*DCVtestkit*: an R package for analysis of quality control dilution curves

## Manuscript Type

Application Note

# Abstract

## Summary

In metabolomic and lipidomic workflows, a linear response in dilution series generated from pooled quality control (QC) samples is used to assess the quality of the measurement of molecules measured before further analysis. Most of the currently used software only allow the analyst to repetitively plot, view and analyse the dilution curves one at a time, a tedious and time-consuming process. In addition, summary statistics of dilution curves are limited to the Pearson Correlation Coefficient and/or Coefficient Of Determination , which are insufficient to fully understand the shape of the dilution curves. DCVtestkit aims to provide additional summary statistics for dilution curves, taken from previous publications but which are not implemented in the current software tools. It also helps to reduce the analyst’s workload by analysing many dilution curves automatically, reporting the statistical results in Excel and recording the dilution plots in a pdf file. In addition, it can also create an interactive trellis displayed as a HTML folder for more exploratory analyses.

## Availability and implementation

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

## Supplementary information

Supplementary data are available at *Bioinformatics* online.

## Issue Section

Data and text mining

# Introduction

Dilution curves have been used in many areas of biomedical and pharmaceutical sciences. In the field of metabolomics and lipidomics, dilution curves for each molecule measured, usually in a quality control sample generated by pooling equal aliquots from each study sample, are plotted as proposed by Croixmarie *et al.* (2009) to check if the sample matrix is the same among all samples. Molecules that do not show a linear dilution curve can be rejected before further statistical analysis, in spite of their relevance to the study, as their quantitation in different sample groups might be compromised. A cut off Pearson Correlation Coefficient value of is commonly used to balance the risk of accepting or removing signals related to relevant molecules. As such, many laboratories use commercial software to plot these dilution curves and report their corresponding Pearson Correlation Coefficient value. These include either mass spectrometry vendors’ software or, more in general, programs such as Excel and GraphPad Prism. Nevertheless, these tools have limitations.

While the Pearson Correlation Coefficient and/or Coefficient Of Determination are popular indicators of how well the dilution points fits a linear line, Sonnergaard (2006) estimated that these are not effective standalone numeric parameter to estimate linearity. Using a modified version of Anscombe’s quartet, the paper showed that a high coefficient of determination does not reflect well a curve relationship. Both Croixmarie *et al.* (2009) and Sands *et al.* (2021) support this view, by indicating that an obvious non-linear response at the high and/or low ends of the quality control sample-based dilution curve can still yield a high correlation coefficient. This is in agreement with Sanchez (2021) findings of several saturated calibration curves with values greater than .

Fortunately, researchers who have studied deeply model validation and misspecification have provided alternative graphical and statistical methods to improve the reliability of linearity assessments. For example, Van Loco *et al.* (2002) highlighted the presence of alternative statistics, such as the Lack-of-fit test and quality coefficient, that can be used to make better judgement on the shape of the dilution curve. While the progress in finding better metrics and algorithms to evaluate the linearity of dilution curves is commendable, these findings are rarely implemented in more general-purpose software.

In addition, with the advancement of the technology used for molecular identification coupled with a high emphasis on high throughput analysis and automation, today’s targeted metabolomics/lipidomics can include more than eight hundred analytes measured by a single method. This number can also increase further in an untargeted approach. In these cases, using general-purpose programs to plot dilution curves for every single molecule may be tedious. Moreover, the reports generated by these programs are static dilution plots coupled with a statistical summary table. Having the analyst to look at five hundred dilution curves individually is time consuming and inefficient for data evaluation.

To address the issues mentioned above, we have created an R package, *DCVtestkit*, to assist analysts, not only to plot these dilution curves for many molecules easily but also providing additional statistical results, other than the Pearson Correlation Coefficient or coefficient of determination , for these curves. These alternative statistical results may be helpful in better describing the shape of the dilution curve. The R package also provides several ways of reporting the dilution data, including an interactive html file. This allows the analyst to group, filter and sort the dilution plots, to only look at the relevant ones, such as the saturated dilution curves, and identify problematic molecules quickly.

# General Workflow

Here, we introduce the workflow used by our R package, *DCVtestkit* in [Figure 1](#fig-DCVtestkit-Workflow).

|  |
| --- |
|  |

Figure 1: *DCVtestkit* Workflow

Beginning with the Input Data phase, the workflow starts with two tables: a table (Transition Signal Data) containing transition signals (y-axis for dilution curve) for each sample and a table (Dilution Annotation) containing dilution curve related information for each sample, such as concentration (x-axis for dilution curve) and dilution batches. As, the two tables share the same sample list, it can be merged together into one table (Dilution Table) using the function create\_dilution\_table.

