Paper SM07

Beyond Outliers: Unsupervised Anomaly Detection in Clinical Trial Data

Helena Yaben, Novo Nordisk A/S, Bagsværd, Denmark
Kian Norouzi, Novo Nordisk A/S, Bagsværd, Denmark

ABSTRACT

Reviewing clinical trial data during conduct is crucial for ensuring patient safety, data quality, and regulatory compliance. However, manual inspection for identifying anomalies can be resource-intensive and limited, especially with the exponential growth in data volume seen the recent years.

To address this challenge, we have developed a Python pipeline which uses a combination of univariate and multivariate unsupervised machine learning algorithms to automatically identify anomalies in SDTM datasets. By employing feature selection, we limit the number of variables used in the multivariate anomaly detection, ensuring the model remains explainable. The multivariate approach detects not only outliers but also abnormal data points that are within the acceptable range but appear unusual when considering other variables.

Our package generates a user-friendly report that highlights anomalous subjects, visits, and findings, along with a detailed explanation of the detected pattern. Results are delivered in a tabular format for integration with visualization tools or custom analysis.

INTRODUCTION

Clinical trials are pivotal in healthcare progression, providing a structured environment to assess innovative treatments. The data collected during these trials, including vital signs, body measurements, and laboratory results, forms multi-dimensional time-series data. The quality and completeness of these data directly impact the reliability of the research outcomes. However, suboptimal data quality can lead to erroneous conclusions about a new treatment, endangering participants, and future patients, and causing significant resource wastage. Traditional methods for identifying poor-quality data, such as data visualisation and comparison with known normal ranges, often fail to detect anomalies that may appear normal in one or two dimensions but become outliers when considering additional variables or the temporal aspect of the data. The globalisation of clinical trials, the increasing complexity of protocols, and the pursuit of more ambitious research activities have led to a significant surge in the volume of data generated. The integration of new data sources, such as wearables and electronic patient diaries, has further amplified this trend. The current rapid growth in data volume and dimensionality surpasses the capacity of manual inspection to effectively identify anomalous patterns. Moreover, defining normal ranges for individual variables requires more comprehensive efforts to cover all collected variables and to customise them to the specific characteristics of each trial.

Recognising the limitations of current data cleaning techniques, there is a shift across the industry from traditional Clinical Data Management (CDM) towards Clinical Data Science (CDS). This transition encourages the use of automation and Artificial Intelligence (AI) to effectively identify anomalous patterns and facilitate reliable, data-driven decisions. However, a significant number of Clinical Data Managers (CDMs) lack the necessary skills to understand and apply advanced analytic techniques. Consequently, there is a rising demand for frameworks that produce explainable and user-friendly outputs, enabling CDMs to easily understand and utilise the information when managing data discrepancies. This approach ensures that the benefits of advanced data analytics are accessible to a wider range of stakeholders, thereby improving the effectiveness of data cleaning and thereby quality. In conclusion, managing the increasing volume and complexity of data necessitates more automated and advanced methods for data cleaning to ensure optimal data quality for statistical analyses. Furthermore, a human-centred

approach is crucial for generating interpretable outputs that serve a diverse range of stakeholders, facilitating data-driven decisions throughout the screening, diagnosis, and treatment phases.

At Novo Nordisk the existing data review process, which combines automated edit checks with visualisation dashboards, faces several challenges in the context of increasing numbers of trials and data points. Firstly, edit checks are primarily used for our Electronic Data Capture (EDC) system and do not cover other data sources such as lab results, diaries, and devices. Secondly, the current implementation of edit checks is one-dimensional, meaning that we can define acceptable ranges for each entry, but these values cannot be conditional and dependent on other entries. Thirdly, our visualisation tools are limited by the number of dimensions that can be included in the plots, which is typically two. Lastly, while we can apply univariate and bivariate outlier detection methods to flag outliers, this still requires the end user to inspect the individual data points and determine why they have been flagged as anomalous.

In this research project, we aim to develop a multivariate and time-sensitive model for automatically detecting anomalies in an explainable and understandable manner. This solution will serve as a recommendation platform for our CDM's, providing data-driven suggestions for querying individual data points. Balancing multivariate analysis and explainability are crucial. Univariate methods, while useful, are insufficient for assessing the quality of clinical trial data points. For instance, a weight measurement may appear normal when viewed in isolation, but when considered alongside waist circumference or longitudinal changes over several visits, it may be flagged as anomalous. Conversely, incorporating hundreds of features into an unexplainable machine learning method would also be inefficient. Our CDM's need to understand the explanation behind a given data point has been flagged for them to evaluate and describe the argument in a query text. Therefore, we need a robust feature selection method to limit the number of explanatory variables included in the ML model, improving its explainability. We are also using various univariate and bivariate methods to further enhance the model's interpretability.

