

PP09: Unlocking the Potential of Historical Clinical Trials: A Hybrid Statistical and Machine Learning Approach for Laboratory Data Quality Assessment

Soon Jye Kho & Bhargav Koduru, Medidata Solutions
soonjye.kho@3ds.com bhargav.koduru@3ds.com



Abstract

Laboratory data is a key indicator of patient safety and sometimes efficacy, yet data inconsistencies frequently arise in multi-site settings, leading to several data quality issues. Outlier detection has been used to detect abnormal values; however, abnormalities vary and heavily depend on context, e.g., some medications lower the glucose level; thus, a low glucose level on a patient treated with the medication is expected and not an outlier. This paper presents an innovative approach, further enhanced with 39 standardized historical clinical trials (eligible for analysis through Medidata's data collaboration program) to develop a statistical and machine learning-based tool for visual data quality exploration and root cause analysis. The tool deploys a suite of algorithmic checks to detect dynamic outliers, such as univariate, bivariate, and time-series outliers, and knowledge-based outliers, such as invalid unit conversion and unit mislabeling. Coupled with the developed visualization tool, users can efficiently assess laboratory data quality.

Introduction

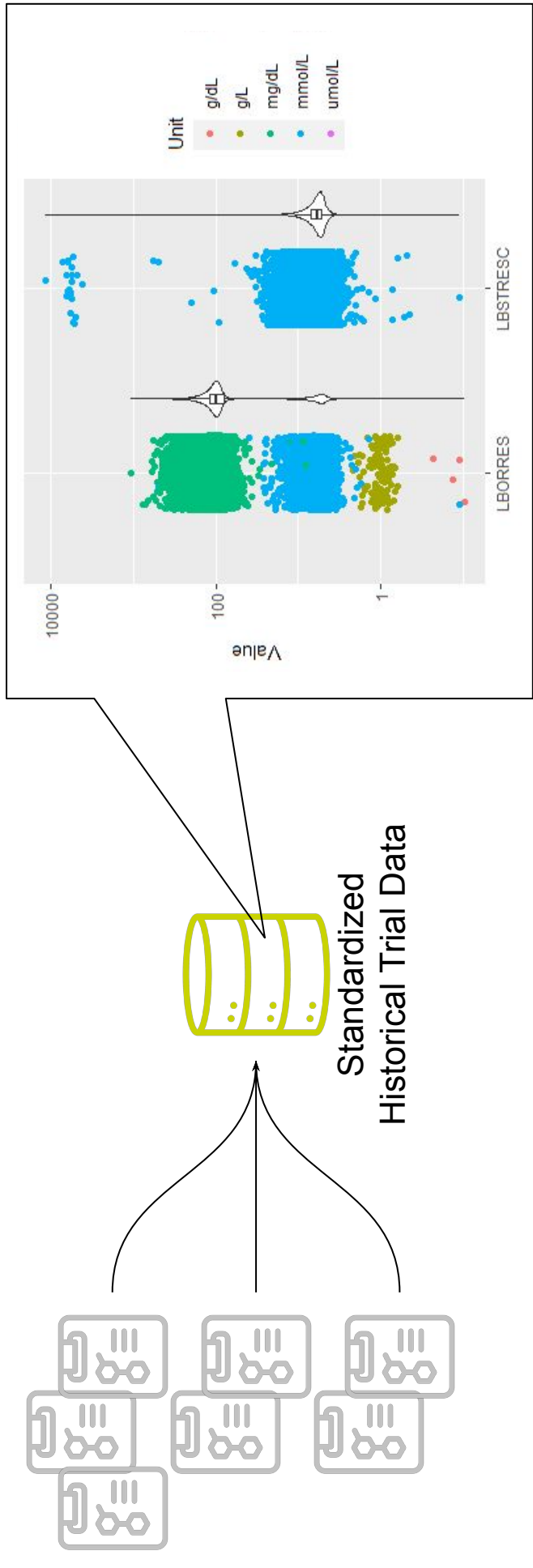
Within the context of clinical trial data standardization, the standardization of laboratory data is of the utmost importance and can represent a challenge given the large variability in terms of the manner and consistency in which the data is collected. That is, patients on a given trial may have routine laboratory studies done at a variety of different laboratories, each of which may have its own preferred units of expression and "normal" ranges. In order to confirm that standardization has been completed properly for the purposes of analysis or FDA submission, thorough quality control (QC) of any data processes needs to be completed, preferably beyond rule-based checks. We have developed a Lab Data Quality tool to facilitate QC through (1) observation of outliers within the data and (2) investigation of outlier's root cause in order to correct standardization issues and inform the appropriate parties of true outliers.

