Understanding the Individual Contributions to Multivariate Outliers in Assessments of Data Quality

Richard C. Zink, Ph.D.
Senior Director, Data Management and Statistics
TARGET PharmaSolutions Inc.
rzink@targetpharmasolutions.com

Thanks to:

- Laura Castro-Schilo
- Jianfeng Ding

TARGET PharmaSolutions

Introduction

- Two aspects to data quality
  - Identifying anomalous data or observations
  - Investigating why the unusual data may have occurred
- Identification
  - Exceed a statistically-relevant threshold
  - Meaningful Δ, though not significant
  - In general, when results do not align with our expectations
- Investigation
  - Natural differences among patients or technique
  - Carelessness
  - Misconduct

Mahalanobis Distance

- Mahalanobis distance[1] is often recommended to identify unusual observations from many variables
- Considers the pairwise correlation among variables
- Distance is often calculated from the centroid, or the multivariate mean
- Two considerations
  - Outliers: Patients extreme in one or more covariates that cause them to be far away from the centroid
  - Inliers [2]: Patients close to the centroid that could be considered too good to be true

Challenge

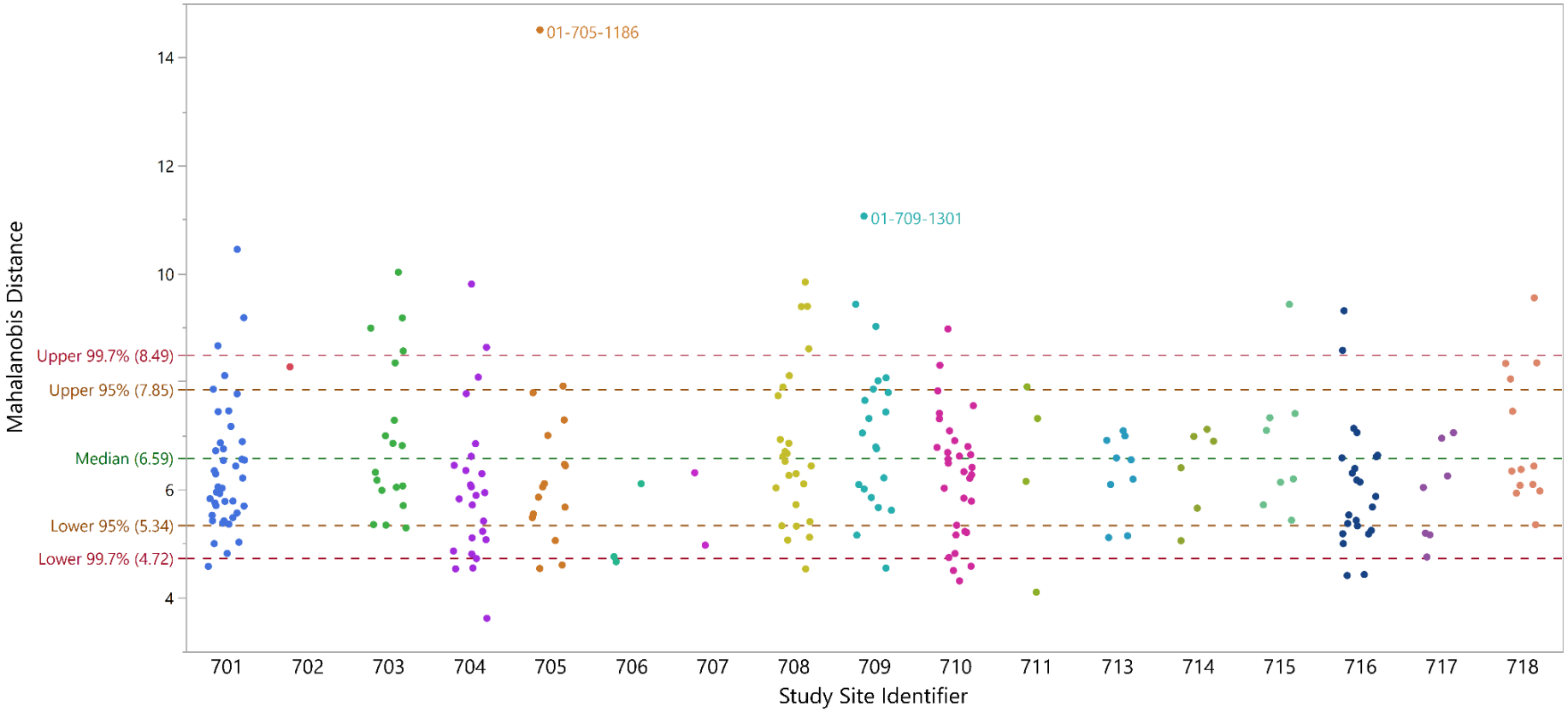
- Efficiently identify inliers and outliers
- Potentially create thresholds of varying severity to triage numerous signals
- Investigate what contributes to the outlier by partitioning the distance among the covariates
- Quickly and easily…
- … and without a lot of statistics or numbers

