*lancer*: an R package for linearity assessment and visualisation of multiple curves

## Manuscript Type

Application Note

# Abstract

## Summary

Linearity assessment plays a significant role in the validation of instrumentation and experimental procedures. Linearity can be tested by applying several graphical and numerical approaches. Summary statistics of these curves are mostly limited to the Pearson Correlation Coefficient, which is insufficient to fully test the for linearity. In addition, spreadsheet software only allow the analyst to repetitively plot, view and analyse the linearity of curves one at a time, a tedious and time-consuming process. While plotting of many curves at once and calculation of additional summary statistics for assessing linearity of curves can be done using R, implementing this from scratch can be arduous task for the analyst. As such, we created the R package *lancer* so that curves can be analysed efficiently with a few functions. To our knowledge, this is currently not implemented in other software or programming languages. *lancer* furthermore reports the statistical results in tables/spreadsheets and records the plots in a pdf file. In addition, *lancer* can also create interactive trellis plots, displayed as a HTML folder, for exploratory analysis, helping analysts when dealing with large datasets.

## Availability and implementation

*lancer* is available on GitHub <https://github.com/SLINGhub/lancer>. The documentation and tutorials can be accessed from <https://slinghub.github.io/lancer/>

## Supplementary information

Supplementary data are available at *Bioinformatics* online.

## Issue Section

Data and text mining

# Introduction

Linearity assessment is an important performance test of an analytical instrument or method when a linear response is assumed. It is applied in many fields in analytical sciences, such as assay development (Ross and Sweep (2003) and Hsieh and Liu (2008)), calibration/dilution studies (Rodríguez *et al.* (1993) and Sands *et al.* (2021)) and laboratory tests. After analysis, curves representing response values versus concentration are plotted for each analyte of interest. As the accuracy of an analytical method is linked to linearity, a non-linear behaviour must be recognized and addressed accordingly.

Visual inspection of curves is useful but must be accompanied by statistical tests as decision parameters. A commonly used test is the Pearson Correlation Coefficient. However, Sonnergaard (2006) suggests that it is an ineffective standalone numeric parameter for accessing linearity. While Van Loco *et al.* (2002), Sanchez (2021) and Logue and Manandhar (2018) have indicated other metrics for linearity evaluation, these metrics are generally not implemented in most software. Furthermore, using spreadsheet software to individually plot numerous curves is time-consuming. While R (R Core Team (2022)) can plot numerous curve easily, it is hard to implement for a novice analyst.

The R package *lancer* addresses these issues by assisting analysts with useful functions to plot many curves with additional metrics that better describe the curves’ characteristics. [Fig. 1](#fig-visualisation) shows that *lancer* also provides an interactive viewer to group, filter and sort the curves, for examination of problematic cases, such as curves generated by saturated signals.

|  |
| --- |
| Fig. 1. *lancer*’s interactive visualisation of curves |

# Approach

Using response curves in metabolomic/lipidomic studies from Croixmarie *et al.* (2009) and Sands *et al.* (2021) as an example, Supplementary Figure 1 depicts the workflow of *lancer*. The workflow starts with two tables:

* Curve Batch Annotation, describing the curve(s), such as concentration (x-axis) and response curve batches, if any.
* Curve Signal Data, containing response values (y-axis) for each sample and curve.

Using a common column Sample Name, the two tables are merged into one table (Curve Table) via create\_curve\_table.

Next, each curve’s summary statistics are calculated via summarise\_curve\_table. Besides the Pearson Correlation Coefficient, an additional statistical test is the Mandel’s Fitting Test ( in [Equation 1](#eq-mandel-test)) as described by Andrade and Gómez-Carracedo (2013). The test can give evidence that a quadratic model fits better than a linear model, indicating a non-linear curve.

Another statistical test is Percent Residual Accuracy ( in [Equation 2](#eq-pra)) from Logue and Manandhar (2018). Ranging from to , according to which a perfectly linear curve gives a value of .

The software also calculates the concavity of the fitted quadratic model, to identify if the curve is dominantly non-linear at high (concavity ) or low (concavity ) concentrations.

Supplementary Figure 2 gives the summary statistics of three simulated curves as examples, where one curve is linear and two non-linear curves with a plateau at high and low concentrations, denoted as saturation and noise regime curve respectively as defined by Galitzine *et al.* (2018). The corresponding Pearson Correlation Coefficient values (r\_corr) are (, and respectively) even for the non-linear curves. However, non-linear curves are detected by much lower Mandel’s Fitting Test values (mandel\_p\_val) ( and respectively vs for the linear curve) and Percent Residual Accuracy values (pra\_linear) ( and respectively vs for the linear curve).

After calculating the summary statistics for each curve, *lancer* uses the function evaluate\_linearity to group the curves according to the workflows proposed in Supplementary Figure 3. Workflow 1 uses the Pearson Correlation Coefficient and Percent Residual Accuracy to determine if the curve is linear or not, labelled as Good or Poor Linearity respectively. Workflow 2 goes one step further, using Mandel’s Fitting Test and the fitted quadratic model’s concavity to label a non-linear curve as Saturation or Noise Regime, if it plateaus at high or low concentrations respectively. Non-linear curves that do not follow these trends are then labelled as Poor Linearity.

A benchmark workflow using only Pearson Correlation Coefficient value of is compared with Workflow 2 on simulated data sets of 200 linear curves (labelled as Linear), non-linear curves that plateau at high and low concentrations (labelled as Saturation and Noise Regime respectively). Supplementary Figure 4 shows that Workflow 2 better identifies the saturation and noise regime curves than the benchmark workflow. While Workflow 2 correctly classifies a lower number of linear curves than the benchmark workflow, its percentage of correctly classified linear curves, 181/200 (90.5%), is high. See <https://lancer-simulation.netlify.app> for report details. While the threshold values of Pearson Correlation Coefficient and Percent Residual Accuracy are based on the interpretation of Y. H. Chan (2003) and Logue and Manandhar (2018), they remain subjective and arbitrary. Nevertheless, *lancer* allows optimization of these threshold values according to the analyst’s preference.