The entire pipeline, from setting up a new trial to reading metadata and the Study Data Tabulation Model (SDTM) data, preparing the master table, applying different methods, detecting outliers and anomalies, and visualising the results, has been developed in an object-oriented and modularised way. This approach not only makes the entire library applicable across various clinical trials, but also allows some of the library's modules to be used for other descriptive analysis and reporting purposes involving SDTM data.

METHODS

This study introduces a structured approach for detecting anomalies in clinical trial data, which includes stages of data preparation, pre-processing, the application of anomaly detection techniques, and the subsequent communication of findings to relevant end-users such as CDM's. The primary technique utilised in this research is a reconstruction-based anomaly detection method, influenced by multidirectional time-series data imputation strategies for handling missing values. The emphasis is predominantly on the data cleaning process, specifically addressing anomalies that may occur from random errors during data collection or entry.

To ensure the proposed system's applicability across various clinical trials, the SDTM data model is employed. The research primarily focuses on anomalies found within the Vital Signs (VS) and Laboratory (LB) SDTM domains, which encompass results for an array of numerical continuous variables.

DATA PREPARATION

In line with a pre-defined Risk-Based Data Management (RBDM) approach, only variables within the Vital Signs (VS) and Laboratory (LB) domains that correspond to at least one primary, secondary, or exploratory endpoint of trials are considered. After identifying these pertinent variables, a master table is constructed, with each record corresponding to a single visit (VISITNUM) of one subject (USUBJID). All values collected during a specific visit are then added as columns to this master table. The following considerations are crucial during the master table preparation:

- The same TESTCD can be associated with multiple STRESU values, necessitating the STRESU column to identify results for a specific variable in particular units.
- There are instances where TESTCD and STRESN columns are insufficient. For example, plasma glucose can be measured under both fasting and non-fasting conditions. In such cases, the FAST column in LB can determine the test's condition.

- It is common to collect a variable multiple times for each subject during a single visit in certain cases. To obtain a unique value per subject, visit, and variable, it is essential to define an appropriate aggregation function. Depending on the clinical trial requirements and the type of assessment, aggregation functions such as the minimum, maximum, mean, or median of the visit values can be utilised.
- Specific visits may be excluded from the analysis, either because the relevant data is not typically collected during these visits (e.g., phone visit) or because the visit is inconsistent across subjects (e.g., unscheduled visits).

To streamline the process of identifying the visits and variables for consideration in the pipeline's subsequent steps, and the aggregation function for each variable, a Python script has been developed. This script automatically derives the information and generates an Excel workbook consisting of four sheets for any given trial:

1. Trial endpoints: Derived from the Trial Summary (TS) dataset of SDTM, this represents the primary, secondary, and exploratory endpoints for the trial.
2. Visits: Utilising the Trial Visits (TV) dataset of SDTM, this sheet contains the visit number, name, and planned study day for each visit. An additional column, INCLUDE, allows the user to manually specify whether data for a specific visit should be included in the analysis.
3. Vital signs: This sheet contains all the column sets in the VS dataset that provide information about the collected variable and the context of collection. It allows the user to define which combination should be used. The user can specify the aggregation function to be implemented using the added column AGG_FUNC.
4. Laboratory data: This sheet contains information similar to the Vital Signs but is based on the LB dataset.

As previously mentioned, the final master table used for further steps comprises USUBJID, VISITNUM, ID, and STRESN, which contain the subject, visit, variable ID, and the collected value in a standard format. In cases where multiple records exist for a variable in one visit of a subject, the chosen aggregation function is applied. The results from the VS and LB datasets are also concatenated in the final dataset.

