Received: 4 October 2017 | Accepted: 10 October 2017

DOI: 10.1002/prp2.369

ORIGINAL ARTICLE

Data visualizations to detect systematic errors in laboratory assay results

Jörn Lötsch[1,2]

[1]Institute of Clinical Pharmacology, Goethe - University, Frankfurt am Main, Germany

[2]Fraunhofer Institute of Molecular Biology and Applied Ecology - Project Group Translational Medicine and Pharmacology, IME-TMP, Frankfurt am Main, Germany

Correspondence
Jörn Lötsch, Goethe - University, Frankfurt am Main, Germany.
Email: j.loetsch@em.uni-frankfurt.de

Funding information
European Union Seventh Framework Programme, Grant/Award Number: FP7/2007 – 2013; Landesoffensive zur Entwicklung wissenschaftlich - ökonomischer Exzellenz (LOEWE), LOEWE-Zentrum für Translationale Medizin und Pharmakologie

Abstract

The measurement of concentrations of drugs and endogenous substances is widely used in basic and clinical pharmacology research and service tasks. Using data science-derived visualizations of laboratory data, it is demonstrated on a real-life example that basic statistical exploration of laboratory assay results or advised standard visual methods of data inspection may fall short in detecting systematic laboratory errors. For example, data pathologies such as generating always the same value in all probes of a particular assay run may pass undetected when using standard methods of data quality check. It is shown that the use of different data visualizations that emphasize different views of the data may enhance the detection of systematic laboratory errors. A dotplot of single data in the order of assay is proposed that provides an overview on the data range, outliers and a particular type of systematic errors where similar values are wrongly measured in all probes.

KEYWORDS

data quality check, data science, R programming language

The measurement of concentrations of drugs or endogenous substances in biological materials plays a major role in pharmacologic research. The reliability of the measurements is of crucial importance. Therefore, quality control is routinely implemented in the workflow of analytical laboratories. Standards of biomedical data reporting comprise a variety of measures for assay error detection[1] including summary statistics for plausibility checks and data visualizations.[2] Nevertheless, further improvements of the detection of assay errors are desirable. In the present report, a real-life example is given that laboratory errors may pass undetected with advised methods of data exploration. A simple solution employing the application of data science-based visualizations, which may enhance the detection of laboratory errors, is proposed.

The data originate from a current project on biomarker concentration assessment in the plasma of patients and healthy controls. All subjects had consented into biomarker assessments and the study followed the Declaration of Helsinki on Biomedical Research Involving Human Subjects including approval from the Ethics Committee of the Medical Faculty of the Goethe-University, Frankfurt, Germany. However, for reasons of nondisclosure, the present technical observation will be reported using anonymized data rescaled with a constant numerical factor. Three different plasma-derived biochemical markers are reported, arbitrarily named "Lab1", "Lab2" and "Lab3". While one of the markers was assessed probably without laboratory errors (named "Lab1"; Figure 1), the two other markers carry systematic assay errors at different degrees as verified in assay repetitions. Specifically, in the measurements of the second marker, in a particular assay run, the laboratory produced always the same value ("Lab2", data marked with a red ellipse in Figure 1), whereas the third marker ("Lab3") was usually measured at a concentration of zero except for one day when the laboratory produced highly variable values above the lower limit of quantification (Figure 1 right panel).

