Paper AR04

Finding anomalies and outliers in clinical trial time series data

Pekka Tiikkainen, Bayer Pharma, Espoo, Finland

ABSTRACT

Correcting errors in clinical trial data as the trial is on-going is vital for an accurate safety and efficacy read-out. In this talk, we will show how time series can be used to identify trial sites and individual subjects with potential errors in their clinical parameters (e.g., vital signs or laboratory measurements). Each subject-level time series is summarized as scalars such as mean and entropy. These features alone are useful in identifying individual subjects with suspicious data. 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). Results of the algorithm are available to anyone in our development organization who are involved in ensuring the quality of our trial data. Actions resulting from the results can range from sending queries to planning complete study audits.

INTRODUCTION

Ensuring the high quality of data is one of the main tasks of a clinical trial sponsor. This work is done when the study is on-going, and it is an absolute requirement for a solid safety and efficacy readout at the end of the study. If issues are identified early enough, the study site could be further trained to avoid problems later in the study. In the extreme case of fraud, a site can be completely excluded from the study to avoid bias in study results.

There are various ways how sponsors try to ensure that the trial sites record results as defined in the trial protocol. Source data verification aims to catch transcription errors by comparing results in original documentation with what has been entered into the electronic data capture system. At central monitoring, protocol compliance is controlled to ensure that, for example, subject visits take place at correct intervals, or that inclusion and exclusion criteria have been met.

Central statistical monitoring screens subject measurements (e.g., laboratory results or vital signs) for any anomalies such as individual measurement outliers. In this paper, we present an internally developed tool for the identification of anomalous time series collected during the study. Anomalous time series are monitored both on the level of an individual subject and a trial site. The rest of the paper describes how the time series are processed and analyzed. Examples of typical anomalies are also given.

PROCESSING TIMESERIES

The source of our clinical data is our harmonized clinical trial database Raven. It contains both clinical subject-level data and operational data for on-going Bayer studies. The following is an overview of how we define the time series of interest and how we process them.

PARAMETERS

Parameters from the following data domains are in scope: LB (laboratory measurements), EG (ECG parameters), VS (vital signs) and QS (patient questionnaires). We consider all parameters which fulfill the criteria for time series (see below) and make no distinction between critical and non-critical parameters. An alternative approach would have been to only include parameters deemed critical for safety and study endpoints. We decided not to do this because bias in non-critical parameters can be indicative of general misconduct at a site which might result in also the critical data being compromised.

Further, each parameter is considered twice: once using the actual measurements and once with baseline-adjusted values. The latter can be interesting for identifying trend outliers.

DEFINING TIME SERIES

For each parameter, one or more time series can be defined. A time series must have at least three time points and there must be at least 30 subjects with data for all time points. Subjects with missing data in more than a third of the time points are excluded.

In an on-going study, some subjects might have only taken the first visits while some have already finished the study. This is the reason why we might define more than one time series per parameter. For example, one time series would include almost all planned time points and would be useful for comparing sites and subjects which have largely finished the trial. Another time series might focus on the first few visits to also include subjects which have only lately been enrolled. Figure 1 gives a simplified example of a set of two time series defined for a parameter.

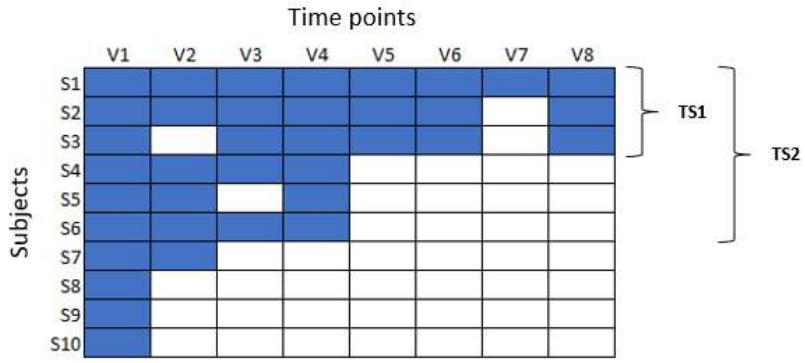

Figure 1. Simplified example of time series definition. Blue cells denote subjects with a measurement at a given time point. Two time series have been defined. TS1 includes subjects (S1-S3) which have finished the study while the second time series (TS2) focuses on the first four time points and includes also subjects S4-S6 in addition to those in the TS1. Remaining subjects have data for too few time points to be considered in either of the time series.

