Data in Brief 14 (2017) 291–294

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Simulation data for an estimation of the maximum theoretical value and confidence interval for the correlation coefficient

Paolo Rocco *, Francesco Cilurzo, Paola Minghetti, Giulio Vistoli, Alessandro Pedretti

Department of Pharmaceutical Sciences, Università degli Studi di Milano, via G. Colombo, 71, I-20133 Milan, Italy

ARTICLE INFO

Article history:
Received 13 June 2017
Received in revised form
14 July 2017
Accepted 18 July 2017
Available online 23 July 2017

Keywords:
Numerical simulation
Fisher transform
Skin permeability
Maximum theoretical correlation coefficient
Confidence interval around correlation coeffcient

ABSTRACT

The data presented in this article are related to the article titled "Molecular Dynamics as a tool for in silico screening of skin permeability" (Rocco et al., 2017) [1]. Knowledge of the confidence interval and maximum theoretical value of the correlation coefficient r can prove useful to estimate the reliability of developed predictive models, in particular when there is great variability in compiled experimental datasets. In this Data in Brief article, data from purposely designed numerical simulations are presented to show how much the maximum r value is worsened by increasing the data uncertainty. The corresponding confidence interval of r is determined by using the Fisher $r \rightarrow Z$ transform.

Specifications Table

Subject area	*Chemistry*
More specific subject area	*Chemometrics*
Type of data	*Table*

DOI of original article: http://dx.doi.org/10.1016/j.ejps.2017.06.020
* Corresponding author.
E-mail address: paolo.rocco@unimi.it (P. Rocco).

http://dx.doi.org/10.1016/j.dib.2017.07.045

How data was acquired	*Numerical simulation*
Data format	*Raw, Analysed*
Experimental factors	*Not applicable*
Experimental features	*Reduced set (Reduced_ser.pdf) modified by randomly generated errors.*
Data source location	*Not applicable*
Data accessibility	*Data is contained in this article and files: Reduced_set.pdf, simulation_data.xlsx*

Value of the data

- When there is great variability in a compiled experimental dataset, considerations on the confidence interval for the correlation coefficient r and on the maximum theoretical value achievable for r can offer hints as to what to expect from a predictive model based on that set.
- Numerical simulations used to generate a dataset of arbitrary average uncertainty and to estimate a confidence interval around the correlation coefficient r and its maximum theoretical value are easily applicable to all experimental datasets
- The here proposed data can be easily utilized to derive the range of r that can be pursued when the variability of a given dataset is known
- Along with well-known statistical parameters (such as r, r^2, q^2, F, SE, etc), the here proposed confidence interval of r can become a meaningful parameter to better evaluate the reliability of a given model and to understand whether there is still room for statistical improvements.

1. Data

Data presented here represent maximum theoretical average values and confidence interval for the correlation coefficient r and the determination coefficient r^2 as obtained through numerical simulation (Table 1). The values of r and r^2 correspond to different simulated levels of random error (ε) in the experimental data set.

Original data, on which data in Table 1 are based, are contained in the files *Reduced_set.pdf* and *simulation_data.xlsx. Reduced_set.pdf* contains a set of 80 permeability coefficients k_p [1] assembled as the intersection of Flynn's set [2] and the Fully Validated data set [3]. The file *simulation_data.xlsx* contains data from the numerical simulation described below.

2. Experimental design, materials and methods

Given a set of experimental data, y_i, we can assume that a perfect estimator ϕ for the set is known (in [1], y_i correspond to pk_p values). ϕ is a mathematical function, which correlates a set of variables $\{x_{ij}\}$ with the experimental value y_i, where x_{ij} represents the j-th molecular property of the i-th molecule (Eq. (1)).

Table 1
Maximum theoretical average values and 95% confidence interval for r and r^2 from numerical simulation, at different simulated levels of random error (ε) in the experimental data set.