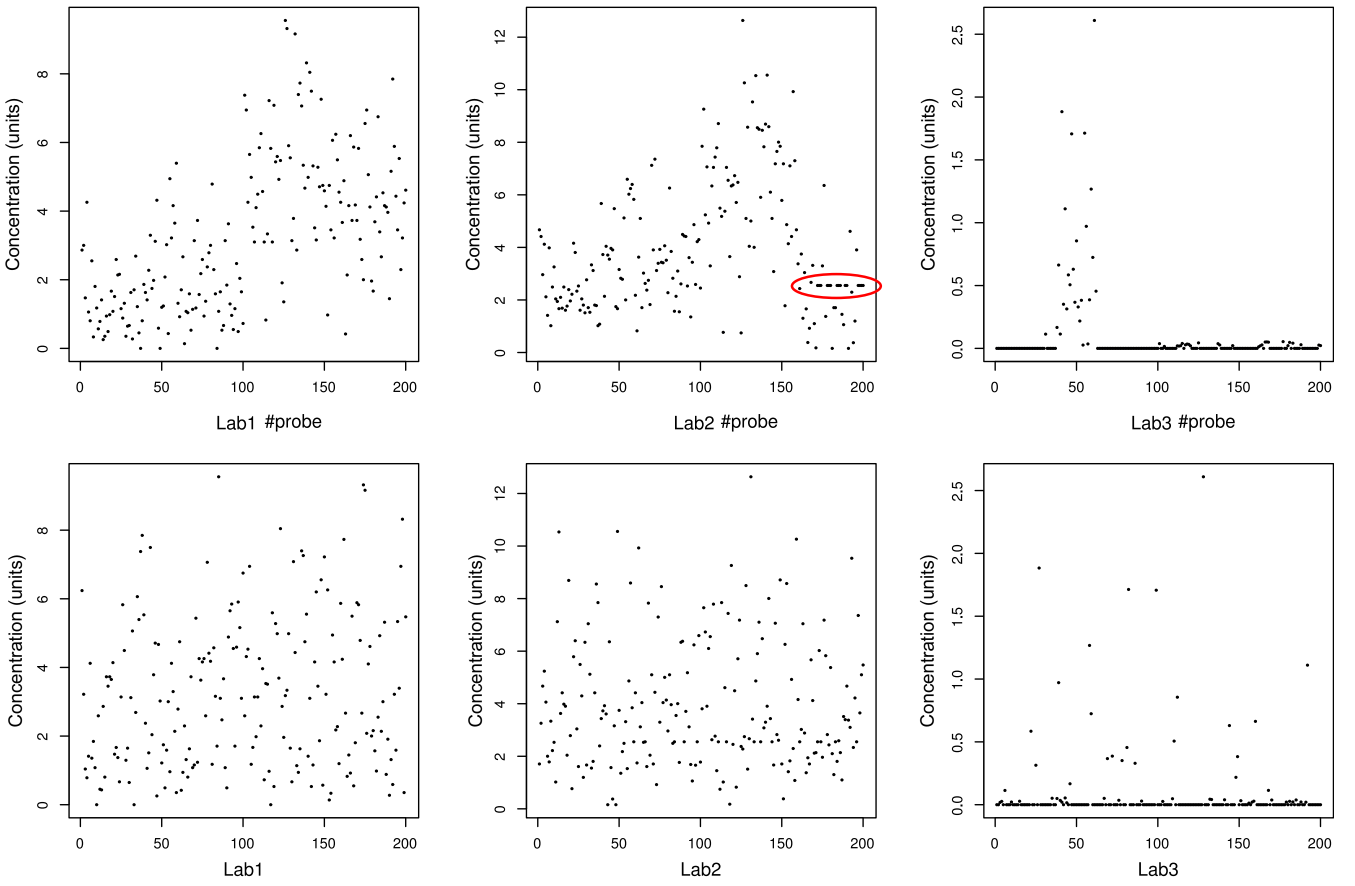

FIGURE 1 Dotplot of plasma concentrations of three different biochemical markers (arbitrarily named "Lab1", "Lab2" and "Lab3"). The dots display the single data, sorted in order of consecutive assay (upper line). Two different clinical phenotypes are included with a distribution of n = 100/100. In the parameter "Lab2" a short temporal window (red ellipse) was detected during which all measured concentrations had wrongly the same numerical value. In "Lab3" all measurements were zero except for one assay day during which highly variable results were produced. The detection of these errors became impossible when the temporal succession of assay was destroyed (bottom line)

TABLE 1 Descriptive statistical analysis of the three laboratory parameters, originating from an actual scientific project but presently arbitrarily named "Lab1", "Lab2" and "Lab3"

Parameters	Lab1	Lab2	Lab3
N	200	200	200
Mean	3.22	3.95	0.1
Standard deviation	2.19	2.43	0.33
Median	3.06	3.33	0
Trimmed mean	3.04	3.7	0.01
Median absolute difference	2.5	2.25	0
Minimum	0	0.16	0
Maximum	9.55	12.64	2.61
Range	9.55	12.48	2.61
Skewness	0.62	0.91	4.72
Kurtosis	−0.34	0.32	25.02
Standard error	0.15	0.17	0.02

The calculations were made using the "describe" command of the R library "psych" (Revelle W, Northwestern University, Evanston, Illinois, https://CRAN.R-project.org/package=psych) on the R software package (version 3.4.1 for Linux; http://CRAN.R-project.org/[5]).

