# Reproducible, flexible and high throughput data extraction from primary literature: The **metaDigitise** R package

## Supplementary Materials

Joel L. Pick, Shinichi Nakagawa & Daniel W.A. Noble

February 27, 2018

## S1 Comparison with existing software

| Function | metaDigitise | GraphClick[1] | DataThief[2] | DigitizeIt[3] | WebPlotDigitizer[4] | metagear[5] | digitize[6] |
|---|---|---|---|---|---|---|---|
| Scatterplots | ✓ | ✓ | ✓ | ✓ | ✓ | ✓[7] | ✓ |
| Mean/error plots | ✓ | ✓ | ✓ | × | × | ✓[7] | × |
| Boxplots | ✓ | × | × | × | × | × | × |
| Histograms | ✓ | × | × | × | ✓[7] | × | × |
| Graph rotation[8] | ✓ | ✓ | ✓ | ✓ | ✓ | × | × |
| Groups | ✓ | ✓ | × | ✓ | ✓ | × | × |
| Entry of metadata | ✓ | × | × | × | × | × | × |
| Summarising data | ✓ | × | × | × | × | × | × |
| Multiple image processing | ✓ | × | × | × | × | × | × |
| Reproducable[9] | ✓ | ✓ | ✓ | × | ✓ | × | × |
| Automated point detection | × | ✓ | × | ✓ | ✓ | ✓ | × |
| Line extraction | × | ✓ | ✓ | ✓ | ✓ | × | × |
| Zoom | × | ✓ | ✓ | ✓ | ✓ | × | × |
| Log axis | × | ✓ | ✓ | ✓ | ✓ | × | × |
| Dates | × | × | ✓ | × | ✓ | × | × |
| Asymmetric error bars | × | × | ✓ | × | × | × | × |
| Freeware | ✓[10] | ✓[11] | ✓[11] | ×[11] | ✓[11] | ✓[10] | ✓[10] |

[1] Arizona-Software (2008) [2] Tummers (2006) [3] Bormann (2012) [4] Rohatgi (2017) [5] Lajeunesse (2016) [6] Poisot (2011)

[7] Only automated, no manual extraction.

[8] Or handles rotated graphs.

[9] Allows saving, re-plotting and editing of data extraction.

[10] R package.

[11] Standalone software.

Table S1: Comparison of functionality between different digitisation softwares.

# S2 Derivation of mean, standard deviation and sample size from different plot types

## S2.1 Mean/Error Plots

The standard deviation is calculated depending on the type of error presented. The user can choose from standard deviation (SD, $\sigma$), standard error (SE) or 95% confidence intervals (CI95). Standard deviation is calculated from standard error as

$$\sigma = SE\sqrt{n} \tag{S1}$$

and from 95% confidence intervals as

$$\sigma = \frac{CI}{1.96}\sqrt{n} \tag{S2}$$

## S2.2 Box Plots

The mean ($\mu$) and SD are calculated sing the maximum ($b$), upper quartile ($q_3$), median ($m$), lower quartile ($q_1$) and minimum ($a$) as

$$\mu = \frac{(n+3)(a+b) + 2(n-1)(q_1 + m + q_3)}{8n} \tag{S3}$$

following Bland (2015) and

$$\sigma = \frac{b-a}{4\Phi^{-1}(\frac{n-0.375}{n+0.25})} + \frac{q_3 - q_1}{4\Phi^{-1}(\frac{0.75n-0.125}{n+0.25})} \tag{S4}$$

where $\Phi^{-1}(z)$ is the upper zth percentile of the standard normal distribution, following Wan et al. (2014).

## S2.3 Histograms

For each bar, the user click two point (the top of the bar). Using these points, a midpoint (m; mean x coordinates) and a frequency (f; mean y coordinates, rounded to the nearest integer) is calculated for each bar. The sample size, mean and SD are calculated as:

$$n = \sum_{i=1}^{n} f_i \tag{S5}$$

$$\mu = \frac{\sum_{i=1}^{n} m_i f_i}{n} \tag{S6}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (m_i f_i - \mu f_i)^2}{n-1}} \tag{S7}$$

# S3 Tutorial

# References


Arizona-Software (2008) *GraphClick Software, Version 3.0.*

Bland, M. (2015) Estimating Mean and Standard Deviation from the Sample Size, Three Quartiles, Minimum, and Maximum. *International Journal of Statistics in Medical Research*, **4**, 57–64.

Bormann, I. (2012) *DigitizeIt Software, Version 2.0.* Braunschweig, Germany.

Lajeunesse, M.J. (2016) Facilitating systematic reviews, data extraction, and meta-analysis with the metagear package for R. *Methods in Ecology and Evolution*, **7**, 323–330.

Poisot, T. (2011) The digitize package: extracting numerical data from scatterplots. *The R Journal*, **3**, 25–26.

Rohatgi, A. (2017) *WebPlotDigitizer Software, Version 4.0.* Austin, Texas, USA.

Tummers, B. (2006) *DataThief Software, Version 3.0.*

Wan, X., Wang, W., Liu, J. & Tong, T. (2014) Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC Medical Research Methodology*, **14**, 135.