TIME SERIES FEATURES

A set of features is calculated for each time series. The features Mean, Standard deviation and Unique value count are very simple and require no further explanation here. Autocorrelation is calculated with a lag of one timepoint and values close to minus one pinpoint to a zig-zagging behavior in the time series. Figure 2 illustrates the features for a sample time series.

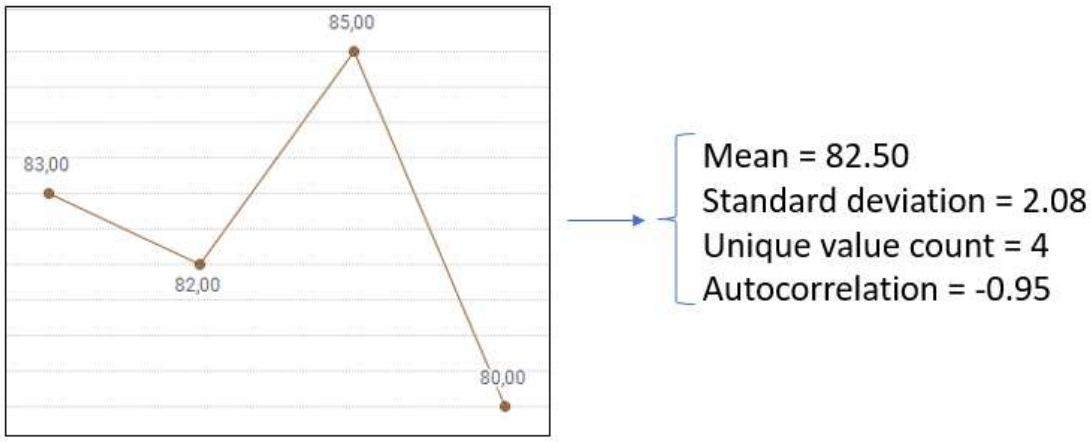

Figure 2. Example of a time series with four time points and the features calculated from it.

There is one feature which needs more explanation: site co-clustering - a measure of how similar the time series is to other subjects from the same site vs. subjects from all other sites. It is used to identify sites whose time series for a particular parameter are more similar to each other than could be expected by chance. In the extreme case, this could be indicative of sample splitting, i.e., a type of fraud where samples are collected from only one individual but assigned to several subjects. Figure 3 illustrates how co-clustering is calculated for a single time series.

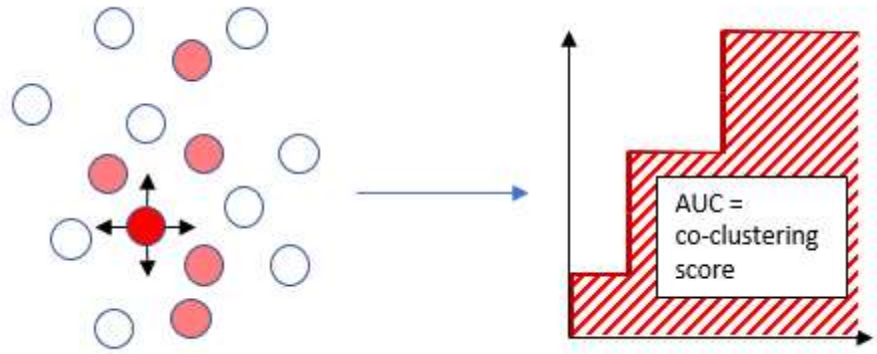

Figure 3. Calculating a site co-clustering feature for a time series. Spheres on the left represent individual time series (red = time series for which we are calculating the feature (query), pink = time series of other subjects from the same site, white = time series of subjects from other sites). First the time series are ranked based on their distance to the query. A ROC curve is calculated for the ranked time series and a Area Under Curve (AUC) is calculated for the curve. The AUC is then used as the co-clustering feature. A feature of value of 1 means that fellow subjects from the same site are all closer to the query than subjects from other sites. A value of zero is the opposite of this and the value 0.5 means that there is no difference between subjects from the query's own site and the subjects from other sites.

