Diplomarbeit

zum Thema

Meta-Regression and Robustness

ausgeführt am

Institut für Statistik und Wahrscheinlichkeitstheorie

der Technischen Universität Wien

unter der Anleitung von

A.o.Univ.Prof. Dipl.-Ing. Dr.techn. Peter Filzmoser

durch

Klaudius Kalcher

Matr.Nr. 0226160

1140 Wien, Mooswiesengasse 22

Wien, am 12. Oktober 2009 Klaudius Kalcher

Contents

Introduction

Meta-analysis is a tool to synthesize information from multiple primary studies of similar design [63] which, even though first attempts have been made over a century ago [52], has only recently seen wider dissemination in practical research [54]. The number of meta-analyses performed has considerably increased over the last years, which may be explained by the ever increasing number of primary studies [14], the advancement in methodology, and the availability of comprehensive guidelines and software packages opening the field of meta-analysis to a larger community of researchers and practitioners alike. Although the number of published meta-analyses has risen from about 400 in the year 2000 [39] to about 2300 in 2006 [63] and meta-analytic papers are among the most cited in their respective fields [50], the number of methodological papers has remained more or less constant with about 10 publication per year [63].

During the course of clinical research in the field of psychiatry, meta-analyses on the treatment of depression [8, 19, 35, 36, 44, 45, 61, 64, 69, 70], schizophrenia [9, 15, 34, 41, 40, 42, 59] and bipolar disorder [16] were studied, prompting the need for more thorough investigation of statistical procedures employed in meta-analytic research scenarios. Various efforts of professional associations to establish more coherent procedures and improve the overall quality of meta-analytic research [46] have accompanied the development of meta-analytic methodology ever since its introduction to the scientific community. Still, problems and inconsistencies abound [21] and hamper the ability of both producers and consumers of meta-analyses to perform and correctly interpret meta-analytic datasets. This has even lead to the point where it has been argued that the majority of meta-analytic findings in clinical research may be false [29].

Indeed, the meta-analytic setting adds to the already well-known potential pitfalls of basic statistical methods employed in clinical research [5] several new aspects to consider during generation of the research question, data collection, analysis, and interpretation. Among these are the questions whether meta-analysis is an adequate tool, whether meta-regression can and should be employed [66], how to address data quality issues like publication bias and heteroscedasticity, whether a common effect size of primary studies can be assumed, and how to define this effect size in the first place.

Some of the issues raised by meta-analytic problems can adequately be identified and dealt with, e.g. heterogeneity of true effect sizes between primary studies, others can be identified, but there exists no canonical solution in current modeling approaches. Examplary for the latter stand publication bias and heteroscedasticity in meta-regression. This situation lends itself to the exploration of robust methods in the meta-analytic context. However, due to the scarcity of robust methods for weighted data, robustness is difficult to achieve in meta-analysis. In particular, there is no established method of robust regression on weighted data, making robust meta-regression currently impossible.

Important problems arising in meta-analysis are highlighted by re-analysis of a published dataset on clinical efficacy of second-generation anti-depressant medication. Consecutively, a first approach to robust meta-regression is presented in the form of a newly developed estimator, termed `meta.lts`, based on the least trimmed squares (LTS) estimator. Its properties were discussed and demonstrated on several simulated datasets.

Various caveats encountered during the course of the meta-analytic investigations were documented, a special focus was set on illustrating issues of meta-regression, and the usefulness of `meta.lts` as a robust meta-analytic tool was demonstrated.

Methods

1 Meta-Analytic Methods

1.1 Fixed-Effect Models

1.1.1 General Idea

The idea of meta-analysis is to synthesize the results of multiple primary studies using quantitative statistical methods. Thus, it will be first considered using tools already established for simple analyses using the results of primary studies as individual observations. One obvious difference between the two is that in meta-analysis results of the primary studies are only estimates rather than measurements of a true effect size θ and thus subject to a certain amount of variability. Usually, this variability is asserted in the primary studies and reported in the form of measures of precision which in meta-analysis are used to improve the quality of overall results.

Most of the currently employed methods in meta-analysis can be labeled as either fixed-effect or random-effects models. The fixed-effect model assumes that a true effect size θ is common to all n primary studies and that the deviation of the location estimates $\hat{\theta}_i$ of the i-th study is due to its sampling error ε_i only.

$$\hat{\theta}_i = \theta + \varepsilon_i \quad i = 1, \ldots, n \tag{1}$$

However, the confidence one can put into these estimates depends on their respective precisions. Those can be explained by two factors—variances within primary studies and their sample sizes. The larger the sample size and the smaller the

within-study variability, the higher the estimation accuracy of the true effect size by a single study.

In order to use reported information about the precision of the primary studies location estimates in the calculation of the overall effect size and its variability can be weighted accordingly. Commonly, the studies are weighted proportionally to their precision—that is, the inverse of the variance [71]

$$W_i = \frac{1}{S_i^2} \tag{2}$$

where S_i^2 denotes the sample variance of the estimate of the true effect size by the i-th study.

The estimate of the true effect size can be computed as the weighted mean of the effect sizes of the individual studies (cf. [72])

$$\hat{\theta} = \frac{\sum_{i=1}^{k} W_i \hat{\theta}_i}{\sum_{i=1}^{k} W_i}. \tag{3}$$

and the variance of $\hat{\theta}$ can be estimated by

$$S_{\hat{\theta}}^2 = \frac{1}{\sum_{i=1}^{k} W_i}. \tag{4}$$

This model provides a compelling interpretation: if the variances within the primary studies are homogeneous, the fixed-effect estimate of the true effect size is equivalent to an estimate of the effect size using a pooled sample of all the primary studies' data points.

1.1.2 Underlying Assumptions

The main assumption of fixed-effect models is that the true effect size is the same across all the primary studies. In particular, this would be the case if all primary

studies had drawn their samples from the same population, in which case the fixed effect model provides the optimal estimate. On the other hand, this assumption precludes the possibility that the primary studies differ in (possibly unknown) covariates that have an influence on the effect size.

Note that homogeneity of variances is not a prerequisite for valid application of a fixed-effect model, since this is taken into account by the weighting procedure. Independence of errors, however, is necessary for unbiased effect size estimation in the fixed-effect model, as well as normal distribution for the calculations presented above to be applied correctly.

1.1.3 Problems

In practical meta-analysis, studies usually come from distinct sources. While they share several key characteristics allowing for quantitative synthesis, almost always there are relevant differences in sample properties between the studies, e.g. differences in study populations, treatment, design and execution [18]. Consequently, it cannot be assumed that all the individual samples share a common effect size, which means that the application of a fixed-effect model is not justified. Different methods of assessing heterogeneity between studies exist [53, 25], and including information on between study variation in the analytic process leads to the random-effects model.

1.2 Random-Effects Models

1.2.1 General Idea

The inhomogeneity of studies encountered in practical meta-analysis is addressed by random-effects models. Instead of assuming a common true effect size for all studies the effect sizes θ_i of the n different studies are taken to follow a distribution

of true effect sizes with mean μ and variance σ^2. Thus, the goal of the analysis is not to estimate a single true effect size common to all studies, but the distribution of true effect sizes [27]. Interesting statistics to be derived naturally include the expected value of the distribution, but it should not be forgotten that other metrics might be just as relevant.

In this model, two sources of variability can be identified—variance due to the distribution of true effect sizes σ^2 and variance due to sampling error $\sigma^2_{\varepsilon_i}$. Accordingly, the overall variance can be partitioned to reflect these two sources. The model employed allows for the effect size estimator $\hat{\theta}_i$ of the i-th study to be formulated as

$$\hat{\theta}_i = \theta_i + \varepsilon_i \quad \varepsilon_i \sim \mathrm{N}(0, \sigma^2_{\varepsilon_i}) \tag{5}$$

$$\theta_i = \mu + \delta_i \quad \delta_i \sim \mathrm{N}(0, \sigma^2). \tag{6}$$

The variance of the effect size estimations can then be partitioned in true variation in effect sizes and sampling error.

$$\mathrm{Var}(\hat{\theta}_i) = \sigma^2 + \sigma^2_{\varepsilon_i} \tag{7}$$

These differences to fixed-effect models manifest in different weighting of the studies compared to fixed-effect models. This will be discussed in more detail later on.

1.2.2 Underlying Assumptions

Random-effects models assume the true effects of the primary studies to be different but drawn from a common distribution in contrast to fixed-effect models which assume that all studies share a common effect size. For modeling purposes it makes no difference whether the variation between true effect sizes is due to (not

explicitely modeled) covariates or entirely random. If such covariates exist, there is of course a distinct interest in modeling them. For this purpose meta-regression can be employed.

The data points within the primary studies, however, are explicitly considered to be sampled from different distributions. Therefore, both location and scale parameters of these distributions can vary between studies.

1.2.3 Difference in Weighting Compared to Fixed-Effect Models

Since every primary study included in a random-effects model describes its own sampling population, it contains unique information about the distribution of the true effect sizes and cannot be replaced by information of other studies, however precise those might be.

To reflect the uniqueness of information of even the smaller studies, these are assigned relatively more weight than in fixed-effect models, while larger and more precise studies are assigned relatively less weight. One might say that the weights are more homogeneous. Weights are assigned analogously to equation (2). However, the variance in random-effects models is composed of between-studies variation of true effect sizes τ^2 and within-study sampling errors σ_i^2. Thus, the weight assigned to the i-th primary study is

$$U_i = \frac{1}{S_i^2 + T^2} \tag{8}$$

where T^2 and S_i^2 denote estimates of τ^2 and σ_i^2 respectively. S_i^2 is usually reported by the primary studies, and only in few cases has to be imputed. τ^2, in contrast, has to be estimated by the meta-analyst. A method commonly employed is the DerSimonian and Laird [11] method:

$$T^2 = \frac{Q - \nu}{C} \tag{9}$$

where

$$Q = \sum_{i=1}^{k} W_i \hat{\theta}_i^2 - \frac{\left(\sum_{i=1}^{k} W_i \hat{\theta}_i\right)^2}{\sum_{i=1}^{k} W_i}, \tag{10}$$

$$\nu = k - 1, \tag{11}$$

where k denotes the number of studies,

$$C = \sum_{i=1}^{k} W_i - \frac{\sum_{i=1}^{k} W_i^2}{\sum_{i=1}^{k} W_i}, \tag{12}$$

and W_i denotes the weights as computed in the fixed-effect model in equation (2).

The mean μ of the true effect sizes is estimated by

$$\hat{\mu} = \frac{\sum_{i=1}^{k} U_i \hat{\theta}_i}{\sum_{i=1}^{k} U_i}, \tag{13}$$

and the variance of $\hat{\mu}$ is estimated by

$$S_{\hat{\mu}}^2 = \frac{1}{\sum_{i=1}^{k} U_i}. \tag{14}$$

These estimates take into account both the between-studies variance and the sampling error and thus provide realistic results in the most common situation in meta-analysis, where different studies are sampled from different populations and/or under different circumstances (e.g. different investigators, institutions, etc.).

1.2.4 Improvements

Unlike the fixed-effect model, the random-effects model has sparked various attempts to improve it [63]. While the general theory of the random-effects model is not restricted to specific distribution families, the DerSimonian and Laird estimator for τ^2 assumes normal distribution of true effect sizes.

This approach has two fundamental drawbacks: on one hand the assumption of normal distribution is not always correct, on the other hand, even when normal distribution can correctly be assumed, the DerSimonian and Laird estimator is not the statistically optimal estimator [10] and has been shown to introduce potentially considerable bias [7]. Notably, even before publication of the orignal DerSimonian and Laird estimator [11], a better estimator had beed presented by Paule and Mandel [51], but unfortunately, maybe due to its lack of an explicit formula, it has not reached widespread use in the meta-analytic research community.

The Paule and Mandel estimator is defined as

$$T^2_{\mathrm{PM}} = \frac{\widetilde{Q} - B}{\widetilde{C}} \tag{15}$$

where

$$\widetilde{Q} = \sum_{i=1}^{k} \widetilde{W}_i \hat{\theta}_i^2, \tag{16}$$

$$B = \sum_{i=1}^{k} \widetilde{W}_i S_i^2 - \frac{\sum_{i=1}^{k} \widetilde{W}_i^2 S_i^2}{\sum_{i=1}^{k} \widetilde{W}_i}, \tag{17}$$

and

$$\widetilde{C} = \sum_{i=1}^{k} \widetilde{W}_i - \frac{\sum_{i=1}^{k} \widetilde{W}_i^2}{\sum_{i=1}^{k} \widetilde{W}_i}. \tag{18}$$

It thus bears a resemblence to the DerSimonian and Laird estimator, but instead of using W_i as computed in the fixed-effect model in equation (2), alternative weights $\widetilde{W}_i$ are used which are defined as

$$\widetilde{W}_i = \frac{1}{\tau^2 + S_i^2}. \tag{19}$$

Because the τ^2 are needed to calculate the weights in the formula, there is no explicit solution, but instead an iterative algorithm is needed. This is done by numerically solving the equation

$$F(\tau^2) = \sum_{i=1}^{k} \widetilde{W} \left(\frac{\sum_{i=1}^{k} \widetilde{W}_i \hat{\theta}_i}{\sum_{i=1}^{k} W_i} \right)^2 - (k-1) = 0 \tag{20}$$

with respect to τ^2. Because $F(\cdot)$ is strictly monotonic decreasing, this formula has a unique solution if and only if $F(0) \geq 0$, which can be computed iteratively, using $\tau_0^2 = 0$ as starting value for τ^2. Accordingly, if $F(0) < 0$, equation (20) has no solution, and T^2_{PM} is set to zero.

The Paule and Mandel estimator can be regarded as being better than the DerSimonian and Laird estimator mainly for two reasons. First, it does not require the assumption of normal distribution, and second, if normal distribution can be assumed, it has even been proven to be the restricted maximum likelihood estimator, and thus, in this sense, statistically optimal. The only advantage of the DerSimonian and Laird estimator is the availability of an explicit solution. If an iterative computation is feasible, however, there is no sound reason not to employ the Paule and Mandel estimator.

As a matter of fact, an algorithm for iterative computation actually is available and, due to the increase in computational power since 1982—when it has originally been published—the estimator can sensibly be employed today.

Even if an explicit formula is needed, a two step DerSimonian and Laird estimator [10], which gives solutions closer to the (maximum likelihood) optimum, can be used. This estimator is calculated by

$$T^2_{\mathrm{DL2}} = \frac{\left[\sum_{i=1}^{k} W_i^* \left(\hat{\theta}_i - \hat{\mu}\right)^2\right] - \left[\sum_{i=1}^{k} W_i^* S_i^2 - \frac{\sum_{i=1}^{k} (W_i^*)^2 S_i^2}{\sum_{i=1}^{k} W_i^*}\right]}{\sum_{i=1}^{k} W_i^* - \frac{\sum_{i=1}^{k} W_i^2}{\sum_{i=1}^{k} W_i^*}} \tag{21}$$

with $\hat{\mu}$ as in equation (13) and

$$W_i^* = \frac{1}{T_{\mathrm{DL}}^2 + S_i^2}. \tag{22}$$

While this two-step DerSimonian and Laird estimator leads to a better solution, it still bears the drawback of requiring the normality assumption.