The Lab Data Quality tool uses both statistical and machine learning methods to identify potentially problematic outlier observations to end users. Some checks dynamically establish a set of reference ranges by the data distribution and use it to detect outliers; while others identify patterns from historical laboratory data to detect data points with abnormal patterns, enabling users to perform a root cause analysis by comparing it to past observations. Once standardized and assured of its high quality, the resulting data sets are available for further analyses for various industry purposes.

Most importantly, these methods involve taking a first step into the broader potential of using historical clinical trial data for the purposes of data quality improvement. By using historical data to inform how an algorithm expects a particular piece of data to be represented, we can catch erroneous data (as distinct from true extreme values) on entry or during standardization review to be corrected prior to further analysis.

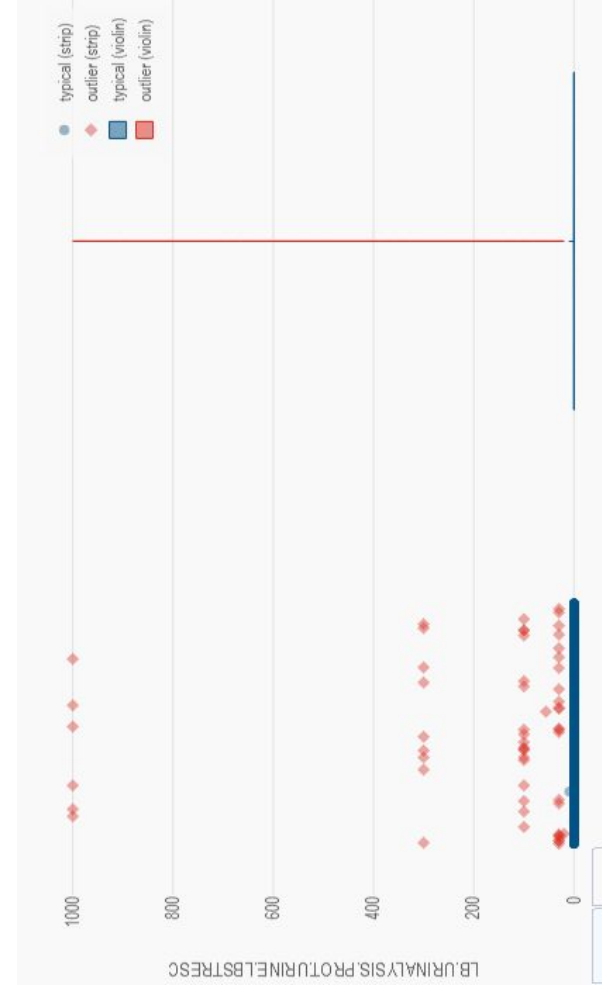
Data Standardization

Historical clinical trials data were harmonized and standardized into SDTM format before integrated into a huge single source of dataset. Data points from the same lab tests were pooled to have a consistent and comparable distribution of the measurements.