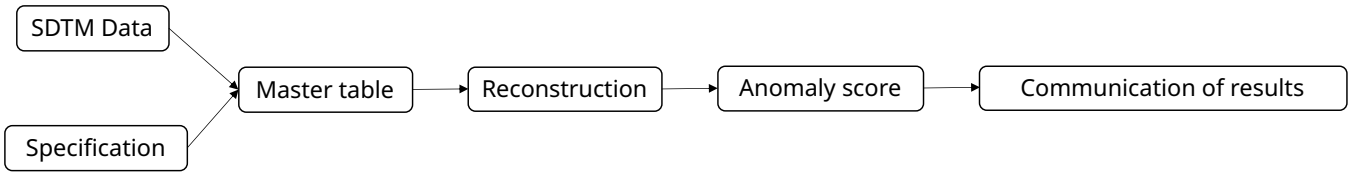

Figure 1: This diagram illustrates the anomaly detection process. The only manual intervention required in this pipeline is to complete the Excel specification file and define parameters such as the number of predictors, the number of iterations, and the anomaly.

DATA PRE-PROCESSING

Before implementing the anomaly detection method, below pre-processing steps are undertaken:

1. Log-transformation: Variables with skewed distributions are subjected to log-transformation to attain a greater degree of symmetry. This is due to the normality assumption inherent in some anomaly detection methods.

2. Min-Max Normalisation: Each variable is transformed to a [0,1] scale using a min-max normalisation formula. This normalisation technique ensures that all values operate on the same scale, thereby facilitating learning and enabling fair comparisons between columns with differing original scales.

ANOMALY DETECTION

CDM's are tasked with ensuring the completeness, validity, and consistency of clinical trial data through careful examination of individual data points. When a data point appears inconsistent with expectations, a query is initiated to the medical investigator requesting confirmation, clarification, or correction of the unexpected datapoint. Within the framework of anomaly detection, this process involves assigning an anomaly score to each value in the dataset and determining whether a query should be raised based on this score. However, providing a unique anomaly score per subject or subject visit does not offer the level of detail necessary for Data Cleaning. The requirement to assign a unique anomaly score to each data point in an unsupervised manner limits the range of applicable methods. While

distance-based, density-based, or ensemble methods can detect anomalies in multidimensional data, their anomaly score remains consistent across all elements within each input vector, rendering them unsuitable. Reconstruction-based methods, which compute anomaly scores based on the reconstruction error for each input vector, emerge as the most suitable choice for this use case, as they can calculate individual anomaly scores for each vector element.

Like many machine learning algorithms, traditional reconstruction models such as Autoencoders (AE) require input data to be devoid of missing values. However, clinical trial datasets often exhibit a degree of missingness across variables, subjects, and visits. Basic univariate imputation techniques, such as replacing missing values with the mean or median of each variable, could lead to suboptimal model performance, particularly when a significant portion of the data is missing. Therefore, it is essential to employ more sophisticated methods that can leverage the inherent multivariate and time-series relationships in the data for more accurate imputation of missing values. Several algorithms have been proposed for imputing missing entries in multidimensional clinical datasets, which can also account for the temporal nature of the data. These methods learn the patterns in the data, compute a prediction for each element of the input matrix, and apply a mask to preserve the predictions for initially missing entries. The learning process typically involves minimising the prediction error for the originally observed entries. An anomaly score can then be computed for each matrix element, derived from the reconstruction error.

In this paper, we adopt a method inspired by two distinct missing data imputation techniques - Multivariate Imputation by Chained Equations (MICE) [1] and Multidirectional Time Series Imputation (MD-MTSI) [2]. This method not only imputes missing values in our dataset but also predicts values for existing data points. The prediction then facilitates anomaly detection through the calculation of the reconstruction error.

