Extracting data from graphs: A case-study on animal research with implications for meta-analyses

Stevie Van der Mierden1 | Loukia Maria Spineli2 | Steven R. Talbot1 | Christina Yiannakou1 | Eva Zentrich1 | Nora Weegh1 | Birgitta Struve1 | Talke Friederike Zur Brügge1 | André Bleich1 | Cathalijn H.C. Leenaars1

1Institute for Laboratory Animal Science, Hannover Medical School, Hannover, Germany
2Midwifery Research and Education Unit, Hannover Medical School, Hannover, Germany

Correspondence Cathalijn Leenaars, Institute for Laboratory Animal Science, Hannover Medical School, Hannover, Germany. Email: leenaars.cathalijn@mh-hannover.de

Funding information Deutsche Forschungsgemeinschaft, Grant/Award Numbers: 2591, BI 953/11-1, TO 542/6-1, JI 276/1-1, LE 490 2356/5-1, GA427/12-1; Federal State of Lower Saxony (R2N)

Abstract
Systematic reviews with meta-analyses are powerful tools that can answer research questions based on data from published studies. Ideally, all relevant data is directly available in the text or tables, but often it is only presented in graphs. In those cases, the data can be extracted from graphs, but this potentially introduces errors. Here, we investigate to what extent the extracted outcome and error values differ from the original data and if these differences could affect the results of a meta-analysis. Six extractors extracted 36 outcome values and corresponding errors from 22 articles. Differences between extractors were compared using overall concordance correlation coefficients (OCCC), differences between the original and extracted data were compared using concordance correlation coefficients (CCC). To test the possible influence on meta-analyses, random-effects meta-analyses on mean difference comparing original and extracted data were performed. The OCCCs and CCCs were high for both outcome values and errors, CCCs were >0.99 for the outcome and >0.92 for errors. The meta-analyses showed that the overall effect on outcome was very small (median: 0.025, interquartile range: 0.016–0.046). Therefore, data extraction from graphs is a good method to harvest data if it is not provided in the text or tables, and the original authors cannot provide the data.

KEYWORDS
concordance correlation coefficient, data extraction, meta-analysis, systematic review

1 INTRODUCTION
Systematic reviews (SR) are powerful tools; they aim to answer a specific research question by identifying, appraising and combining all available relevant evidence objectively.1 When used appropriately, SRs can, inter alia, increase the precision of found intervention effects, identify areas with a lack of adequate evidence, and generate new research questions.2 SRs differ from narrative reviews by having a structured approach using multiple...
well-described phases; searching papers, selecting papers, extracting data, appraising papers, and synthesizing evidence. Of these phases, data extraction is one of the most sensitive to reviewer errors. While extracting data, one depends on how this data is presented in the original paper. Ideally, all relevant data is reported completely in the text or tables, but frequently it is only reported graphically. In those cases, it is the best practice to request the data from the original authors. Unfortunately, it is not always possible to acquire this information in a reasonable time or at all, for example, when authors move to different institutes. In those cases, the data can be extracted directly from the graphs.

Data extraction from graphs is subject to human error, and particularly challenging if the graphs are of low quality. As a result, the data used in the systematic review could be incorrect, and this could potentially influence the results of meta-analysis, and consequently the conclusion of the SR. Previous research has investigated the quality of data extracted from graphs. A paper of Kadic et al. compared the data from 62 graphs extracted by two extractors. They found high intraclass correlation coefficients (ICC) between the extracted and the original data of 0.924 and 0.926. Burda et al. had 12 reviewers extract data from two graphs. They found a high ICC between the different extractors (>0.95), but the percentage difference between the original and extracted data varied from 0.29% to 8.92%. Crandom et al. compared data from twenty artificially created graphs extracted by ten reviewers. They found that a mean of 69.91% of all extracted data points were of sufficient accuracy compared to the original values, with the percentages per graph ranging from 0% to 100%.

The literature thus provides some indications of the goodness of extracting data from graphs, but these studies have their limitations concerning the number of extractors, number of graphs, or ecological validity (using artificially created graphs). Furthermore, the preceding studies did not focus on dispersion measures related with the effect size, such as standard errors or confidence intervals, with only Burda et al. providing data on error bars from one of their two graphs. Dispersion measures are important for meta-analyses as the weight of the study used for calculating the overall effect estimate depends, partly, on the variance of the study. Therefore, we wanted to investigate data extraction from multiple published graphs using multiple reviewers, extracting both outcome values and the corresponding errors (hereafter, standard errors). Although the process of extracting the standard error is often similar to extracting outcome values, some factors that might affect the extracted standard errors differently. For example, standard errors are smaller compared to the outcome values and thus extraction may lead more easily to a mistake, it also means that differences between the extracted and original values are proportionally larger for standard errors compared to the outcome values, which may have a bigger influence on meta-analysis than expected based on extracted outcome values. To the best of our knowledge, this is the first study that also focuses on extracting standard errors from graphs and investigates the influence of using extracted instead of the original data on systematic reviews. As far as we know this is also the first study investigating data extraction from graphs in the field of animal sciences.

