Received: 19 September 2020 | Accepted: 23 February 2021

DOI: 10.1111/acps.13293

Acta Psychiatrica Scandinavica WILEY

SYSTEMATIC REVIEW

Systematic evaluation of the 'efficacy-effectiveness gap' in the treatment of depression with venlafaxine and duloxetine

Carolin Schneider[1,2] | **Johanna Breilmann**[1] | **Benedikt Reuter**[2] | **Thomas Becker**[1] | **Markus Kösters**[1]

[1]Department of Psychiatry II, Ulm University, Bezirkskrankenhaus Günzburg, Germany

[2]Department of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany

Correspondence
Markus Kösters, Department of Psychiatry II, Ulm University, Bezirkskrankenhaus Günzburg, Germany.
Email: markus.koesters@uni-ulm.de

Funding information
The study was funded by the young scientists' programme of the German network 'Health Services Research Baden-Württemberg' of the Ministry of Science, Research and Arts in collaboration with the Ministry of Employment and Social Order, Family, Women and Senior Citizens, Baden-Württemberg.

Abstract

Objective: Evidence of larger drug effects in highly standardized studies (efficacy) compared to clinical routine (effectiveness) is discussed as *efficacy-effectiveness gap*. This study aimed to quantify effect size differences of RCTs and non-RCTs in the treatment of depression with venlafaxine and duloxetine and to identify effect modifying predictors.

Methods: A comprehensive systematic review and meta-analysis was conducted, including all prospective trials, which evaluated the treatment effects of duloxetine or venlafaxine in patients with depression. The primary outcome was the pre-post effect size after acute therapy, which were compared between RCTs and non-RCTs. Moreover, an exploratory analysis of predictors in a mixed meta-regression model within an information-theoretic approach was performed.

Results: 171 RCTs and 74 non-RCTs were included. The pre-post effect size differed significantly between RCTs and non-RCTs (−3.04 vs. −2.62, $\Delta = 0.41$, $p = 0.012$, high heterogeneity). Study characteristics were very similar between RCTs and non-RCTs. Most important variables to predict effect sizes were 'depression severity', 'dose' and 'number of participants'.

Conclusion: Despite differences in effect sizes between RCTs and non-RCTs, study design is not clearly an important predictor for the effect sizes. Our results question the common assumption that non-RCTs are generally better suited to describe a drug's effectiveness in clinical practice than RCTs. Future studies and their reporting should put more emphasis on the description of external validity, in order to allow better assessments of clinical relevance.

KEYWORDS

meta-analysis, depression, psychopharmacology, antidepressives

Summations

- RCTs showed higher effect sizes than non-RCTs, but the two study designs were very similar in most study characteristics.

- RCTs and non-RCTs did not differ in external validity, questioning the common assumption of non-RCTs to be generally better suited to describe a drug's effectiveness in clinical practice.
- Improved descriptions and assessments of external validity are needed, in order to design more clinically relevant trials or to improve clinical practice to reach similar effects as in RCTs.

Limitations

- Our analyses are restricted to two antidepressants.
- As a result of the large number of included studies and varying study designs, high unexplained heterogeneity remained in our models.
- The subscale 'external validity' had very low variance and is therefore of limited value for the explanation of effect size differences.

1 | INTRODUCTION

A chasm between theoretical research results and practical implementation exists in many scientific disciplines.[1-5] In psychiatric research, the gap between a drug's effect in highly standardized clinical trials and those in clinical real-world practice is often discussed. The discrepancy between drug effects in randomized controlled trials (RCTs, 'efficacy') and under clinical routine conditions ('effectiveness') is referred to as the *efficacy-effectiveness gap*.[6-8]

The antidepressant efficacy of duloxetine and venlafaxine, two selective serotonin norepinephrine reuptake inhibitors (SNRIs), has been shown in various trials. A large network meta-analysis confirmed the superior efficacy of several antidepressants over placebo, including duloxetine and venlafaxine.[9] Likewise, the efficacy of duloxetine and venlafaxine could be demonstrated in earlier meta-analyses with aggregated data[10-12] and with individual patient data,[13] while there is no meta-analytic study of the drugs' effects under clinical routine conditions.