All three estimators presented can be formulated as special cases of the general method-of-moments estimator for τ^2 [33]

$$T_{\mathrm{MM}}^2 = \frac{\left[\sum_{i=1}^k A_i \left(\hat{\theta}_i - \frac{\sum_{i=1}^k A_i \hat{\theta}_i}{\sum_{i=1}^k A_i}\right)^2\right] - \left[\sum_{i=1}^k A_i S_i^2 - \frac{\sum_{i=1}^k A_i^2 S_i^2}{\sum_{i=1}^k A_i}\right]}{\sum_{i=1}^k A_i - \frac{\sum_{i=1}^k A_i^2}{\sum_{i=1}^k A_i}} \tag{23}$$

The actual estimator resulting from this definition is determined by the choice of A_i, underlining the above mentioned similarity between the Paule and Mandel and the DerSimonian and Laird estimators. The former results from using the weights $A_i = \widetilde{W}_i$ in equation (23), whereas the latter results from setting $A_i = W_i$:

$$T^2_{\mathrm{DL}} = \frac{\left[\sum_{i=1}^k W_i \left(\hat{\theta}_i - \frac{\sum_{i=1}^k W_i \hat{\theta}_i}{\sum_{i=1}^k W_i}\right)^2\right] - \left[\sum_{i=1}^k W_i S_i^2 - \frac{\sum_{i=1}^k W_i^2 S_i^2}{\sum_{i=1}^k W_i}\right]}{\sum_{i=1}^k W_i - \frac{\sum_{i=1}^k W_i^2}{\sum_{i=1}^k W_i}} \tag{24}$$

$$= \frac{\left[\sum_{i=1}^k W_i \hat{\theta}_i^2 - 2\sum_{i=1}^k W_i \left(\hat{\theta}_i \frac{\sum_{j=1}^k W_j \hat{\theta}_j}{\sum_{j=1}^k W_j}\right) + \sum_{i=1}^k \left(W_i \left(\frac{\sum_{j=1}^k W_j \hat{\theta}_j}{\sum_{j=1}^k W_j}\right)^2\right)\right]}{C}$$

$$- \frac{\left[\sum_{i=1}^k \frac{1}{S_i^2} S_i^2 - \frac{\sum_{i=1}^k \left(\frac{1}{S_i^2}\right)^2 S_i^2}{\sum_{i=1}^k \frac{1}{S_i^2}}\right]}{C} \tag{25}$$

$$= \frac{\left[\sum_{i=1}^k W_i \hat{\theta}_i^2 - 2\frac{\sum_{j=1}^k W_j \hat{\theta}_j}{\sum_{j=1}^k W_j} \sum_{i=1}^k W_i \hat{\theta}_i + \left(\frac{\sum_{j=1}^k W_j \hat{\theta}_j}{\sum_{j=1}^k W_j}\right)^2 \sum_{i=1}^k W_i\right]}{C}$$

$$- \frac{\left[\sum_{i=1}^k 1 - \frac{\sum_{i=1}^k \frac{1}{S_i^2}}{\sum_{i=1}^k \frac{1}{S_i^2}}\right]}{C} \tag{26}$$

$$= \frac{\left[\sum_{i=1}^k W_i \hat{\theta}_i^2 - 2\frac{\left(\sum_{i=1}^k W_i \hat{\theta}_i\right)^2}{\sum_{i=1}^k W_i} + \frac{\left(\sum_{i=1}^k W_i \hat{\theta}_i\right)^2}{\sum_{i=1}^k W_i}\right] - (k-1)}{C} \tag{27}$$

$$= \frac{\left[\sum_{i=1}^k W_i \hat{\theta}_i^2 - \frac{\left(\sum_{i=1}^k W_i \hat{\theta}_i\right)^2}{\sum_{i=1}^k W_i}\right] - \nu}{C} \tag{28}$$

$$= \frac{Q - \nu}{C} \tag{29}$$

□

Besides these method-of-moments estimators, more general methods have been proposed that do not assume normal distribution of the means of the primary studies. One of them is the non-parametric maximum likelihood approach by Aitkin [2]. Unfortunately, this estimator has not yet reached the state of practical application and has never been employed in a real meta-analysis.

1.3 Meta-Regression

As explained in the previous section some of the between-studies variance may be explained by study-level covariates. Meta-regression is a tool to assess the relation between one or more study-level covariates and the observed effect size in a study.

1.3.1 The Idea of Meta-Regression

In analogy to classical regression analysis, where the variance of the dependent variable is partitioned into variance that can be explained by regressors and residual variance, meta-regression partitions the between-studies variance of effect sizes into variance that can be explained by study-level covariates and residual variance [66]. As in the estimation of effect sizes it is appropriate to take into account the differences in precision of the studies. This can be accomplished by weighting the studies in meta-regression. With this in mind most of the tools of classical regression can be employed.

1.3.2 Random-Effects vs. Fixed-Effect Models

In principle, both fixed-effect and random-effects models can be used for meta-analysis. A fixed-effect model in meta-analysis means that the true effect size dependent on the covariates is common for all primary studies and the residual variance is sampling error only.

Conversely, the random-effects model assumes that there is no common effect size for all studies, but the true effect sizes of the individual studies follow a distribution whose location parameter is dependent on the covariates [4]. Again, the random-effects model is more realistic than the fixed-effect model in practical meta-analysis since in general, while it can be assumed that the primary studies share common

characteristics necessary for joint analysis, their true effect sizes still reveal differences that cannot be explained by known covariates.

The main difference in calculation between the two types of models manifests in different weighting—in random-effects models, the weights are more homogeneous than in fixed-effect models (cf. section 1.2.3).

1.3.3 Model Assumptions

The principal assumptions of linear regression are independence, normal distribution, and homoscedasticity of residuals. In general, it can be said that all three also apply to meta-regression. Some specific aspects, however, should be highlighted for meta-regression.

Independence, for instance, is certainly necessary in meta-regression. A priori, it seems reasonable to assume it in meta-analysis because the data points represent studies conducted more or less independently from each other. In practice, it is difficult to assess the exact extent of dependencies between studies. For once, it can be taken for granted that researchers conducting primary studies already know the results of previous studies in the same field, possibly imposing a limitation on the independence assumption. Furthermore, multiple studies in the same field may be conducted by the same or related investigators, at the same locations or under circumstances introducing dependencies in any other way.

The assumption of normal distribution of the residuals is also inherited by meta-regression from OLS. In addition, the DerSimonian and Laird estimator for between studies variance also assumes normal distribution. The most important sources of violation of this assumption in practical meta-analysis is probably publication bias. Publication bias is the effect that studies underperforming the expectations or the bulk of available studies are sometimes omitted from publication. This leads to an asymmetric and thus not normal distribution of residuals.

While inhomogeneous variances within the different studies are less problematic in overall effect size estimation, this is not the case in meta-regression. Heteroscedasticity is as much of a problem in meta-regression as it is in ordinary least squares (OLS) regression since the former are directly based on the latter. In addition, in meta-regression studies with low weights (low precision) can gain unrequited influence if their covariate levels are in a range of high between-studies variance. This can occur in covariate ranges where all studies are of low weight, e.g. because only few observations in these ranges can be obtained at the level of individual studies.

1.3.4 Data Quality Issues

In addition to possible violations of model assumptions as discussed above, other aspects of data quality also influence the overall character of the data and in particular the adequacy of certain modeling approaches. Prominent among these are the presence of outliers in general, heterogeneous quality of individual studies in meta-analysis [30], and incorrect data points due to errors in the data acquisition and transmission processes in particular.

Heterogeneity between the quality of individual observations is of particular interest in meta-analysis. Each data point originates from a different primary study and thus may substantially differ from other studies in the quality of the study itself, in the reporting of the results [56], and in the acquisition of data for the purpose of meta-analysis. While some aspects such as differences of within-study variation of effect sizes between the primary studies are taken into account by the meta-analytic methods, others such as loss of precision due to data extraction methods are harder to evaluate. In particular, some meta-analyses use data obtained directly from the authors of the primary studies and data extracted by measuring of figures in published papers with a ruler. This approach is problematic, especially since there is no established method of modeling the resulting differences in data reliability.

Another problem especially pertinent in meta-analysis is data integrity. Typing and related errors can occur at different levels from primary study execution to publishing to retrieving of data by meta-analysts. Since data extraction is often performed by hand from published articles, human error is of special relevance in the context of meta-analysis.

2 Robustness

2.1 General Idea

The estimators discussed, like all statistical models, are based on some model assumptions. Among these are randomness, independence, and normal distribution. On one hand, these assumptions allow to draw valid conclusions, and justify the calculations presented above. On the other hand, under certain circumstances, even small violations of the assumptions can corrupt the conclusions.

In practice, the situation always deviates from the ideal situation to some degree and it is important to check model assumptions. However, the nature of these deviations can be complex. Three cases will be considered in detail.

First, individual data points can simply be wrong (typing errors, transmission errors, measuring errors, etc.). Second, some of the data points, although correct in principle, could be inhomogeneous with respect to the majority of the sample. These are referred to model outliers. Third, the index by which the effect sizes are measured might introduce a bias which cannot simply be removed at the level of meta-analysis, and thus the influence of this bias should be limited by the statistical methodology. In the context of parametric models, these three cases are dealt with in different ways, including the so called robust statistical methods.

The most common way to address the issue of incorrect data points is to identify and discard them from the analysis. However, while there are some unambiguous

situations, in general, on the basis of the reported data alone, it is impossible to clearly label data points as correct or incorrect. Such points may or may not have a distorting influence on the results. If they do, robust methods can contain their influence.

In contrast to the concept of incorrect data points, the concept of model outliers is only meaningful in the context of a specific statistical model. But having chosen a specific model, any outliers with respect to this model are always identifiable. In non-robust models, outliers of this kind might easily distort the results, and thus lead to invalid conclusions. Robust methods are by nature all but immune to this source of distortions.

The third case is hardest to grasp. First of all, whether an index introduces a bias depends on the nature of the problem. As an example, both HAMD score reduction and response rates (encouraged by Melander et al. [45]) are meaningful indices of antidepressant efficacy but might lead to different conclusions (cf. figure 6). Second, the nature of the influence of index choice is very hard to assess. If possible, taking this into account during the modeling phase is the best way of handling it. However, sometimes this cannot be done. Thus, it is desirable to reduce the potential influence of bias by way of robust statistical analysis.

To sum up, in all three cases described, methods of robust statistics help to improve data analysis. In the next sections, outliers occuring in meta-analysis will be grouped into two categories discussed seperately: outliers in the space of actual data points and outliers in the weight space due to inaccurate estimation of their variance. But first, two measures of robustness will be introduced.

2.2 Breakdown Point

The breakdown point of a statistical estimator T is defined as the smallest amount of contamination $y_1, \ldots, y_n$ sufficient to arbitrarily distort its result on any sample $x_1, \ldots, x_n$, with n going to infinity. In formal terms, this can be written as

$$\varepsilon^*(T) = \lim_{n\to\infty} \sup_{\{x_1,\ldots,x_n\}\in\Re^n} \min\left\{\frac{k}{n}; \max_{I\subseteq\{1,\ldots,n\},|I|\leq k} \sup_{\{y_1,\ldots,y_n\}\in\Re^n} |T(z_1,\ldots,z_n)| = \infty\right\} \tag{30}$$

where

$$z_i = \begin{cases} x_i & \text{if } i \notin I \\ y_i & \text{if } i \in I \end{cases} \tag{31}$$

2.3 Influence Function

In order to assess the robustness of an estimator it is important to be able to measure the impact of a small amount of contamination of the sample on the estimator. The tool to employ here is the influence function of an estimator T at a location x which can be defined as

$$\mathrm{IF}(x; T, F) = \lim_{n\to\infty} \frac{T\left(\left(1-\frac{1}{n}\right)F + \frac{1}{n}\delta_x\right) - T(F)}{\frac{1}{n}} \tag{32}$$

where F denotes a distribution function and δ_x the Dirac-measure in x.

2.4 Extensions for Meta-Analytic Settings

Both the breakdown point and the influence function are defined for unweighted data only. For use on weighted data, as in meta-analysis, a generalization is necessary. However, defining the two measures of robustness in this context is far from trivial.

The influence function, at first glance, seems easier to adapt: the infinitesimal contamination of data can be defined as a contamination with infinitesimally small weight, which can be written as

$$\mathrm{IF}(x; T, F) = \lim_{w/n \to 0} \frac{T\left((1 - w/n)F + w/n\delta_x\right) - T(F)}{w/n} \tag{33}$$

In contrast, a generalization of the breakdown point does not immediatly come to mind. Several options are to be considered, including using a proportion of weights instead of a proportion of data points. This definition, however, might lead to problems, since with weighted data, even when the number of observations goes to infinity, it cannot be outruled that the majority of the weights is held by a finite number of observations. Thus, the limit as proposed in equation (30) cannot be interpreted meaningfully on weighted data since it is possible that almost all of the observations carry only negligible weight.

3 Robust Meta-Regression

3.1 The Weighted Ordinary Least Squares Estimator

The dominant approach to meta-regression is based on the Ordinary Least Squares (OLS) estimator to model the linear relationship between a p-dimensional regressor $\mathbf{x} = (1, x_1, \ldots, x_p)^\top$ and a dependent random variable y by

$$y = \boldsymbol{\beta}^\top \mathbf{x} + \varepsilon. \tag{34}$$

with $\varepsilon \sim \mathrm{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$. The regression coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ are estimated from a sample of size n, composed of

$$\mathbf{X} = (\mathbf{1}, \mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p}) = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} \tag{35}$$

and

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}. \tag{36}$$

by minimizing the sum of the squared residuals

$$\sum_{i=1}^{n} r_i^2(\boldsymbol{\beta}) \tag{37}$$

with $r_i^2(\boldsymbol{\beta})$ being the i-th squared residual, i.e.

$$r_i^2(\boldsymbol{\beta}) = y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j. \tag{38}$$

It is thus an optimization of the target function

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} r_i^2(\boldsymbol{\beta}). \tag{39}$$

In meta-regression, a weighted version of this estimator (wOLS) is used, with weight w_i assigned to the i-th observation, leading to

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} w_i r_i^2(\boldsymbol{\beta}). \tag{40}$$

The solution can be calculated using the closed form

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{y} \tag{41}$$

where $\boldsymbol{W}$ is the matrix of weights in the form

$$\boldsymbol{W} = \begin{pmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_n \end{pmatrix}. \tag{42}$$

Despite the wide application of wOLS and its elegant closed form solution, its breakdown point of 0 sparks interest in other, more robust estimators.

3.2 The Least Trimmed Squares Estimator

Following the idea of fitting the model to the majority of the data and downweighting outliers, the LTS estimator [57] is based on the OLS estimator, but instead of fitting the model to all data points, a subset is selected such that the sum of squared residuals of an OLS estimation within this subset is minimal. The target function can be written as

$$\hat{\boldsymbol{\beta}}_{\text{LTS}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{h} (r^2(\boldsymbol{\beta}))_{(i)} \tag{43}$$

Here, only part of the ordered squared residuals $r^2(\boldsymbol{\beta})_{(i)}$ are used for the optimization. More specifically, the regression line is fitted to the h/n proportion of

the data points with the smallest squared residuals. Conversely, a proportion of $\alpha = 1 - h/n$ is ignored in the fitting process, hence the estimator is termed as α-trimmed. The choice of h determines breakdown point and efficiency of the estimator and is bounded by

$$\left\lceil \frac{n}{2} \right\rceil \leq h \leq n \tag{44}$$

where $\lceil \cdot \rceil$ denotes the ceiling function.

There is no explicit formula for computing this estimator, instead, an iterative algorithm is used. More recently, an efficient implementation of LTS regression even for large datasets has been introduced [58], making this estimator a practical choice for robust regression. However, the utility of LTS in meta-regression is limited by the lack of adaptation to weighted data.

3.3 The `meta.lts` Estimator

The `meta.lts` estimator adds to the classic LTS estimator by introducing an implementation of weighting adapted to meta-analysis (cf. code listing `meta.lts`). The basic principle remains the same as in classic LTS, but weights are used in a number of steps of the algorithm. First and most obvious is the use of weighted least squares in the estimation of the regression model on the chosen subset of data in each step. Let

$$\widetilde{r}_i^2(\boldsymbol{\beta}) = w_i r_i^2(\boldsymbol{\beta}) \tag{45}$$

be the weighted squared residuals and $\widetilde{r}_{(i)}^2(\boldsymbol{\beta})$ the ordered weighted squared residuals. The target function of `meta.lts` is thus

$$\hat{\boldsymbol{\beta}}_{\texttt{meta.lts}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{h} \tilde{r}^2_{(i)}(\boldsymbol{\beta}) \tag{46}$$

Apparently, the only difference of the final version of `meta.lts`, as presented here, to the classic LTS estimator is the use of the weights w_i in the target function. Still, the use of weighting in LTS estimation is not limited to this. There are several other parts of the algorithm where the use of weighting was considered and, ultimately, implemented. First, it was chosen to trim the largest weighted rather than unweighted squared residuals. Second, weights were involved in determining the proportion of data points to be trimmed, i.e., the choice of h, which will be explained in section 3.3.2. And third, weights were taken into account to improve the suitability of the initial sample as a starting point for the iterative search of the optimal solution.

3.3.1 Weighted Residual Ordering

A basic aspect of the `meta.lts` estimator is that, just as in the case of the classic LTS estimator, the data points with the highest squared residuals are excluded from the estimation of the regression model. In the context of weighted data, the question arises whether the raw squared residuals or rather the weighted squared residuals should be ordered for subsequent trimming.