Data

- Phase II clinical trial to examine the safety and efficacy of the xanomeline transdermal therapeutic
- Patients with mild to moderate Alzheimer's disease
- 254 patients randomized 1:1:1 to either high or low dose xanomeline or placebo
- 17 clinical sites
- 44 covariates measured at baseline: age, vital signs, laboratory measurements and a subset of items from questionnaires
- Data available from CDISC[3]

TARGET PharmaSolutions

Mahalanobis Distance



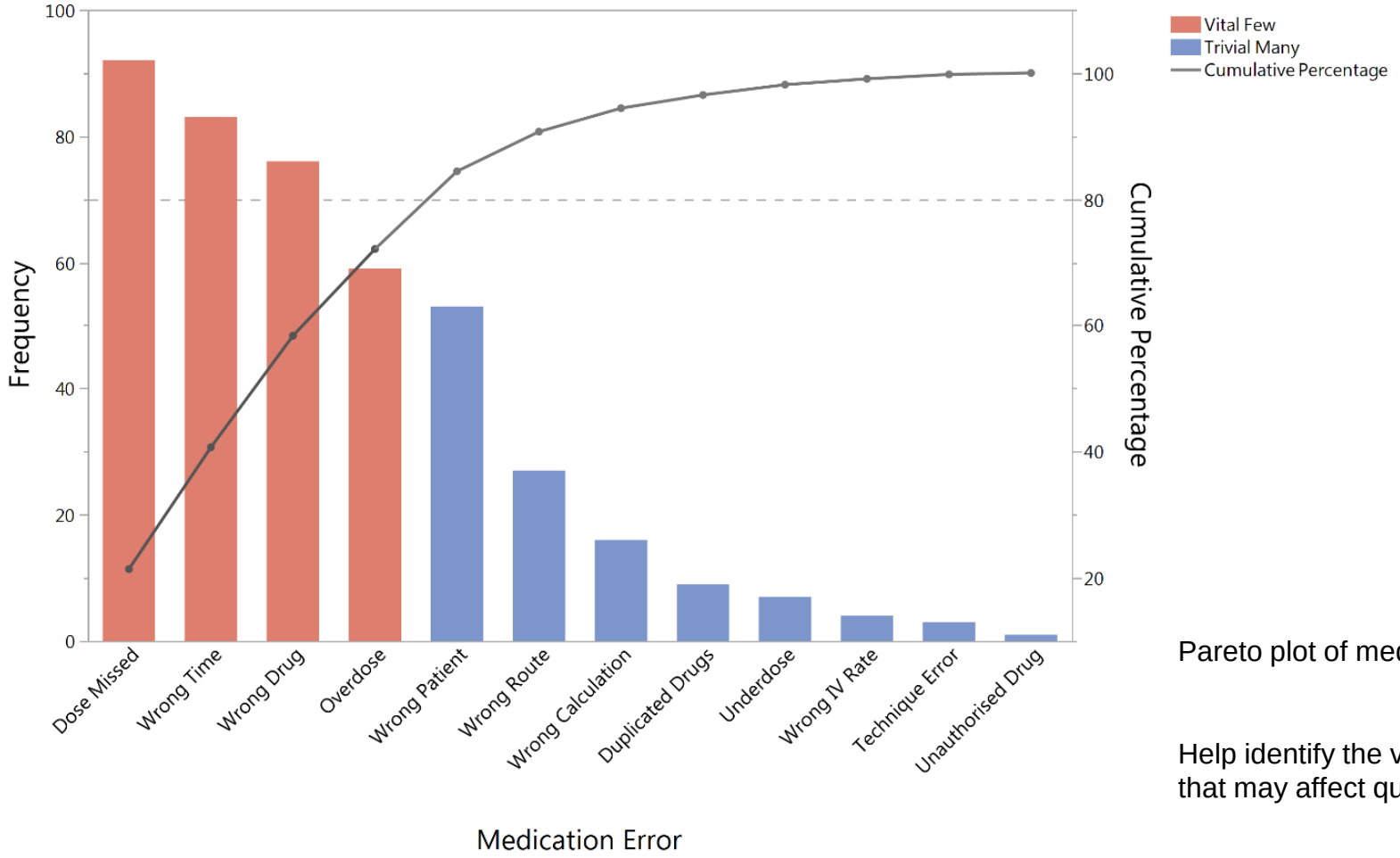
Investigation

- Identified the outliers and inliers
- Question that we constantly received
  - What covariates caused outliers to be unusual?
- Partition the squared-distance into components that reflect the degree to which each covariate contributes
- Call these components c_{ij}, ith person, jth covariate
- Values aren't particularly meaningful
- Compute proportions p_{ij} of the squared-distance attributable to a covariate

Investigation

- Some math:
  - $C = (Y_{cs}E)D_{\sqrt{\lambda}}^{-1}E^T$ defines contributions, though no straightforward interpretation
  - $t_i^2 = \sum_{j=1}^{p} c_{ij}^2$
  - Define contribution proportions $p_{ij} = \frac{c_{ij}^2}{t_i^2}$, the proportion of the squared Mahalanobis distance of observation i that is attributable to covariate j
- Contribution proportions have more straightforward interpretation and can leverage Pareto plots for summary