With this study, we aim to answer the following questions: how comparable the extracted outcome data and corresponding standard errors are, first, among different extractors, and second, with the original data; and third, what are the potential consequences of the differences between the extracted and original data on the meta-analysis results.

2 METHODS

2.1 Graph selection

The graphs came from a previous review that dealt with corticosterone concentrations in mice. Baseline corticosterone concentrations plus the standard error of the mean (SEM) were only available in graphs in several of the included publications. For these publications we contacted the authors to request the original data; 22 authors responded and provided the requested data. In the end, we had data from 28 figures. Twenty-six of the figures were bar plots, and two were line plots with circle symbols for the outcome values. As some figures presented more than one relevant corticosterone concentration (e.g., for male and female mice separately), we had the data for 36 corticosterone concentrations with their corresponding standard errors.

2.2 Extraction method

Six individuals (S.V.D.M., C.Y., E.Z., N.W., B.S., T.Z.B.), extracted the data from the graphs. The order of the graphs for each extractor was randomized on the publication level to mitigate any potential learning or fatiguing effect. Five of them (C.Y., E.Z., N.W., B.S., T.Z.B.) were volunteers and had no previous experience extracting data from graphs. They extracted data from the figures without knowledge of the author-provided data. S.V.D.M. had seen the author-provided data prior to extraction. In contrast, S.V.D.M. had previous experience with extracting
data from graphs. All other extractors received a short demonstration of the method by S.V.D.M., as well as a protocol with instructions. All graphs were taken from PDF files. Images in PDF files can be vectorized or rasterized, in our case six of the studies had vectorized graphs, with the rest having rasterized graphs.

The graphs were cropped from the PDFs by Windows’ built-in snipping tool, this also converts the vectorized graphs into rasterized images, and pasted directly in the image editor program. The free and open source image editor software GIMP (GNU Image Manipulation program, 2.10.1030) was used for measuring the number of pixels in a graph. The Measure Tool was used, which can measure distances in a figure in pixels. In general, the number of pixels from the top of one line to the top of the second line was measured, that is, from the top of the x-axis to the top of the tick mark on the y-axis, and for the data bars, from the top of the x-axis to the top of the bar, etcetera. An exception: two measurements per outcome value were done for line plots: once from the top of the x-axis to the top of the circle, and once form the bottom of the x-axis to the bottom of the symbol. The average was taken for the final value. For the error bars of the line plot first the distance between either the top of the circle to the top of the error bar, or the bottom of the circle to the bottom of the error bar (whichever was easier to measure) was measured, and the difference of the two measurements made previously for the outcome value was added. We chose the top of the lines and symbols instead of the center because this was less ambiguous. When lines are more than a couple of pixels in height it is often not immediately clear what the center is. Using the top of the line was easier and less ambiguous to identify. Similarly, for symbols it was not always immediately obvious what the center of the symbol was, the tops or bottoms of the symbol are easier to find and measure. The extractors were instructed to zoom in on the graphs. Different monitors were used, but all monitors had a resolution of 1920 × 1080 pixels.

The calculations were done automatically in Excel. The excel sheet was prepared in such a way that the extractors only had to fill in the pixels of the y-axis, the value of the y-axis, the pixels for the outcome value and the pixels for the SEM.

The outcome value was calculated as follows:

1. the length of the y-axis in pixels for a certain value was measured and the value of one pixel was calculated by dividing the y-axis value by the number of pixels;
2. the number of pixels from the x-axis to the data point (i.e., top of the bar or the circle in a line figure) was measured, and then multiplied by the value from step 1 to calculate the outcome value;
3. the number of pixels for the error bar was measured and multiplied by the same value to calculate the SEM.

### 2.3 | Statistics

Our sample comprised 36 concentrations with the corresponding SEM, all provided by the original authors and extracted from their figures by the six extractors.