In practice, using weighted squared residuals means that the probability to include studies with higher weights is increased compared to the model using raw squared residuals. At first glance, this might seem to introduce a bias towards studies with lower weights. However, it merely counteracts the tendency of outliers with high weights to mask their influence on the model estimation.

Furthermore, using raw squared residuals would introduce an inconsistency between subset selection and model estimation on the selected subset by using different

target criteria. Thus, using weighted squared residuals for the ordering leads to a more consistent overall model.

3.3.2 Weighted Trimming

When it comes to excluding a certain number of data points, either a proportion of absolute numbers of studies or of weights could be used. As a tentative approach, the former was implented (cf. code listing `meta.lts.unweighted`) with a choice of h as in LTS presented in formula (44). Since in meta-analysis, weights are regarded as proportional to the evidence provided by the individual studies excluding a fixed number of studies can lead to high variation in the proportion of available evidence actually taken into account. Taking this into consideration, it might be more appropriate to use a fixed amount of evidence, regardless of how many data points this corresponds to. The corresponding choice to trim away a certain proportion α of weights was implemented in the final version of `meta.lts` (as presented in code listing `meta.lts`). In this version, termed α-trimmed `meta.lts`, h was defined as

$$h = \max \left\{ j : \sum_{i=1}^{j-1} \widetilde{w}_{(i)} < (1-\alpha) \sum_{j=1}^{n} w_j \right\} \tag{47}$$

where $\alpha \in [0, 0.5]$ is the trimming parameter and $\widetilde{w}_{(i)}$ the weight corresponding to $\widetilde{r}^2_{(i)}$, that is, the i-th element of the weight vector ordered by ascending weighted squared residuals.

How exactly to determine the amount of evidence provided by each observation depends on whether a fixed-effect or random-effects model is used, as explained in section 1. Both models can be incorporated into a meta-regression using `meta.lts` by using the appropriate weights in the analysis.

3.3.3 Initialization

Being an iterative algorithm, the initialization of `meta.lts` can have a profound influence on the result. As a consequence, calculations are performed using different starting configurations to increase the probability of finding the optimal solution. Further improvement can be achieved by choosing them to be contextually adequate instead of purely random.

In the case of `meta.lts`, to find a conductive initial regression model, $p + 1$ points (p being the number of covariates) are selected randomly from the whole dataset using weighted sampling. Through these points, a first regression hyperplane is drawn and the resulting model is used for calculating initial residuals for all data points. Subsequently, a first subset is chosen using the methods described above.

4 Analysis and Simulations

4.1 A Real Dataset

A recent meta-analysis on the efficacy of anti-depressant medication analyzed 35 studies of different drug types. The dataset of this paper was of special interest because it included both published and unpublished studies, relevant data were available without explicit request and the presence of outliers made it particularily suited for application of robust methods.

4.1.1 Structure

The dataset consists of 35 studies conducted on the efficacy of 4 drugs of 2 different types: the Selective Serotonin Reuptake Inhibitors (SSRI) Fluoxetine and Paroxetine with 5 and 16 primary studies respectively and the Serotonin-Norepinephrine Reuptake Inhibitors (SNRI) Venlafaxine and Nefazodone with 6 and 8 primary

studies. For each of the primary studies, available data included number of participants, mean Hamilton Depression Rating Scale (HDRS or HAMD) baseline score, mean HAMD change, Cohen's d, and a 95% confidence interval for the Cohen's d, all of those for both drug and placebo groups seperately.

The goal of the study was to establish a relation of baseline severity and antidepressant efficacy. In this context, it is worth noting that the mean HAMD baseline scores for both drug and placebo groups ranged from approximately 23 to 30 except for one outlying Fluoxetine study with a mean baseline score of 17. The mean change in HAMD scores varied in the range of 3 to 11 in the placebo groups and 6 to 14 in the drug groups. The numbers of participants in the placebo groups were between 10 and 163 and those in the drug groups between 13 and 403. For more details cf. tables 1, 2, and 3).

4.1.2 Methods Used

Methods applied to the example dataset included several linear models, using both fixed-effect and random-effects approaches. Analyses were performed on the complete dataset as well as on subsets. Two different effect size measures for the dependent variable were compared. The estimation of a mean effect size was conducted using a random-effects model and the result illustrated by a forest plot [3].

4.2 Simulated Data

4.2.1 Different Groups

A sample dataset consisting of three distinct groups was generated. One group consisting of 40 observations with high weights constituted the most meaningful subset of the data and were considered to represent the true relationship to be

modeled. Two different groups of 30 lower weighted observations each were added as outliers. The important point was that the combined weights of the outlier groups did not exceed those of the main group for the `meta.lts` estimator to detect the latter as representative for the relationship to be modeled.

This dataset was utilized for two purposes. Firstly, the dependency of `meta.lts` results on the type of initialization algorithm was assessed. Secondly, possible improvements of initialization by weighted sampling was examined. Lastly, a comparison was established between OLS, LTS and `meta.lts` regression performance.

4.2.2 Breakdown Point

To illustrate the conceptual differences between the breakdown points of OLS, classic LTS and `meta.lts` regression, datasets consisting of a group of data points with homogeneous weights and three points with excessively high weights were generated. On this dataset, OLS, LTS, `meta.lts` without weighted trimming, `meta.lts` without weighted sampling, and final `meta.lts` models were calculated. In particular, the effect of the outliers in weight space was explored.

4.2.3 Publication Bias

Finally, simulations to assess the practical problem of publication bias were performed. As explained above, publication bias describes the effect that underperforming studies sometimes remain unpublished. This can lead to an overestimation of the studied effects.

Multiple scenarios were considered. In all of them, 15 points with large weights and 85 points with low weights were generated according to a linear model where the residual variance of each point was proportional to its weight. To model publication bias, data points with negative residuals were considered as unpublished and thus

excluded from the model with a probability increasing with the absolute value of their residuals.

The scenarios were set apart by using different true linear model parameters as well as introducing varying amounts of heteroscedasticity into the model. The latter was studied because of its frequent occurence in real datasets. First, homoscedastic datasets were simulated such that the publication bias constituted the only violation of the meta-analytic model assumptions. Second, heteroscedasticity was introduced by increasing the residual variance proportially to the independent variable. Third, the residual variance was set to be inversely proportional to the independent variable. And last, it was proportional to the squared values of the independent variable. All of these were simulated with various slopes—both positive and negative—of the true linear relationship.

On the resulting datasets, the performance of weighted OLS, LTS and `meta.lts` were compared.

ID	Label	Drug	Difference Between Drug and Placebo Groups			
			HAMD Score Difference		Cohen's d Difference	
			Value	SD	Value	SD
1	19	Fluoxetine	7.00	12.31	0.81	0.41
2	25	Fluoxetine	-1.60	12.18	-0.20	0.40
3	27	Fluoxetine	2.60	13.51	0.27	0.13
4	62 (mild)	Fluoxetine	0.07	8.00	-0.03	0.18
5	62 (moderate)	Fluoxetine	3.13	11.11	0.41	0.18
6	203	Venlafaxine	4.50	11.56	0.55	0.15
7	301	Venlafaxine	4.45	11.12	0.57	0.26
8	302	Venlafaxine	3.02	14.47	0.29	0.22
9	303	Venlafaxine	0.21	11.26	0.03	0.22
10	313	Venlafaxine	1.51	11.64	0.19	0.18
11	206	Venlafaxine	9.40	14.85	1.02	0.27
12	03A0A-003	Nefazodone	1.57	12.04	0.23	0.22
13	03A0A-004A	Nefazodone	0.00	10.76	0.00	0.19
14	03A0A-004B	Nefazodone	1.90	11.46	0.24	0.19
15	030A2-004/0005	Nefazodone	0.16	10.88	0.04	0.23
16	030A2-0007	Nefazodone	2.50	12.37	0.31	0.22
17	CN104-002	Nefazodone	2.60	11.24	0.33	0.25
18	CN104-005	Nefazodone	4.00	11.22	0.50	0.21
19	CN104-006	Nefazodone	1.10	10.52	0.14	0.22
20	01-001	Paroxetine	3.00	11.43	0.37	0.45
21	02-001	Paroxetine	5.50	13.66	0.58	0.25
22	02-002	Paroxetine	5.10	12.48	0.57	0.30
23	02-003	Paroxetine	2.50	14.75	0.24	0.29
24	02-004	Paroxetine	5.10	9.60	0.75	0.36
25	03-001	Paroxetine	6.10	9.59	0.91	0.31
26	03-002	Paroxetine	1.80	9.94	0.26	0.29
27	03-003	Paroxetine	-0.10	11.87	-0.01	0.30
28	03-004	Paroxetine	3.70	11.04	0.47	0.31
29	03-005	Paroxetine	5.90	14.21	0.58	0.26
30	03-006	Paroxetine	6.10	11.53	0.74	0.27
31	PAR 09	Paroxetine	0.90	10.11	0.14	0.20
32	UK 06	Paroxetine	-0.20	9.70	0.14	0.41
33	UK 12	Paroxetine	2.40	10.74	0.37	0.55
34	UK 09	Paroxetine	4.30	14.33	0.31	0.37
35	PAR 07	Paroxetine	2.20	15.50	0.21	0.59

Table 1: Overview of the example dataset.

ID	N	Baseline	HAMD Change		Cohen's d			
		HAMD Score	Mean	SD	Score	SD	Confidence Interval	
1	22	28.6	12.5	8.68	1.44	0.33	0.79	2.09
2	18	26.2	7.2	8.67	0.83	0.30	0.24	1.41
3	181	27.5	11	9.57	1.15	0.10	0.96	1.34
4	299	17	5.89	5.77	1.02	0.07	0.88	1.16
5	297	24.3	8.82	7.81	1.13	0.07	0.98	1.27
6	231	25.6	11.2	8.18	1.37	0.09	1.19	1.55
7	64	25.4	13.9	7.85	1.77	0.20	1.36	2.17
8	65	25	11.9	10.26	1.16	0.17	0.84	1.49
9	69	23.6	10.1	7.95	1.27	0.16	0.94	1.59
10	227	25.7	11	8.21	1.34	0.09	1.16	1.52
11	46	28.2	14.2	9.79	1.45	0.22	1.02	1.89
12	101	25.4	9.57	8.32	1.15	0.13	0.90	1.41
13	153	23.4	8.9	7.61	1.17	0.11	0.97	1.38
14	156	25.3	11.4	8.09	1.41	0.11	1.18	1.63
15	74	23.4	10	7.63	1.31	0.16	0.99	1.63
16	175	25.7	12.3	8.66	1.42	0.11	1.20	1.63
17	57	23.3	10.8	7.94	1.36	0.19	0.99	1.73
18	86	24.5	12	7.95	1.51	0.16	1.20	1.83
19	80	23.8	10	7.46	1.34	0.16	1.03	1.65
20	24	28	13.5	8.08	1.67	0.34	0.99	2.34
21	51	26.6	12.3	9.61	1.28	0.19	0.89	1.66
22	36	25	10.9	8.86	1.23	0.23	0.78	1.69
23	33	28.6	9.7	10.43	0.93	0.21	0.50	1.35
24	36	28.9	12.7	6.79	1.87	0.29	1.29	2.44
25	40	24.9	10.8	6.75	1.60	0.25	1.11	2.09
26	40	24.9	8	7.02	1.14	0.21	0.72	1.55
27	41	25.7	9.9	8.39	1.18	0.21	0.76	1.59
28	37	27.6	10.4	7.82	1.33	0.23	0.86	1.79
29	40	26.1	10	10.10	0.99	0.20	0.60	1.39
30	39	29.7	9.1	8.20	1.11	0.21	0.69	1.52
31	403	25.2	9.1	7.11	1.28	0.07	1.15	1.41
32	19	23.7	6	6.19	0.97	0.31	0.38	1.57
33	19	22.8	9.1	7.40	1.23	0.33	0.57	1.88
34	20	26.8	8.8	11.00	0.80	0.28	0.26	1.35
35	13	30.5	13.1	10.92	1.20	0.42	0.38	2.03

Table 2: Data for the drug groups in the primary studies.

ID	N	Baseline	HAMD Change		Cohen's d			
		HAMD Score	Mean	SD	Score	SD	Confidence Interval	
1	24	28.2	5.5	8.73	0.63	0.24	0.17	1.10
2	24	25.8	8.8	8.54	1.03	0.27	0.50	1.56
3	163	28.2	8.4	9.55	0.88	0.09	0.69	1.06
4	56	17.4	5.82	5.54	1.05	0.17	0.71	1.38
5	48	24.3	5.69	7.90	0.72	0.17	0.39	1.05
6	92	25.3	6.7	8.17	0.82	0.12	0.58	1.06
7	78	24.6	9.45	7.88	1.20	0.15	0.91	1.50
8	75	24.4	8.88	10.21	0.87	0.14	0.60	1.14
9	79	24.6	9.89	7.98	1.24	0.15	0.94	1.54
10	75	25.4	9.49	8.25	1.15	0.15	0.85	1.45
11	47	28.6	4.8	11.16	0.43	0.16	0.12	0.74
12	52	25.9	8	8.70	0.92	0.17	0.59	1.26
13	77	23.5	8.9	7.61	1.17	0.15	0.88	1.47
14	75	25	9.5	8.12	1.17	0.15	0.87	1.47
15	70	24	9.84	7.75	1.27	0.16	0.94	1.59
16	47	26.4	9.8	8.83	1.11	0.19	0.74	1.49
17	57	23.1	8.2	7.96	1.03	0.17	0.70	1.36
18	90	23.3	8	7.92	1.01	0.13	0.75	1.27
19	78	23.5	8.9	7.42	1.20	0.15	0.90	1.49
20	24	27.4	10.5	8.08	1.30	0.30	0.71	1.88
21	53	25.9	6.8	9.71	0.70	0.16	0.39	1.01
22	34	24.9	5.8	8.79	0.66	0.19	0.27	1.04
23	33	28.9	7.2	10.43	0.69	0.20	0.29	1.08
24	38	27.3	7.6	6.79	1.12	0.21	0.70	1.54
25	38	24.8	4.7	6.81	0.69	0.19	0.33	1.06
26	40	25.6	6.2	7.05	0.88	0.19	0.50	1.26
27	42	27	10	8.40	1.19	0.21	0.78	1.60
28	37	27	6.7	7.79	0.86	0.20	0.46	1.25
29	42	26.8	4.1	10.00	0.41	0.16	0.08	0.73
30	37	28.7	3	8.11	0.37	0.17	0.02	0.71
31	51	24.5	8.2	7.19	1.14	0.18	0.77	1.50
32	22	24.2	6.2	7.47	0.83	0.27	0.31	1.35
33	10	22.3	6.7	7.79	0.86	0.44	0.00	1.73
34	21	25.5	4.5	9.18	0.49	0.24	0.01	0.97
35	12	28.3	10.9	11.01	0.99	0.41	0.19	1.79

Table 3: Data for the placebo groups in the primary studies.

Results

5 Real Dataset

5.1 Extension of Original Analysis

5.1.1 Quadratic Model

The first model was calculated along the lines taken by Kirsch et al. [36] and is similar to model 1b in the paper. It includes both linear and quadratic terms of the baseline HAMD severity, the group as a dummy variable (0 for placebo vs. 1 for drug) and interaction between group and baseline severity. The results are summarized in R output 1. The values calculated do not exactly correspond to those in the paper, one reason being omissions in the reporting by Kirsch et al. In addition, inspection of figure 1 in comparison with the corresponding figure in the paper showed a discrepancy between the data reported in the table 1 and the points plotted in figures 2 and 3 by Kirsch et al. [36]. The points in question are highlighted in figure 1 by the use of different colors.

Besides the differences noted, figure 1 qualitatively replicates the results reported in the paper, in particular that the "improvement from baseline operated as a ∩-shaped curvilinear function in relation to baseline severity, with those at the lowest and highest levels experiencing smaller gains, whereas those in-between experienced larger gains" [36].