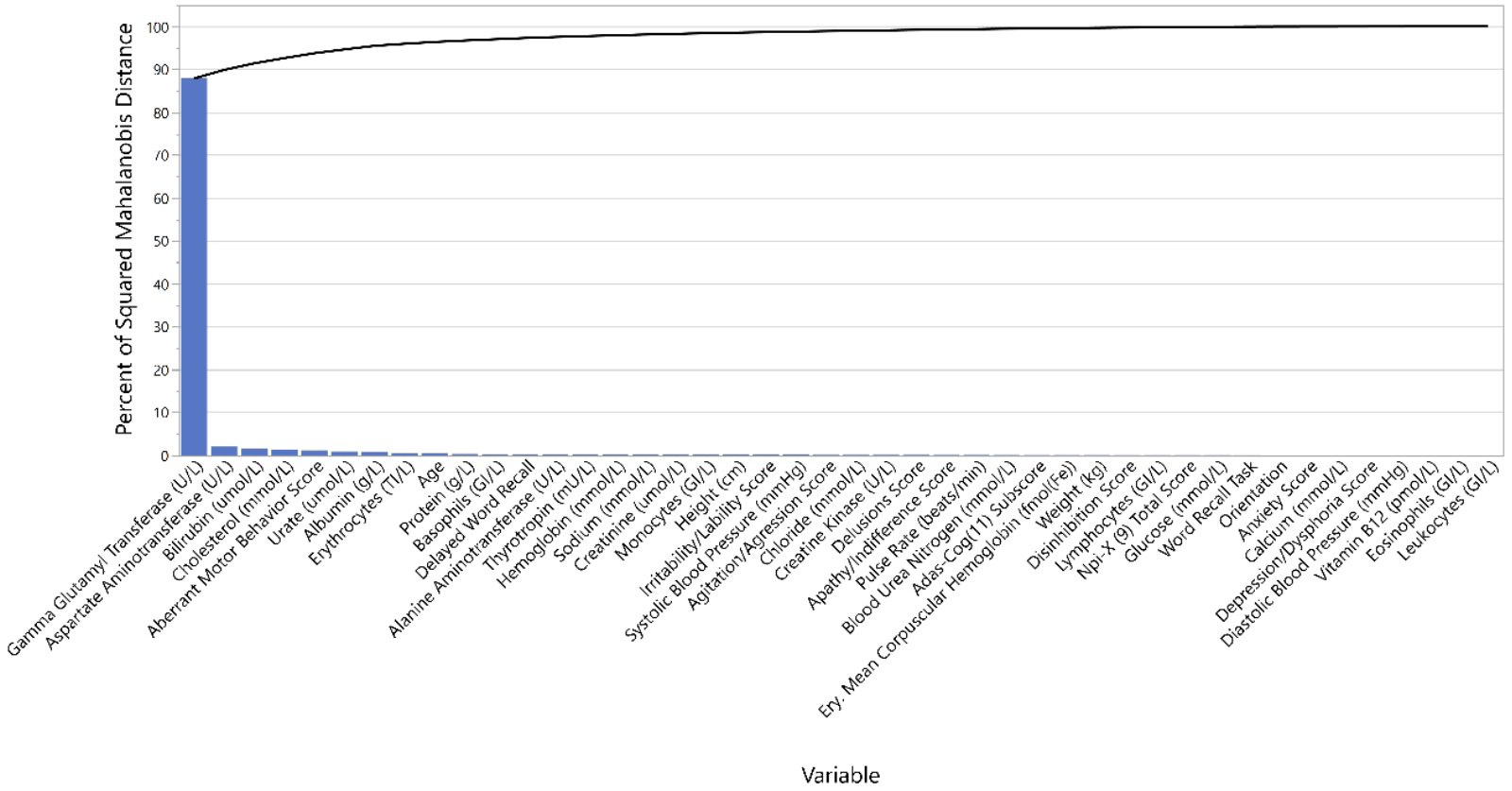
Pareto Plot - Medication Errors



Pareto plot of medication errors

Help identify the vital few issues that may affect quality.