The original data had a large range of values due to reporting in different units (e.g., pg/ml and ng/ml). As we were not interested in the actual corticosterone concentrations, but in the method of extracting, we did not convert into a common unit, but used the values directly as extracted. For the Bland–Altman plots the data was standardized (see below).

The overall concordance correlation coefficients (OCCC) were calculated among the different extractors and the concordance correlation coefficients (CCC) were calculated between the extracted and the original data for each individual extractor. These methods quantify how well the measurement of a different method agrees with the measurements made by the “gold standard” and calculates both accuracy and precision. In this case the original author-provided values were considered the “gold standard.” The OCCC is based on the weighted average of pairwise CCCs.

A Bland–Altman plot is a visual representation of the agreement between two sets of measurements. The x-axis shows the mean for each pair of measurement, the y-axis shows the difference. The plots can also show systematic biases: if the mean differences differ from 0 it means that one set of measurements is systematically higher or lower than the other set. For the Bland–Altman plots both the outcome and SEM values were standardized, the former by dividing by the corresponding SEMs, the latter by the following formula:

\[
\frac{\text{SEM} - \min(\text{SEM})}{\max(\text{SEM}) - \min(\text{SEM})}
\]

This method ensures that all values lie between 0.0 and 1.0. To prevent dividing by 0 errors, 0.001 was added to each normalized SEM value. The Bland–Altman plots were created with these transformed values, but instead of the difference between the original and extracted data, the ratio of the original to the extracted data was used to limit heteroscedasticity affecting the visualization.

Random-effects meta-analyses on the mean difference between the original and extracted data per extractor were performed using the two-step inverse-variance approach. These meta-analyses assess the overall differences between the original and extracted data by taking
the differences for each graph and combining the differences in a weighted manner. The between-trial variance was estimated using the restricted maximum likelihood (REML) estimator for being recommended from a recent simulation study. As the original and extracted data should be similar the data is correlated. Since we do not know the “true” correlation coefficient between the original and the extracted data, we used different positive correlations in the range of 0.00–0.95 as a sensitivity analysis to investigate the effect of the correlation coefficient on the overall differences between original and extracted data. The standard deviation was calculated using the formula 4.15 in Introduction to Meta-analysis.

All statistical calculations were done in R using RStudio. The OCCC and CCC were calculated using the EpiR package. The meta-analyses were performed using the metafor package.

3 | RESULTS

The OCCC among the different extractors is shown in Table 1 for both the concentration and the SEM values, with the corresponding accuracy and precision. The OCCC for concentrations was high with 0.999. The overall accuracy was perfect for the concentration, meaning the OCCC was only reduced due to the precision. The precision was still high with 0.999. The OCCC for SEM was reduced by both the precision and accuracy, but both were still larger than 0.96 and the OCCC was high with 0.95.

### TABLE 1  Overall concordance correlation coefficients among all extractors for outcome and SEM values

|              | Outcome | SEM  |
|--------------|---------|------|
| Overall CCC  | 0.999   | 0.947|
| Overall precision | 0.999 | 0.969|
| Overall accuracy | 1.000 | 0.977|

### TABLE 2  The concordance correlation coefficient between data extracted by each extractor and the original data for both the outcome and SEM values

| Extractor | Outcome | SEM  |
|-----------|---------|------|
| 1         | 0.9998  | 0.9434|
| 2         | 0.9991  | 0.9731|
| 3         | 0.9998  | 0.9429|
| 4         | 0.9986  | 0.9416|
| 5         | 0.9998  | 0.9975|
| 6         | 0.9999  | 0.9257|

The CCCs between data extracted by each extractor and the original data for both the outcomes and SEM values were generally high as shown in Table 2. The lowest CCCs for the outcome values were 0.99 and the lowest CCCs for SEM was 0.93.

Figure 1 shows Bland–Altman plots (BA) of the outcome and SEMs values separated by both extractor and study. The BA for the outcome values indicated no overall bias, and no large discrepancies between the extractors, that is, the points seemed to be distributed randomly along the line of no difference (a ratio of 1). It did reveal that differences tended to be clustered by study. For example, all points below the lower-bound CIs belonged to the same study (discussed below).

The BA for SEMs showed there was some bias (range from 1.18 to 2.02), partly caused by one study being a larger outlier. With this study removed, the mean ratio ranges from 0.77 to 1.64 (data not shown). Like the BA for concentration, there were no large discrepancies between extractors although it seemed that each extractor had a consistent bias compared to the other extractors.