It is assumed that efficacy and effectiveness trials are distinguished by various characteristics. RCTs are usually highly standardized and follow strict protocols. They are the gold standard in pharmacological research and are highly associated with a drug's 'efficacy'. On the contrary, non-RCTs (ie observational or non-randomized trials) are thought to evaluate a drug's effect in a setting that is closer to clinical practice and are therefore associated with its 'effectiveness'. The high degree of standardization in RCTs aims at increasing internal validity.[6,14] In contrast, effectiveness studies are supposed to have more heterogeneous and more representative patient populations or more feasible and more pragmatic interventions. Thus, effectiveness trials should have higher external validity.[14-17] These differences between RCTs and non-RCTs have been described anecdotally while systematic evaluations of such differences are rare.

Differences in study characteristics and samples question the transferability of results of RCTs to clinical practice. Clinical practice guidelines are mainly based on findings from RCTs (efficacy), but put into action in clinical practice (effectiveness). So far, two studies were published in which treatment outcomes of antidepressants in RCTs and observational studies were compared. A meta-analysis of fluoxetine and venlafaxine studies revealed that response rates in RCTs were significantly higher than in observational studies.[18] In addition, a comparison of meta-analytical data and routine outcome monitoring data of antidepressants and psychotherapy showed that significantly more depressed participants remitted in RCTs than in clinical routine studies.[19]

These first findings indicate divergent treatment effects in efficacy and effectiveness trials. However, besides the studies published several years ago focussing only on venlafaxine,[18,19] there is no up-to-date meta-analytical study investigating possible efficacy and effectiveness effect size differences from RCTs (including unpublished data) and non-RCTs of venlafaxine and duloxetine. Furthermore, only a small number of effectiveness studies were analysed in previous studies (eg Naudet et al. (2011)[18]). Therefore, a larger database, especially including more non-RCTs, is necessary for more reliable results. Moreover, predictors of these effects should be identified in order to gain a better understanding of the effect size differences.

1.1 | Aims of the study

This study aimed at comparing pre-post effect sizes of RCTs and non-RCTs of duloxetine and venlafaxine trials based on a comprehensive meta-analysis. For a more thorough understanding of both study groups, key study characteristics and methodological quality are evaluated. Based on an exploratory approach, predictors that might explain divergencies in effect sizes are to be identified.

2 | MATERIALS AND METHODS

This systematic review and meta-analysis followed an *a priori* defined protocol, which was published at the start of this project.[20] This study is reported in accordance with the PRISMA guideline.[21]

2.1 | Study inclusion and exclusion criteria

The present systematic review and meta-analysis included RCTs, non-randomized controlled trials and observational prospective studies. Eligible studies examined the efficacy or effectiveness of duloxetine or venlafaxine in the acute treatment of unipolar depression in adults (≥18 years). Included studies needed to provide sufficient outcome data of the depression scale. No restrictions were defined for study type, patients' comorbidities, trial duration, language, study or publication year, publication status, doses or dosing regimen. Exclusion criteria were as follows: studies in which duloxetine or venlafaxine were prescribed as adjunctive therapy, pooled analyses, double publications and studies for which no full-text publication was available (eg conference abstracts).

2.2 | Search strategy

For the present study, a sensitive search strategy was developed (Supplement material S1). Studies were selected with a systematic literature search in the electronic databases EMBASE, Medline, PsycLit, PSYNDEXplus and the Cochrane Central Register of Controlled Trials. In order to include unpublished studies, additional hand searches were performed on the manufacturer databases (eg lillytrialguide.com) and on study registry websites (eg clinicaltrials.gov). In addition, references to included studies were screened for potentially relevant publications. The last update of the search was carried out in January 2020.

2.3 | Data extraction and management

Studies were independently selected by two researchers based on titles and abstracts with the aid of the defined inclusion and exclusion criteria. After this screening, potentially relevant publications were examined for eligibility based on the full-text publication by one researcher and reviewed by a second one. Disagreements were resolved by the consultation of a third team member.

Data were extracted in a standardized Microsoft Excel sheet, which was developed based on trial reporting guidelines[22,23] and on instruments assessing the quality of non-randomized trials.[24] The extraction sheet comprised items for methods (study type, number of study centres, number of interventions, study duration), patient sample (number of participants, dropouts), interventions (type of drug, dose, dose regimen), outcome measures (means and standard deviations of baseline (pre-score) and endpoint (post-score)), change score of the outcome of depression symptoms and pre-post correlation (if reported). For depression scores, preference was given to the Hamilton Depression Rating Scale (HAMD)[25] and after that to the Montgomery-Asberg Depression Rating Scale (MADRS)[26] and other depression scales. In case of incompletely reported data, the respective authors were contacted. If essential endpoints were not available, studies were excluded from effect size estimations. For the assessment of the methodological quality of included studies, the Downs and Black (1998)[27] scale was deployed, which includes subscales for reporting, external validity, bias and confounding.