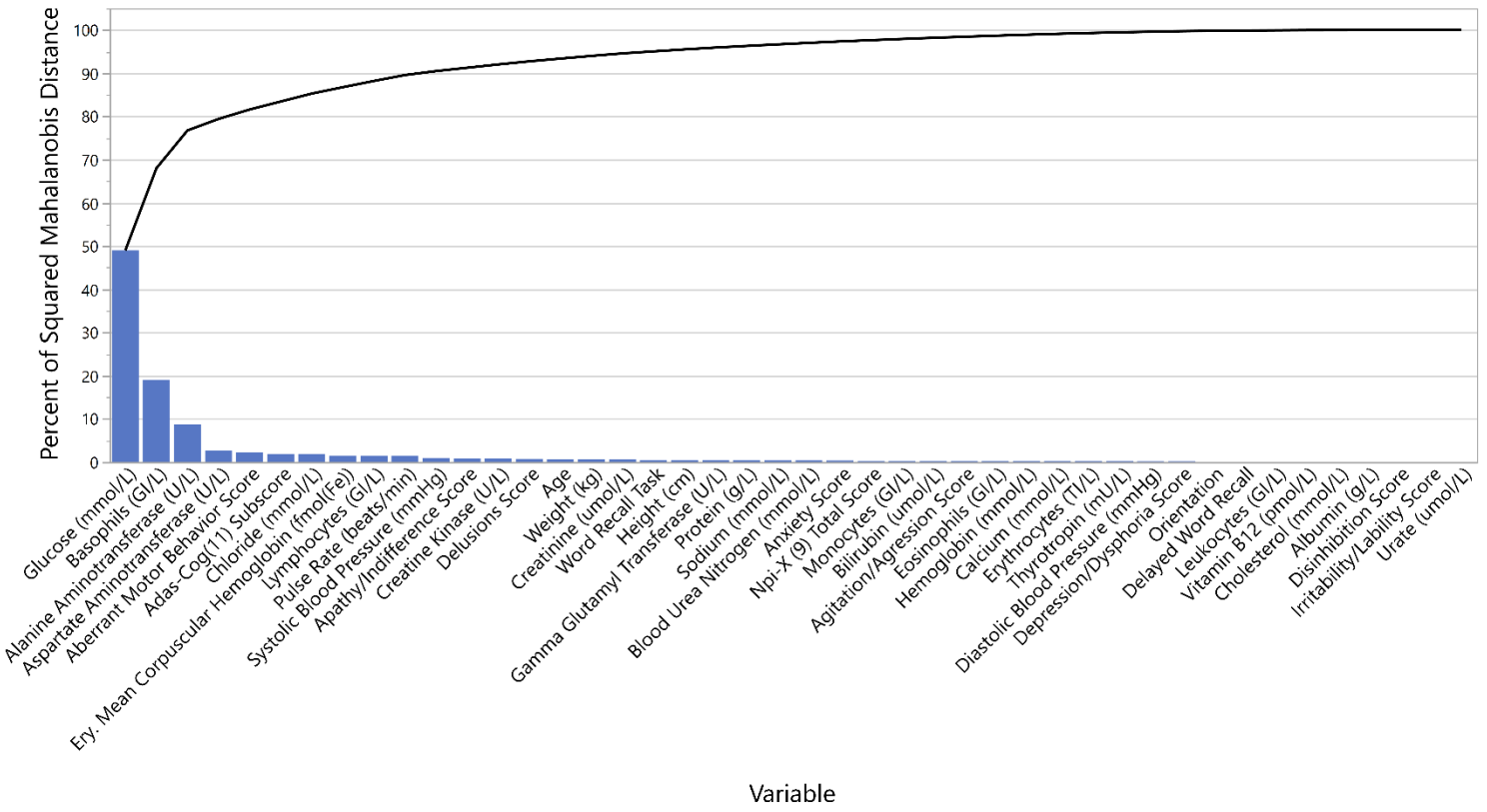
Investigation



Pareto plots per person – 01-705-1186
Black line reflects cumulative proportion of error
Covariates ordered according to contribution