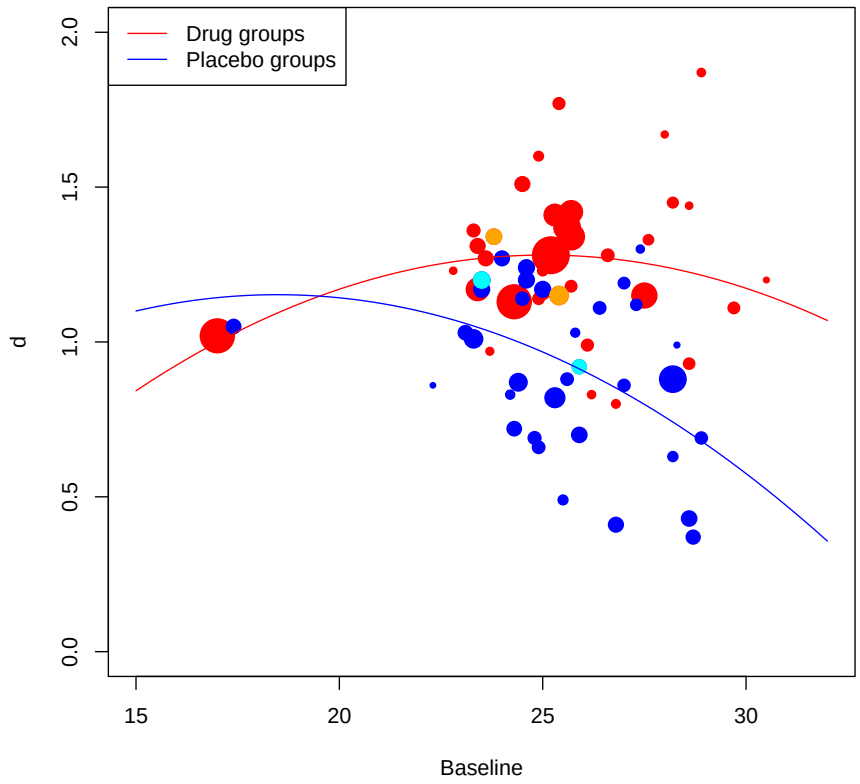

Figure 1: Mean standardized improvement as a function of initial severity and treatment group as reported by Kirsch et al. [36] in figure 2. Blue and light blue points represent placebo groups, red and orange points represent drug groups. The area of the plotted points is proportional to their weight in the fixed effect model. Light blue and orange indicate points plotted according to data as reported in table 1 by [36] that do not correspond exactly to points plotted in figure 2 by [36]. The red and blue lines represent prediction of d score for drug and placebo groups based on the model presented in R output 1.

5.1.2 Confidence Intervals

In figure 1, high variance of Cohen's d scores (cf. section 5.2.2) can be observed, in particular when compared to the variance explained by the model. Figure 2 shows the same model as figure 1 with added 95% confidence bands. Note that

R output 1 Quadratic model analogous to model 1b in Kirsch et al. [36].

```
lm(formula = d ~ group + baseline + I(baseline^2) + group:baseline,
    weights = w)

Residuals:
     Min       1Q   Median       3Q      Max
-2.70312 -0.93918  0.04522  1.04329  2.40091

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.334039   1.270684  -0.263   0.7935
group          -1.113263   0.515794  -2.158   0.0346 *
baseline        0.160888   0.101398   1.587   0.1174
I(baseline^2)  -0.004353   0.002060  -2.113   0.0384 *
group:baseline  0.057050   0.020442   2.791   0.0069 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.217 on 65 degrees of freedom
Multiple R-squared: 0.5208,     Adjusted R-squared: 0.4913
F-statistic: 17.66 on 4 and 65 DF,  p-value: 7.41e-10
```

considering the confidence bands, the direction of the true curve at the extreme ends of the baseline severity scores cannot clearly be identified. Furthermore, at baseline severity levels of 20–30, where all but one studies are located, the confidence bands show that both true curves cannot be distinguished from straight lines, in particular from a horizontal line in case of the drug group.

Following this observation, the next model considered was a purely linear model as seen in figure 3. In analogy to figure 3 by Kirsch et al., a linear model was also fitted to all data points but the one with a mean baseline HAMD score of 17. Both models can be compared in figure 3, the model incorporating all data points being shown in orange and dark blue and the model without the outlier (analogous to figure 3 in the paper) in red and light blue. Both models were plotted with their respective confidence bands. These do not support any difference between the two models on the placebo groups. Considering the drug groups, the regression line

of the model without the outlier is more horizontal and lies within the confidence bands of the model that includes the outlier for HAMD baseline scores of 23 to over 30 and outside the confidence bands for baseline scores under 23.

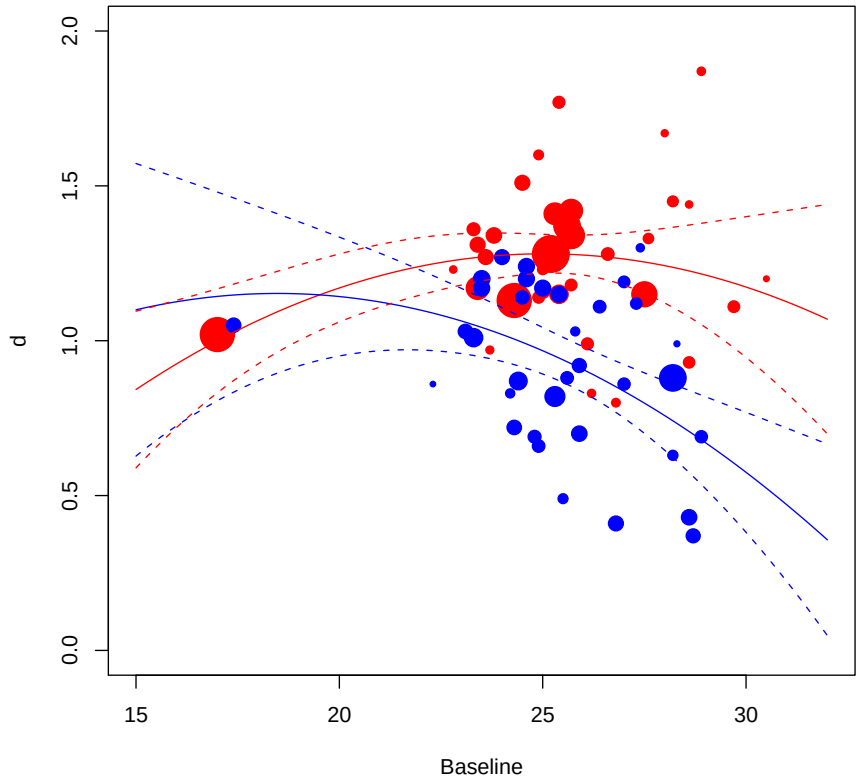

Figure 2: Same as figure 1 with 95% confidence bands added.

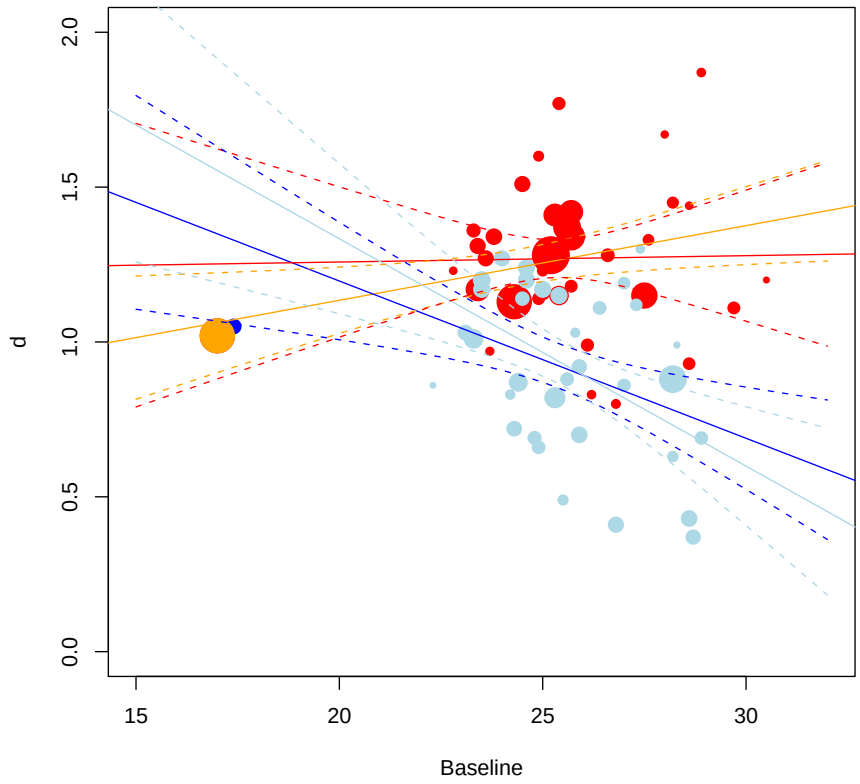

Figure 3: Linear least squares regression of mean standardized improvement as a function of initial severity and treatment group using fixed effect weights. Red and light blue lines correspond to regression excluding the outlying study with the baseline HAMD score of 17. Note that the red and light blue lines correspond to the regression lines in figure 3 by [36] with minor variations due to the data inconsistencies highlighted in figure 1. The area of the plotted points is proportional to their weight in the fixed effect model.

5.1.3 Random Effects

The next step to a more formally correct model was to apply random effects weighting instead of the fixed effects weights used in the previous section. The results are summarized in the R output 2 and in figure 4. Compared to the fixed effect model, in the random effects model the weights are smaller and more homogeneous.

R output 2 Linear model using random effects weights.

```
lm(formula = d ~ group + baseline + group:baseline,
    weights = w.random.effects)

Residuals:
      Min        1Q    Median        3Q       Max
-1.635963 -0.500561  0.004967  0.606901  1.706309

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.86720    0.39693   2.185  0.03246 *
group           1.29473    0.59584   2.173  0.03338 *
baseline        0.01574    0.01566   1.005  0.31854
group:baseline -0.06474    0.02344  -2.762  0.00744 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8072 on 66 degrees of freedom
Multiple R-squared: 0.4268,     Adjusted R-squared: 0.4007
F-statistic: 16.38 on 3 and 66 DF,  p-value: 4.621e-08
```

The regression lines of the placebo groups are similar to those of the fixed effect model, while those of the drug groups are slightly more horizontal. For both groups, the confidence bands are wider. As a consequence, for placebo and for drug groups, the regression lines of the models with and without the outlier remain within the confidence bands of each other. Therefore, there is no statistical evidence of the models being different.

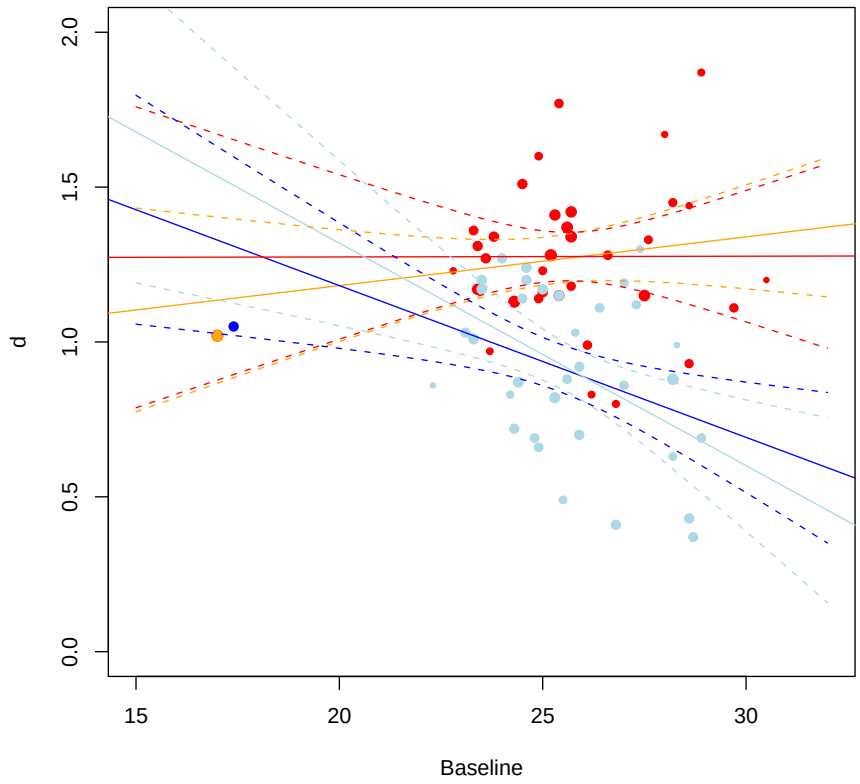

Figure 4: Similar to figure 3, but using weights according to a random effects model. The area of the plotted points is proportional to their weight in the random effects model. Note that in contrast to the fixed effect model as shown in figure 3 the regression line for the drug group using the model without the outlier remains well within the confidence band of the model including the outlier, illustrating the smaller influence of large studies in the random effects model.

5.2 Other Analyses

5.2.1 Forest Plot

The forest plot, a commonly used method to illustrate the estimation of the mean effect size, in this case (see figure 5) shows the differences between the drug and placebo groups of each study. Here, due to the extremely high variance estimates it

cannot support a significant difference between drug and placebo treatment. The high variance observed is inherent to the effect size measure employed and limits the precision of any assessment of difference between groups.

Other effect size measures, in particular response rates (cf. Melander et al. [45]), might yield better results. However, response rates could not be used in the analysis here since this would require either their direct publication or the availability of individual patient data (IPD).

Mean effect size	Standard deviation	95% confidence interval	
2.70	1.93	-1.10	6.48

Table 4: Mean effect size and scale estimation of the random effects model.

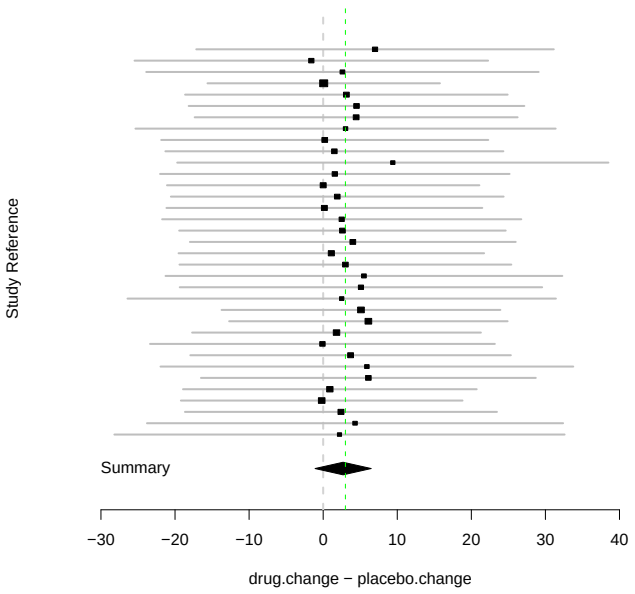

Figure 5: Forest plot of the differences in HAMD change between drug and placebo groups. Grey lines represent 95% confidence intervals based on the (conservative) estimation of the variance of the difference as the sum of the variances of the HAMD changes of the two groups. Note that the weights used correspond to the weights of the fixed effect model since the between-studies variance pales in comparison to the large within-study variances. The DerSimonian and Laird estimator for τ^2 yields a negative result and T^2 is thus set to 0.

5.2.2 Different Effect Size Measures

The two effect size measures reported were raw HAMD scores and Cohen's d. Figure 6 shows a direct comparison between the models resulting from the use of these different effect size measures. Despite the close relationship between the two, Cohen's d being the normalized mean HAMD score change, the models results show different relationships between the baseline HAMD scores and their respective effect sizes within each group.

While in the drug groups Cohen's d was constant over all observed baseline scores, raw HAMD change increased at higher levels of baseline. On the other hand, in the placebo groups, high baseline scores lead to lower values of d scores, but had no influence on raw HAMD change. Although these differences occur only as a consequence of data presentation, they might induce discordant interpretation of the relation between result of treatment and baseline HAMD score and effect size choice is thus to be handled with care.