As the extracted on original data are similar, Figure 2 presents the mean differences for the different extractors between the extracted and original data plus the 95% confidence intervals (CI) under a range of different positive correlation coefficients. Compared to the mean original concentration value, 82.07, the differences are small (median: 0.025, interquartile range: 0.016–0.046). The considerable variability on the side of the included trials was propagated as relatively wide 95% CIs for all extractors, especially, for low values of correlation coefficient. The mean differences themselves are largely unaffected by the correlation coefficient.

Figure 3 shows the forest plot of the random-effects meta-analysis between the extracted and original data for all graphs for extractor 3 and 0.95 correlation coefficient. Extractor 3 was chosen randomly as an example. The forest plots for the other extractors (at 0.95 correlation coefficient) are presented in Figure S1. In all graphs, the 95% CI of the mean difference between the original and extracted values included 0. For extractor 3, the summary MD was 0.05 with 95% CIs: −0.48 to 0.58. Even though MD differed to some extent across the trials, between-trial variance was estimated to be zero due to the substantial overlapping of the 95% CI of the included trials. This pattern was observed in all extractors, regardless of the correlation coefficient. Consequently, the results are identical with a fixed-effect model. The results based on the random effects or the fixed effect models are presented in Figure S2. However, we considered the random-effects model to be conceptually the proper model for meta-analysis since the included studies have been conducted under diverse protocols and hence, clinical and
methodological heterogeneity are imminent. As the between-trial variance is zero, the prediction intervals are the same as the confidence intervals.

4 | DISCUSSION

In this study, we investigated the concordance of the extracted outcome and SEM data between different extractors, between the extracted and original outcome and error data as well as the potential consequences of the differences between the extracted and original data on a meta-analysis.

When we compared the extracted outcome and SEM values between the different extractors, the differences were very small, with the OCCC for the SEM being slightly lower than for the outcome values. The likely reasons for the slightly lower OCCC for SEM may be that error bars were smaller than the data bars, and did not start on the x-axis, which made them more difficult to measure.

One of the possible contributions to the high OCCC overall could be that the extractors first received a visual demonstration and a data extraction protocol. The training and protocol taught extractors how to extract data.

**FIGURE 1** Bland–Altman plots of the outcome (A) and standard error of the mean values (B) for the different studies and different extractors. The blue line indicates the mean ratio for either all outcome values or errors. The orange lines indicate the 95% confidence intervals. The dashed line indicates no bias (ratio of 1) [Colour figure can be viewed at wileyonlinelibrary.com]
(e.g., how does one measure symbols in a line graph?), and hence, prevented making decisions on a graph-by-graph basis. Furthermore, the extractors in the current study were also not blinded to the research question and therefore the data extraction for this paper might have been more of a conscious effort for precision compared to a reviewer who extracts numerous graphs as part of a large systematic review project. A future study could address naturalistic data extraction while following an existing review protocol.

The data extraction for this study was based on published graphs from different journals that were identified in a previous systematic research, and as such represented a real-life scenario; but even so some considerations have to be kept in mind when extrapolating the results. Besides the two above-mentioned considerations of using a protocol and the extractors not being blinded to the research question, one other major consideration was the fact that most graphs were bar plots. Cramond et al. used different figure types commonly found in pre-clinical and clinical science and found that bar plots generally had the highest accuracy, followed by box plots and dot plots, indicating that the OCCCs might be lower if there are more dot plots to extract data from. However, with at least two extractors extracting the data, we do not anticipate that the plot type would lead to large differences between the original and the extracted data, as discrepancies between the extractors are easy to spot and correct. It is strongly advised to perform the data extraction by at least two extractors anyway as this may decrease considerably the number of errors introduced by data extraction. Future research specifically focusing on different graph types from published articles may give a definitive answer.

We compared the extracted outcome and SEM data with the original data using the CCC. The CCC for the outcome data was very high, with even the lowest
These values were even higher than found by Kadic et al. It should be noted that Kadic et al. used the intraclass correlation coefficient, which should produce similar results to the CCC but is not the same. The Bland–Altman plot for the outcome value showed there were no large discrepancies between the different extractors, that is, the extractors did not introduce biases overall in the extracted values.

The CCC for the SEM data was slightly lower compared to the outcome values, but still very high. The lowest CCC was 0.926, with the lower confidence interval limit being 0.885. We are not aware of any data allowing for direct comparisons, but the study by Burda et al., also showed that the differences between the original and extracted data were higher for the error values (in their case the confidence intervals) than for the outcome values. Our Bland–Altman plots did show an extractor-bias for SEM (i.e., all points seem to be transposed higher or lower compared to other extractors). Regardless, even with these biases, the absolute differences between the original and extracted values were very small.

