Beyond the forest plot: The drapery plot

Gerta Rücker & Guido Schwarzer

Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center – University of Freiburg, Freiburg, Germany

Present Address: Albert-Ludwigs-Universität Freiburg, Freiburg, Germany

Correspondence
Gerta Rücker, Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center – University of Freiburg, Freiburg, Germany.
Email: ruecker@imbi.uni-freiburg.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: RU 1747/1-2

In the era of the “reproducibility crisis” and the “P-value controversy” new ways of presentation and interpretation of the results of a meta-analysis are desirable. One suggestion that has been made for single studies almost six decades ago and taken up now and then is the P-value function. For a given outcome, this function assigns a P-value to each possible hypothetical value, given the data. Moreover, the P-value function simultaneously provides two-sided confidence intervals for all possible alpha levels. An application to meta-analysis, while suggested early, has not been widely established. We introduce the drapery plot that presents the P-value function for all individual studies and pooled estimates in a meta-analysis as curves and the prediction range for a single future study. We also present a scaled variant with the test statistic on the y-axis. Both plots visualize the full information of a pairwise meta-analysis. We see a drapery plot as a complementary figure to a forest plot. It may be even an alternative in meta-analyses with many studies where forest plots tend to become very large and complex.

KEYWORDS
drapery plot, forest plot, meta-analysis, P-value, P-value curve, reproducibility

1 | INTRODUCTION

Scientists have proclaimed a “reproducibility crisis” that manifests itself, among others, in selective publication of statistically significant results, often based on the conventional, but arbitrary significance level of .05.1-3 In the biomedical literature an excessive use of P-values has been observed and criticized, leading to a “P-value controversy.”4 Wasserstein et al took the step to call for abandoning statistical significance.5 While these and other authors mainly question the concept of statistical significance, a journal has completely banned P-values (Basic and Applied Social Psychology6).

It is widely agreed that new ways of presentation and interpretation of the results of pre-clinical experiments, trials, and observational studies are desirable.7 One suggestion is the P-value function which was introduced almost six decades ago,8-10 taken up now and then,11,12 and has recently been picked up again.13,14 For a given outcome, this function assigns a P-value to each possible hypothetical value, given the data. As the P-value function also simultaneously provides two-sided confidence intervals for all possible alpha levels, it is also called the confidence interval function or confidence curve.12,15 An application of P-value functions in meta-analysis was also suggested,15 but has not been widely established in the field.

In this paper we suggest a graph, the drapery plot, that presents P-value curves for (a) all individual studies in a meta-analysis as well as (b) the pooled estimates and (c) contains a prediction region for a single future study. Like a forest plot, the drapery plot thus visualizes the main information of a pairwise meta-analysis.

In Section 2 we present the definition of the P-value function, discuss its properties, and explain how a meta-
analysis can be visualized as a drapery plot. We describe the interpretation of a drapery plot using two examples in Section 3 and discuss strengths and limitations in Section 4, and concluding with a recommendation in Section 5.

## 2 METHODS

The P-value is a frequentist concept, defined in the context of a statistical hypothesis test. Let $\theta$ be the estimand, for example, a treatment effect. To test whether $\theta$ differs from a hypothetical value $\theta_0$, two hypotheses are formulated, the null-hypothesis $H_0$: $\{\theta = \theta_0\}$ and an alternative hypothesis, $H_1$: $\{\theta \neq \theta_0\}$ (two-sided test). A one-sided test can also be formulated comparing $H_0$: $\{\theta \leq \theta_0\}$ and $H_1$: $\{\theta > \theta_0\}$ (or vice versa). In this paper, we focus on the two-sided setting.

Whether to reject the null-hypothesis is decided based on an estimator $\hat{\theta}$ informed by the data. To construct a statistical test, an assumption about the statistical distribution of the estimator under the null-hypothesis $H_0$ is needed. For example, the assumption of a normal distribution for a mean difference (MD) leads to a test statistic

$$\frac{\theta_{MD} - \theta_0}{\text{SE}_{MD}}$$

where $\theta_{MD}$ denotes the observed mean difference and $\text{SE}_{MD}$ its SE.

To keep it simple, we assume throughout this paper that the test statistic is approximated by the standard normal distribution. For a fixed value $\theta_0$, the two-sided P-value $p$ is defined as the probability under the null-hypothesis $H_0$: $\{\theta = \theta_0\}$ that the absolute value of the test statistic is greater or equal than its observed value. Under the normality assumption, we may write this as

$$p = 2 \left(1 - \Phi \left( \frac{|\theta - \theta_0|}{\text{SE}} \right) \right),$$

where $\Phi$ is the probability function of the standard normal distribution. If $p$ is large (eg, 0.8), the observed data does not contradict the null-hypothesis. The smaller the P-value, the less compatible with the null-hypothesis we think the data. A common convention for deciding whether the null-hypothesis can be rejected is to prespecify a threshold, the significance level (denoted by $\alpha$) and to reject the null-hypothesis if $p < \alpha$. Frequently used values for $\alpha$ are 0.05 and 0.01.