$\varepsilon \rightarrow$	0.10	0.15	0.20	0.25	0.30
Maximum average r	0.97	0.94	0.90	0.86	0.81
Confidence interval around r	0.96–0.98	0.91–0.96	0.85–0.94	0.78–0.91	0.70–0.89
Maximum average r^2	0.95	0.88	0.81	0.74	0.66
Confidence interval around r^2	0.92–0.97	0.83–0.93	0.72–0.88	0.60–0.84	0.50–0.79

The correlation, based on a perfect estimator, yields a correlation coefficient r = 1.

$$y_i = \phi(x_{ij}) \quad i = 1, 2\ldots n,\ j = 1, 2\ldots m \tag{1}$$

For every y_i, we introduce an error $\varepsilon \cdot c_{ik} \cdot y_i$, where $\{c_{ik}\}$ is a set of normally distributed pseudo-random numbers with zero average and unitary standard deviation (obtained by applying the Box-Muller transform [4] to a set of a linearly distributed random numbers); ε corresponds to the standard deviation of the errors, normalized by y_i.

For the k-th simulation, Eq. (1) becomes Eq. (2):

$$y_{ik} = y_i + \varepsilon c_{ik}\, y_i = \phi_k(x_{ij}) \quad i = 1, 2\ldots n,\ j = 1, 2\ldots m, k = 1, 2\ldots l \tag{2}$$

Since ϕ_k, by definition, is a perfect estimator, the values of r obtained for Eq. (2) in the simulation are the maximum theoretical correlation coefficients achievable given the uncertainty introduced (ε).

For different values of ε, the numerical simulation is repeated 99 times (l=99) obtaining 99 correlation coefficients r_k (*simulation_data.xlsx*). Table 1 shows how much r and r^2 worsen when ε increases and confirms that the formula: maximum $r^2 \cong (1-\varepsilon)$ is an approximate but yet reasonable way to estimate the worsening effect of ε.

As for the confidence interval around *r*, it can be estimated, for each value of *r*, by using Fisher $r \rightarrow Z$ transform [5]:

$$Z = \frac{1}{2}[\ln(1 + r) - \ln(1 - r)] \tag{3}$$

We apply Fisher $r \rightarrow Z$ transform to the r_k values, obtaining 99 Z_k values. Unlike *r*, *Z* tends to a normal distribution as the number of data becomes large. Therefore, the standard deviation S_z can be calculated by Eq. (4):

$$S_z = \sqrt{\frac{1}{98}\sum_{k=1}^{99}(Z_k - \bar{Z})^2} \tag{4}$$

The 95% confidence interval around *Z* is then calculated as $(\bar{Z} - 1.96 \cdot S_z,\ \bar{Z} + 1.96 \cdot S_z)$, and the 95% confidence interval around r is obtained from it, through the reverse transform (Eq. (5)):

$$r = \frac{e^{2Z} - 1}{e^{2Z} + 1} \tag{5}$$

The confidence intervals around r and r^2 for different values of ε are shown in Table 1.

Transparency document. Supplementary material

Transparency data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2017.07.045.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2017.07.045.

References

[1] P. Rocco, F. Cilurzo, P. Minghetti, G. Vistoli, A. Pedretti, Molecular dynamics as a tool for in silico screening of skin permeability, Eur. J. Pharm. Sci. 106 (2017) 328–335. http://dx.doi.org/10.1016/j.ejps.2017.06.020 (2017 Jun 13).

[2] G.L. Flynn, Physicochemical determinants of skin absorption, in: T.R. Garrity (Ed.), Principles of Route to Route Extrapolation for Risk Assessment, Elsevier, New York, 1990, pp. 93–127.

[3] B.E. Vecchia, A.L. Bunge, Evaluating the transdermal permeability of chemicals, in: R.H. Guy, H.J. (Eds.), Transdermal Drug Delivery, CRC Press, New York, 2003, pp. 38–39.
[4] G.E.P. Box, E. Mervin, A. Muller, Note on the generation of random normal deviates, Ann. Math. Stat. 29 (1958) 610–611.
[5] R.A. Fisher, Statistical Methods for Research Workers, Fifth Edition, Oliver and Boyd, Edinburgh (1934) 160–197.