The Bland–Altman plots did show there was a consistent bias for certain studies. For example, the study by Berry et al. had both a low ratio for outcome values (<0.5) and a high ratio for the error (>15). This was not due to low image quality but simply because the graph showed two interventions. The corticosterone concentration of the second intervention was much higher compared to the first, meaning that the outcome and standard error of interest were very small, and although the actual concentrations differed, the bars had the same height for both data points due to the chosen scales. The other outlier for the outcome values was the study by Oishi et al. There the graph was a relatively small line-graph. The other study with line graphs (Otsuka et al.) showed only a small difference, with ratios between 0.97 and 1.15. The figure for Otsuka et al. was larger compared to the figure by Oishi et al.

For errors, the study by Berry et al. had a large ratio because the standard error bars were very small (as described above). Three of the extractors had a low ratio (<0.44) for the graph by Enayati et al., and one extractor had a high ratio (<0.8) for the graph by Kember et al. There were no immediate reasons apparent for these ratios. The relatively small standard error bars compared to the outcome values bars was the most striking part of this figure. Based on the Bland–Altman plots we think that the largest differences between the original and

### FIGURE 3 Random effect meta-analysis between the original and extracted (by extractor 3) data for each graph ordered by precision (from most to least precise). Each row represents one extracted corticosterone concentration plus the corresponding standard error. The values on the right indicate the mean difference between the original and extracted data, followed by the 95% confidence intervals in brackets.

| Study       | Mean difference [95% CI] |
|-------------|-------------------------|
| Berry       | 0.19 [-0.82, 1.19]      |
| Enayati     | 0.19 [-0.90, 1.29]      |
| Enayati     | 0.23 [-1.85, 2.29]      |
| Enayati     | 0.26 [-1.18, 1.64]      |
| Trefit      | 0.01 [-1.39, 1.41]      |
| Kember      | -0.45 [-2.67, 1.77]     |
| Berry       | -0.74 [-2.88, 1.41]     |
| Zhu         | 0.18 [-2.34, 2.68]      |
| Ros-Simó    | 0.22 [-2.45, 2.89]      |
| Ros-Simó    | 0.36 [-1.24, 4.95]      |
| Kurata      | 0.50 [-1.02, 2.03]      |
| Innos       | 0.51 [-0.72, 1.75]      |
| Kember      | 0.04 [-0.88, 1.86]      |
| Kember      | 0.27 [-0.41, 0.94]      |
| Moussaliev  | 0.03 [-0.74, 0.79]      |
| Kinoshita   | 0.15 [-1.76, 2.06]      |
| Kember      | -0.12 [-0.89, 0.65]     |
| Pastor      | -0.59 [-1.19, 0.00]     |
| Breuilaud   | 0.43 [-0.76, 1.63]      |
| Ros-Simó    | 0.46 [-0.92, 1.84]      |
| Chourbaji   | -0.32 [-1.92, 1.28]     |
| Moussaliev  | -0.24 [-1.68, 1.20]     |
| Otsuka      | -0.05 [-1.13, 0.97]     |
| Otsuka      | -0.09 [-1.10, 0.91]     |
| Chourbaji   | -0.25 [-1.27, 0.77]     |
| Parkhurst   | -0.11 [-1.32, 1.09]     |
| Philbert    | 1.84 [-0.26, 3.94]      |
| Van der Sluis| -0.09 [-1.47, 1.29]    |
| Renguist    | 0.02 [-0.23, 0.27]      |
| Chourbaji   | 0.65 [-1.30, 2.60]      |
| Innos       | 0.14 [-0.40, 0.68]      |
| Jiang       | -0.10 [-1.81, -0.15]    |
| Chourbaji   | 1.89 [-0.37, 4.15]      |
| Cuffe       | -3.36 [-18.14, 10.45]   |
| Oishi       | -1.40 [-8.13, 5.33]     |

**RE Model**

0.05 [-0.48, 0.58]
extracted data are affected by the size of the figures and
the size of the (error)bars, that is, when having a small fig-
ure and/or small (error)bars it is easier to make a mistake
with extracting, and a mistake leads to a relatively large
difference between the original and extracted data.