### 2.1 The P-value function

The concept of the P-value function goes back to Birnbaum (1961), Miettinen (1985), and Poole (1987).\textsuperscript{8-10} Given the data and the distribution assumption, the P-value depends on the hypothetical value $\theta_0$. The idea of the P-value function $F$ is to consider the P-value as a function of $\theta_0$, given the data, that is, under the normality assumption:

$$F: \theta_0 \mapsto 2 \left(1 - \Phi \left( \frac{|\theta - \theta_0|}{\text{SE}} \right) \right).$$

We interpret values of $\theta_0$ resulting in large P-values as well compatible with the data and values with small P-values as less compatible with the data.

#### 2.1.1 Range and symmetry

P-values range between 0 and 1. Because $F$ depends only on the absolute distance $|\theta - \theta_0|$, the P-value function is symmetric with respect to the observed estimate, $\hat{\theta}$. It takes its unique maximum 1 if $\theta_0$ is exactly equal to the observed effect estimate, $\hat{\theta}$. With $\theta_0$ increasing from $-\infty$ to $\hat{\theta}$, $|\theta - \theta_0|$ strictly decreases and thus $F(\theta_0)$ strictly increases, while it strictly decreases for $\theta_0 > \hat{\theta}$, approaching zero for $\theta_0 \to +\infty$.

#### 2.1.2 P-value function and confidence intervals

Consider an arbitrary value $\alpha \in (0, 1]$, say $\alpha = 0.05$. We may look for those hypothetical parameters $\hat{\theta} - \theta_0$ and
\( \hat{\theta} + \theta_{a} \) that the P-value function maps to \( \alpha \), that is, \( F(\hat{\theta} - \theta_{a}) = F(\hat{\theta} + \theta_{a}) = \alpha \). This means that the null-hypotheses \( H_{0}: \{ \theta = \hat{\theta} - \theta_{a} \} \) and \( H_{0}: \{ \theta = \hat{\theta} + \theta_{a} \} \) are both rejected exactly with probability \( \alpha \), given \( \hat{\theta} \). In other words, the values \( \hat{\theta} - \theta_{a} \) and \( \hat{\theta} + \theta_{a} \) are the limits of the symmetric two-sided \( (1 - \alpha) \) confidence interval for \( \hat{\theta} \). Thus, the P-value function simultaneously displays confidence intervals for \( \hat{\theta} \) at any level. It is for this reason that it was often termed confidence curve\(^{8,12,16-18} \) or confidence interval function.\(^{15} \)

### 2.1.3 A scaled version of the P-value function

The P-value function maps the hypothetical parameters to their P-values. Alternatively, the test statistic may be displayed directly. Applying the quantile function to the P-values provides

\[
G: \theta_{0} \rightarrow \Phi^{-1}\left(1 - \Phi\left(\frac{|\hat{\theta} - \theta_{0}|}{\text{SE}}\right)\right) = \Phi^{-1}\left(\Phi\left(-\frac{|\hat{\theta} - \theta_{0}|}{\text{SE}}\right)\right) = -\frac{|\hat{\theta} - \theta_{0}|}{\text{SE}}
\]

which is piecewise linear in \( \theta_{0} \in (-\infty, \hat{\theta}) \) and \( \theta_{0} \in (\hat{\theta}, + \infty) \).

### 2.2 Visualization of a pairwise meta-analysis

Pairwise meta-analysis combines the results of similar studies reporting the same outcome of interest. Typically a weighted average of study-specific estimates is calculated under a fixed effect (also called common effect) or random effects model. The random effects model assumes that the true study means are drawn from a distribution. Typically, a normal distribution is assumed with a variance parameter that is estimated from the data. The pooled estimate is defined as the mean of the between-study distribution. Its variance depends both on the standard errors of individual studies and the estimated between-study variance.

An established and useful visualization of the results of a meta-analysis is the forest plot, showing the confidence intervals of the primary studies along with a confidence interval for each pooled estimate. It is recommended to add a prediction interval, which gives a region where the estimate from a single future study is expected, based on the random effects model.\(^{19} \)

In this paper we propose a **drapery plot** as a complementary figure to a forest plot. The drapery plot shows the P-value functions of all primary studies in one graph, together with a P-value curve for each pooled estimate and a shaded prediction region. In contrast to the forest plot, the drapery plot does not only show results for an (arbitrary) confidence level, but for any possible confidence level. We will elaborate on this feature in the next section.