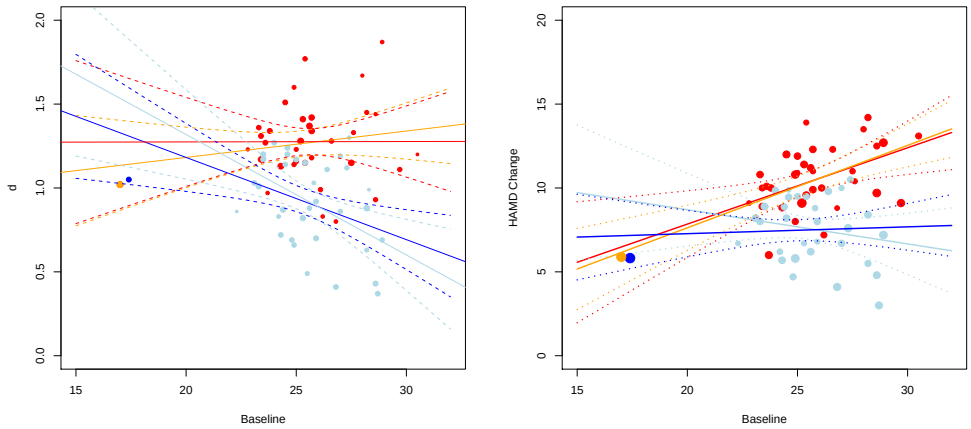

Figure 6: Linear least squares regression of mean standardized improvement (left) and absolute improvement in HAMD scores (right) as a function of initial severity and treatment group using random effect weights. Note how the percieved influence of baseline severity changes substantially when using different effect size measures. Regression lines and 95% confidence bands for models including and excluding the outlying study.

5.2.3 Groupwise Analysis

In general, data aggregation is a factor that can strongly influence the results of statistical analysis. In particular, it is problematic when aggregating data to a level that does not allow for meaningful interpretation anymore. Accordingly, such data cannot be used as a covariate in regression analysis.

In regression, aggregating data into groups is particularily dangerous when the variance within the groups is large in relation to the variance between groups. When doing so, most of the relevant information on the covariance between independent and dependend variables is effectively eliminated. Fitting a meaningful regression model is thus made practically impossible.

The effect of the above can be high, even to the point of completely reversing the results. To illustrate this, simulated data in figure 7 show a reversal of slope in regression induced by data aggregation.

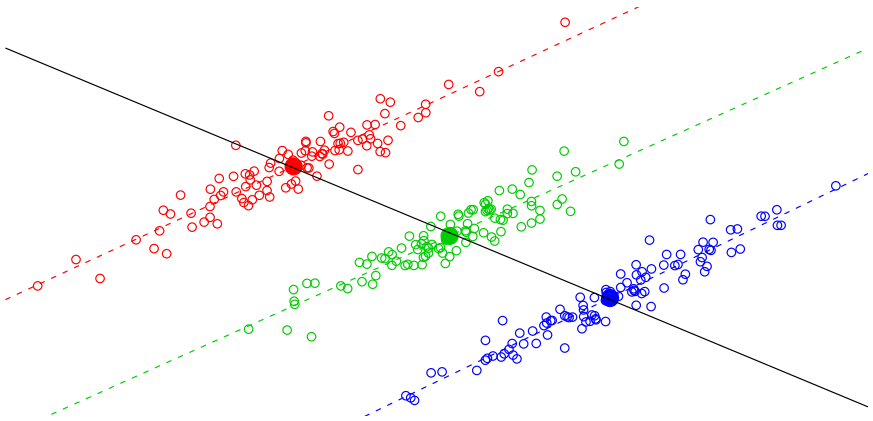

Figure 7: Simulated data demonstrating the potential reversal of percieved trend when aggregating data. The dashed red, green and blue lines represent the regression lines within the respective subgroups while the solid black line corresponds to the regression of the aggregated data (the solid red, green and blue points represent the respective group means). Note that while all of the subgroups show the same upwards trends, the aggregation reverses the slope.

A related problem is that of jointly modeling heterogeneous groups of data points as being drawn from one homogeneous population. Information on differences between groups is thus lost, possibly leading to corrupted regression results. As an example, in figure 8 the different models resulting from regression grouped by active agent of the study can be seen. Note that in contrast to the pooled regression model in figure 4 where the regression line of the drug group is horizontal, the individual regression lines of drug groups in 8 all have positive slopes.

Considering figure 8, one might be tempted to identify a marked distinction between the regression lines of the SSRI and SNRI subgroups. It should be noted, however, that on one hand these differences are not significant, and on the other hand, to draw inferences about patient level covariates in a study level analysis is principally flawed (this will be discussed in more detail in section 7.2). In short, the relationship observed here should be considered as coincidental.

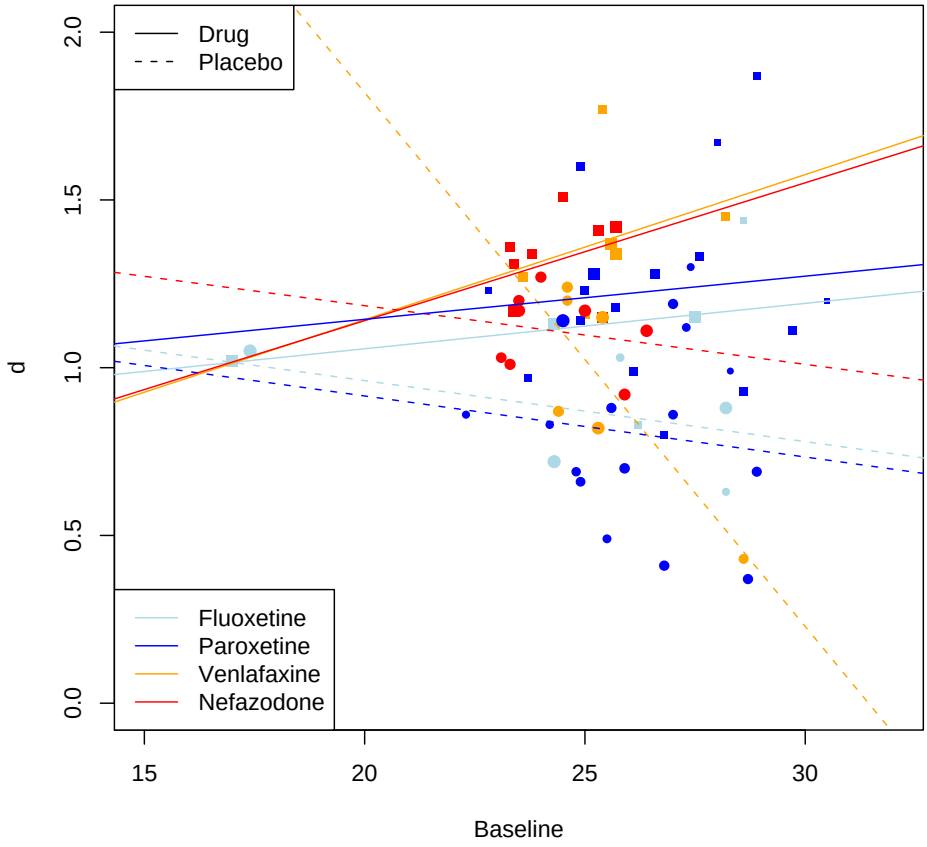

Figure 8: Individual regression lines for subgroups of studies using the same active agent. Squares represent drug groups, circles represent placebo groups, their respective sizes being proportional to their weights in random effect analysis. Solid lines represent regression of drug groups, dashed lines the corresponding regression of placebo groups.

6 Simulations

6.1 Three Groups

6.1.1 Initialization

The first simulation result concerns the assessment of the influence of weighting on the initialization of `meta.lts`. The black lines in figure 9 show the initial models of all iterations, the green line shows the final model. The plot on the left shows the result using unweighted sampling, the plot on the right weighted sampling. Using weighted sampling, a greater proportion of the initial models lie in the vicinity of the final model.

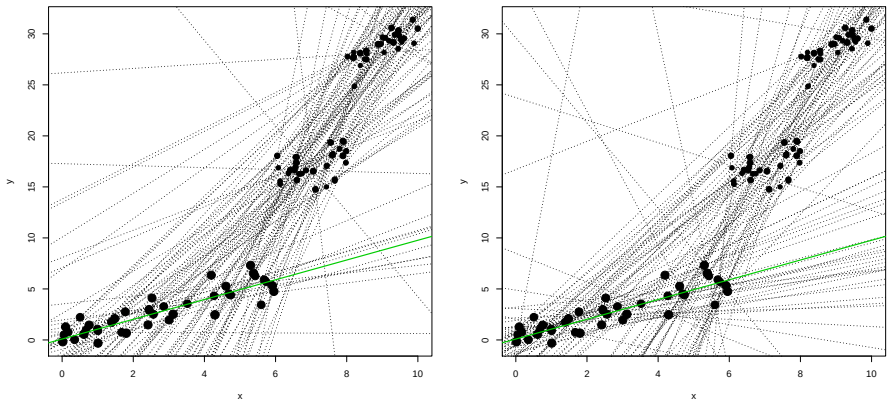

Figure 9: Improvement of starting configurations using weighted sampling (right) compared to unweighted sampling (left).

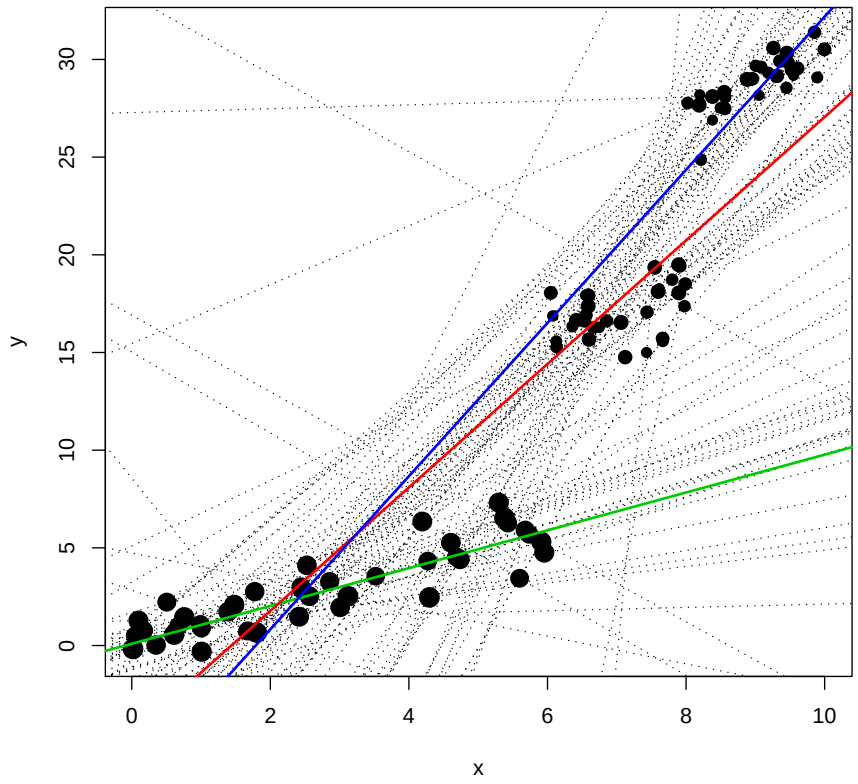

Figure 10: Comparison of OLS (red), LTS (blue) and `meta.lts` (green) regression models on simulated data.

6.1.2 Three Estimators

Figure 10 compares final results of OLS, LTS (with $h = n/2$) and `meta.lts` (with $\alpha = 1/2$) regression on a simulated dataset comprising three groups with different weights (cf. section 4.2.1). The `meta.lts` regression line is closest to the group with high weights, while the other two are drawn towards the groups with smaller weights.

6.2 Weight Space Outliers

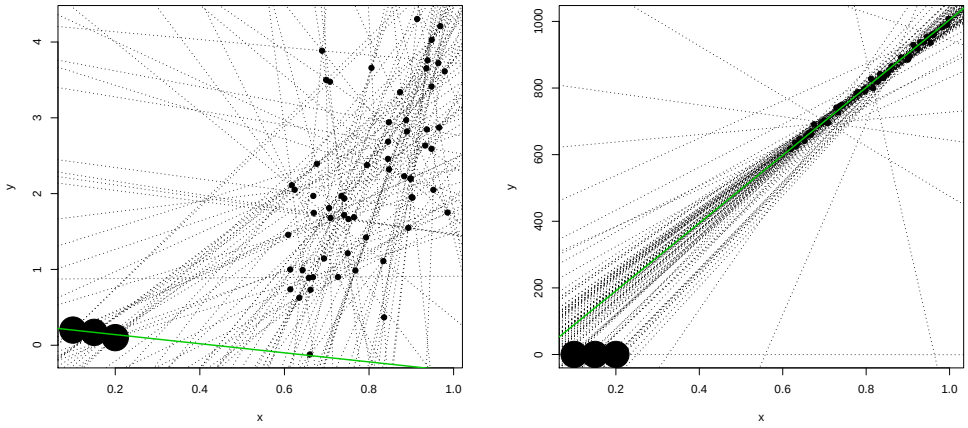

Figure 11: Illustration of the complexity of the concept of the breakdown point on weighted data. The `meta.lts` regression line (green) is drawn to the three heavyweight points with the majority of the weights in the left plot, and to the majority of data points in the right plot.

The two plots in figure 11 show the effect of outliers in weights space on the `meta.lts` regression estimate in cumulated datasets consisting of a group of data points with homogeneous weights and three points with excessively high weights. The `meta.lts` regression line in the left plot drawn in green is fitted only to the three heavyweight points, downweighting the rest of the data. Conversely, in the right plot, where only the location of the 60 data points with a cumulative weight of less than 1/1000 of the total weights has changed, the regression line is fitted to the 60 points with a very small proportion of weights. Furthermore, it can be seen that at least one of the initial regression lines was drawn through two of the three heavyweight points, thus ruling out the possibility of the result being purely due to initialization issues.

6.3 Publication Bias

6.3.1 Homoscedastic Case

Three types of models were fitted to a simulated dataset of 100 data points with varying weights and residual variances proportional to these weights. Representing publication bias, observations with negative residuals were excluded from the model fitting with probability of omission increasing with higher absolute residual values.

All three types of models that were fitted show essentially the same estimates of intercept and slope (see figure 12). The two estimates using weights—OLS and `meta.lts`—had approximately the same variance of their slope estimate while the LTS estimator which did not use the weights had a higher variance. The median of the estimates of all three types of models on 1000 simulated datasets matches the true slope (see figure 13).

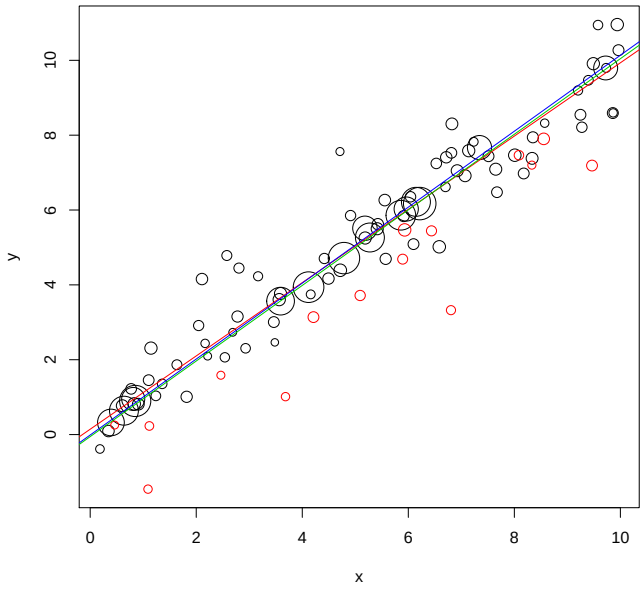

Figure 12: Comparison of OLS (red line), LTS (blue) and `meta.lts` (green) estimation on simulated homoscedastic data with publication bias. The red points are considered as unpublished and are thus not included in the fitting of the regression models.

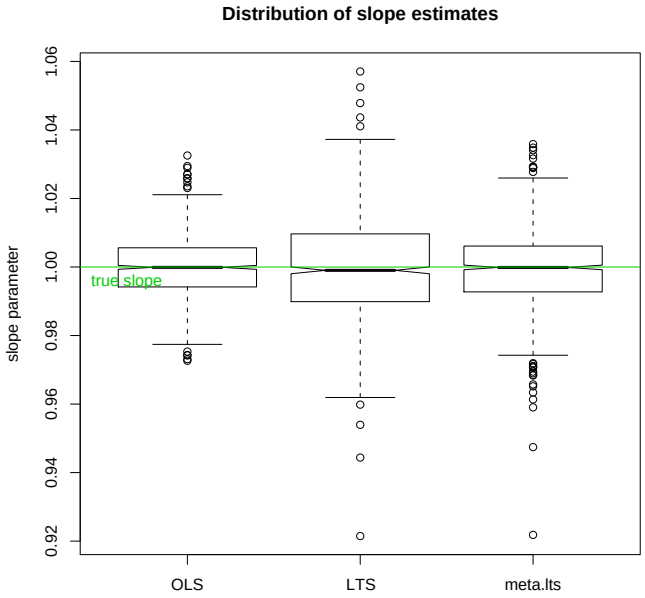

Figure 13: Comparison of OLS, LTS and `meta.lts` estimates of the slope vs. true slope on simulated homoscedastic data with publication bias.