TARGET PharmaSolutions

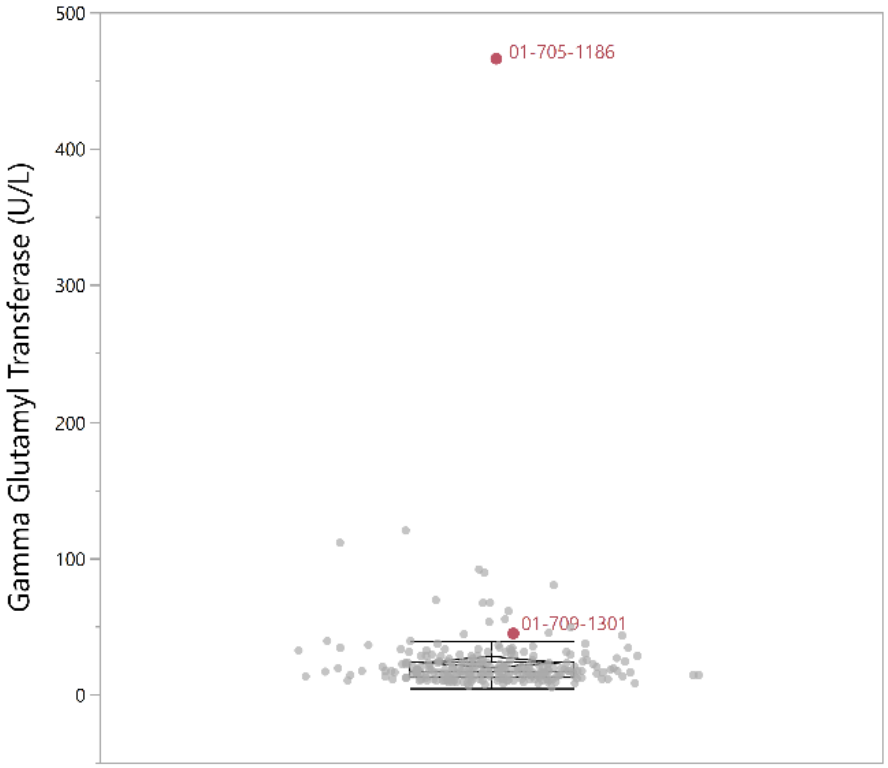
Investigation

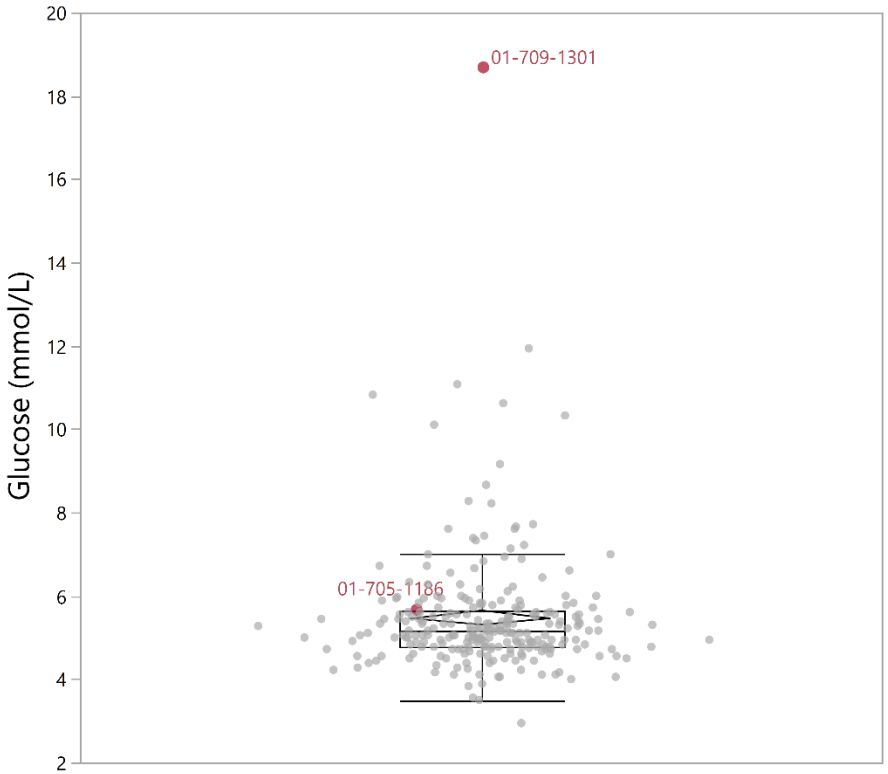


Pareto plots per person – 01-709-1301
Black line reflects cumulative proportion of error
Covariates ordered according to contribution

