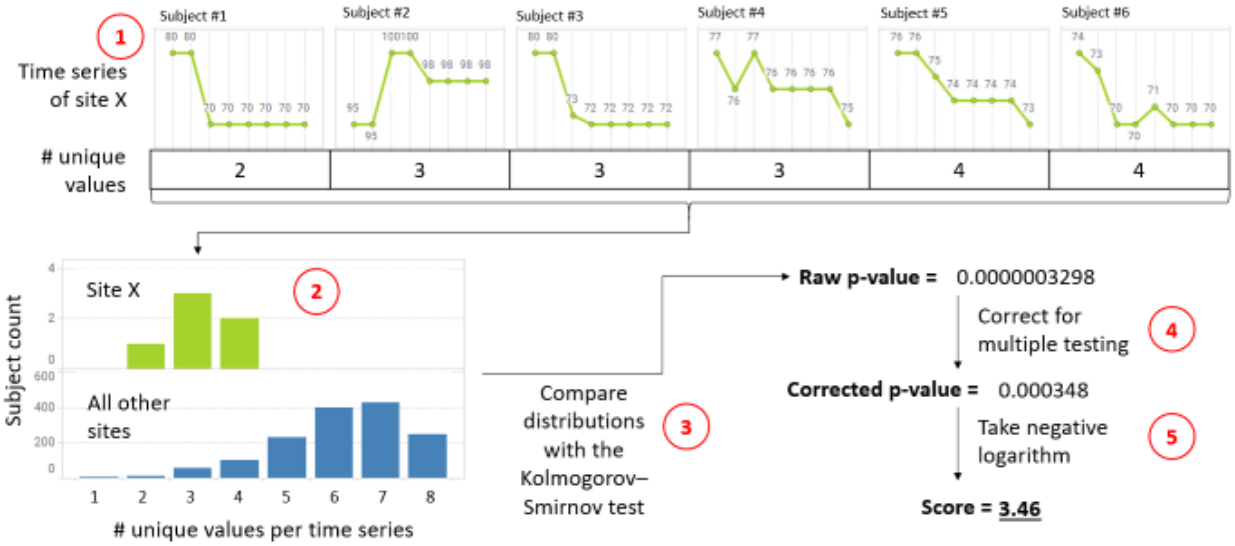

Figure 2. Example of site flagging with CTAS. The timeseries has eight time points and the question is whether the site (six subjects) has reported fewer unique values per time series than other sites in the study. Part 1) has the individual subject time series and the unique value counts. It is clear from the histograms (2) that the site is biased when compared to other sites. To quantify the bias, the distributions are compared with the Kolmogorov-Smirnov test (3) which gives us a raw p-value. As we perform several tests per study, the p-value must be corrected (4). Finally, the negative logarithm of the corrected p-value is taken to come up with the final score for the site (5).

An earlier version of CTAS was originally internally developed at Bayer [9]. The tool was shared with IMPALA partners for co-development with the vision that having more developers would lead to improved code quality and more ideas for new features. In essence, the collaborative approach is a testament to the commitment of both Bayer and the IMPALA Consortium to advancing the field of clinical trial data analysis and promoting best practices within the industry.

After few months of co-development within IMPALA, the CTAS R package source code was published in December 2023 and is hosted on the IMPALA Consortium's GitHub repository [2]. The consortium openly invites all interested partners to test the package and provide feedback, whether it's suggestions for new features or reporting of bugs. This collaborative approach will only serve to enhance the CTAS R package and its effectiveness in ensuring data quality in clinical trials.

In summary, CTAS is an AI-powered tool that identifies anomalous sites by using statistical significance testing. Additionally, the team is exploring the integration of a modeling approach for site flagging, which for instance would automatically adjust for local demographics in the site scoring process.

It is important to stress that while advanced AI/ML methods should also be evaluated, they should demonstrate significant improvement in performance over simpler approaches. Otherwise, they would only lead to increased complexity.

ANOMALY DETECTION PIPELINE WITH CTAS

To transition CTAS into a production environment, merely possessing the R package is insufficient. The initial step involves pre-processing in-house data to conform to the format that CTAS accepts, as illustrated in Figure 3. This process entails the selection and derivation of clinical parameters, as well as the definition of other CTAS arguments. Detailed documentation regarding the input data format and arguments can be found on the CTAS Github repository page [2].

Following the data processing by CTAS, the results necessitate visualization, and any detected signals require review and tracking. Currently, the CTAS package does not incorporate tools for these downstream activities. It is presumed that each user will utilize their own standard tools and processes for these tasks.

Figure 3. High-level overview of a production pipeline for anomaly detection with CTAS.