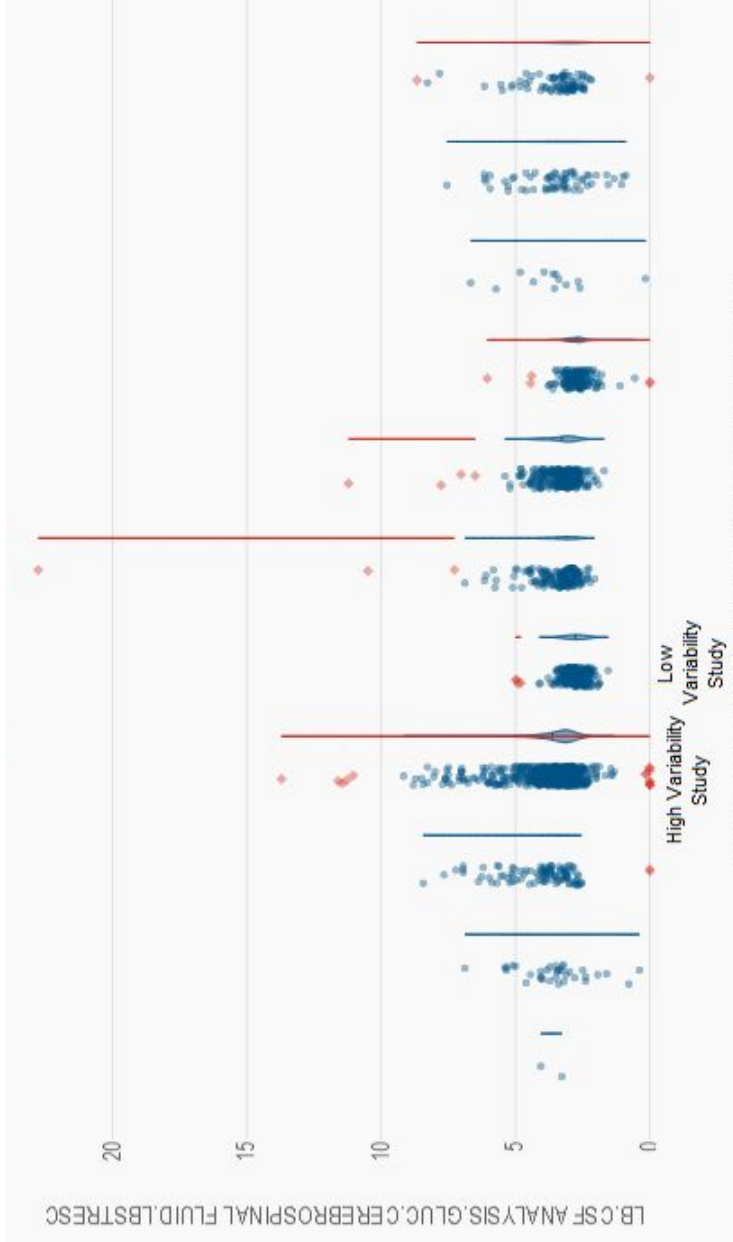


After being translated into a consistent format (SDTM), blood glucose levels collected from multiple studies and sites were converted into a common unit (LBSTRESQ) for data integration, increasing the amount of data points that could be utilized for several purposes.