TARGET PharmaSolutions

Mahalanobis Distance

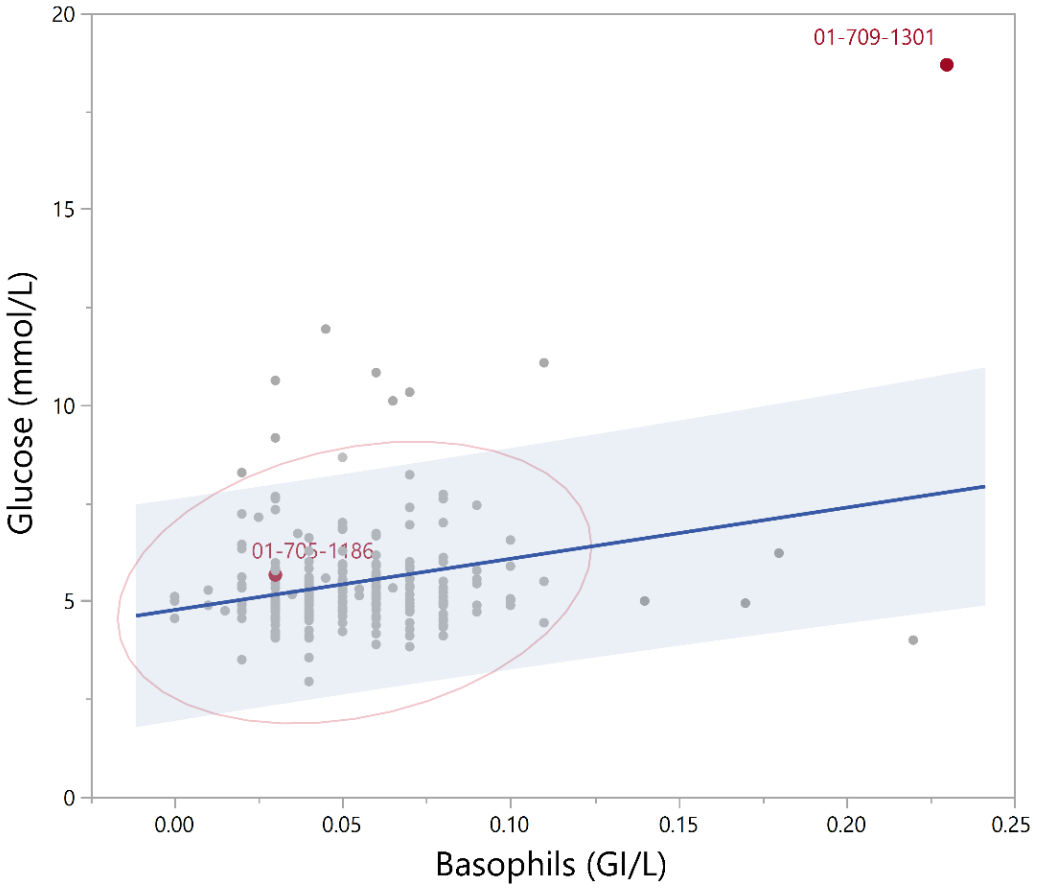




Box plots to assess meaningful covariates in context of entire population.