FLAGGING SITES

Flagging sites with a systematic bias in their time series is an important part of study monitoring. If identified early enough, the site can be offered further training if the site has had trouble in interpreting the study protocol. In the extreme case, if the bias is due to intentional misconduct, the site can be closed and excluded from analysis.

METHOD

For each time series and feature, the site's distribution of feature values is compared to the distribution of feature values for subjects enrolled in other sites in the study. Kolmogorov-Smirnov test is used for the comparison. The greater the difference in the two distributions is, the smaller the p-value given by the test is. The p-value is corrected for multiple testing and converted into its negative logarithm which is the site's "biasness" score for the parameter and feature. Figure 4 gives an example on how the method is used to identify a site with relatively few unique values per time series.

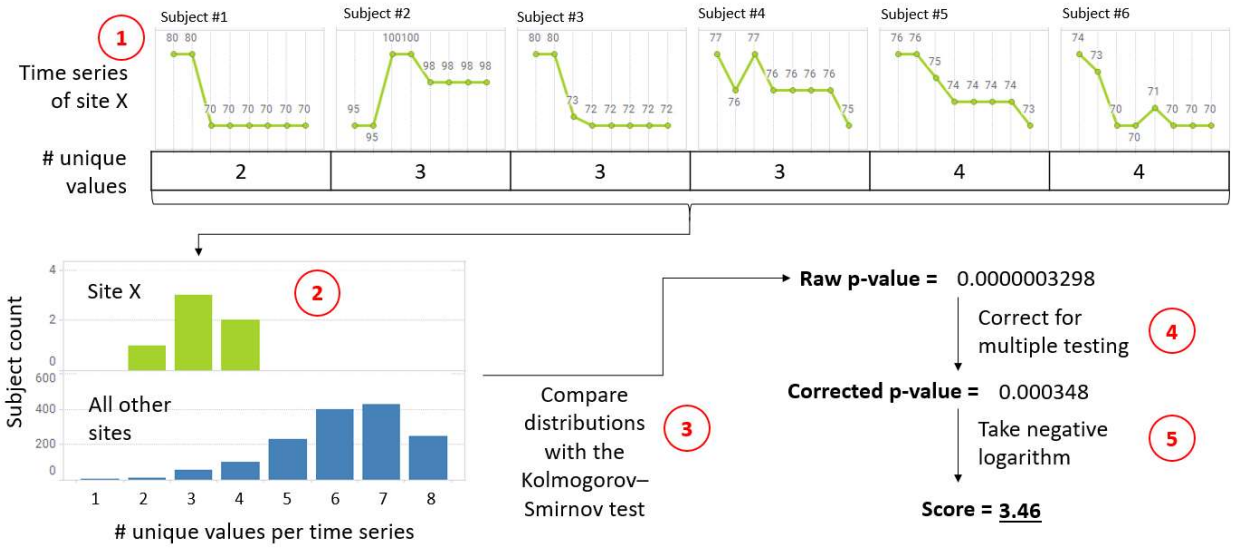

Figure 4. Example of site flagging. The timeseries has eight time points and the question is whether the site (six subjects) has reported fewer unique values per time series than other sites in the study. Part 1) has the individual subject time series and the unique value counts. It is clear from the 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).

EXAMPLES OF SITES WITH BIAS

Figures 5 and 6 give examples of actual cases of bias from a Bayer-sponsored study identified with our method. Given the output of the method alone, it is not possible to say what exactly is the reason behind the results – this would need further investigation such as site inspections.