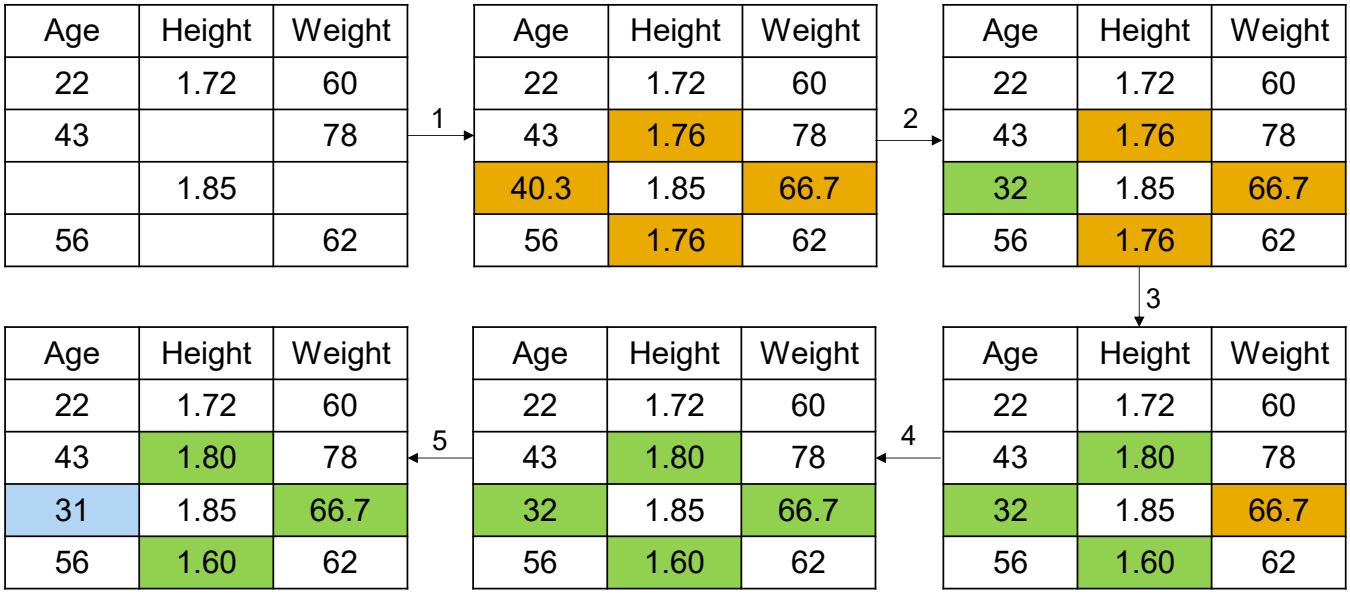

Figure 2: Initial Steps of the MICE Algorithm. This illustration depicts the initial stages of the Multiple Imputation by Chained Equations (MICE) algorithm. 1) Missing values are initially imputed with the mean of each variable. 2) The variable 'Age' is modelled as a function of 'Height' and 'Weight', and initial imputations for missing entries are replaced with predictions. 3) The variable 'Weight' is modelled as a function of 'Age' and 'Height', and initial imputations for missing entries are replaced with predictions. 4) The variable 'Height' is modelled as a function of 'Age' and 'Weight', and initial imputations for missing entries are replaced with predictions, concluding the first iteration. 5) The second iteration begins, where the variable 'Age' is modelled as a function of 'Height' and 'Weight', and imputations from the previous iteration are refined with predictions from the new model.

Despite their differing approaches, both MICE and MD-MTSI share the commonality of imputing data for a specific variable using information from other variables in the dataset. In MICE, an iterative process involving a series of prediction models is used to impute missing values in a given dataset. Initially, missing entries are replaced using a simple approach such as the mean or median of each variable. The algorithm then models the first variable as a function of the remaining variables in the dataset, using the initially imputed data for training. The trained model generates predictions for the originally missing entries, replacing the initially imputed values. This process is sequentially repeated for each variable in the dataset until all variables are imputed. In the second iteration, the imputed data from the previous iteration is used to train a new model for each variable in the same sequential manner. This iterative process continues until a predetermined number of iterations are completed or convergence

is achieved.

MD-MTSI, on the other hand, first expands the original dataset with additional columns to enable multidirectional imputation. For each variable, nine new columns, including the values recorded at the three preceding and three subsequent timestamps, are added. Four additional columns representing spatiotemporal information are also included.

In this approach, each variable is individually modelled as a function of the remaining columns using a tree-based ensemble of regressors, specifically the Light Gradient-Boosting Machine (LGBM). LGBM is a boosting algorithm renowned for its ability to effectively handle large datasets. It includes built-in regularisation and bagging options, making it less susceptible to overfitting compared to simpler tree-based regressors. Furthermore, due to the ability of tree-based models to handle missing values, there's no need for initial imputation. Similar to MICE, predictions from the trained models are used to fill in missing entries, while entries originally observed remain unchanged. However, unlike MICE, the MD-MTSI approach does not subject imputations to iterative updates; instead, a single imputation is performed. The method's accuracy is enhanced by including both preceding and succeeding values for each variable, along with basic statistics and other spatiotemporal features, as predictors, making it more precise than simpler methods.

MD-MTSI has demonstrated superior performance when compared to 3D-MICE, a variant of MICE specifically designed for multivariate time-series data. Furthermore, 3D-MICE has shown improved results when compared to the traditional MICE method.