6.3.2 Heteroscedastic Case

In analogy to the simulation of homoscedastic data, 1000 datasets for the heteroscedastic case were also created. Heteroscedasticity was induced by increasing residual variance proportionally to the independent variable. Again, the three types of models were fitted. In contrast to the homoscedastic case, results varied between the different models.

As expected from a theoretical point of view, the OLS estimator is the most strongly influenced by the bias (figure 14). Its slope estimate is substantially higher than the true slope in the great majority of the simulated datasets (figure 15). The LTS estimator also overestimated the slope, albeit less so. Still, more than three quarters of the simulated datasets resulted in an estimate above the true slope. The `meta.lts` estimator, finally, most consistently delivered the estimates closest

to the true slope. Even so, the deviation of the median of the estimates from the true slope was statistically significant.

To examine the influence of different patterns of heteroscedasticity, several other simulations were performed. A selection of four cases is presented (figure 16). In the top left subfigure, the case of linearily increasing residual variance as described above is shown, the true slope was set to 0.5 instead of 1. To the right, a negative slope was considered. In the bottom left subfigure, instead of variance increasing with higher values of the independent variable, it was simulated as increasing with lower values. The last subfigure, at the bottom right, shows the case of residual variance increasing proportionally to the square of the independent variable.

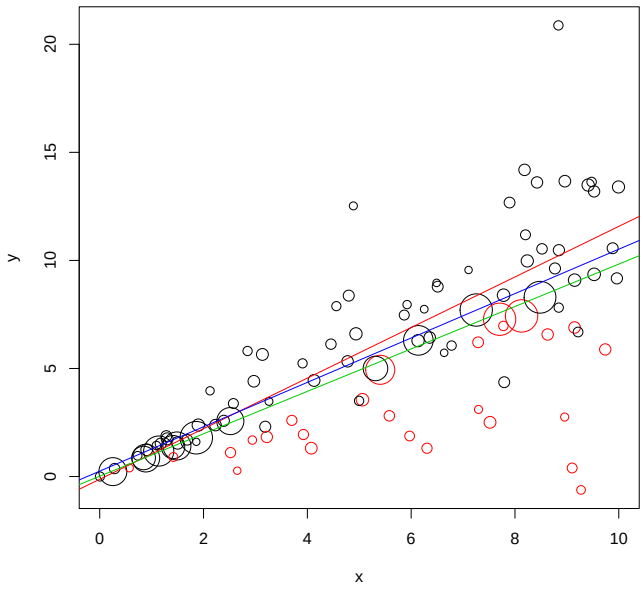

Figure 14: Comparison of OLS (red line), LTS (blue) and `meta.lts` (green) estimates on simulated heteroscedastic data with publication bias. The red points are considered as unpublished and are thus not included in the fitting of the regression models.

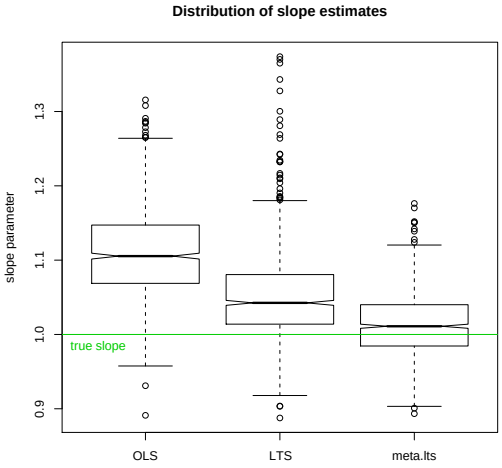

Figure 15: Comparison of OLS, LTS and `meta.lts` estimates of the slope vs. true slope on simulated heteroscedastic data

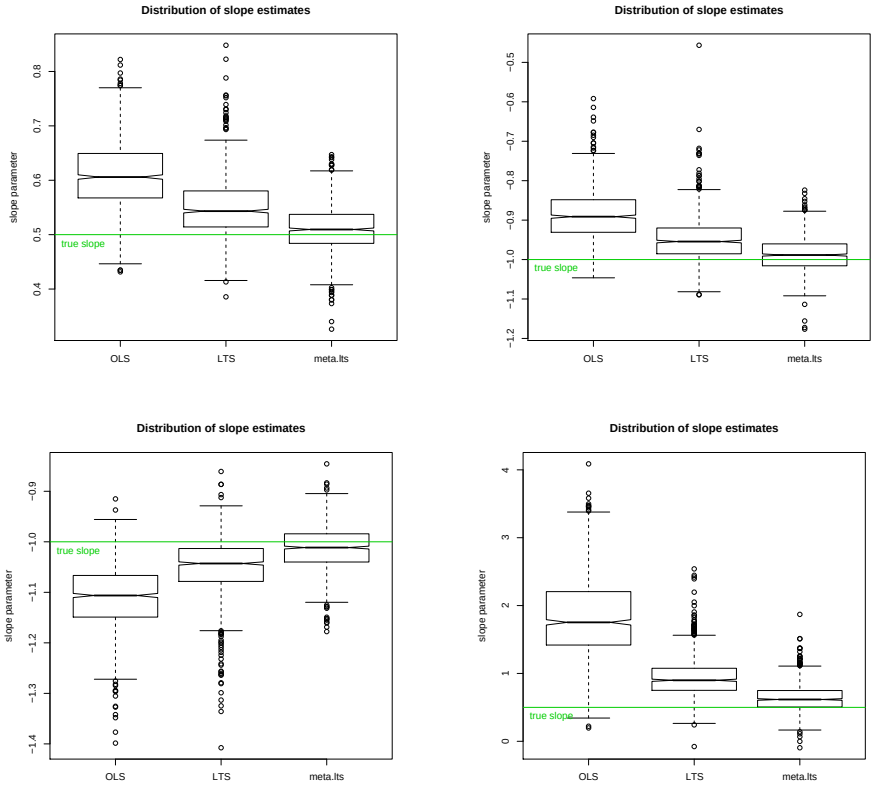

Figure 16: Comparison of influence of different slopes and types of heteroscedasticity.

Discussion

7 Example Dataset

7.1 Data Quality

The present study dealt with three main aspects of meta-analysis: data quality, choice of model and model fitting. Furthermore, it presented a new estimator for meta-regression and performed simulations to further determine its properties. First, aspects of data quality will be discussed, in particular how to deal with false data points as well as imputation, and existing guidelines will be highlighted.

Data quality is of utmost importance in statistical analysis. Even in the most assiduously performed research, problems regarding data quality cannot completely be outruled. Sources of low data quality include technical problems in general, transmission errors, low precision of measurement instruments, and others. In meta-analysis in particular, the quality of the different primary studies might vary considerably and effectively cannot be influenced by the meta-analyst.

To incorporate measures of study quality into meta-analysis, methods based on the calculation of a quality score and downweighting of low quality studies have been presented [68]. However, it has been shown that approaches of this kind are not robust and, far from improving overall results, actually introduce additional bias [17, 32]. This leads to the conclusion that genuine robust statistical methods are pertinent.

7.1.1 False Data Points

Even the seemingly trivial possibility of typing errors is a relevant practical problem. Indeed, such a situation can be found in [36], where four data points in the figures in the paper do not match the data reported in the corresponding table. This can be seen in figure 1. Although this does not in any relevant way seem to influence the results of the particular analysis presented, it stresses the point that data quality issues have to be dealt with at every level of statistical analysis.

Failing to effectively handle data quality issues properly might entail distorted and even qualitatively different results. Because of their ability to inherently deal with contaminated data, the use of robust statistical methods should be encouraged. Unfortunately, robust methods, which might provide a sensible way to deal with these problems, like the proposed `meta.lts` estimator, are not yet widely available in software packages for meta-analysis.

7.1.2 Imputation

When synthesising data from primary studies, meta-analysts are often confronted with incomplete reporting. Thus, the need for data imputation arises [28, 56]. This often concerns the standard deviations of the effect sizes reported by primary studies. Missing standard deviations are of particular concern in meta-analysis since the assignment of weights depends on this information. It has been shown that, in general, imputing standard deviations can lead to higher quality of overall analysis [13].

The methods used for imputation, however, should be documented in meta-analytic reporting and the data points actually imputed should be marked as such. This procedure guarantees that the analyses performed can be retraced and the amount

of influence of—as well as potential bias introduced by—the imputation method can be assessed.

7.1.3 Established Guidelines

There are several important sets of guidelines, most prominently the internationally recognized guidelines by PRISMA (formerly known as QUOROM), and expert opinions stressing more specific aspects of meta-analysis. The most important points of the PRISMA Statement are how to assess data quality, how to identify primary studies used or excluded, and structure of the reporting of meta-analysis.

In order to allow the reader to assess data quality, a high level of transparency should strongly be encouraged. Three aspects can be stressed on this behalf: the process of data generation should be documented extensively (cf. QUOROM guidelines [46], PRISMA guidelines [47, 43]), imputed data should clearly be identified and the imputation process explained, and finally the data used in the calculation should be made available to the reader without the necessity of explicit requests. This last point is particularily important (cf. section) since the need for explicit requests hampers double-checking of findings by peers and thus reduces the overall quality of scientific research.

On the other side, consensus guidelines for the reporting of primary studies have been published, focusing especially on their suitability for future meta-analysis. The CONSORT statement [48, 49] established checklists of essential elements that should be reported to allow the reader to assess the reliability and relevance to the study results. A similar checklist was compiled by the STARD initiative [6] with a focus on potential bias and generalizability of studies of diagnostic accuracy.

7.2 Appropriate Model

At the beginning of any meta-analysis arises the question of whether to just determine a common or mean effect size or even further to assess the influence of covariates on effect size variation between studies. The choice of a model depends on the research question to be answered by the meta-analysis. Furthermore, the availability and quality of data may restrict the possibility to answer certain questions. In the case of a planned regression analysis, it is important to discriminate between covariates more or less common for all patients in a primary study, these are termed study-level covariates, and those covariates representing distinct properties of the individual subject, these are termed patient-level covariates.

Irrespective of whether the analysis to perform is a simple effect size estimation or meta-regression, meta-analysts must choose between a fixed-effect or random-effects approach to model a common or mean effect size of all primary studies. The difference between the two is that in the fixed-effect model all primary studies are considered to have a common effect size whereas in the random-effects model effect sizes may differ. Instead of modeling a common effect size, the random-effects model determines the mean of a distribution of effect sizes.

In general, it is appropriate to choose random-effects modeling. It can rarely be assumed that all studies involved share the exact same effect size unless they were all performed by the same institution, the same investigator, at the same time and place, on the same population etc. [1]. It should explicitly be discouraged to test for homogeneity of effect sizes and confuse the failure of the test to reject the null hypothesis with evidence that the effects are homogeneous [26] and subsequently use a fixed-effect model, since the power of such tests is generally very low [23, 31], in particular when one study make up a large proportion of the total information [22]. Instead, whether to employ a fixed-effect or random-effects model should solely be decided on grounds of a priori knowledge to avoid bias in the results [24].

An important limiting factor in meta-analysis is the availability of data appropriate for the planned analysis. As mentioned above, some research questions can only be treated adequatly on data of high granularity, e.g. individual patient data (IPD) as opposed to study means and variances. However, quite often only aggregated data are available [63] or collecting individual patient data may be uneconomical relative to the research question under consideration [62] and, as can be seen in figure 7, trends in IPD can be obfuscated by aggregation. While this means that analyses pertaining to individual patient properties cannot adequatly be performed on aggregated data, it does not imply that the use of aggregated data must a priori be seen as a problem. Whether the level of granularity is adequate for the analysis depends on the research question, therefore it is pertinent to clearly define the latter before starting the actual analysis. In meta-regression, this means that data aggregated to study level has the most appropriate granularity to explain between-study variance (cf. [20]) whereas IPD is necessary to explain between-patient variance [38].

Similar to the question whether to use IPD or study-level data is whether to commonly analyze all data points or take into account subgroups of the data. As can be seen in figures 8 and 4, individual analyses of subgroups can fundamentally change the percieved relation between independent and dependent variables. On one hand, seperately analyzing groups leads to smaller sample sizes in each group and therefore reduces precision of estimation and power of statistical tests, on the other hand, seperate analysis can be indispensable if the groups behave differently.

It is self-evident that the use of different effect size measures leads to different results. The nature of this difference, however, may not be clear from the outset. Therefore, it is important to impartially consider the different possibilities and evaluate which aspects of the data are highlighted by the different effect size measures. In this context, it might even be reasonable to juxtapose different results obtained in this manner, despite the risk of data dredging usually involved in analyzing one

set of data in different ways. An example for this procedure can be seen in figure 6.

Various criteria can be summoned to justify the choice of a certain effect size measure. Some criteria might arise from the clinical context (response rates vs. remission rates), others from mathematical considerations. In the example of antidepressant efficacy, it is worth noting that HAMD scores are ordinally scaled, and thus regression on these is formally incorrect. Despite this, resulting models might sometimes provide insightful interpretations, but in general, it is preferable to choose a mathematically correct model by employing a suitable effect size measure. If no such measure is immediately available, it can often be obtained by transformation of the available data (even of ordinally scaled data).

Ultimately, there is no right effect size measure common for all meta-analyses. Rather, the choice of effect size should be determined taking into consideration the context in general and the field of research in particular. It might be helpful to consider the choices of effect size measures employed in earlier literature published on the subject matter.

7.3 Data Dredging

Data dredging is a common problem in meta-analysis [67]. One of the reasons might be a lack of discrimination between exploratory data analysis and inferential statistics. Two main goals of exploratory analysis are generation of hypotheses and validation of model assumptions for subsequent inferential statistics. Inferential statistics, in contrast, deals with testing hypotheses.

Commonly, a certain amount of exploratory analysis is an integral part of any statistical analysis, primarily to generate hypotheses by inspecting different aspects of the data, and secondly to check model assumptions based on visual impression. In meta-analysis, however, using exploratory analysis in the first sense is problematic

because of the difficulties involved in testing on an independent sample hypotheses generated in this manner [67]. This would require new data, which can be hindered by prohibitive cost, ethical issues (convincing results regarding efficacy of some treatment may prohibit withholding this treatment from patients, thus eliminating the possibility to recruit participants for a new study), or in some cases even be wholly impossible (e.g. if the year of the study is a covariate). Nonetheless, even in meta-analysis, exploratory analysis can be justified as far as validation of model assumptions is concerned, and furthermore in some cases even be employed in hypothesis generation if due diligence is taken to avoid data dredging, an example being generation of hypotheses where testing with a single new study is tractable.

Inferential statistics are used to determine, by means of statistical tests, whether the data support a predetermined null hypothesis or not. In contrast to exploratory analysis where it is perfectly acceptable to examine a multitude of different models, when using multiple models in inferential statistics, adjustments for multiple testing must be employed. Since this adjustment lowers the power of tests, it advisable to keep the number of tests low by a priori eliminating implausible models. Second, the use of excessively complex models, tending towards overfitting, can also give an impression of relationships between variables, thus opening the path to overinterpretation and data dredging. A third pitfall possibly leading to data dredging is using models without due consideration of their assumptions, among which commonly are independence, normal distribution and homoscedasticity of residuals.

Even after the formal analysis itself, data dredging can occur during interpretation of results [37]. To avoid this, the reliability of the data should be assessed, the process of choosing and fitting the models should be made transparent and thoroughly discussed. To this end, regression diagnostics, especially plots, should be used. Furthermore, the plethora of available tools to aid the consumer of meta-analysis in interpretation, such as confidence bands in graphs and other information regarding

the stability of results (cross validation, bootstrapping, etc.), should be employed wherever possible.