In the meta-analyses, the overall mean difference
between the original and the extracted data for the out-
come data was virtually zero. The between-trial variance
was also virtually zero. However, between-trial variance
would have been non-zero had we used Bayesian
methods to estimate the between-trial variance (e.g., by
assigning a weakly-informative prior on the between-trial
standard deviation, such as a half-normal prior distribu-
tion with variance 1), since there was some observed sta-
tistical heterogeneity across the studies (Figure 3). We
believe that within settings that are comparable to ours,
it would be highly unlikely that using extracted data
instead of the original data would have influenced the
results of a meta-analysis in a way that may compromise
the conclusions delivered to the end-user.

For this study, we used GIMP to extract the data from
graphs, but there is other software available specifically
for extracting data from graphs like WebPlotDigitizer and
DigitizeIt. The underlying processes between GIMP
and the software for establishing the data values, that is,
measuring the number of pixels and calculating the out-
come on the value per pixel, are the same and we thus do
not expect that the choice of software will influence the
extracted outcome values. The specialized software might
have some features that can optimize the workflow
(e.g., WebPlotDigitizer makes it possible to extract all
data points in a graph and export all values in one go, in
GIMP we had to export the data after each data point),
possibly decreasing the time spent extracting data.
WebPlotDigitizer and DigitizeIt also have functions to
automatically extract data, but their usability varies based
on the graph types. For example, in WebPlotDigitizer
there is a function to extract from bar graphs but it
requires manual modification of parameters or adjusting
of points and thus not necessarily increases extraction
speed when extracting graphs with only a few bars. How-
ever, as these functions are actively being developed their
usability will likely increase and we encourage
researchers in the future to look into these functions.

Images in PDF files can be vectorized or rasterized.
Vectorized images make it possible to zoom in without
alasing, possibly making extracting more precise or ac-
urate. Of our studies, six had vectorized graphics. When
copying the graphs to GIMP these graphs were turned
into rasterized images. However, we do not think this
rasterizing has affected the outcomes because the figures
were already zoomed in before rasterizing; meaning that
aliasing was not an issue as aliasing only becomes
apparent when zooming in even further. Furthermore,
the graph for which we found the biggest difference
between the original and extracted concentration and
error was a vectorized graph. This was not due to being
rasterized or vectorized but because the outcome and
error bars were very small. We therefore think that the
advantages of vectorized graphics are small compared to
the scale of the graph (i.e., a large image with large
(error)bars is more important than being vectorized).

 Naturally, for data extraction in systematic reviews,
the best option remains to extract the data directly from
the text or tables in a paper, and we would urge scientists
to publish all data either in the text or tabulated, even
when also shown in figures. If the data are not directly
available, reviewers should contact the authors to request
the data. This ensures that the correct data are used for
meta-analyses. However, our study strongly indicated
that even if the original data cannot be acquired, using
data carefully extracted from graphs is a good alternative
and is unlikely to have a significant impact on the results
of a meta-analysis. The data and code that support the
findings of this study are available in GitHub and
OSF.

ACKNOWLEDGEMENTS
Open access funding enabled and organized by
ProjektDEAL.

CONFLICT OF INTEREST
The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS
Conceptualization: SvdM, AB, CL. Formal Analysis: SvdM,
LMS, SRT. Investigation: SvdM, CY, EZ, NW, BS, TFzB.
Methodology: SvdM, LMS, SRT, AB, CL. Writing - Original
Draft: SvdM, CL. Writing - Review & Editing: All authors

ORCID
Stevie Van der Mierden https://orcid.org/0000-0001-7080-1872
Loukia Maria Spineli https://orcid.org/0000-0001-9515-582X
Eva Zentrich https://orcid.org/0000-0002-5309-7024
Talke Friederike Zur Brügge https://orcid.org/0000-0003-0563-0389
Cathalijn H.C. Leenaars https://orcid.org/0000-0002-8212-7632

REFERENCES
1. de Vries RB, Wever KE, Avey MT, Stephens ML, Sena ES,
Leenaars M. The usefulness of systematic reviews of animal
experiments for the design of preclinical and clinical studies.
ILAR J. 2014;55(3):427-437.
2. Egger M, Davey-Smith G, Altman D. Systematic reviews in health care: meta-analysis in context. Hoboken, NJ: John Wiley & Sons; 2008.

3. Buscemi N, Hartling L, Vandermeer B, Tjosvold L, Klassen TP. Single data extraction generated more errors than double data extraction in systematic reviews. J Clin Epidemiol. 2006;59(7):697-703.

4. Horton J, Vandermeer B, Hartling L, Tjosvold L, Klassen TP, Buscemi N. Systematic review data extraction: cross-sectional study showed that experience did not increase accuracy. J Clin Epidemiol. 2010;63(3):289-298.