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 boxplot 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

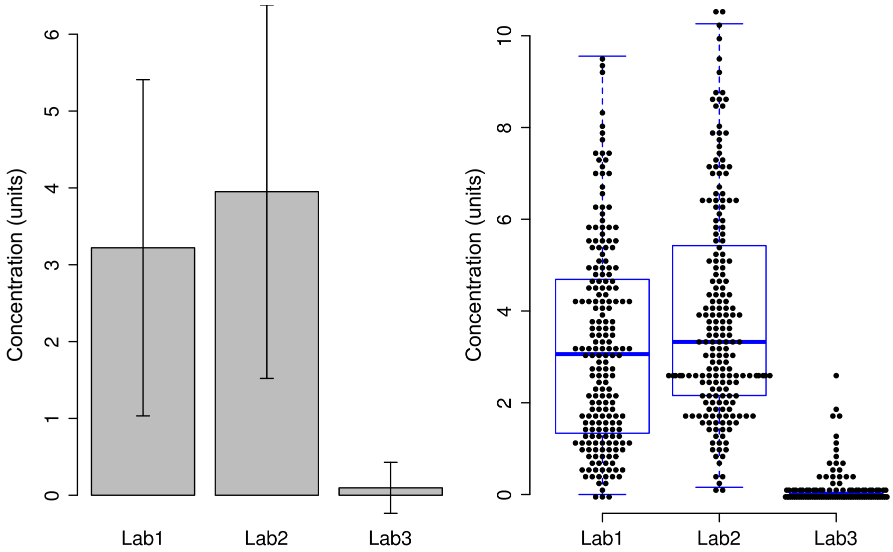

FIGURE 2 Graphical presentation of plasma concentrations of three different biochemical markers (arbitrarily named "Lab1", "Lab2" and "Lab3"). Left panel: Bar plot with means and standard deviations (error bars). Right panel: Boxplots overlaid with the original data observations. Quartiles and medians (solid horizontal line within the box) were used to construct a "box and whisker" plot. The whiskers add 1.5 times the interquartile range (IQR) to the 75th percentile or subtract 1.5 times the IQR from the 25th percentile and are expected to include 99.3% of the data if normally distributed

order of their assay (Figure 3 top). For "Lab3" the plot clearly emphasized high values during a limited assay set contrasting with the other values that were usually zero. However, the shorter period of similar values in a single assay run for "Lab2" was barely detectable in the heatmap. A comparably informative visualization was provided by the probability density function of the measured values described as the Pareto density estimation (PDE; Figure 3 bottom). This is a kernel density estimator that has been developed with the focus to be particularly suitable for the discovery of groups in the data.[3] An implementation can be found in the R package "AdaptGauss" (https://cran.r-project.org/package=AdaptGauss[4]). In the present example, the PDE (Figure 3 bottom) clearly showed a dominance of zero values in "Lab3" and emphasized the rather broad variance suggesting accidental measurements in the rest of this parameter. In contrast to the heatmap, this information was conferred without a necessity of preserving the assay succession.

The best visualization of potential systematic laboratory errors was provided the simplest plot, that is, a visualization showing single data points in the order of assay (Figure 1). This visualization provided a quick overview on the range and outliers. Moreover, the data