8 Robust Estimation

8.1 Existing Methods

8.1.1 Assessment of Heterogeneity

Three methods to calculate the between study variance τ^2 have been presented: the one-step [11] and two-step [10] DerSimonian and Laird estimators as well as the Paule and Mandel [51] estimator.

The one-step DerSimonian and Laird estimator, which is the most widely employed random-effects estimator in clinical research, is problematic because of its assumption of normality and lack of robustness. Furthermore, its only advantage over the other two estimators lies in its easy computability, which nowadays has become all but irrelevant. Even when its model assumptions are not violated in any way it is, from several points of view, not the best estimator available [60].

The two-step DerSimonian and Laird estimator, under favorable conditions, leads to better results than the one-step version. Still, conserving the main properties of the one-step estimator, it basically has the same problematic model assumptions and lack of robustness.

Finally, the Paule and Mandel estimator has no such restrictive model assumptions and yields better results under normality, that is, the only case where the other two estimators could correctly be applied at all. Therefore, out of these three, it is the estimator that should be used in practice. Still, it is far from perfect since it is not robust against outliers and low data quality.

8.1.2 Correcting for Publication Bias with Trim and Fill

The methodological literature is rife with attempts to deal with publication bias. Among these, a method termed trim and fill [12] has gained popularity, not least because it is based on the funnel plot, a widely used graphical tool to reveal the presence of publication bias. The latter is a scatter plot, relating sample sizes (or similar measures of study precision) to the reported effect sizes. The variances of these effect sizes are expected to be higher with lower study sample size (precision), and the pattern of the plot should be symmetric as long as there is no publication bias.

Trim and fill tries to deal with publication bias by imputing artificial data, compensating for any asymmetry in the funnel plot. The number of artificial data points added is chosen to reflect the presumed number of unobserved studies. One method to estimate this number is

$$L_0^+ = \left\lceil \frac{4T_n - n(n+1)}{2n-1} \right\rceil \tag{48}$$

where T_n is the Wilcoxon statistic

$$T_n = \sum_{i=1}^{n} I_{[\hat{\mu},\infty)}(\hat{\theta}_i) R_i \tag{49}$$

with R_i being the rank of the i-th observation in the sample $(|\hat{\theta}_i - \hat{\mu}|)_{i=1,\ldots,n}$ with μ being calculated as in equation (13). The estimation is repeated iteratively, in each step removing the most extreme L_0^+ positive observations from the totality of studies until the algorithm converges. Then, L_0^+ studies are imputed by mirroring the L_0^+ most extreme positive ones with respect to the estimated mean effect size [65].

The trim and fill method, however, like the funnel plot on which it is based, assumes the effect sizes of the individual studies coming from the same distribution, and does not allow for either heterogeneity as assumed in the random effects model or differences in effect sizes due to covariates. Both of these lead to the possibility of assymmetry in the funnel plot even in absence of publication bias. The application of trim and fill would thus impute datapoints that were not missing in the first place.

Futhermore, problems can arise if the sample sizes are not random, but rather the result of power analysis. In this case, the a priori expectation of high effect sizes leads to smaller studies and thus an asymmetry in the funnel plot resulting from the correlation between effect size magnitude and sample size. As in the case described above, trim and fill is not appropriate.

These considerations have been confirmed by Terrin et al. via simulation studies on publication bias correction methods in the presence of heterogeneity [65].

8.2 The `meta.lts` Estimator

To date, the common methodology to fit meta-regression models is weighted OLS. A problem with wOLS, however, is its lack of robustness or stability against outliers and data errors, both in the data and the weight space. In the non-weighted case, a multitude of robust estimators is available, among them the commonly used LTS estimator. Still, there is no commonly known robust regression estimator for weighted data. As a consequence, it would be sensible to extend an already tried and tested robust estimator for use with weighted data. This is, however, not a trivial task due to the interrelationship of residuals and weights in the optimization process.

Because of the possibility of efficient implementation, LTS was chosen as the estimator to extend for the meta-regression scenario. This approach also bears the

advantage that the influence of weights can immediatly be inherited from the wOLS estimator, all the while preserving the robustness of LTS. As intended by the idea of meta-analytic modeling the use of weights conveys smaller variance of the estimator. The special appeal of the `meta.lts` estimator is not only combining the individual advantages of wOLS and LTS, but in particular to provide a robustness concept adapted to the meta-analytic setting.

8.2.1 Design Choices

The central question about design choices is where and how to use weighting in an extended version of the original LTS algorithm. The primary focus was not only to adaptLTS to weighted data, but more specifically to the meta-analytic setting. There are several distinct parts of the algorithm where weights can be taken into account. First, the regession line that, in the original LTS estimator, is fitted to a part of the data using OLS is, in `meta.lts`, fitted using wOLS. This implies that, when the proportion to be trimmed is set to zero, the `meta.lts` estimator corresponds to the wOLS estimator just as LTS corresponds to OLS. Second, the LTS algorithm involves ordering the data by the size of the squared residuals, and at this point, it might be considered to use weighted squared residuals in the weighted case. Third, instead of trimming an α-proportion of the number of data points, an α-proportion of the weights could be used. The latter was implemented in the final version of `meta.lts`.

Two important design considerations were to find a balance between the influence of the weights and the influence of the squared residuals in the fitting of the regression line. Skewing this balance towards the weights might lead to the regression line being fitted to a very small proportion of actual data points, whereas tipping the balance in favor of the number of observations would devalue the confidence put into the higher-weighted data points in the meta-analytic model.

Using weighted squared residuals in the process of ordering the data points for trimming is, unlike using wOLS to fit the regression line, not the only choice immediately coming to mind. In practice, it has the effect of increasing the tendency to trim away higher weighted points due to their higher priority in the weighted ranking. At first sight, this seems counter-intuitive because in the meta-analytic setting, the higher weighed points are considered as the most reliable. Nonetheless, it was chosen in the final implementation due to two factors. First, using unweighted squared residuals would lead to conflicting target functions between the fitting of the regression line itself and the selection of data points to fit the line to. This would entail inconsistency of the model, possibly resulting in instability of the estimator. Second, it serves to reduce the existing bias against the trimming of higher weighted data points introduced by the weighting of the least-squares estimation that pulls the regression line towards these points, thus reducing their squared residuals and consequently their probability to be trimmed when using unweighted residual ordering.

Once an ordering of the data points is achieved, the actual trimming can then be performed in two ways. Either a fixed proportion of the number of data points or of the sum of the weights can be used for the fitting of the regression line. Both approaches have advantages and disadvantages and whether to trim one way or the other is not a clear-cut decision in the design of a robust estimator for meta-regression on the basis of LTS. When using the first method, a fixed number of data points can be guaranteed to be used in the model fitting process. These, however, might be of low weight, and thus carry only a relatively small part of the overall evidence as defined in the meta-analytic framework. In contrast, when trimming a proportion of the sum of weights, the model is fitted to the majority of the available evidence. But then, the data points retained might constitute a minority of the total number of data points, which is not in itself undesirable, but might lead to instable model estimation if this minority is too small a proportion of

all data points. Nonetheless, the central importance of weights in the interpretation of the meta-analytic model leads to a preference to fit the model to the majority of the weights rather than of the number of data points. Therefore, this option was chosen in the implementation of the `meta.lts` estimator. As an interesting venue for future work, the ambiguity of this decision might be resolved in part by finding a middle way, e.g. retaining both a minimum proportion of weights and of data points.

Another difficult question in the design of the estimator was the initialization. Since `meta.lts` uses an iterative algorithm, choice of a good initial model is important to find the global optimum. Four ideas were taken into consideration. First, random α-trimmed subsets were used as initial sample for wOLS estimation. This often lead to the finding of only local minima of the target function, even when large numbers of initial samples were computed. As an attempt to start with a model supposedly closer to the global optimum, it was then considered using classic LTS estimations as starting values. This, however, also lead to local minima, since the global minimum of the LTS target function is not necessarily close to the global optimum of the `meta.lts` target function. Already close to the final implementation was the third attempt which randomly selected two points as support for a line used as starting direction for the first iteration. A final improvement upon this could be achieved by using weighted sampling in the selection of the initial data points, which increases the probability of starting the first iteration in a direction close to the final solution constituting the global optimum.

8.2.2 Application of `meta.lts`

In the meta-analytic setting, due to data quality issues, heteroscedasticity inherent in many datasets, and the risks of publication bias and data dredging, robust estimation is highly desirable. However, when employing robust methods, one has

to bear in mind the strengths and weaknesses of robust estimation, which are primarily that the robust estimators discard some of the information but in exchange lead to more stable estimates. The presented `meta.lts` method in particular has, compared to classical robust estimators, the drawback of the complex issues involving the breakdown point, but in contrast to the latter makes efficient use of information on study precision represented by the weights.

When using `meta.lts`, weights can be calculated according to any meta-analytic modeling approach, in particular the fixed-effect or random-effects models. In most practical cases, it is recommended to use the random-effects weighting due to its realistic assessment of between-study variance and more robust results.

9 Simulation

9.1 Scenarios

The simulations performed in this study belong to three distinct scenarios. The first two were designed to highlight theoretical properties of the `meta.lts` estimator, and the third was meant as a more practical example about the common problems of heteroscedasticity and publication bias.

In the first scenario, each dataset was generated from three distinct distributions representing three groups of data points, such that one of the groups with less than half of the data points comprised more than half of the total weight. This simulation highlighted the main property of `meta.lts` to fit a model to the majority of weights, rather than number of data points. The difficulties in initialization were illustrated, as well as the advantages of weighted sampling to find an initial model matching the structure of the subset of observations with the majority of weights. Furthermore, this dataset allowed a comparison of `meta.lts` with wOLS and LTS

to showcase the advantage of the former in using weights in datasets containing outliers.

The second simulation shows the influence of outliers in weight space rather than in data space. Similarly to the first simulation, it was not meant to replicate a practical case but rather illustrate the complexity of accurately defining a breakdown point for weighted data. The conflict between the majority of data points and the majority of weights is accentuated by the spatial disparity of the two groups.

The third example was motivated by practical considerations and designed to show the influence of publication bias in the context of heteroscedastic data, a compex yet realistic scenario. Heteroscedasticity is encountered in real meta-analytic settings, as seen in the example dataset, often occuring on the basis of reduced availability of subjects in extreme groups and the resulting high variance in the necessarily small studies on these groups. Publication bias, as explained in section 4.2.3, is all the more relevant on heteroscedastic data because of the increasing likelihood of studies remaining unpublished where variance of results is high. The problem is further aggravated by the necessity of pharmaceutical companies to submit several independent positive findings regarding the efficacy of a newly developed drug to comply with drug approval procedures in many countries. On the other hand, the recently established practice of registering every pharmaceutical study before execution is likely to alleviate the problem in the future. Regardless of future change, however, meta-analyses have to deal with it and the simulations show how different estimators cope with these sources of bias.

9.2 Properties of `meta.lts`

Two distinct theoretical properties of `meta.lts` illustrated by the simulations were choice of initialization and measures of robustness in the context of weighted data. As discussed above, the selection of a `meta.lts` starting model is optimized with

respect to the majority of weights. When two points are selected with weighted sampling as opposed to non-weighted sampling, the probability of finding a direction corresponding to the inherent structure of data points is increased. Figure 9 shows the initial configurations with dotted lines and demonstrates the improvement achieved by the use of weighted sampling: in the right subfigure, a larger proportion of the initial regression lines are situated close to the global optimum.

The issue of the breakdown point is, as mentioned in section 2.4, a more complex one. When an estimator is designed to be fitted to the majority of weights instead of the majority of data points, the classical concept of the breakdown point inevitably yields a breakdown point of zero. Majorities of number of data points and weights can be mutually exclusive and therefore if one wants to catch the idea of a breakdown point, the classical definition cannot be applied. Instead, a breakdown point must be defined as a minimal proportion of weights needed to arbitrarily distort the estimator. However, the simulations showed that the presence of three points with the majority of weights does not necessarily attract the regression line.

9.3 Comparison

The simulation on data with publication bias examined differences between the three estimators considered in homoscedastic and heteroscedastic case. In absence of heteroscedasticity, no relevant difference can be seen between wOLS, LTS and `meta.lts`, and the robustness of the latter two was of no importance. The only observed difference concerned the variance of the estimates, the weighted estimators achieving a higher precision than LTS.

The situation was entirely different when heteroscedasticity was introduced. Violation of the model assumptions of least-squares optimization lead to a marked distortion of the wOLS estimates, while the robust estimates by LTS and `meta.lts` were less adversely influenced. A comparison of the robust estimators not only

replicates the difference in variance as seen in the homoscedastic case, but also a significantly better estimation of the true slope by `meta.lts` in the median.

The precise forms in which heteroscedasticity was introduced had no relevant influence on the general pattern of results, besides an exacerbation of the distorting effect on the estimates with increasing heteroscedasticity.

10 Conclusions

10.1 Robustness in Meta-Regression

The distinction between fixed-effect and random-effects models was introduced (section 1) as well as the idea of robustness (section 2), and the usefulness of robust methods in meta-regression was established (section 3). A new estimator for robust meta-regression was presented (section 3.3) and compared with classical methods on real (section 5) and simulated datasets (section 6). Overall, the results presented encourage further examination and development of the `meta.lts` estimator.

10.2 Potential Problems of `meta.lts`

To date, the `meta.lts` estimator cannot unconditionally be recommended for use in clinical research, mainly because its formal properties are not fully understood yet. In spite of the positive results provided by the simulations presented here, a comprehensive theory on the breakdown point for weighted data is needed to conclusively answer these open questions. Furthermore, a consensus on the extent of trimming is necessary for wider acceptance by clinical researchers of robust methods in general and `meta.lts` in particular.

10.3 Outlook

Besides the above mentioned theoretical considerations on robustness in the context of weighted data, future work on this topic might include efficient implementation and integration into widely disseminated software packages, in particular R (www.r-project.org) [55]. This would not only include the creation of a dedicated package but also the adaptation of the interface to the generally accepted standards of the respective environment.

Furthermore, an implementation of the `meta.lts` estimator on a multivariate setting should certainly be considered. In clinical research practice, meta-analysis on multivariate data is still rare due to the novelty of the meta-analytic concept and because the prohibitively low number of data points available in many of today's research questions makes untenable the examination of a multitude of covariates. However, due to increasing automation—and therefore data generation—in medicine, this can realistically be expected to change in the future.

Code Listings

```
‘meta.lts‘ <-
function (x, y, weights = rep(1, length(y)), alpha = 1/2, nsamp = 100,
    iter = 100, do.plot = TRUE, ...)
{
    if (!is.vector(x) | !is.vector(y) | !is.vector(weights)) {
        error("x, y, and wt must be vectors")
    }
    h = ceiling(length(y) * (1 - alpha))
    global.optimum = Inf
    if (do.plot) {
        plot(x, y, cex = pmax(1, sqrt(weights)/10), pch = 19)
    }
    for (i in 1:nsamp) {
        starting.sample = sample(1:length(y), 2, prob = weights)
        starting.model = lm(y ~ x, subset = starting.sample)
        if (do.plot) {
            abline(starting.model, lty = 3)
        }
        y.hat = predict(starting.model, data.frame(x = x))
        residuals.squared = (y.hat - y)^2
        residuals.squared.order = order(residuals.squared)
        best <- residuals.squared.order[1:h]
        for (j in 1:iter) {
            mod = lm(y ~ x, weights = weights, subset = best)
            y.hat = predict(mod, data.frame(x = x), weights = weights)
            residuals.squared = weights * (y.hat - y)^2
            residuals.squared.order = order(residuals.squared)
            cs = cumsum(weights[residuals.squared.order])
            in.local.best = c(TRUE, (cs < sum(weights) * (1 -
                alpha))[1:(length(cs) - 1)])
            if (sum(residuals.squared[residuals.squared.order[in.local.best]])
                < sum(residuals.squared[best])) {
                best <- residuals.squared.order[in.local.best]
            }
            else {
                break
            }
        }
        if (sum(residuals.squared[best]) < global.optimum) {
            global.optimum = sum(residuals.squared[best])
            res = list(model = mod, best = best, x = x, y = y,
                weights = weights, alpha = alpha, nsamp = nsamp,
                iter = iter)
        }
    }
    if (do.plot) {
        abline(res$model, col = 2, lwd = 2)
    }
    return(res)
}
```

```
‘meta.lts.unweighted‘ <-
function (x, y, weights = rep(1, length(y)), alpha = 1/2, nsamp = 100,
    iter = 100, do.plot = TRUE, ...)
{
    if (!is.vector(x) | !is.vector(y) | !is.vector(weights)) {
        error("x, y, and wt must be vectors")
    }
    h = ceiling(length(y) * (1 - alpha))
    global.optimum = Inf
    if (do.plot) {
        plot(x, y, cex = sqrt(wt), pch = 19)
    }
    for (i in 1:nsamp) {
        starting.sample = sample(1:length(y), 2, prob = weights)
        starting.model = lm(y ~ x, subset = starting.sample)
        if (do.plot) {
            abline(starting.model, lty = 3)
        }
        y.hat = predict(starting.model, data.frame(x = x))
        residuals.squared = (y.hat - y)^2
        residuals.squared.order = order(residuals.squared)
        best <- residuals.squared.order[1:h]
        for (j in 1:iter) {
            mod = lm(y ~ x, weights = weights, subset = best)
            y.hat = predict(mod, data.frame(x = x), weights = weights)
            residuals.squared = weights * (y.hat - y)^2
            residuals.squared.order = order(residuals.squared)
            in.local.best = 1:h
            if (sum(residuals.squared[residuals.squared.order[in.local.best]])
                < sum(residuals.squared[best])) {
                best <- residuals.squared.order[in.local.best]
            }
            else {
                break
            }
        }
        if (sum(residuals.squared[best]) < global.optimum) {
            global.optimum = sum(residuals.squared[best])
            res = list(model = mod, best = best, x = x, y = y,
                weights = weights, alpha = alpha, nsamp = nsamp,
                iter = iter)
        }
    }
    if (do.plot) {
        abline(res$model, col = 2, lwd = 2)
    }
    return(res)
}
```

List of Figures

List of Tables

References

[1] A. E. Ades, G. Lu, and J. P. T. Higgins. The interpretation of random-effects meta-analysis in decision models. *Med Decis Making*, 25(6):646–654, 2005.