5. Kadic AJ, Vucic K, Dosenovic S, Sapunar D, Puljak L. Extracting data from figures with software was faster, with higher interrater reliability than manual extraction. J Clin Epidemiol. 2016;74:119-123.

6. Burda BU, O’Connor EA, Webber EM, Redmond N, Perdue LA. Estimating data from figures with a Web-based program: considerations for a systematic review. Res Synth Methods. 2017;8(3):258-262.

7. Cramond F, O’Mara-Eves A, Doran-Constant L, Rice AS, Macleod M, Thomas J. The development and evaluation of an online application to assist in the extraction of data from graphs for use in systematic reviews. Welcome Open Res. 2018;3:1-25.

8. Berry A, Bellisario V, Capoccia S, et al. Social deprivation stress is a triggering factor for the emergence of anxiety-and depression-like behaviours and leads to reduced brain DBNF levels in C57BL/6J mice. Psychoneuroendocrinology. 2012;37(6):762-772.

9. Breuillaud L, Rossetti C, Meylan EM, et al. Deletion of CREB-regulated transcription coactivator 1 induces pathological aggression, depression-related behaviors, and neuroplasticity genes dysregulation in mice. Biol Psychiatry. 2012;72(7):528-536.

10. Courny S, Hirtz M, Molteni R, Riva M, Gass P, Hellweg R. The impact of environmental enrichment on sex-specific neurochemical circuitries–effects on brain-derived neurotrophic factor and the serotonin system. Neuroscience. 2012;220:267-276.

11. Cuffe J, O’sullivan L, Simmons D, Anderson S, Moritz K. Maternal corticosterone exposure in the mouse has sex-specific effects on placental growth and mRNA expression. Endocrinology. 2012;153(11):5500-5511.

12. Enayati M, Solati J, Hosseini M-H, Shahi H-R, Saki G, Salari A-A. Maternal infection during late pregnancy increases anxiety-and depression-like behaviors with increasing age in male offspring. Brain Res Bull. 2012;87(2-3):295-302.

13. Gurfein BT, Stamm AW, Bacchetti P, et al. The calm mouse: an specific neurochemical circuitries effects on brain-derived neurotrophic factor and the serotonergic system. Neurosci. 2012;222:267-276.

14. Innos J, Philips M-A, Raud S, Lilleväli K, Köks S, Vasar E. Deletion of the Lsamp gene lowers sensitivity to stressful environments in mice. Behav Brain Res. 2012;228 (1):74-81.

15. Jiang B, Xiong Z, Yang J, et al. Antidepressant-like effects of ginsenoside Rg1 are due to activation of the BDNF signalling pathway and neurogenesis in the hippocampus. Br J Pharmacol. 2012;166(6):1872-1877.

16. Kember R, Dempster E, Lee T, Schalkwyk L, Mill J, Fernandes C. Maternal separation is associated with strain-specific responses to stress and epigenetic alterations to Nr3c1, Avp, and Nr4a1 in mice. Brain Behav. 2012;2(4):455-467.

17. Kinoshita C, Miyazaki K, Ishida N. Chronic stress affects PERIOD2 expression through glycogen synthase kinase-3β phosphorylation in the central clock. Neuroreport. 2012;23(2):98-102.

18. Kurata K, Nagasawa M, Tomonaga S, et al. Orally administered L-ornithine reduces restraint stress-induced activation of the hypothalamic-pituitary-adrenal axis in mice. Neurosci Lett. 2012;506(2):287-291.

19. Moussaieff A, Gross M, Nesher E, Tikhonov T, Yadid G, Pinhasov A. Incenseol acetate reduces depressive-like behavior and modulates hippocampal BDNF and CRF expression of submissive animals. J Psychopharmacol. 2012;26(12):1584-1593.

20. Oishi K, Uchida D, Itoh N. Low-carbohydrate, high-protein diet affects rhythmic expression of gluconeogenic regulatory and circadian clock genes in mouse peripheral tissues. Chronobiol Int. 2012;29(7):799-809.

21. Otuka T, Goto M, Kawai M, et al. Photoperiod regulates corticosterone rhythms by altered adrenal sensitivity via melanotonin-independent mechanisms in Fischer 344 rats and C57BL/6J mice. PLoS One. 2012;7(6):e39090.

22. Pankhurst MW, Gell DA, Butler CW, Kirkcaldie MT, West AK, Chung RS. Metallothionein (MT)-I and MT-II expression are induced and cause zinc sequestration in the liver after brain injury. PLoS one. 2012;7(2):e31185.