2.4 | Data analysis

The primary outcome of this study was the pre-post effect size of the depression scale after acute therapy with duloxetine or venlafaxine. In order to estimate pre-post effect sizes, the standardized mean difference (SMD) of the baseline depression score (pre) and the depression score after acute antidepressant therapy (post) was used. Effect sizes were standardized based on the standard deviations of the baseline depression score. Differences in effect size between efficacy and effectiveness trials were tested using a mixed-effects model with the 'study design' (RCTs vs. non-RCTs) as study-level moderator.

Missing standard deviations were imputed according to a validated method.[28] In order to estimate the variance of the pre-post effect size, the correlations of baseline and post-intervention depression scores were taken into account. This correlation is hardly reported, and missing correlations were imputed from an individual patient data set of duloxetine trials ($n = 6890$; $r = 0.25$ for HAMD and $r = 0.21$ for MADRS, unpublished data). Different versions of the HAMD scale were used in the included studies; if the item number was not reported in the primary study, it was estimated based on the baseline depression score. This estimation was necessary to calculate the depression severity as percentage of a standardized baseline score. Sensitivity analyses were conducted examining the impact of these imputations and estimations (pre-post correlations, standard deviations, HAMD items).

Study characteristics and methodological quality based on the Downs and Black (1998)[27] scale of RCTs and non-RCTs were compared with Welch's t tests and Pearson's chi-squared tests, respectively.

In our protocol,[20] we planned to develop a sum score based on the extracted data to describe proximity to clinical

routine conditions, but with the development of improved meta-analytic methods we found a multivariate predictor analysis to be more adequate than an analysis of an unvalidated sum score. Therefore, we conducted an exploratory analysis of predictors within an information-theoretic approach.[29,30] In detail, mixed meta-regression models with multiple predictors were fitted using maximum-likelihood estimation to predict standardized mean differences. Models were selected by the corrected Akaike information criterion (AICc) as decision criterion. As a result, the 'best model' and top models within two information criterion units of the best model were identified. For this purpose, the Akaike weights, that is the probability of a certain model to be the best model, were reported. Hence, the relative importance of each predictor ('variable importance') was calculated by summing up the weights of all models in which the predictor appeared. The often used cut-off at 0.8 was used as orientation[31] to distinguish the most relevant variables from not so important ones. The inferences of all predictors across all models were estimated by a multimodel inference approach.[30]

Predictors were selected based on an extensive literature search, clinical relevance and plausibility. As a result of this search, 11 predictors were included in the analysis: 'study design' (RCT/non-RCT), 'depression severity' (baseline depression score in % of the scale maximum), 'dose' (% of recommended maximum dose), 'publication year' (year of article publication or trial end year for unpublished studies), 'number of participants' (number of participants at randomization in a duloxetine or venlafaxine arm), 'setting' (inpatients/outpatients/in- and outpatients), 'dropouts' (number of dropouts of duloxetine or venlafaxine treatment arm as percentage of the total number in the relevant treatment arm), 'published' (unpublished/published trial), 'medication' (duloxetine/venlafaxine), 'number of interventions' (number of intervention arms in the study), 'multicenter' (yes/no). To compare duloxetine and venlafaxine doses, the percentage of the recommended maximum dose according to the U.S. Food and Drug Administration was analysed (120 mg/d for duloxetine; 225 mg/d for venlafaxine). The choice of 11 selected predictors resulted in 2^{11} (= 2048) possible models.

Potential publication bias was estimated by funnel plots and by Egger's regression test, both for the total data set and for subgroups (RCTs/non-RCTs). For sensitivity analyses, outliers were identified through visual inspection of Baujat plots[32] and Cook's distance measures. Furthermore, the impact of the imputations of pre-post correlations and missing standard deviations of the depression score as well as the impact of the estimation of HAMD items was evaluated in sensitivity analyses. We added a graphical post hoc analysis to explore the association of the Downs and Black (1998)[27] subscale 'external validity' and effect sizes. Furthermore, we explored the differences between venlafaxine and duloxetine trials post hoc by applying the model selection approach to each drug separately. To test a possible non-linear relationship of 'dose' and effect sizes, we performed an additional meta-regression with the variables identified as most important, including a quadratic polynomial term for 'dose'.