Drapery plots are implemented in function **drapery** of R package **meta**.\(^{20,21} \)

### 3 APPLICATION

We will use two meta-analyses, which are available in R package **meta**,\(^{21} \) as illustrative examples to describe drapery plots. R code to reproduce our analyses and figures is given in Data S1 of this article.

#### 3.1 Smoking and lung cancer

The data were reconstructed based on the Smoking and Health Report to the Surgeon General\(^{22} \) and reanalyzed in Schumacher et al.\(^{23} \) They include the number of lung-cancer deaths in smokers and non-smokers and the number of person-years in each group. The effect measure is the mortality rate ratio of smokers compared to non-smokers.

Figure 1 shows a forest plot with the results for the common effect (blue diamond) and random effects meta-analysis (red diamond). The pooled lung cancer mortality was about 11 times higher for smokers compared to non-smokers, and also the prediction interval, though wide, was far away from 1. There is clear evidence that smokers have a much higher risk of lung-cancer mortality.

Figure 2 (left) shows a drapery plot. Gray curves correspond to primary studies, with study weights from the random effects model represented on a grayscale (studies with higher precision shown in darkgray, studies with low precision in lightgray). Each point estimate can be read off at the peak of the respective curve. The prediction region (lightblue) is broader than the P-value curve of the pooled estimates (red and black curves), indicating heterogeneity. Horizontal dashed lines can be used to identify confidence (prediction) intervals for common alpha levels (0.1, 0.05, 0.01). Figure 2 (right) shows a variant of the drapery plot with a quantile-scaled y-axis, such that all curves become triangles. Here, an alpha value of 0.001 is also printed in the figure. None of the curves crosses the null effect line (Incidence Rate Ratio IRR = 1).
3.2 Effects of elevated CO2 on total biomass of woody plants

This example comes from ecology. It is a meta-analysis on effects of elevated CO2 on total biomass of woody plants, used as an example in Hedges (1999).24

Figure 3 shows the rescaled variant of the drapery plot. Despite the large number of 102 studies with many overlapping P-value curves, we get a good impression of the data. The estimated effect is larger than 1 in the majority of studies; only five studies have an estimate below 1 (green curves). The common effect (blue curve) and random effects estimate (red curve) are very distinct, indicating a small-study effect. Similarly as in a funnel plot, this also leads to visible asymmetry: there are a number of studies with large uncertainty on the right-hand side without counterparts on the left-hand side of the plot. These studies pull the random effects estimate (which is more susceptible to small studies) to the right. The large between-study heterogeneity is visible in the wide prediction region (in lightblue). While the 90% prediction interval does not include the null effect, the 95% prediction interval clearly includes the null effect. A forest plot for these data is available as Figure S1.
4 | DISCUSSION

In this paper, we introduce a graph, the drapery plot, to summarize the results of a pairwise meta-analysis. The graph builds on the well-established P-value function which, in our experience, has been rarely used in practice. Our implementation in R function drapery presents P-value curves for individual studies and meta-analysis estimates as well as a prediction region for a single future study. At present, R function drapery only provides P-value curves under the normality assumption.

Bender et al, reviewing the history of P-value curves, critically discussed various terms for this function occurring in the literature.12 They opted for the term confidence curve, arguing “If the confidence interpretation is applied, the terms confidence curve, confidence interval function and confidence distribution function are appropriate, whereas the term P-value function is misleading as the confidence level is plotted on the y-axis and not the P-value”. Depending on the viewpoint, the function can be interpreted as providing P-values for shifted null hypotheses (varying $\theta_0$) or confidence intervals for all levels, given the data. While we mainly used the term P-value curve in this manuscript due to its current popularity,13,14 the drapery plot provides two y-axes on the P-value and confidence level scale enabling both interpretations.

The classic drapery plot - like the P-value curve - shows the P-values directly on the y-axis (Figure 2, left panel). Berrar showed a kind of drapery plot he called “average confidence curve,” without reference to meta-analysis (his Figure 5f).18 We also implemented a variant of the drapery plot showing the negative test statistic, that is, a transformation of the P-value, on the y-axis (Figure 2, right panel) which has been utilized before.14 The main advantages of this scaled variant of the drapery plot are that (a) changes in P-values and confidence limits are linear and that (b) results for smaller P-values/larger confidence levels are clearly visible. This feature is similar to a funnel plot using the recommended reversed SE on the y-axis instead of the inverse of the variance or SE.25 Accordingly, the scaled version of the drapery plot is the default in R function drapery.