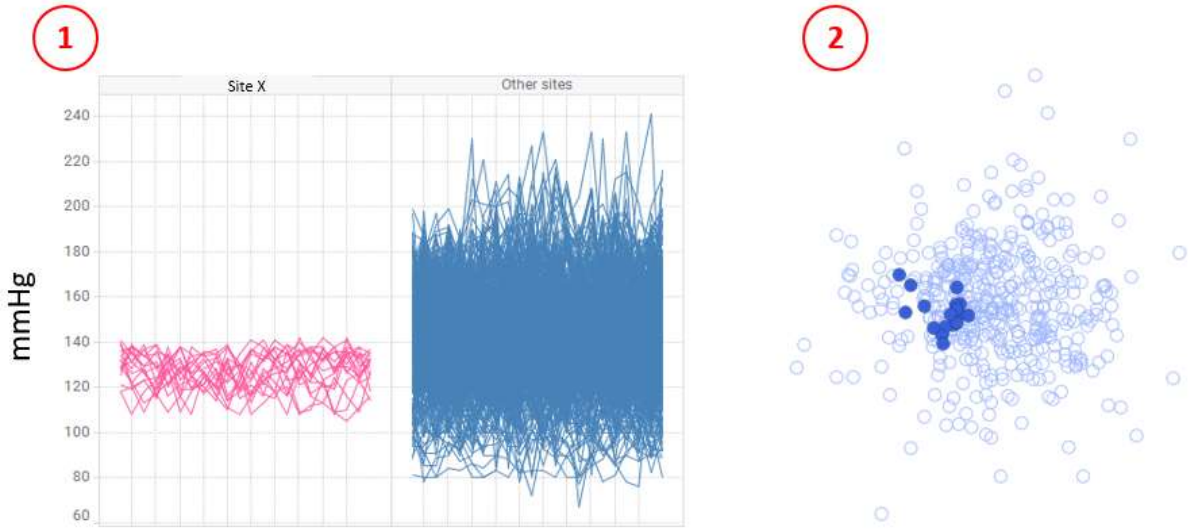

Figure 5. Example of co-clustering of systolic blood pressure profiles. Site X was flagged by the tool for a significant bias in the site co-clustering feature. This is evident in the time series (1) as almost all measurements are within the narrow range of 120 to 140 mmHg whereas measurements from other sites vary over a wider range (blue lines). This is also evident in the similarity plot (2) which visualizes the relative distance between time series. On this plot, almost all Site X subjects (filled circles) are clearly co-clustered.

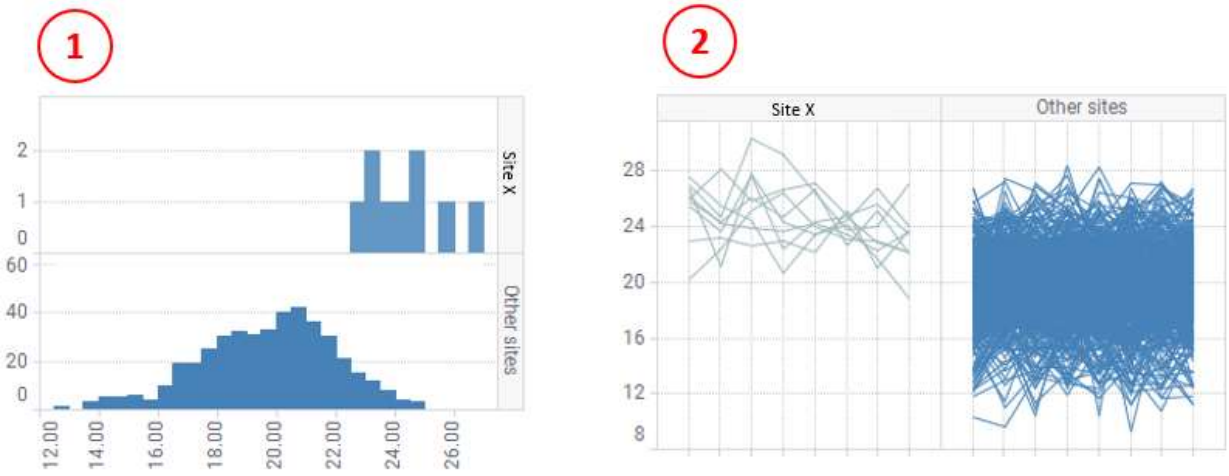

Figure 6. Example of a site with systematically large values for a laboratory assay. The difference in the average results is clear in both the histogram (1) and the individual time series (2).

IDENTIFYING ANOMALIES IN INDIVIDUAL TIME SERIES

In addition to flagging sites, the results are valuable for identifying individual subjects with anomalous time series. Sites which exhibit no systematic bias might still contain individual anomalous subjects and time series.