Our reconstruction-based anomaly detection method is implemented by amalgamating the iterative aspect of MICE with the multidirectional component of MD-MTSI. However, several modifications to the MD-MTSI method are made to address the following situations:

- Considering that ongoing trial data is anticipated to contain discrepancies, utilising features such as the minimum, maximum, and mean of values as predictors could potentially lead to less-than-optimal predictions when at least one discrepant value is present. These discrepancies could be manifested as extremely high or low values for a subject or a variable, which would influence the minimum, maximum, and other statistical measures. Consequently, this could lead to inflated anomaly scores for values that should ideally have lower scores, thereby heightening the risk of false positives.
- The more variables used as predictors, the higher the likelihood of incorporating a discrepant value in the prediction. Therefore, integrating feature selection into the algorithm is crucial to aim for a minimal subset of predictors that yield optimal reconstruction for a given variable. Fewer predictors also enhance the interpretability of the output, which is important when results need to be communicated to end-users in a user-friendly manner.
- Since the data cleaning process takes place on an ongoing basis as data is being captured, values recorded at subsequent timestamps may not be as relevant as those recorded at preceding timestamps. Moreover, a given variable might not be collected at all three preceding or subsequent visits for a timestamp, which can reduce the usability of these features as predictors. However, if the columns including the preceding and subsequent timestamps for each variable were also iteratively imputed as part of the process, this could help achieve more accurate predictions.

Considering these factors, a combined and adapted version of MICE and MD-MTSI is proposed. In this method, the original dataset is first expanded with additional columns that encapsulate temporal and demographic information. For each row in the master dataset, which includes the data collected for a subject at a visit, the following information is added:

1. Demographic variables: Two additional columns are introduced, containing the age and gender of the subject respectively.
2. Temporal variables: For each variable, two new columns are introduced, comprising the values recorded at the preceding visit and the subsequent visit.
3. Spatiotemporal variables: Three additional columns are included, including the study day value associated with the visit, as well as the study day interval between the current visit, previous and the next ones.

In the initial round of our reconstruction model, each column's missing values are filled in with the median value of that column. The subsequent sequential imputation has certain unique characteristics. For the group of ordered outcome variables, the imputation procedure is as follows:

1. Initially, the correlation matrix of the expanded dataset is computed.
2. For each outcome variable:
    a. Using the initially imputed data, a regression model for the specific outcome variable is developed. The predictors in this model are the variables that have the highest correlation with the outcome variable. The number of predictors, denoted as 'c', needs to be initialised.
    b. The developed model is used to generate predictions for the entries that were originally missing, thereby updating the imputations. The entries that were originally observed remain unchanged.
    c. Subsequently, the variable from the previous visit is recalculated using the updated values of the current variable. Since values from the previous visit are invariably missing for the first visit of each subject where the variable is collected, the missing entries are filled in by modelling this variable using the variables with the highest correlation in the dataset. This procedure aids in ensuring a more accurate prediction when the predictor from the previous visit is missing.
    d. The variable from the next visit is also recalculated using the updated values of the current variable. As values from the next visit are always missing for the last visit of each subject where the variable is collected, missing entries are filled in by modelling this variable using the variables with the highest correlation in the dataset. This aids in ensuring a more accurate prediction when the predictor from the next visit is missing.
3. Steps 1 and 2 are repeated for a predetermined number of iterations, or until a point of convergence is reached.
4. The models developed in the final iteration are utilised to make predictions for the observed entries of each outcome variable, and a reconstructed dataset, which has the same shape and missing entries as the original dataset, is returned.

Given the continuous nature of the VS and LB variables, regression models are appropriate. However, it's worth noting that using a model specifically designed for each variable's nature, such as a classifier for categorical variables and a regressor for continuous variables, is a feasible option. This research proposes two key regressors, Linear Regression (LR) and Light Gradient Boosting Machine (LGBM), to assess the performance of our method. These models are selected due to their computational efficiency, which is vital when running models for each variable at every iteration, particularly with large datasets. In addition to their computational efficiency, these models are highly interpretable, providing a clear understanding of each predictor's significance in the model.