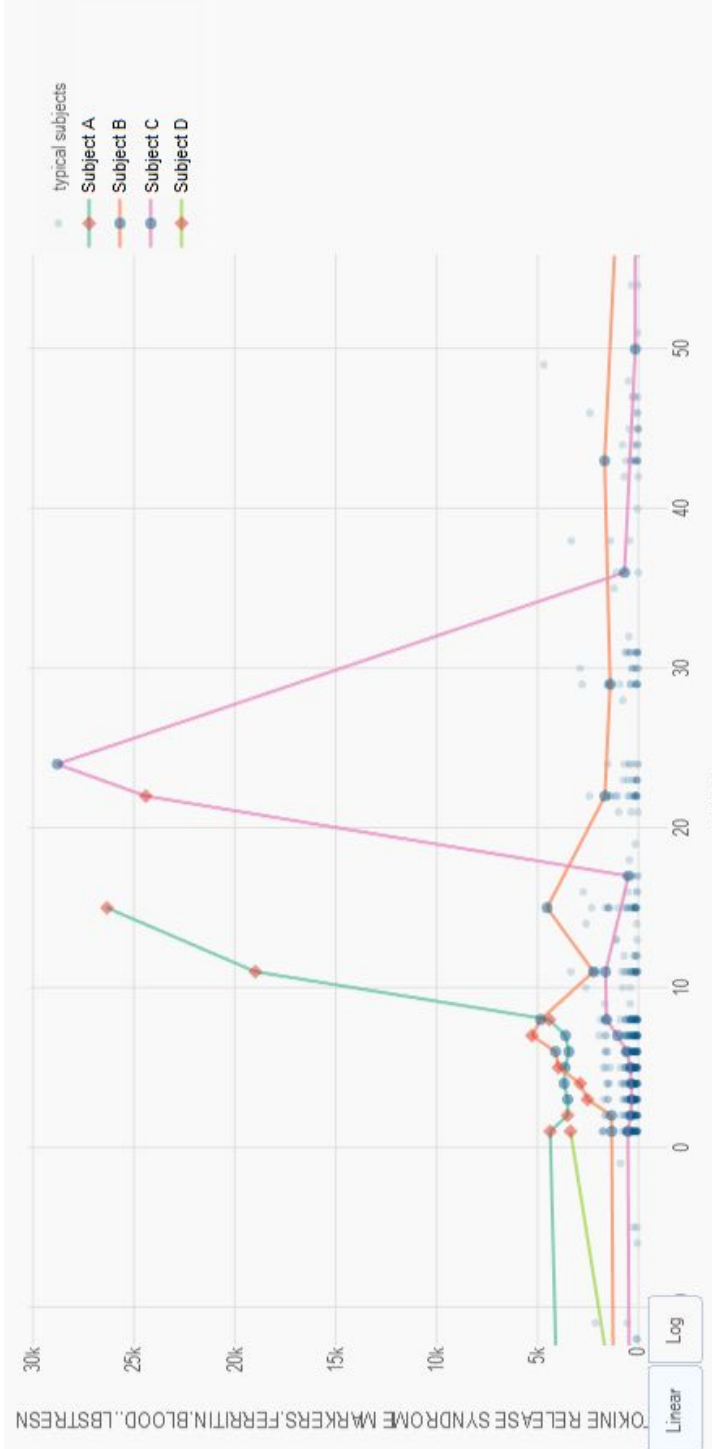
Data Quality Assessment



Univariate Outlier: A few number of data points (~0.8%) are found to have abnormally high levels of total protein in urine samples, where majority data points are close to 0 g/L.