EXAMPLES OF SIGNALS DETECTED WITH CTAS

In the following, examples of real clinical trial anomalies are given. Due to protect intellectual property protection and patient privacy reasons, details such as parameter and subject IDs are not shown.

Figure 4 and Figure 5 illustrate site-level signals detected with the tool. CTAS can also be used to identify subject-level anomalies by identifying extreme time series feature values. Examples of such findings are given in Figure 6 and Figure 7.

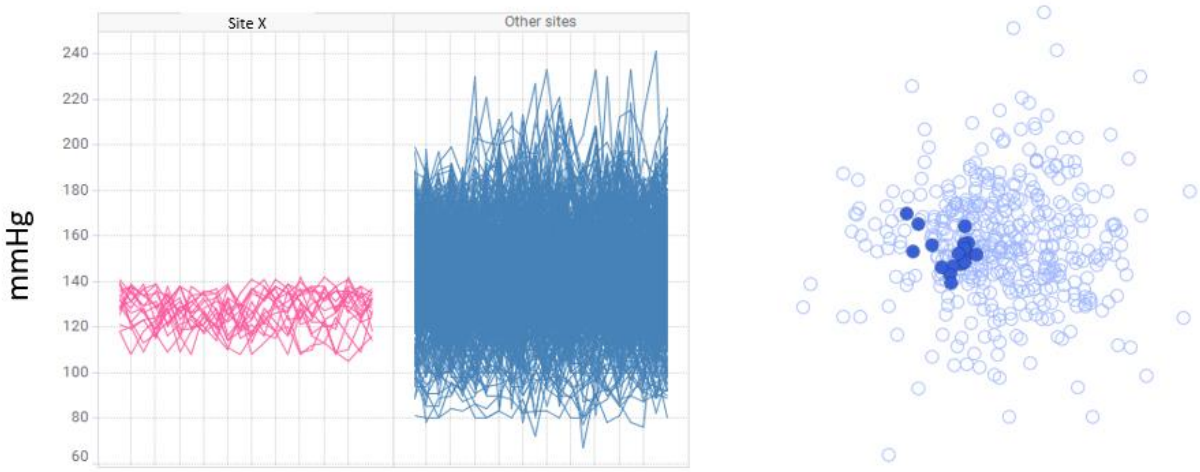

Figure 4. Site with co-clustered blood pressure profiles. All measurements are within a relatively narrow band of measurements (110-140 mmHg) which is also evident on the similarity plot on the right. When compared to the other profiles from the same study, it is extremely unlikely for the profiles to co-cluster so closely by chance. Potential reasons for this can be bias in subject selection, sample "splitting" (all measurements collected from one or few healthy persons) or outright result fabrication.

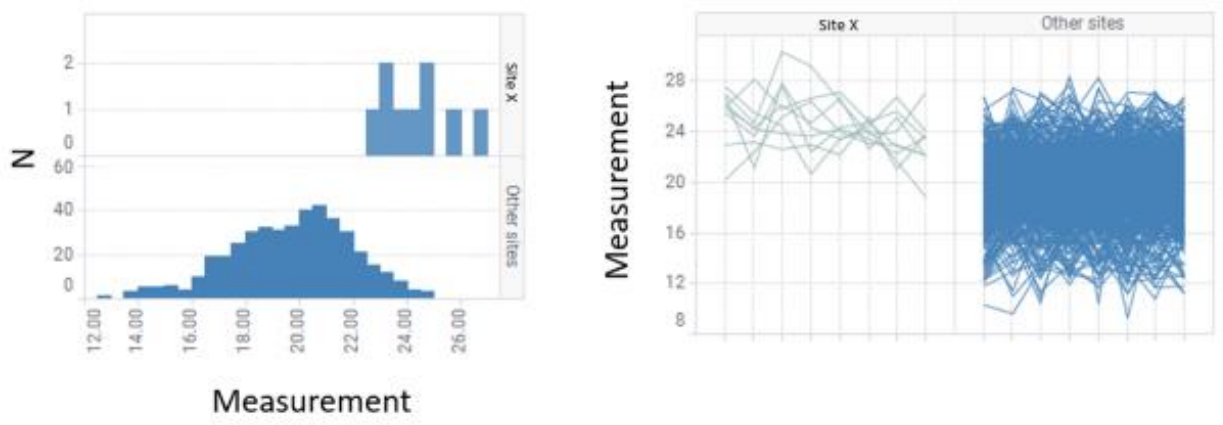

Figure 5. Example of a site with systematically higher results for a laboratory assay. Here the potential reasons can include bias in subject selection, an incorrectly calibrated assay (local lab) or errors in sample handling (central labs).