In LGBM, two primary metrics, gain and split, are employed to assess feature importance. The gain metric signifies the decrease in training loss when data is partitioned based on a particular feature. The significance of a feature can be quantified by aggregating the total gain realised when that feature is utilised for data partitions. Conversely, the split metric quantifies the frequency of a feature's usage in data partitions across all trees in the ensemble. In both instances, a higher metric denotes a more significant feature. Features that attain high scores in both metrics are deemed highly crucial for the model. For regression tasks, the gain metric is an appropriate method for ranking predictor importance in each model. The LGBMRegressor method in the Python package LightGBM flawlessly incorporates the calculation of feature importance.

In LR, the significance of features can be evaluated by scrutinising the estimated coefficients for each feature. A prevalent method for assessing the significance of these coefficients is to use a t-test to test the null hypothesis that a specific coefficient is equal to zero. The t-test computes a t-statistic by dividing the estimated coefficient by its standard error. If the computed t-statistic significantly deviates from zero, it implies that the associated feature likely has a non-zero influence on the dependent variable, rendering it statistically significant in the model. If the t-statistic does not significantly deviate from zero, there is insufficient evidence to reject the null hypothesis, suggesting that the corresponding feature may not be a significant predictor in the model. The OLS method in the Python package Statsmodels incorporates the computation of t-statistics and p-values associated with each model coefficient.

The ability of our method to account for the data's multivariate and longitudinal relationships, coupled with its efficiency, flexibility, and interpretability, renders this method a promising approach for reconstruction-based anomaly detection.

ANOMALY SCORE

Reconstruction-based anomaly detection assumes that models will struggle to accurately reconstruct infrequently encountered data during training. As a result, anomalous data points, which are found in a small fraction of the data, are expected to have significantly higher reconstruction errors. These errors are anticipated to be at the extremes of the error distribution, providing a statistical basis for calculating anomaly scores.

If the Probability Density Function (PDF) of the reconstruction errors is known, an anomaly score can be computed. This score is associated with high reconstruction errors located at the tails of the distribution. To enhance interpretability, the distribution of absolute reconstruction errors can be used instead, with values on the right tail representing poorly reconstructed data and those on the left tail representing well-reconstructed data. An anomaly score within the range [0, 1] can be determined by evaluating the Cumulative Density Function (CDF) at a specific absolute reconstruction error.

Given the potential variability in reconstruction performance across outcome variables due to differing levels of discrepant values and varying extents of missing data, it is crucial to examine the distribution of reconstruction errors separately for each variable. For a specific variable, anomaly scores are computed considering the absolute reconstruction errors across subjects and visits for that variable.

The PDF for a given random variable can be estimated non-parametrically using a Kernel Density Estimator (KDE). The estimated density function can then be integrated to evaluate the CDF at a specific value. In this study, a Gaussian KDE is used to make PDF estimates, providing smooth and robust estimations which are less affected by outliers than other kernels. The method 'gaussian_kde' in the Python package SciPy is utilised for this purpose.

ANOMALY LABEL

After calculating the anomaly scores, the next step is to create a set of binary anomaly labels. This is done by applying a certain threshold to the anomaly scores. If a score surpasses this threshold, the corresponding label is marked as '1', indicating an anomaly. If it doesn't, the label is marked as '0', indicating normal data.

In situations where we have labels for our data, we can determine the best threshold by optimising the performance of the classifier. This could be done by maximising metrics like the Area Under the Receiver Operating Characteristic (ROC) Curve or the F1-Score. However, in situations where we don't have labels for our data, this approach isn't possible.

Given that a specific anomaly score represents the likelihood of encountering a reconstruction error of a certain size or smaller and considering that reconstruction errors for anomalous data are expected to be at the higher end of the distribution, we might choose a threshold within the range of 0.95 to 1. The exact choice of threshold depends on the context. In the context of data cleaning in clinical trials, where false positives can lead to significant waste of resources, a threshold closer to 1 might be preferable. The optimal threshold selection depends on the balance we want to achieve between sensitivity (identifying true positives) and specificity (avoiding false positives) in the given context.

Given that every clinical trial has unique needs in terms of the importance of anomaly detection for certain data and the available resources, it's proposed that the threshold should be a parameter that the user can specify. This allows for customisation to meet the specific needs and characteristics of individual trials.

