DATA SCIENCE - EXTRACTING INSIGHTS
Open Journal

*Review*

# Visualizations of Data Created to Aid in the Detection of Laboratory Data

**Akmal Djamaan*; Reni Mayerni; Betna Dewi**

*Institute for Geophysics, Austin, TX, USA*

**Corresponding author*
**Akmal Djamaan**
Institute for Geophysics, Austin, TX, USA

***Article information***
**Received:** March 23rd, 2024; **Revised:** September 1st, 2024; **Accepted:** September 17th, 2024; **Published:** October 14th, 2024

***Cite this article***
Djamaan A, Mayerni R, Dewi B. Visualizations of data created to aid in the detection of laboratory data. 2024; 2(2).
doi: https://doi.org/10.70705/ppp.dsei.2024.v02.i02.pp151-152

**ABSTRACT**

Research and service activities in fundamental and clinical pharmacology rely heavily on the assessment of pharmacological and endogenous chemical concentrations. Using data science visualizations of laboratory data, it is shown on a real-life example that recommended standard visual techniques of data inspection or basic statistical investigation of laboratory test findings may fail to discover systemic laboratory mistakes. For instance, conventional data quality control approaches could miss data pathologies like an experiment run where all probes provide the same result. We find that sys- tematic laboratory mistakes may be better detected when using data visualizations that highlight diverse perspectives on the data. To better understand the data range, outliers, and a specific kind of systematic error—where identical values are incorrectly obtained in all probes—a dotplot of individual data organized by assay is suggested.

**Keywords**
Data quality check; Data science; R programming language.

## INTRODUCTION

Pharmacologic research relies heavily on the assessment of medication or endogenous substance concentrations in biological materials. It is extremely important that the measurements be reliable. Consequently, analytical laboratories consistently include quality control into their operation. Biomedical data reporting standards include a number of tools for detecting assay errors1, such as summary statistics for plausibility checks and data visualizations.2 However, it would be ideal if the identification of assay mistakes could be further improved. The current research provides a case study that illustrates how recommended techniques of data exploration might allow laboratory mistakes to go unnoticed. A straightforward approach that might improve the identification of laboratory mistakes is suggested, which makes use of visuals derived from data science.

These results are based on an ongoing study that compares the plasma concentrations of patients and healthy controls for biomarkers. The study adhered to the Declaration of Helsinki on Biomedical Research Involving Human individuals and obtained clearance from the Ethics Committee of the Medical Faculty of the Goethe-University in Frankfurt, Germany. All individuals had consented to biomarker evaluations. However, in order to maintain confidentiality, we will describe this technical finding using anonymised data that has been rescaled by a constant numerical factor. A trio of biochemical indicators isolated from plasma, with the arbitrary designations "Lab1," "Lab2," and "Lab3," have been documented. The first marker, "Lab1" (Figure 1), was likely evaluated without laboratory errors; the other two markers, "Lab2" and "Lab3" (Figure 1), both exhibit systematic assay errors to varying degrees, as confirmed by assay repetitions. Specifically, for the second marker, "Lab2" (data marked with a red ellipse in Figure 1), the laboratory consistently produced the same value; for the third marker, "Lab3," the concentration was typically zero, with the exception of one day when the laboratory produced exceptionally variable values higher than the lower limit of quantification (Figure 1 right panel).

A common approach to data quality check is the application of basic descriptive statistics (Table 1). This can provide plausibility checks when comparing the observed values with the expectations of a domain expert, who knows the physiological or pathophysiological value range of the parameters or the magnitude and direction of expected differences among, for example, clinically relevant groups. In the present example, an assay error in "Lab3" would have been suggested from the median and mean values of zero or almost zero, respectively. However, the error in "Lab2" in particular the similar values obtained in a particular assay run would pass undetected as the descriptive statistics appear to be unsuspicious.

Similar results were provided by basic data visualizations, of which the simplest and generally discouraged variant is a bar blot with error bars (Figure 2 left). A more sophisticated variant, a box- plot overlaid with the observed single data (Figure 2 right), again would indicate merely the almost always zero values in "Lab3" whereas the more subtle pathologies of the datasets, that is, same values during a particular assay run or during a whole day for "Lab2" and "Lab3" respectively, were not visualized by these standard plots. A better visualization of systematic errors in laboratory assay results provided the heatmap. If the data values were entered in the problems, such as identical results across all probes in a given experiment run, were glaringly obvious. It is crucial to plot the data in the order of the assay. If this sequence was lost (see Figure 1 bottom line), the short repetition of the same value in "Lab2" would no longer be visible, and any values above zero in "Lab3" would no longer be interpreted as a sign of systematic error in the laboratory, but rather as random fluctuations.

The R software program, for example, is freely available online and may be used to create any kind of visualization (http://CRAN.R-project.org/5). Figure 1 was created using the standard R command plot (LabValues, pch = 20, cex =.1), where "labValues" is a vector of assay results for one parameter in the assay order, and "pch" and "cex" serve as appropriate symbols and sizes for the scatterplot's dots. You can find the whole R script that was used to generate these figures in the report's Appendix S1. Depending on the local standards, it can be modified to work in various environments like Python or MATLAB.

It is shown in this brief paper that data visualization is essential for detecting systematic laboratory mistakes when measuring concentrations in biological materials. In contrast, it is insufficient to just evaluate fundamental statistical parameters. Also, picking one of the recommended plots might not be enough to catch systematic mistakes. We suggest a dotplot of each data ordered by assay to show the whole data set, any outliers, and a specific kind of systematic mistake where identical values are mismeasured in every probe. As a result, data science methods2 that offer various visualizations that highlight diverse perspectives on the data should be used to improve the quality check of biochemical laboratory data in order to potentially increase the identification of systematic mistakes in the lab.

## REFERENCES

First, Landis SC, Amara SG, Asadullah K, and crew. An appeal for open reporting in order to maximize the predictive power of preclinical studies. "Nature" published in 2012, 490(5), 187–191.

2. Data science with Wickham and Grolemund: importing, cleaning, transforming, visualizing, and modeling. O'Reilly Media, 2017; Beijing, Boston, Farnham, Sebastopol, and Tokyo: Publications.

Three, Pareto density estimation (Ultsch A., ed.) is a density estimate tool for finding new information. new developments in data science, information systems, and classification - Proceedings of the 27th Annual Conference of the German Classification Society (GfKL). Leipzig: Springer Verlag, 2003.

4. Thrun MC, Hansen-Goos O, Lo€tsch J., and Ultsch A. Using the AdaptGauss interactive mixture model R package, we were able to identify chemical fingerprints in human thermal pain thresholds. International Journal of Molecular Sciences, 2015, 16, 25897–259111.

R Core Team, which is number five. An Environment and Language for Statistical Computing: R. Vienna, Austria; 2008.