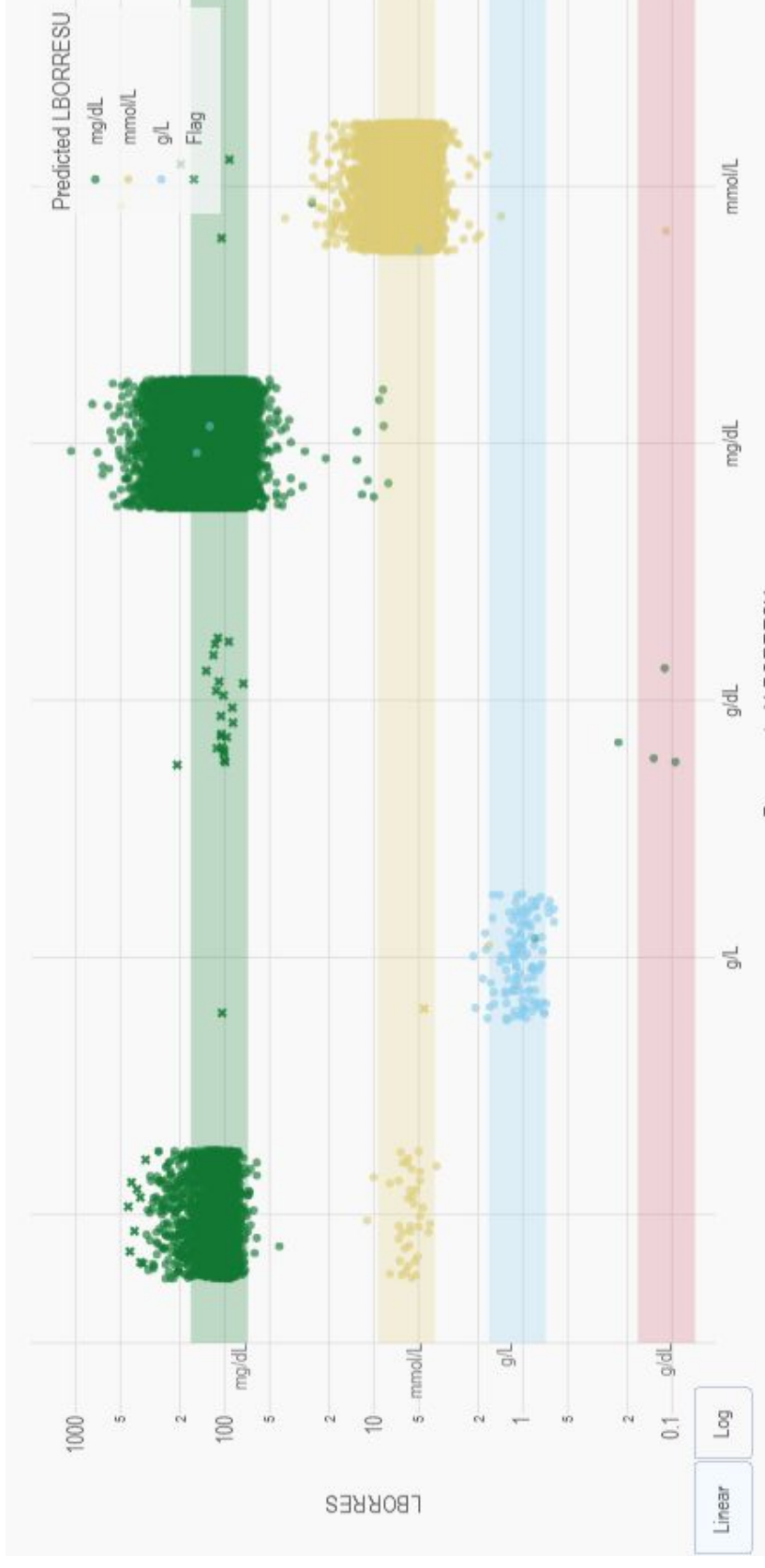
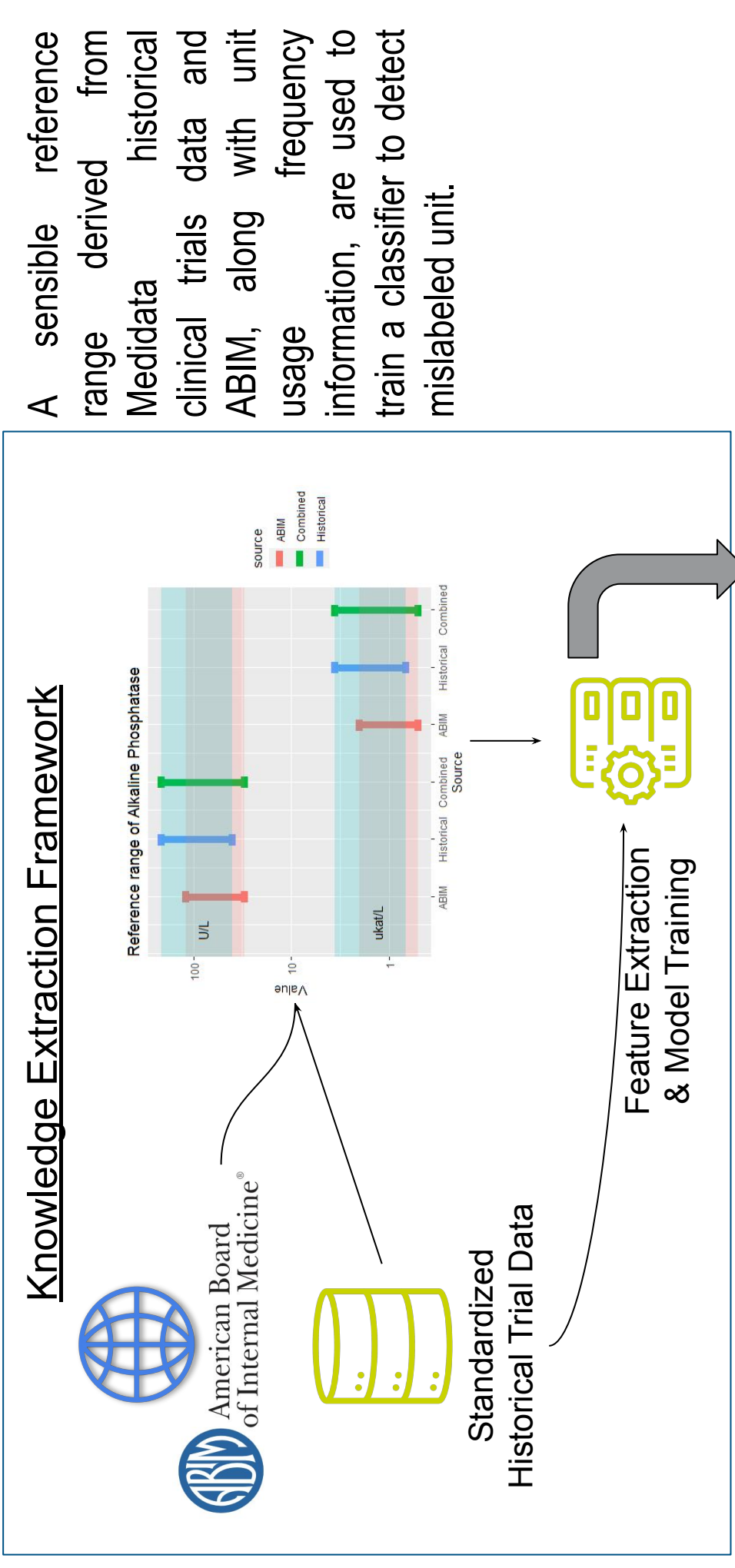
Bivariate Outlier: Bivariate analysis identifies outliers based on different context (e.g. study difference in the example). One study has a higher amount of variability and it has a higher threshold in flagging outliers (>10 mmol/L). Around 1.7% of data points were flagged as outliers, half the number of data points flagged when all studies were pooled together.