Data analyses were performed in R Studio using the 'metafor package' and the 'glmulti package'.[33,34]

3 | RESULTS

3.1 | Characteristics of included studies

In total, the literature search yielded 1,848 records. Additionally, 45 records were identified through other sources (eg hand search, study registries, manufacturer databases). After the removal of duplicates, title and abstract screening as well as full-text revision, 245 studies met inclusion criteria and were analysed (Figure 1). Study characteristics and full references of included studies are given in Supplement material S2 and S3.

The data set included 171 RCTs and 74 non-RCTs (non-randomized controlled trials and observational studies). Among them were 85 duloxetine and 164 venlafaxine trials. Four studies reported the effects for both duloxetine and venlafaxine. Table 1 provides further characteristics of included studies.

3.2 | Study characteristics and methodological quality

Overall, most study characteristics in RCTs and non-RCTs were very similar. We were unable to identify any unpublished non-RCT fulfilling our inclusion criteria. The number of treatment arms was significantly higher for RCTs than for non-RCTs, and the settings of the studies were significantly different between RCTs and non-RCTs (Table 1).

With regard to the Downs and Black (1998)[27] scale, RCTs received significantly better ratings for the subscales 'reporting', and for the internal validity subscales 'bias' and 'confounding'. However, no difference was found for the subscale 'external validity' between the study types. We explored the association of the variable 'external validity' and the effect sizes in a scatter plot (Supplement material S4). The plot indicates a very similar pattern of RCTs and non-RCTs in external validity ratings and revealed that, independent form the study type, the included studies received generally very low external validity ratings. Only nine studies (five RCTs[35-39] and four non-RCTs[40-43]) gained higher external validity scores (≥0.66).

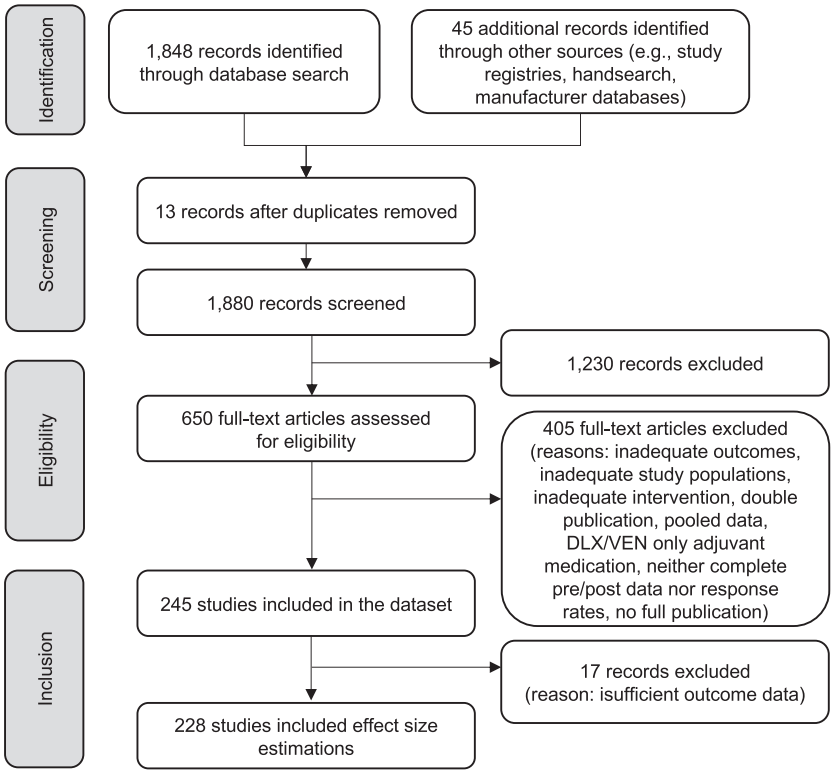

FIGURE 1 Flow chart of included studies.