RESULTS

Explainability is crucial in the context of anomaly detection in clinical trial data. Firstly, CDMs need to understand the reasoning behind flagging a specific data point as anomalous. Secondly, CDMs must be able to utilise the output of anomaly detection when raising queries. The output should provide enough information for CDMs to easily explain to investigators why the entered data appears suspicious and to identify potential sources of errors. The proposed strategy for reconstruction-based anomaly detection provides four key pieces of information for each data point:

- Reconstructed Value: This offers insights into the expected value for a data point within a given context, as well as the magnitude and direction of the deviation from the observed value.
- Anomaly Score: This reflects the likelihood of encountering a lower or equal absolute reconstruction error. It provides a measure of how unusual a particular value is within a given context.
- Anomaly Label: This indicates whether the anomaly score is lower or higher than a threshold, which determines the classification of the data point as anomalous or non-anomalous.
- Feature Importance: Here a list of the most relevant predictors can be returned.

While univariate and bivariate methods may lack robustness in detecting anomalies within multidimensional time-series data, they can serve as valuable tools to supplement the results of reconstruction-based anomaly detection, offering additional insights to CDMs.

A specific data point can be classified according to its position in relation to other data points for the same variable, which can range from extremely low to extremely high. This categorisation can be accomplished non-parametrically using the Inter-Quartile Range (IQR) thresholds suggested by Tukey in 1977. The IQR is defined as the range between the lower quartile (25th percentile) and the upper quartile (75th percentile) of the data set. Data points that fall between an inner fence and its adjacent outer fence are labelled as "outside", while those beyond the outer fences are termed as "far out". The IQR regions proposed by Tukey can be converted into categories that are ideal for interpretation. To enhance the model's explainability, we apply the IQR to each value (Univariate label), as well as to the absolute change of each value compared to its preceding value (Temporal univariate label). These two outlier categories are then included in the report.

The Mahalanobis [3] distance, a simple method often used in clinical trials, helps assess relationships between correlated columns in a dataset. It measures how far an observation is from the centre of a multivariate distribution, considering the correlation between variables and their individual differences. In our report, we apply this method to the target variable and its most correlated variable. The resulting p-value and its corresponding label, indicating whether the data point is normal or anomalous, are then included. Figure 3 serves as an illustration of how the findings can be conveyed to the end users:

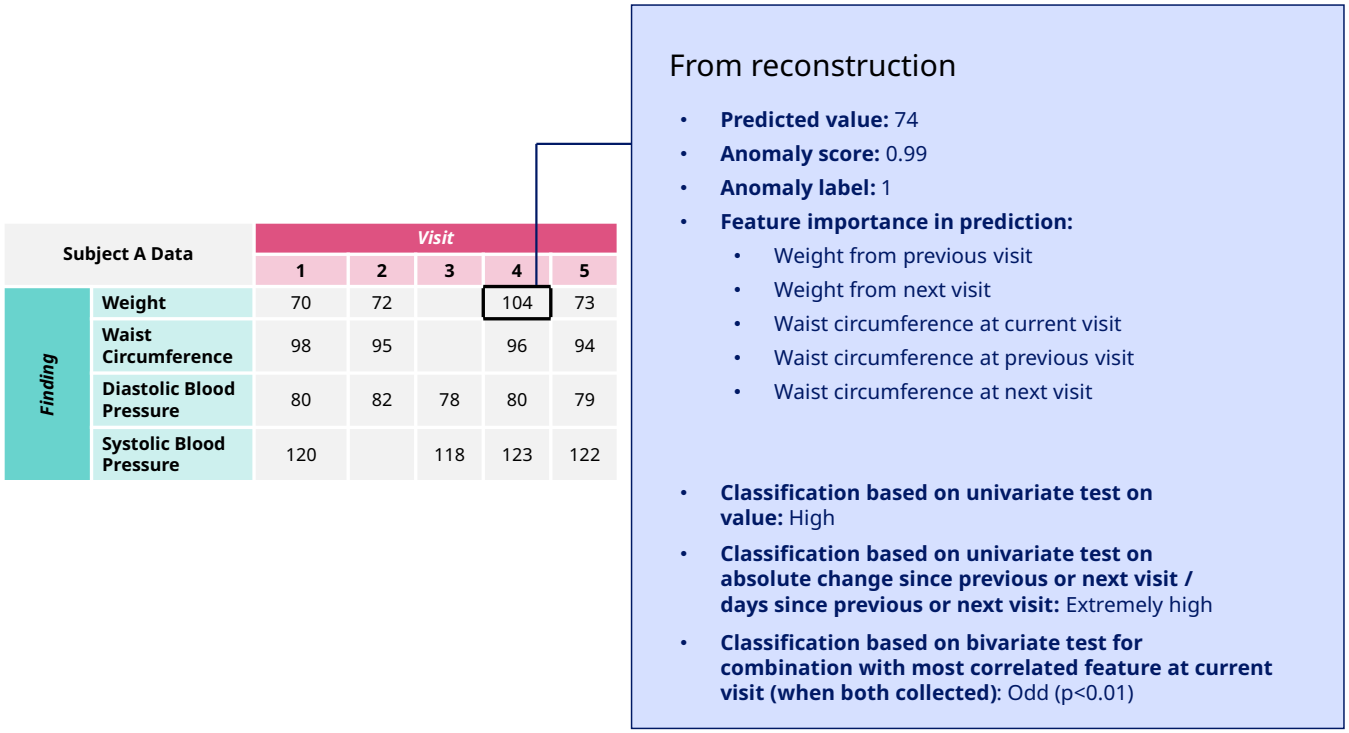

Figure 3 provides an illustration of how the findings can be relayed to end users. The table displays all the data collected for one participant. In this instance, the weight recorded for visit 4 (104 KG) has been flagged as an anomalous data point. In the description section, we initially see the predicted value from our model (74KG), followed by the anomaly score and anomaly label. Subsequently, we identify the most significant features involved in detecting this anomalous data point. In this case, the weight from the preceding and succeeding visits contributes most significantly to the model. Based on the univariate and bivariate methods, we observe that the weight of 104KG has been classified as high in comparison to the other weights recorded in this trial. Furthermore, the absolute change in weight compared to the previous visit has been classified as extremely high relative to other weight changes from previous visits in the trial. The final part pertains to the results of the bivariate test conducted between weight and its most correlated variable (in this case, waist circumference). Here, the p-value is below 0.01, indicating a significant correlation.

CONCLUSION

In this paper, we have proposed a method for utilising AI to identify anomalies in clinical trial data. This approach employs an unsupervised, reconstruction-based anomaly detection method, maintaining its explainability due to the limited number of features selected as predictors and the application of further univariate and bivariate tests.

The entire pipeline has been developed in an object-oriented manner in Python, taking the SDTM data as input. This design facilitates the application of the method across different trials. We have also provided an example of how the results can be communicated to end users, incorporating both visualisation of the data and a textual description. This dual approach aids CDMs in easily evaluating the flagged anomalies.

Our method demonstrates how AI can be leveraged to flag anomalies amidst the exponential growth of data volume. We have shown that methods originally designed for missing data imputation can be repurposed for reconstruction-based anomaly detection in clinical trials, which often present varying degrees of missingness across variables, subjects, and visits.

A crucial aspect of our approach is the balance it strikes between complexity and explainability. Our method, achieves this balance, standing in contrast to other reconstruction methods such as autoencoders, which are often seen as 'black box' solutions.

We are currently in the process of further evaluating the effectiveness of this framework as a tool for automating parts of the query process. This ongoing work aims to provide more comprehensive statistical evidence of the model's performance in the future.

In conclusion, our proposed method offers a route towards building a robust, explainable, and adaptable solution for anomaly detection in clinical trial data.

REFERENCES

[1] Stef Buuren and Catharina Groothuis-Oudshoorn. “MICE: Multivariate Imputation by Chained quations in R.” In: Journal of Statistical Software 45 (Dec. 011). doi: 10.18637/jss.v045.i03.

[2] Xiaofeng Xu et al. “A Multi-directional Approach for Missing Value Estimation in Multivariate Time eries Clinical Data.” In: Journal of Healthcare Informatics Research 4.4 (2020), pp. 365–382. doi: 10.1007/s41666-020-00076-2.

[3] Hamid Ghorbani. “MAHALANOBIS DISTANCE AND ITS APPLICATION FOR DETECTING MULTIVARIATE UTLIERS.” In: Facta Universitatis Series Mathematics and Informatics 34 (Oct. 2019), p. 583. doi: 10.22190/FUMI1903583G.

ACKNOWLEDGMENTS

We would like to acknowledge that this research project was initiated within Novo Nordisk and further developed in academic collaboration with the Technical University of Denmark (DTU). Our sincere gratitude goes to Dr. Morten Mørup for his invaluable insights and mentorship, which greatly enriched the project and the master thesis of co-author Helena Yaben. His profound expertise significantly shaped the direction and outcomes of this research.

CONTACT INFORMATION

Contact the author at:
Author Name: Kian Norouzi
Company: Novo Nordisk
Email: vxmn@novonordisk.com