Time Series Outlier: All patients have relatively stable Ferritin level throughout the trials, except the three patients as shown in the graph. These patients have outlying changes at the beginning, with two patients have abnormally high value as time progressed.

Visualization Dashboard

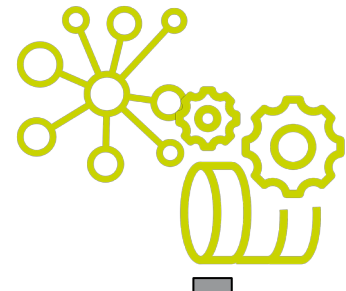
A user-friendly visualization tool has been developed, allowing users to navigate data quality profiles and perform any root cause analysis.



Unit Mislabeling Error: Utilizing the derived reference range and trained classifier, several cluster of Blood Glucose measurements were found to have mislabeled unit. A cluster of data points documented as "g/dL" is predicted to be "mg/dL", as well as several data points documented as "mg/dL" and "mmol/L". During this process, data points with missing unit were relabeled with predicted unit too.

LBTEST	LBORRES	LBORRESU	LBSTRESN	LBSTRESU	Derived Conversion Factor	True Conversion Factor	Expected LBSTRESN
Protein	100	mg/dL	1000	g/L	10	0.01	1
Protein	300	mg/dL	3000	g/L	10	0.01	3
Protein	400	mg/dL	4000	g/L	10	0.01	4
Protein	100	mg/dL	1000	g/L	10	0.01	1

Unit Conversion Error: A few number of data points were converted incorrectly during unit standardization. These data points were detected by comparing with the true conversion factor, retrieved from a knowledge graph housed in Medidata.



Future Potential

Historical clinical trial dataset allows us to learn how data is typically represented. This opens doors to improve data quality across the board in a laboratory setting. The goal with this work is to take the first step into anticipating data entry issues and correcting them on entry rather than later on down the data standardization pipeline. Future work includes enabling users to adjust historical norms based on patient indications to disease history to demographics.

Acknowledgement

Co-Authors: Caleb Strati, Elizabeth Lamont, Katy Grantous, Kelly Burke

Special Thanks: Ian Fleming