[2] M. Aitkin. Meta-analysis by random effect modelling in generalized linear models. *Stat Med*, 18(17-18):2343–2351, 1999.

[3] B. Baujat, C. Mah, J.-P. Pignon, and C. Hill. A graphical method for exploring heterogeneity in meta-analyses: application to a meta-analysis of 65 trials. *Stat Med*, 21(18):2641–2652, Sep 2002.

[4] C. S. Berkey, D. C. Hoaglin, F. Mosteller, and G. A. Colditz. A random-effects regression model for meta-analysis. *Stat Med*, 14(4):395–411, Feb 1995.

[5] J. A. Berlin and E. M. Antman. Advantages and limitations of metaanalytic regressions of clinical trials data. *Online J Curr Clin Trials*, Doc No 134:[8425 words; 84 paragraphs], Jun 1994.

[6] P. M. Bossuyt, J. B. Reitsma, D. E. Bruns, C. A. Gatsonis, P. P. Glasziou, L. M. Irwig, J. G. Lijmer, D. Moher, D. Rennie, H. C. W. de Vet, and S. for Reporting of Diagnostic Accuracy. Towards complete and accurate reporting of studies of diagnostic accuracy: the stard initiative. standards for reporting of diagnostic accuracy. *Clin Chem*, 49(1):1–6, Jan 2003.

[7] D. Bhning, U. Malzahn, E. Dietz, P. Schlattmann, C. Viwatwongkasem, and A. Biggeri. Some general points in estimating heterogeneity variance with the DerSimonian-Laird estimator. *Biostatistics*, 3(4):445–457, Dec 2002.

[8] A. Cipriani, T. A. Furukawa, G. Salanti, J. R. Geddes, J. P. T. Higgins, R. Churchill, N. Watanabe, A. Nakagawa, I. M. Omori, H. McGuire, M. Tansella, and C. Barbui. Comparative efficacy and acceptability of 12

new-generation antidepressants: a multiple-treatments meta-analysis. *Lancet*, 373(9665):746–758, Feb 2009.

[9] J. M. Davis, N. Chen, and I. D. Glick. A meta-analysis of the efficacy of second-generation antipsychotics. *Arch Gen Psychiatry*, 60(6):553–564, Jun 2003.

[10] R. DerSimonian and R. Kacker. Random-effects model for meta-analysis of clinical trials: an update. *Contemp Clin Trials*, 28(2):105–114, Feb 2007.

[11] R. DerSimonian and N. Laird. Meta-analysis in clinical trials. *Control Clin Trials*, 7(3):177–188, Sep 1986.

[12] S. Duval and R. Tweedie. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56(2):455–463, Jun 2000.

[13] T. A. Furukawa, C. Barbui, A. Cipriani, P. Brambilla, and N. Watanabe. Imputing missing standard deviations in meta-analyses can provide accurate results. *J Clin Epidemiol*, 59(1):7–10, Jan 2006.

[14] A. X. Garg, D. Hackam, and M. Tonelli. Systematic review and meta-analysis: when one study is just not enough. *Clin J Am Soc Nephrol*, 3(1):253–260, Jan 2008.

[15] J. Geddes, N. Freemantle, P. Harrison, and P. Bebbington. Atypical antipsychotics in the treatment of schizophrenia: systematic overview and meta-regression analysis. *BMJ*, 321(7273):1371–1376, Dec 2000.

[16] J. R. Geddes, J. R. Calabrese, and G. M. Goodwin. Lamotrigine for treatment of bipolar depression: independent meta-analysis and meta-regression of individual patient data from five randomised trials. *Br J Psychiatry*, 194(1):4–9, Jan 2009.

[17] L. A. Gelfand, D. R. Strunk, X. M. Tu, R. E. S. Noble, and R. J. Derubeis. Bias resulting from the use of 'assay sensitivity' as an inclusion criterion for meta-analysis. *Stat Med*, 25(6):943–955, Mar 2006.

[18] P. P. Glasziou and S. L. Sanders. Investigating causes of heterogeneity in systematic reviews. *Stat Med*, 21(11):1503–1511, Jun 2002.

[19] T. A. Hammad, T. Laughren, and J. Racoosin. Suicidality in pediatric patients treated with antidepressant drugs. *Arch Gen Psychiatry*, 63(3):332–339, Mar 2006.

[20] R. A. Hansen, C. G. Moore, S. B. Dusetzina, B. I. Leinwand, G. Gartlehner, and B. N. Gaynes. Controlling for drug dose in systematic review and meta-analysis: a case study of the effect of antidepressant dose. *Med Decis Making*, 29(1):91–103, 2009.

[21] R. M. Harbord, P. Whiting, J. A. C. Sterne, M. Egger, J. J. Deeks, A. Shang, and L. M. Bachmann. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. *J Clin Epidemiol*, 61(11):1095–1103, Nov 2008.

[22] R. J. Hardy and S. G. Thompson. Detecting and describing heterogeneity in meta-analysis. *Stat Med*, 17(8):841–856, Apr 1998.

[23] L. V. Hedges and T. D. Pigott. The power of statistical tests in meta-analysis. *Psychol Methods*, 6(3):203–217, Sep 2001.

[24] L. V. Hedges and T. D. Pigott. The power of statistical tests for moderators in meta-analysis. *Psychol Methods*, 9(4):426–445, Dec 2004.

[25] J. Higgins, S. Thompson, J. Deeks, and D. Altman. Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and practice. *J Health Serv Res Policy*, 7(1):51–61, Jan 2002.

[26] J. P. T. Higgins, S. G. Thompson, J. J. Deeks, and D. G. Altman. Measuring inconsistency in meta-analyses. *BMJ*, 327(7414):557–560, Sep 2003.

[27] J. P. T. Higgins, S. G. Thompson, and D. J. Spiegelhalter. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A Stat Soc*, 172(1):137–159, Jan 2009.

[28] J. P. T. Higgins, I. R. White, and A. M. Wood. Imputation methods for missing outcome data in meta-analysis of clinical trials. *Clin Trials*, 5(3):225–239, 2008.

[29] J. P. A. Ioannidis. Why most published research findings are false. *PLoS Med*, 2(8):e124, Aug 2005.

[30] J. P. A. Ioannidis, N. A. Patsopoulos, and H. R. Rothstein. Reasons or excuses for avoiding meta-analysis in forest plots. *BMJ*, 336(7658):1413–1415, Jun 2008.

[31] D. Jackson. The power of the standard test for the presence of heterogeneity in meta-analysis. *Stat Med*, 25(15):2688–2699, Aug 2006.

[32] P. Jni, A. Witschi, R. Bloch, and M. Egger. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA*, 282(11):1054–1060, Sep 1999.

[33] R. N. Kacker. Combining information from interlaboratory evaluations using a random effects model. *Metrologia*, 41:132–6, 2004.

[34] G. Kemmler, M. Hummer, C. Widschwendter, and W. W. Fleischhacker. Dropout rates in placebo-controlled and active-control clinical trials of antipsychotic drugs: a meta-analysis. *Arch Gen Psychiatry*, 62(12):1305–1312, Dec 2005.

[35] A. Khan, R. M. Leventhal, S. R. Khan, and W. A. Brown. Severity of depression and response to antidepressants and placebo: an analysis of the food

and drug administration database. *J Clin Psychopharmacol*, 22(1):40–45, Feb 2002.

[36] I. Kirsch, B. J. Deacon, T. B. Huedo-Medina, A. Scoboria, T. J. Moore, and B. T. Johnson. Initial severity and antidepressant benefits: a meta-analysis of data submitted to the food and drug administration. *PLoS Med*, 5(2):e45, Feb 2008.

[37] G. Knapp and J. Hartung. Improved tests for a random effects meta-regression with a single covariate. *Stat Med*, 22(17):2693–2710, Sep 2003.

[38] P. C. Lambert, A. J. Sutton, K. R. Abrams, and D. R. Jones. A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *J Clin Epidemiol*, 55(1):86–94, Jan 2002.

[39] W. L. Lee, R. B. Bausell, and B. M. Berman. The growth of health-related meta-analyses published from 1980 to 2000. *Eval Health Prof*, 24(3):327–335, Sep 2001.

[40] S. Leucht, D. Arbter, R. R. Engel, W. Kissling, and J. M. Davis. How effective are second-generation antipsychotic drugs? A meta-analysis of placebo-controlled trials. *Mol Psychiatry*, 14(4):429–447, Apr 2009.

[41] S. Leucht, C. Corves, D. Arbter, R. R. Engel, C. Li, and J. M. Davis. Second-generation versus first-generation antipsychotic drugs for schizophrenia: a meta-analysis. *Lancet*, 373(9657):31–41, Jan 2009.

[42] S. Leucht, K. Wahlbeck, J. Hamann, and W. Kissling. New generation antipsychotics versus low-potency conventional antipsychotics: a systematic review and meta-analysis. *Lancet*, 361(9369):1581–1589, May 2003.

[43] A. Liberati, D. G. Altman, J. Tetzlaff, C. Mulrow, P. C. Gtzsche, J. P. A. Ioannidis, M. Clarke, P. J. Devereaux, J. Kleijnen, and D. Moher. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that

evaluate health care interventions: explanation and elaboration. *PLoS Med*, 6(7):e1000100, Jul 2009.

[44] H. Melander, J. Ahlqvist-Rastad, G. Meijer, and B. Beermann. Evidence b(i)ased medicine–selective reporting from studies sponsored by pharmaceutical industry: review of studies in new drug applications. *BMJ*, 326(7400):1171–1173, May 2003.

[45] H. Melander, T. Salmonson, E. Abadie, and B. van Zwieten-Boot. A regulatory apologia–a review of placebo-controlled studies in regulatory submissions of new-generation antidepressants. *Eur Neuropsychopharmacol*, 18(9):623–627, Sep 2008.

[46] D. Moher, D. J. Cook, S. Eastwood, I. Olkin, D. Rennie, and D. F. Stroup. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. quality of reporting of meta-analyses. *Lancet*, 354(9193):1896–1900, Nov 1999.

[47] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and P. Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med*, 6(7):e1000097, Jul 2009.

[48] D. Moher, K. F. Schulz, D. Altman, and C. O. N. S. O. R. T. Group (Consolidated Standards of Reporting Trials). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA*, 285(15):1987–1991, Apr 2001.

[49] D. Moher, K. F. Schulz, D. Altman, and C. O. N. S. O. R. T. Group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials 2001. *Explore (NY)*, 1(1):40–45, Jan 2005.

[50] N. A. Patsopoulos, A. A. Analatos, and J. P. A. Ioannidis. Relative citation impact of various study designs in the health sciences. *JAMA*, 293(19):2362–2366, May 2005.

[51] R. C. Paule and J. Mandel. Consensus values and weighting factors. *Journal of research of the National Bureau of Standards*, 87:377–85, 1982.

[52] K. Pearson. Report on certain enteric fever inoculation statistics. *British Medical Journal*, 3:1243–6, 1904.

[53] D. B. Petitti. Approaches to heterogeneity in meta-analysis. *Stat Med*, 20(23):3625–3633, Dec 2001.

[54] M. Petticrew. Systematic reviews from astronomy to zoology: myths and misconceptions. *BMJ*, 322(7278):98–101, Jan 2001.

[55] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[56] C. Robertson, N. R. N. Idris, and P. Boyle. Beyond classical meta-analysis: can inadequately reported studies be included? *Drug Discov Today*, 9(21):924–931, Nov 2004.

[57] P. J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–80, 1984.

[58] P. J. Rousseeuw and K. Van Driessen. Computing LTS regression for large datasets. *Estadistica*, 54:163–90, 2002.

[59] H. Scherk, F. G. Pajonk, and S. Leucht. Second-generation antipsychotic agents in the treatment of acute mania: a systematic review and meta-analysis of randomized controlled trials. *Arch Gen Psychiatry*, 64(4):442–455, Apr 2007.

[60] K. Sidik and J. N. Jonkman. A comparison of heterogeneity variance estimators in combining results of studies. *Stat Med*, 26(9):1964–1981, Apr 2007.

[61] J. R. Sneed, B. R. Rutherford, D. Rindskopf, D. T. Lane, H. A. Sackeim, and S. P. Roose. Design makes a difference: a meta-analysis of antidepressant response rates in placebo-controlled versus comparator trials in late-life depression. *Am J Geriatr Psychiatry*, 16(1):65–73, Jan 2008.

[62] L. A. Stewart and J. F. Tierney. To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. *Eval Health Prof*, 25(1):76–97, Mar 2002.

[63] A. J. Sutton and J. P. T. Higgins. Recent developments in meta-analysis. *Stat Med*, 27(5):625–650, Feb 2008.

[64] M. J. Taylor, N. Freemantle, J. R. Geddes, and Z. Bhagwagar. Early onset of selective serotonin reuptake inhibitor antidepressant action: systematic review and meta-analysis. *Arch Gen Psychiatry*, 63(11):1217–1223, Nov 2006.

[65] N. Terrin, C. H. Schmid, J. Lau, and I. Olkin. Adjusting for publication bias in the presence of heterogeneity. *Stat Med*, 22(13):2113–2126, Jul 2003.

[66] S. G. Thompson. Why sources of heterogeneity in meta-analysis should be investigated. *BMJ*, 309(6965):1351–1355, Nov 1994.

[67] S. G. Thompson and J. P. T. Higgins. How should meta-regression analyses be undertaken and interpreted? *Stat Med*, 21(11):1559–1573, Jun 2002.

[68] D. Tritchler. Modelling study quality in meta-analysis. *Stat Med*, 18(16):2135–2145, Aug 1999.

[69] E. H. Turner, A. M. Matthews, E. Linardatos, R. A. Tell, and R. Rosenthal. Selective publication of antidepressant trials and its influence on apparent efficacy. *N Engl J Med*, 358(3):252–260, Jan 2008.

[70] B. T. Walsh, S. N. Seidman, R. Sysko, and M. Gould. Placebo response in studies of major depression: variable, substantial, and growing. *JAMA*, 287(14):1840–1847, Apr 2002.

[71] A. Whitehead and J. Whitehead. A general parametric approach to the meta-analysis of randomized clinical trials. *Stat Med*, 10(11):1665–1677, Nov 1991.

[72] H. Xu, R. W. Platt, Z.-C. Luo, S. Wei, and W. D. Fraser. Exploring heterogeneity in meta-analyses: needs, resources and challenges. *Paediatr Perinat Epidemiol*, 22 Suppl 1:18–28, Jan 2008.