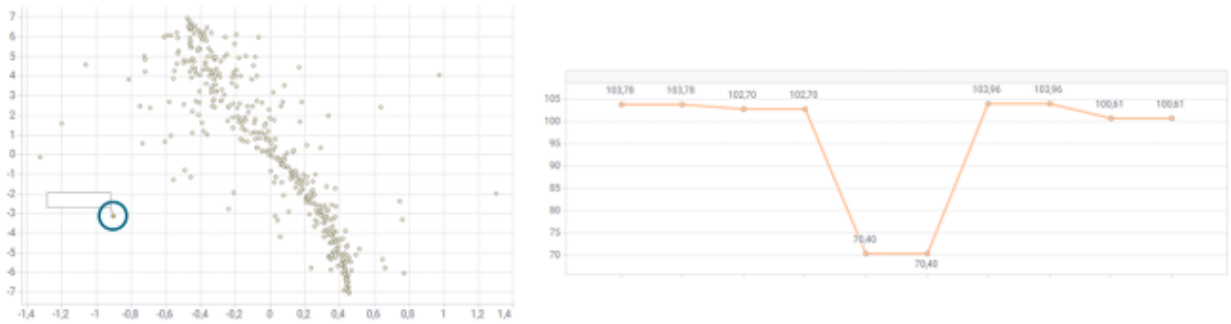

Figure 6. A subject with an "interesting" weight profile. The anomaly was detected visually on the similarity plot (left, highlighted dot). The subject weight is fairly static except for a dramatic drop at two time points (right). This is very likely due to a data entry error which is fortunately easy to fix.

Figure 7. A subject with an extreme standard deviation (red dot on the left). This does not necessarily represent a data quality issue but can be interesting due to a potential safety issue.

CONCLUSION

CTAS represents a significant advancement in central statistical monitoring. However, our vision for CTAS extends beyond its current capabilities and usage.

We envision CTAS as an industry-standard tool, recognized and utilized not just by the industry but also by health authorities worldwide, including the Food and Drug Administration (FDA). We believe that the involvement of health authorities in testing the package would not only validate its effectiveness but also contribute to its further development. Their expertise and regulatory perspective would be invaluable in refining CTAS and ensuring it meets the highest standards of data quality assurance.

In this regard, we invite health authorities to join the IMPALA Consortium. By becoming part of this collaborative network, they would have the opportunity to directly contribute to the development of CTAS. This would also foster a more collaborative environment between regulatory bodies and the consortium, promoting shared learning and mutual growth.

REFERENCES

[1] IMPALA Consortium, https://impala-consortium.org/
[2] Clinical Trial Anomaly Spotter (CTAS) Github Repository, https://github.com/IMPALA-Consortium/ctas
[3] simaerep Github Repository, https://github.com/openpharma/simaerep
[4] RQA Workshop on Data Analytics, https://impala-consortium.org/rqa-workshop-on-data-analytics-write-up-now-available/
[5] Stokman, P. G. et al. (2021) "Risk-based Quality Management in CDM An inquiry into the value of generalized query-based data cleaning", Journal of the Society for Clinical Data Management.
[6] Funning S. et al. (2009) "Quality assurance within the scope of Good Clinical Practice (GCP)—what is the cost of GCP-related activities? A survey within the Swedish Association of the Pharmaceutical Industry (LIF)'s members", The Quality Assurance Journal.
[7] de Viron, S. et al. (2023) "Does Central Monitoring Lead to Higher Quality? An Analysis of Key Risk Indicator Outcomes", Therapeutic Innovation & Regulatory Science.
[8] Guidance for Industry - Oversight of Clinical Investigations — A Risk-Based Approach to Monitoring. FDA Docket Number FDA-2011-D-0597.
[9] Tiikkainen P. (2022) "Finding Anomalies and Outliers in Clinical Trial Time Series Data", PHUSE EU Connect 22, https://phuse.s3.eu-central-1.amazonaws.com/Archive/2022/Connect/EU/Belfast/PAP_AR04.pdf

ACKNOWLEDGMENTS

The first public version of CTAS was the collaborative work of the author and two other IMPALA members Björn Koneswarakantha (Roche) and Frederik Collin (Boehringer Ingelheim).

In addition, the members of IMPALA's Clinical Anomaly Detection Work package team are credited for their input to the slides and this article: Roshan D'Souza (Roche), Björn Koneswarakantha (Roche), Tanja Velecic (Novartis), Michael Pelosi (Astellas), Demi You (Sanofi), Severine Fernez (Sanofi), Bérangère De Witte (Galapagos), Diletta Richi (Galapagos), Frederik Collin (Boehringer Ingelheim) and Sabrina Ward (Amgen).

CONTACT INFORMATION

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

Author Name: Pekka Tiikkainen
Company: Bayer Pharmaceuticals
Email: pekka.tiikkainen@bayer.com