TARGET PharmaSolutions

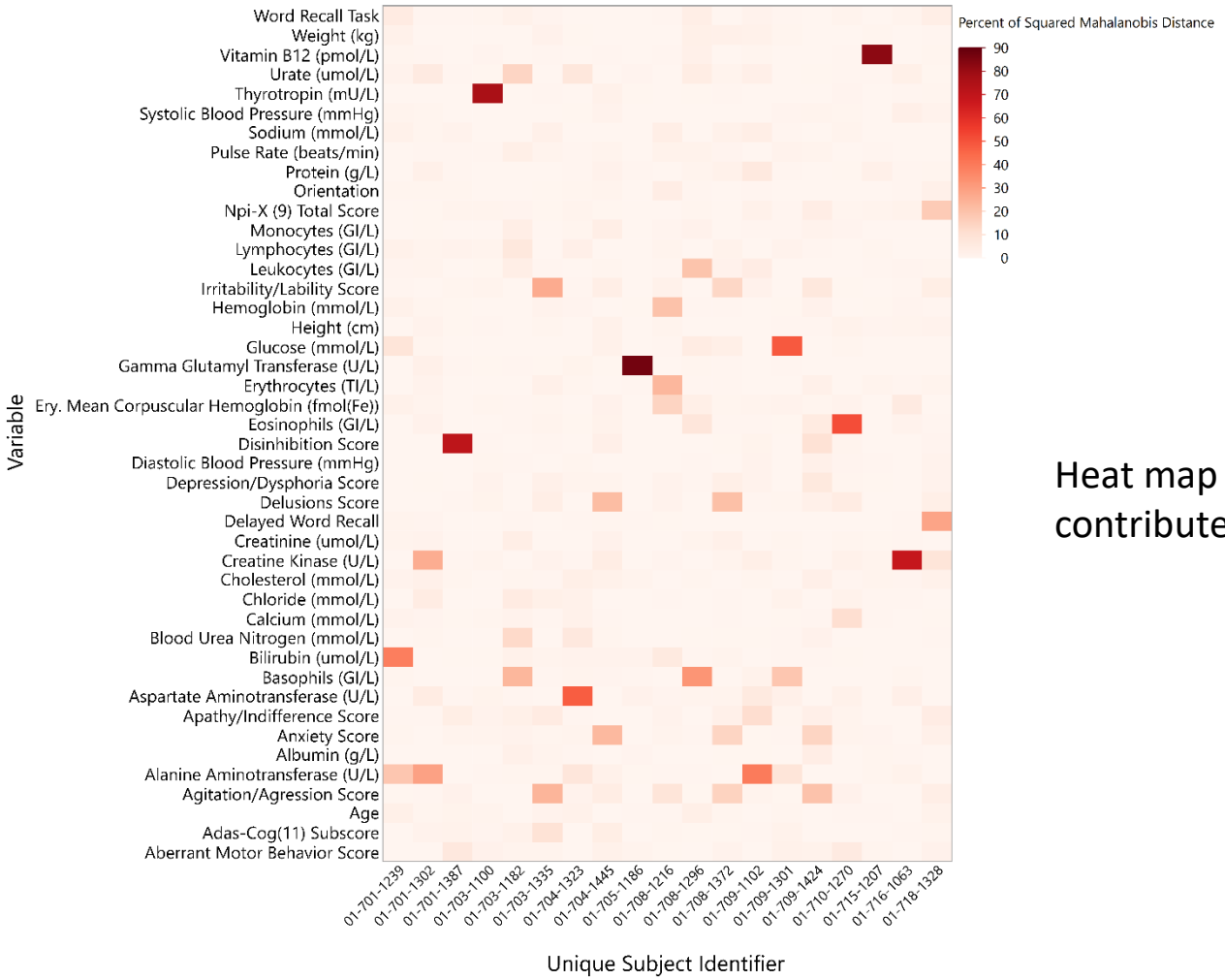
Mahalanobis Distance



Scatter plot with regression line and confidence interval and density ellipse

TARGET PharmaSolutions

Mahalanobis Distance



Heat map to assess which covariates contribute to outlier status.

TARGET PharmaSolutions

Conclusions

- Assessed unusual observations
- Described thresholds of varying severity
- Proposed contribution proportions as a straightforward way to identify influential covariates
- Visualize contribution proportions with Pareto plots
- Use additional graphics to assess outliers in the context of the population
- Limitations
  - Requires normality of original covariates
  - Not useful for inliers

References

1. Mahalanobis PC (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India* 2: 49–55.
2. Evans S. (2001). Statistical aspects of the detection of fraud. In: Lock S & Wells F, eds. *Fraud and Misconduct in Biomedical Research, Third Edition*. London: BMJ Books.
3. CDISC SDTM / ADaM Pilot Project. (2012). Data Package.
4. Zink RC, Castro-Schilo L & Ding J. Understanding the influence of individual variables contributing to multivariate outliers in assessments of data quality. *Pharmaceutical Statistics* (in print).