During the Calculation stage of the workflow, the summary statistics are calculated using the function summarise\_dilution\_table for each dilution curve corresponding to a dilution batch and transition. Optionally, analysts can use the function evaluate\_linearity to group the dilution curves based on the summary statistics. These tabular results can be reported as an Excel sheet using the function write\_summary\_excel.

With the summary statistics results, the workflow can proceed to the Dilution Curve Plotting phase in which static and interactive dilution plots can be created using add\_ggplot\_panel and add\_plotly\_panel respectively. Static plots are output as a pdf file while the interactive plots are output as a HTML folder containing several folders and one index.html file. Analysts can view the interactive plots by double clicking on the index.html file.

# Statistical Summary for Dilution Curves

[Figure 2](#fig-curves) gives a plot of three manually curated curves together with their corresponding summary statistics. The characteristic of three curves are as follows: One is a linear curve. Another is having a plateau at higher concentrations, also known as a saturated curve. The last example is a curve with a plateau at lower concentrations or experiencing a limit of detection (denoted as LOD curves in this paper). It can be observed that the corresponding Pearson Correlation Coefficient values, denoted as r\_corr, are (, and respectively) even for the curves that are non-linear. These results are in agreement with Van Loco *et al.* (2002) as well as Logue and Manandhar (2018) arguments that the Pearson Correlation Coefficient is a poor descriptor of non-linear calibration curves.

To improve the description of the goodness-of-fit for linear dilution curves, a few additional parameters, on top of the popular Pearson Correlation Coefficient , are added in the R package *DCVtestkit*. The first additional parameters are statistical results from Mandel’s Fitting Test (as reported in Andrade and Gómez-Carracedo (2013)) in which a low value gives sufficient evidence that a quadratic model is a better fit than a linear model, indicating the curve may not be linear. Another parameter calculated is the Percent Residual Accuracy (as reported in Logue and Manandhar (2018)) which ranges from to . If the curve is linear, the value should be close to . Finally, the software also fits a quadratic model on the dilution curve and calculates the concavity, trying to verify if the dilution curve is predominantly non-linear at high (concavity ) or low (concavity ) concentrations.

Returning to [Figure 2](#fig-curves), both the saturated and LOD curves give a much lower Mandel’s Fitting Test values, denoted as mandel\_p\_val, ( and respectively vs ). Correspondingly, ranging from to , the Percent Residual Accuracy values, denoted as pra\_linear, are much lower in the saturated and LOD curves compared to the linear curve ( and respectively vs ).

|  |
| --- |
| Figure 2: Plots and statistical results of three curves: Linear curve, Saturated curve and a curve with Limit Of Detection (LOD) |

# Dilution Curve Grouping

*DCVtestkit* uses the function evaluate\_linearity to group the dilution curves. At this moment, two workflows are proposed. Workflow uses both the Pearson Correlation Coefficient and Percent Residual Accuracy to determine if the dilution curve is linear (labelled as Good Linearity) or not (labelled as Poor Linearity). Workflow goes one step further by using the Mandel’s Fitting Test and the concavity of the fitted quadratic model to check if the non-linear dilution curve plateaus at low concentrations (labelling them as limit of detection) or high concentrations (labelling them as saturation). Non-linear dilution curves that do not follow these trends are then labelled as Poor Linearity.

|  |
| --- |
|  |

Figure 3: Dilution Curve Grouping Workflows

Workflow was tested on 200 simulated linear curves (labelled as Linear), 200 curves that plateau at low concentrations (labelled as Limit of Detection) and 200 curves that plateau at high concentrations (labelled as Saturated). A benchmark workflow using only Pearson Correlation Coefficient with a cut off value of is used to compare with Workflow .

|  |
| --- |
|  |

Figure 4: Simulation Results

[Figure 4](#fig-Simulation-Result) showed that Workflow is able to identify the saturated and limit of detection curves better than the benchmark workflow. On the other hand, while Workflow is able to identify less linear curves than the benchmark workflow, the results are comparable score of 181/200 (90.5%)

While the proposed workflows’ threshold values of Pearson Correlation Coefficient and Percent Residual Accuracy are based on the interpretation given by Y. H. Chan (2003) and Logue and Manandhar (2018), respectively, these cut off values remain subjective and arbitrary. Moreover, it is also possible for the dilution curve to exhibit a sigmoidal relationship, or non-linear, at both low and high concentrations, which the proposed workflow would be unable to identify.