Thompson suggested a very similar plot based on a completely different consideration in the framework of Fuzzy Set Theory, called “fuzzy number plot,”26 see also Kossmeier et al.27 Here, the y-axis represents the “membership grade,” in this framework meaning the grade (between 0 and 1) to which an element belongs to a (fuzzy) set. Grade 1 means that the fuzzy set reduces to a one-point interval of length 0, here the observed point estimate. Intervals of positive length around the point estimate correspond to grades between 0 and 1, with longer intervals corresponding to lower grades of.
membership. Thompson’s view is that “...in the absence of fuzzy set theory (ie, simply plotting confidence intervals in the same manner of fuzzy numbers), the metric of the vertical axis is unclear. Using fuzzy set theory, an established and interpretable membership grade is assigned to the vertical axis.” [26, p. 965]. However, the membership grade is just a linear transformation of the test statistic, and we think the test statistic a concept more widely known to statisticians. For examples of fuzzy number plots, see the figures in the work by Thompson.26

Bender et al12 also discussed the scaling of the y-axis and recommended an inverse (top-down) scale for the y-axis, however, they did not consider our linearization transformation. Another potential transformation of the y-axis is the double-square root transformation which also emphasizes smaller P-values and has been used, for example, to compare P-values from different meta-analysis models.28 However, this transformation does not have the linearization property.

We recommend the drapery plot as a complementary figure to a forest plot which is the most popular graph to summarize a meta-analysis. The distinctive feature/advantage of the drapery plot is that confidence intervals for individual studies and pooled estimates can be read off directly for any confidence level. In contrast, a forest plot can only display one confidence level at a time. For example, looking at Figure 2, we immediately notice (a) that all individual studies are statistically significant at the alpha level 0.001 and (b) that the prediction region is well above an incidence rate ratio of 2.

Furthermore, we argue that a drapery plot could be a suitable alternative to a forest plot in meta-analyses with very many studies. Figure 3 shows that while a drapery plot can become busy in this situation, the overall figure is not large and fits on a single page. Accordingly, this graph could be used in an article or presentation instead of (or in addition to) a very long forest plot spanning over several pages.

Our implementation of drapery plots is very flexible and enables the user, for example, to label all or some studies or to choose the appearance of curves for individual studies (line color, width, and type) based on study-level information. In Figure 2, some study labels have been rotated individually to avoid overlapping study labels. In Figure 3, study labels have been suppressed, however, curves of studies with an estimated ratio of means below 1 are printed in green. Similarly, study labels or curves could be modified based on other study-level information like P-values of treatment effects or whether a treatment effect or confidence interval falls outside the prediction interval.

Nevertheless, a limitation of the drapery plot is that each individual study is not easily recognized or identified if there is a large number of studies like in our second example. This is a property shared by other types of plots such as scatter plots: like a scatter plot with a large number of points, a drapery plot gives a good impression of trends in the data, while it is not thought to inform readers about each single point. In fact, the drapery plot shares also another feature of the funnel plot: like in a funnel plot (and by contrast to a forest plot), there is no arbitrariness with respect to the order of the studies, as both axes have a specific meaning. Based on a drapery plot, heterogeneity, small study effects, and outliers can also be identified.

There are other R packages for creating P-value curves. Infanger and Schmidt-Trucksäss released a special R package pvaluefunctions, not directly intended for application to meta-analysis.14 The R package concurve allows visualizing the results of a meta-analysis, based on R package metafor29,30; however, it does not offer a simple way to draw P-value curves for all primary studies.

## 5 | CONCLUSION

In summary, we see drapery plots as a valuable addition to the toolbox of visualization methods for meta-analysis, either as a complementary figure for sensitivity analyses or as a potential alternative in meta-analyses with very many studies.

### ACKNOWLEDGMENTS

G.R. acknowledges the support by the Deutsche Forschungsgemeinschaft (DFG), grant number RU 1747/1-2. We would like to thank the anonymous reviewers for their helpful comments on the text and valuable suggestions to improve our implementation of drapery plots. G.R. receives payment from Wiley as an Editor in Chief of Research Synthesis Methods.

### CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

### AUTHOR CONTRIBUTIONS

G.R. proposed P-value curves for meta-analysis, developed an R function, and wrote the manuscript. G.S. refined the R code, implemented the function into R package meta, and revised the manuscript. Both authors have read and approved the final version of the manuscript.

### DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.
ENDNOTE

1As of today (Apr 3rd, 2020) the function drapery is only available in the GitHub version of meta, available on https://github.com/guido-s/meta. We will submit an update of meta to CRAN in due time.

REFERENCES

1. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. Lancet. 1991;337:676-872.
2. Rothstein HR, Sutton AJ, Borenstein M, eds. Publication Bias in Meta Analysis: Prevention, Assessment and Adjustments. Chichester, UK: Wiley; 2005.
3. Kleinert S, Horton R. How should medical science change? Lancet. 2014;383(9913):197-198. https://doi.org/10.1016/S0140-6736(13)62678-1.
4. Wellek S. A critical evaluation of the current “P-value controversy”. Biom J. 2017;59(5):854-872.
5. Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. Am Stat. 2016;70(2):129-133.
6. Trafimow D, Marks M. Editorial. Basic Appl Soc Psychol. 2015;37:1-2. https://doi.org/10.1080/01973533.2015.1012991.
7. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond “p < 0.05”. Am Stat. 2019;73(sup1):1-19.
8. Birnbaum A. Confidence curves: an omnibus technique for estimation and testing statistical hypotheses. J Am Stat Assoc. 1961;56(294):246-249.
9. Miettinen OS. Theoretical Epidemiology. New York: Wiley; 1985.
10. Poole C. Beyond the confidence interval. Am J Public Health. 1987;77(2):195-199.
11. Kulldorff M, Graubard B, Velie E, Lang JM, Cann CI, Rothman KJ. The P-value and P-value function. Epidemiology. 1999;10(3):345-347.
12. Bender R, Berg G, Zeeb H. Tutorial: using confidence curves in medical research. Biom J. 2005;47(2):237-247.
13. Fraser DAS. The p-value function and statistical inference. Am Stat. 2019;73(sup1):135-147.
14. Infanger D, Schmidt-Trucksäss A. P value functions: an underused method to present research results and to promote quantitative reasoning. Stat Med. 2019;38(21):4189-4197.
15. Sullivan KM, Foster DA. Use of the confidence interval function. Epidemiology. 1990;1(1):39-42.
16. Scherb H, Brüské-Hohlfeld I. The exact confidence curve. Informatik, Biometrie und Epidemiologie in Medizin und Biologie. 1993:24:152-163.
17. Blaker H. Confidence curves and improved exact confidence intervals for discrete distributions. Can J Stat. 2000;28(4):783-798.
18. Berrar D. Confidence curves: an alternative to null hypothesis significance testing for the comparison of classifiers. Mach Learn. 2017;106:911-949.
19. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. J R Stat Soc. 2009;172:137-159.
20. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2020.
21. Balduzzi S, Rücker G, Schwarzer G. How to perform a meta-analysis with R: a practical tutorial. Evid Based Ment Health. 2019;22(4):153-160.
22. Bayne-Jones S, Burdette WJ, Cochran WG, et al. Smoking and Health. Report of the Advisory Committee to the Surgeon General of the Public Health Service of the United States. Public Health Service Publication 1103. Washington, DC: Department of Health, Education and Welfare; 1964.
23. Schumacher M, Rücker G, Schwarzer G. Meta-analysis and the Surgeon General’s report on smoking and health. N Engl J Med. 2014;370(2):186-188.
24. Hedges LV, Gurevitch J, Curtis PS. The meta-analysis of response ratios in experimental ecology. Ecology. 1999;80(4):1150-1156.
25. Sterne JAC, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. J Clin Epidemiol. 2001;54:1046-1055.
26. Thompson CG. Graphing effects as fuzzy numbers in meta-analysis. J Mod Appl Stat Methods. 2017;15(1):49.
27. Kossmeier M, Tran US, Voracek M. Charting the landscape of graphical displays for meta-analysis and systematic reviews: a comprehensive review, taxonomy, and feature analysis. BMC Med Res Methodol. 2020;20:26.
28. Wiksten A, Rücker G, Schwarzer G. Hartung-Knapp method is not always conservative compared to fixed effect meta-analysis. Stat Med. 2016;35:2503-2515. https://doi.org/10.1002/sim.6879.
29. Chow ZR, Vigotsky AD. concurve: Computes and Plots Consonance (Confidence) Intervals, P-Values, and S-Values to Form Consonance and Surprisal Functions. R Package Version 2.1.0, 2019. https://CRAN.R-project.org/package=concurve.
30. Viechtbauer W. Conducting meta-analyses in R with the metafor package. J StatSoftw. 2010;36(3):1-48.

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: Rücker G, Schwarzer G. Beyond the forest plot: The drapery plot. Res Syn Meth. 2021;12:13–19. https://doi.org/10.1002/jrsm.1410