Although *lancer* can export the results in Excel or pdf, an interactive interface can generate a better overview. [Fig. 1](#fig-visualisation) shows a HTML folder, exported by *lancer*. Clicking on the index.html file inside the folder opens an interactive plot that can group, filter and sort curves. This allows room for exploratory data analysis, such as identifying curves with linearity issues or understanding the consequence of changing the Pearson Correlation Coefficient threshold to another value. Such information is hard to achieve with static plots alone. An interactive viewer example created by *lancer* can be found at <https://lancer-interactive-example.netlify.app>

# Conclusion

Linearity is one of the most important parameter of an analytical method to be evaluated. Our R package, *lancer*, can access linearity efficiently using functions that quickly plot and report curve summary statistics, which better describe the shape of the curve. It also provides an interactive plot for exploratory data analysis. *lancer* is available on GitHub <https://github.com/SLINGhub/lancer> while the documentation and tutorials can be accessed from <https://slinghub.github.io/lancer>.

# Acknowledgements

These should be included at the end of the text and not in footnotes. Please ensure you acknowledge all sources of funding, see funding section below.

Details of all funding sources for the work in question should be given in a separate section entitled ‘Funding’. This should appear before the ‘Acknowledgements’ section.

# Funding

The following rules should be followed:

* The sentence should begin: ‘This work was supported by …’ -
* The full official funding agency name should be given, i.e. ‘National Institutes of Health’, not ‘NIH’ (full RIN-approved list of UK funding agencies)
* Grant numbers should be given in brackets as follows: ‘[grant number xxxx]’
* Multiple grant numbers should be separated by a comma as follows: ‘[grant numbers xxxx, yyyy]’
* Agencies should be separated by a semi-colon (plus ‘and’ before the last funding agency)
* Where individuals need to be specified for certain sources of funding the following text should be added after the relevant agency or grant number ‘to [author initials]’.

An example is given here: ‘This work was supported by the National Institutes of Health [AA123456 to C.S., BB765432 to M.H.]; and the Alcohol & Education Research Council [hfygr667789].’

Oxford Journals will deposit all NIH-funded articles in PubMed Central. See Depositing articles in repositories – information for authors for details. Authors must ensure that manuscripts are clearly indicated as NIH-funded using the guidelines above.

# References

Andrade,J.M. and Gómez-Carracedo,M.P. (2013) [Notes on the use of Mandel’s test to check for nonlinearity in laboratory calibrations](https://doi.org/10.1039/c2ay26400e). *Analytical Methods*, **5**, 1145.

Croixmarie,V. *et al.* (2009) [Integrated comparison of drug-related and drug-induced ultra performance liquid chromatography/mass spectrometry metabonomic profiles using human hepatocyte cultures](https://doi.org/10.1021/ac900333e). *Analytical Chemistry*, **81**, 6061–6069.

Galitzine,C. *et al.* (2018) [Nonlinear Regression Improves Accuracy of Characterization of Multiplexed Mass Spectrometric Assays](https://doi.org/10.1074/mcp.RA117.000322). *Molecular & Cellular Proteomics*, **17**, 913–924.

Hsieh,E. and Liu,J. (2008) [On Statistical Evaluation of the Linearity in Assay Validation](https://doi.org/10.1080/10543400802071378). *Journal of Biopharmaceutical Statistics*, **18**, 677–690.

Logue,B.A. and Manandhar,E. (2018) [Percent residual accuracy for quantifying goodness-of-fit of linear calibration curves](https://doi.org/10.1016/j.talanta.2018.07.046). *Talanta*, **189**, 527–533.

R Core Team (2022) [R: A language and environment for statistical computing](https://www.R-project.org/) R Foundation for Statistical Computing, Vienna, Austria.

Rodríguez,L.C. *et al.* (1993) [Estimation of Performance Characteristics of an Analytical Method Using the Data Set Of The Calibration Experiment](https://doi.org/10.1080/00032719308019900). *Analytical Letters*, **26**, 1243–1258.

Ross,H.A. and Sweep,C.G.J. (2003) [An improved procedure for testing for assay linearity](https://doi.org/10.1258/000456303321016204). *Annals of Clinical Biochemistry: International Journal of Laboratory Medicine*, **40**, 75–78.

Sanchez,J.M. (2021) [The inadequate use of the determination coefficient in analytical calibrations: How other parameters can assess the goodness-of-fit more adequately](https://doi.org/10.1002/jssc.202100555). *Journal of Separation Science*, **44**, 4431–4441.

Sands,C.J. *et al.* (2021) [Representing the metabolome with high fidelity: Range and response as quality control factors in LC-MS-based global profiling](https://doi.org/10.1021/acs.analchem.0c03848). *Analytical Chemistry*, **93**, 1924–1933.

Sonnergaard,J.M. (2006) [On the misinterpretation of the correlation coefficient in pharmaceutical sciences](https://doi.org/10.1016/j.ijpharm.2006.06.001). *International Journal of Pharmaceutics*, **321**, 12–17.

Van Loco,J. *et al.* (2002) [Linearity of calibration curves: Use and misuse of the correlation coefficient](https://doi.org/10.1007/s00769-002-0487-6). *Accreditation and Quality Assurance*, **7**, 281–285.

Y. H. Chan (2003) [Biostatistics 104: Correlational analysis](http://www.smj.org.sg/article/biostatistics-104-correlational-analysis). *Singapore Medical Journal*, **44**, 614–619.