One way is to visually inspect similarity plots for subjects with few near neighbors. Figure 7 gives an example of this for an anomalous weight profile. Another approach is to compare time series features and review time series with extreme values for one or more of these. Please see Figure 8 for an example on this.

Please note that outliers identified with the two approaches often correlate, e.g., subjects with unusually high standard deviations also tend to be outliers on the similarity plot.

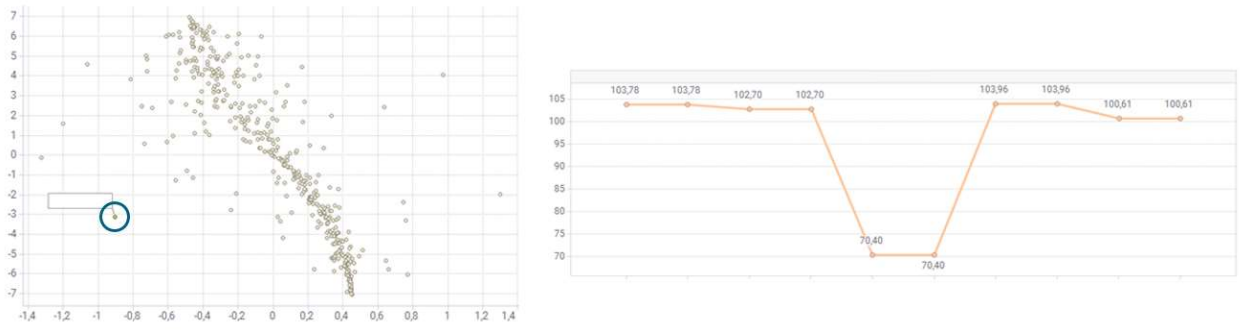

Figure 7. An anomalous weight profile identified based on its distance from other profiles on the similarity plot (left). The profile is given on the right with a sudden drop in weight followed by a return to the previous values. In this case, the reason is probably a data entry error at site.

Figure 8. Identifying an individual time series outlier based on a time series feature. In this case the subject with most variable bilirubin profile (1) has been selected and highlighted with the other subjects (2). In this case, it is possible that the site has collected the data correctly but this might be interesting for someone performing medical review to identify safety issues, for example.

IMPLEMENTATION

The tool has been implemented in R and it is run daily on a cloud service and results are written back into our in-house database. A custom dashboard has been designed with TIBCO Spotfire® for visualizing and inspecting the results. All the examples above are screenshots from the dashboard.

CONCLUSION

In this paper, we have presented a custom tool which regularly screens the Bayer internal clinical database for subjects and sites with time series outliers. The results are updated daily and made available company-wide via an interactive dashboard.

The current version is the first one and we have already identified some new features for future releases. For instance, users have wished for automated E-mail alerts whenever an anomaly is detected in the studies they work with.

IDEAS FOR FURTHER DEVELOPMENT

There can be several reasons for a bias at a site. The tool is meant for detecting bias caused by actions taken by the site (e.g., non-compliance, incorrectly calibrated measurement devices, and fraud).

The background characteristics of the general population in the site's home country can also lead the site being flagged. For example, the population can have a generally lower blood pressure than the global average. In the

current version, we always compare the site to all the other sites globally. This would lead to sites from a such country being flagged for biased blood pressure time series even though they would have done nothing wrong.

In a future version, we are planning to see if comparing a site to other sites from the same part of the world would alleviate this problem. For example, we would compare Finnish sites to other (Northern) European sites.

ACKNOWLEDGMENTS

Development of the tool would not have been possible without the invaluable feedback from the following Bayer colleagues: Cornelia Fischer, Siavash Forootan, Raju Kacchu, Jana Polley, Claudia Prange, Tomomi Terada and Holger Schimanski.

RECOMMENDED READING

For more discussion on fraud in clinical trials and concrete examples of confirmed cases, please see “George and Buyse, Clin Investig (Lond). 2015; 5(2): 161–173.” (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4340084/).

CONTACT INFORMATION

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

Pekka Tiikkainen
Bayer Oy
P.O. Box 73
02151 Espoo
Finland
Email: pekka.tiikkainen@bayer.com