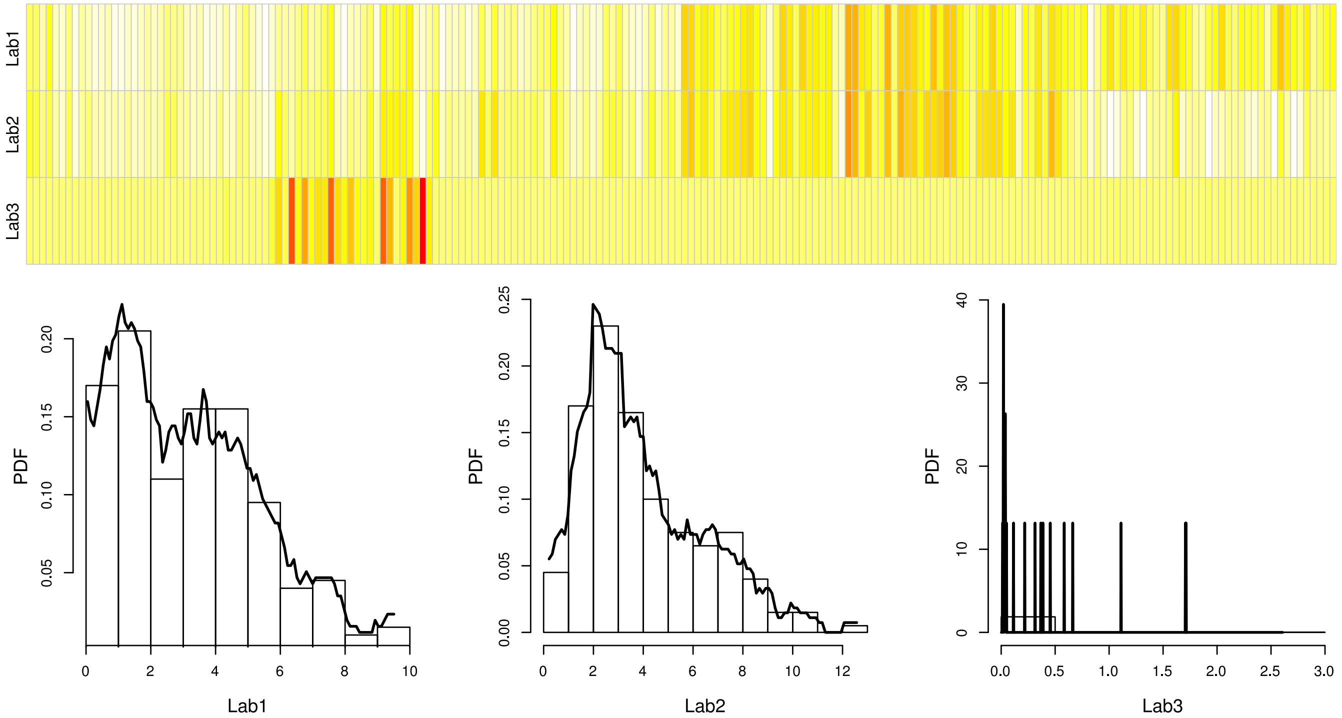

FIGURE 3 Plot of single data and their distribution plasma concentrations of three different biochemical markers (arbitrarily named "Lab1", "Lab2" and "Lab3"). Top: Matrix heatmap showing the data as color-coded from yellow to red, with red indicating higher values. Bottom: Probability density function (PDF) estimated by means of the Pareto density estimation (PDE[3]; bottom, black lines), overlaid on a standard histogram plot of the data

pathologies such as the same values in all probes obtained during a particular assay run were clearly evident. Importantly, the data need to be plotted in the order of assay; if this succession was destroyed (Figure 1 bottom line), the short repetition of the same value in "Lab2" disappeared, and the values above zero in "Lab3" do not anymore hint at a systematic laboratory error but appear simply as noise or outliers.

All visualizations can be easily implemented in free computer software such as in R software package (http://CRAN.R-project.org/[5]). For example, Figure 1 was obtained with the standard R command *plot (LabValues, pch = 20, cex = .1)*, where "labValues" is a vector of assay results for a single parameter in the order of assay, and "*pch*" and "*cex*" provide a suitable symbol and size of the dots in the scatterplot. The complete R script used for creating the present figures is provided in the Appendix S1 of this report. It can be adapted to other environments such as MATLAB or Python according the local standards.

In this short report, it is demonstrated that data visualization of concentration measurements in biological materials is crucial to detect systematic laboratory errors. By contrast, assessing basic statistical parameters is not sufficient. Moreover, a single choice among commonly advised plots might not suffice to detect systematic errors. A dotplot of single data in the order of assay is proposed that provides an overview on the data range, outliers and a particular type of systematic errors where similar values are wrongly measured in all probes. Thus, data science methods[2] providing different visualizations that emphasize different views of the data may enhance the detection of systematic laboratory errors and should be employed to improve the quality check of biochemical laboratory data.

DISCLOSURE

None declared.

ORCID

Jörn Lötsch http://orcid.org/0000-0002-5818-6958

REFERENCES

1. Landis SC, Amara SG, Asadullah K, et al. A call for transparent reporting to optimize the predictive value of preclinical research. *Nature*. 2012;490:187-191.
2. Wickham H, Grolemund G. *R for data science: import, tidy, transform, visualize, and model data*. Beijing, Boston, Farnham, Sebastopol, Tokyo: O'Reilly Media; 2017.
3. Ultsch A, ed. Pareto density estimation: a density estimation for knowledge discovery. innovations in classification, data science, and information systems- Proceedings 27th Annual Conference of the German Classification Society (GfKL). Berlin: Springer; 2003.
4. Ultsch A, Thrun MC, Hansen-Goos O, Lötsch J. Identification of molecular fingerprints in human heat pain thresholds by use of an Interactive mixture model R toolbox (AdaptGauss). *Int J Mol Sci*. 2015;16:25897-25911.
5. R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2008.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Lötsch J. Data visualizations to detect systematic errors in laboratory assay results. *Pharmacol Res Perspect*. 2017;e369. https://doi.org/10.1002/prp2.369