Nevertheless, while the proposed workflow is not a perfect solution, *DCVtestkit* allows the analyst to change and optimise the threshold values, to adjust the determinants of linearity and saturation of dilution curves according to the user’s needs. More information on how to perform such customisation can be found in the program documentation titled: [“Customise Linearity Evaluation Workflow”](https://slinghub.github.io/DCVtestkit/articles/lineval-customization.html). Furthermore, it is also possible to output these statistical summaries in Excel without the classification results.

# Results Distribution

Lastly, *DCVtestkit* can export the statistical results of each dilution curve in Excel. Furthermore, it can report the list of dilution plots as static charts with the statistical summary results included in a single pdf file.

As some may find the final pdf file too complex for further interpretation, the R package also provides an interactive viewer, as a HTML folder, to look at dilution plots in a trellis display framework, as introduced in Becker *et al.* (1996). Clicking on the index.html file inside the folder will open the interactive viewer that allows the user to group, filter and sort these trellis charts in a meaningful way, based on Hafen *et al.* (2013)’s definition of “cognostics” or computer guiding diagnostics. This allows the user to do exploratory data analysis, such as identifying groups of molecules with linearity issues or finding out the effects of changing the Pearson Correlation Coefficient to another value. Such tasks can be challenging and time consuming if only the Excel and pdf files full of static plots were provided.

|  |
| --- |
|  |

Figure 5: Viewing Several Interactive Dilution Curves

An example of an interactive viewer created by *DCVtestkit* can be accessed from <https://dcvtestkit-interactive-example.netlify.app/>. Such interactive visualisation was also utilised by Hafen *et al.* (2019) in the exploration of proteomics, metabolomics and lipidomics data in a multi-omics approach. The HTML folder can be compressed as a zip file and shared as a pdf or Excel file. This interactive viewer can also be displayed as a report created from Quarto (https://quarto.org/), an open-source scientific and technical publishing system. The report can be viewed at <https://dcvtestkit-quarto-example.netlify.app/> R script and result outputs can be found in the Supplementary Material or by accessing this GitHub page <https://github.com/JauntyJJS/DCVtestkit_Output>

An example of an interactive viewer created by *DCVtestkit* can be accessed from <https://dcvtestkit-interactive-example.netlify.app/>. Such interactive visualisation was also utilised by Hafen *et al.* (2019) in the exploration of proteomics, metabolomics, and lipidomics data in a multi-omics fashion. The HTML folder can be compressed as a zip file and distributed to others like a pdf or Excel file. This interactive viewer can also be displayed in as a report created from Quarto (<https://quarto.org/>), an open-source scientific and technical publishing system. The report can be viewed at <https://dcvtestkit-quarto-example.netlify.app/>. R script and result outputs can be found in the Supplementary Material or by accessing this GitHub page <https://github.com/JauntyJJS/DCVtestkit_Output>.

# Conclusion

A linear response in quality control dilution curves is one of the requirements in metabolomics and lipidomics to assess the suitability of the molecules measured for further analysis. Despite its importance in testing data quality, the shortage of software tools to analyse dilution curves effectively and efficiently prevents researchers from fully analysing data. An R package called *DCVtestkit* is created to rectify this issue by allowing many dilution curves to be analysed quickly and reliably. This is done by the use of alternative statistics, other than the Pearson Correlation Coefficient, to better understand the linearity of the dilution curves, as well as reporting of statistical results and dilution plots that can be distributed to others. It is also able to provide an interactive trellis plot for exploratory data analysis. The R package is available on GitHub <https://github.com/SLINGhub/DCVtestkit> while the documentation and tutorials can be accessed from <https://slinghub.github.io/DCVtestkit/>. As *DCVtestkit* is an opensource software, it is accessible to collaboration works which can help improve its usability and performance in the future.

# 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.

Becker,R.A. *et al.* (1996) [The Visual Design and Control of Trellis Display](https://doi.org/10.1080/10618600.1996.10474701). *Journal of Computational and Graphical Statistics*, **5**, 123–155.

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.

Hafen,R. *et al.* (2013) [2013 IEEE symposium on large-scale data analysis and visualization (LDAV)](https://doi.org/10.1109/LDAV.2013.6675164). IEEE, Atlanta, GA, USA, pp. 105–112.

Hafen,R. *et al.* (2019) [MODE: The multi-omics data exploration platform phase i final technical report](https://doi.org/10.2172/1630300).

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.

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.