23. Pastor R, Reed C, Meyer PJ, McKinnon C, Ryabinin AE, Phillips TJ. Role of corticotropin-releasing factor and corticosterone in behavioral sensitization to ethanol. J Pharmacol Exp Therapeut. 2012;341(2):455-463.

24. Philbert J, Pichat P, Palme R, Belzung C, Griebel G. The CRF1 receptor antagonist SSR125543 attenuates long-term cognitive deficit induced by acute inescapable stress in mice, independently from the hypothalamic pituitary adrenal axis. Pharmacol Biochem Behav. 2012;102(3):415-422.

25. Renquist BJ, Murphy JG, Larson EA, et al. Melanocortin-3 receptor regulates the normal fasting response. Proc Natl Acad Sci USA. 2012;109(3):E1489-E1498.

26. Ros-Simó C, Valverde O. Early-life social experiences in mice affect emotional behaviour and hypothalamic-pituitary-adrenal axis function. Pharmacol Biochem Behav. 2012;102(3):434-441.

27. Trent S, Denney A, Richardson H, et al. Steroid sulfatase-deficient mice exhibit endophenotypes relevant to attention deficit hyperactivity disorder. Psychoneuroendocrinology. 2012;37(2):221-229.

28. van der Sluis RJ, van Puijvelde GH, Van Berkel TJ, Hoekstra M. Adrenalectomy stimulates the formation of initial atherosclerotic lesions: reversal by adrenal transplantation. Atherosclerosis. 2012;221(1):76-83.

29. Zhu W-L, Shi H-S, Wei Y-M, et al. Green tea polyphenols produce antidepressant-like effects in adult mice. Pharmacol Res. 2012;65(1):74-80.

30. GIMP—GNU Image Manipulation Program [computer program]. Version 2.10.10.

31. Excel [computer program].

32. Barnhart HX, Haber M, Song J. Overall concordance correlation coefficient for evaluating agreement among multiple observers. Biom. 2002;58(4):1020-1027.

33. Milligan GW, Cooper MC. A study of standardization of variables in cluster analysis. J Classif. 1988;5(2):181-204.

34. Deeks JJ, Higgins JPT, Altman DG. Chapter 10: analysing data and undertaking meta-analyses. In: Higgins JPT, Thomas J, Chandler J, et al., eds. Cochrane Handbook for Systematic Reviews of Interventions. Hoboken, NJ: Cochrane; 2019.
35. Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *J Edu Behav Stat*. 2005;30(3):261-293.

36. Langan D, Higgins JP, Jackson D, et al. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Res Synth Methods*. 2019;10(1):83-98.

37. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Effect sizes based on means. In: Borenstein M, Hedges LV, Higgins JPT, Rothstein HR, eds. *Introduction to Meta-Analysis*. Chichester, UK: John Wiley & Sons Ltd; 2009:21-32.

38. *R: A Language and Environment for Statistical Computing* [computer program]. Version 4.0.0, 2020.

39. RStudio. *Integrated Development Environment for R* [computer program]. Boston, MA: RStudio, PBC; 2020.

40. *epiR: Tools for the Analysis of Epidemiological Data* [computer program]. Version 1.0-142020.

41. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw*. 2010;36(3):1-48.

42. Li T, Higgins JPT, Deeks JJ. Chapter 5: collecting data. In: Higgins JPT, Thomas J, Chandler J, et al., eds. *Cochrane Handbook for Systematic Reviews of Interventions*. Hoboken, NJ: Cochrane; 2019.

43. Nickerson CA. A note on “A concordance correlation coefficient to evaluate reproducibility”. *Biometrics*. 1997;53:1503-1507.

44. Rohatgi A. WebPlotDigitizer; 2020. https://automeris.io/WebPlotDigitizer

45. Bormann. DigitzeIt, 2015; 2020. www.digitizeit.de

46. Van der Mierden S, Spineli L. R Script for “Extracting data from graphs: a case-study on animal research with implications for meta-analyses”; 2020. https://github.com/SvdMierden/graph_data_extraction

47. Van der Mierden S, Yiannakou C, Zentrich E, Weegh N, Struve B. Dataset for “Extracting data from graphs: a case-study on animal research with implications for meta-analyses”; 2020. https://osf.io/dmuwn/

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article**: Van der Mierden S, Spineli LM, Talbot SR, et al. Extracting data from graphs: A case-study on animal research with implications for meta-analyses. *Res Syn Meth*. 2021;12:701–710. https://doi.org/10.1002/jrsm.1481