3.3 | Effect sizes

RCTs (SMD = −3.04, 95% CI = [−3.20; −2.88]; 165 studies; I^2 = 96.43%) showed significantly higher (Δ = 0.41, p = 0.012) pre-post effect sizes than non-RCTs (SMD = −2.62, 95% CI = [−2.90; −2.35], 64 studies, I^2 = 97.94%) (Figure 2; for forest plots showing all included studies see Supplement material S5). Heterogeneity was very high in both groups.

3.4 | Mixed model meta-regression for predictor analyses

Model selection according to AICc identified the 'best model' and 4 other models within two information criterion units of the best model (Table 2). Model 1 had the highest weight with a probability of 5% to be the 'best model'. All 5 models contained the variables 'depression severity', 'dose', 'number of participants', 'dropouts', 'study design' and 'setting' while the predictor 'publication year' occurred in 3, and 'medication' in 2 out of 5 models.

Through the examination across all models, the variables with highest importance across all models were identified (Figure 3). More specifically, the variables 'depression severity', 'dose' and 'number of participants' were most important and exceeded the cut-off for important variables, followed by 'dropouts' and 'study design' with importance values slightly below the cut-off of 0.8.

Multimodel inference of all predictors across all models revealed that the variables 'depression severity' and 'dose' were significant predictors for effect sizes of the depression score ($p \leq 0.05$) with importance values close to 1 (Table 3). Studies with more severely depressed patients showed larger pre-post effect sizes, while studies with higher doses yielded smaller effect sizes. Multimodel inference shows no statistically significant influence for the variables 'study design', 'dropouts' and 'number of participants' even though these variables show high importance values and occur in the best models. Studies with a higher rate of dropouts, larger numbers of participants and non-RCTs were associated with smaller effect sizes.

3.5 | Sensitivity analyses

Through Cook's distance analyses and Baujat plots, eight studies[44-51] were identified as outliers and removed from analyses for a sensitivity analysis. The meta-regression model was only slightly affected by the sensitivity analyses. The most notable change affected the variable 'setting' with a decrease in the importance score by 0.1 when outliers were excluded and 0.2 when studies with imputed SD or with an unclear number of HAMD items were excluded. In addition, the latter analysis resulted in a slightly decreased importance of 'study design' (Δ = 0.1). Moreover, 'study design' was not included in all top models when outliers and when studies with unknown number of HAMD items were excluded. None of the subgroup analyses affected the pre-post effect sizes markedly (Supplement

TABLE 1 Characteristics of included studies.

	Total ($n = 245$)	RCTs ($n = 171$)	non-RCTs ($n = 74$)	Test statistics
Total number of participants	63,416	49,880	13,536	
Drug, no. of studies (%)[a]				
Duloxetine	85 (34)	54 (31)	31 (43)	
Venlafaxine	164 (66)	121 (69)	43 (57)	
Sample size per arm, median (range)	84 (7–3543)	114 (8–821)	50 (7–3543)	$t(78) = -0.44$, $p = 0.659$
Sex (female) in %	65	65	66	$t(109) = -0.57$, $p = 0.572$
Age of participants, mean (SD)	46 (9.87)	45 (9.09)	47 (11.47)	$t(109) = -1.12$, $p = 0.267$
Publication year, median (range)	2008 (1990–2019)	2007 (1990–2019)	2010 (1997–2018)	$t(150) = -2.27$, $p = 0.247$
Published, no. (%)	223 (90)	149 (85)	74 (100)	χ^2 (1) = 10.74, $p = 0.001$
HAMD-17 at baseline, mean (SD) ($n = 124$)	22.66 (3.14)	22.81 (3.43)	22.32 (2.38)	$t(103) = 0.92$, $p = 0.360$
Number of treatment arms, median (range)	2 (1–8)	2 (1–8)	1 (1–5)	$t(203) = 13.43$, $p < 0.001$
Dose, median % of maximum dose (range)	89 (14–178)	75 (14–178)	100 (17–178)	$t(126) = -1.19$, $p = 0.235$
Dosing regimen, number (%)				χ^2 (2) = 1.90, $p = 0.386$
Flexible	130 (52)	87 (50)	43 (58)	
Fixed	104 (42)	78 (45)	26 (35)	
Unclear	15 (6)	10 (6)	5 (7)	
Trial duration in weeks, mean (SD)	8.54 (3.24)	8.38 (2.56)	8.93 (4.44)	$t(94) = -1.01$, $p = 0.317$
Number of study centres, median (range)	11 (1–836)	18.5 (1–100)	1 (1–836)	$t(49) = -0.22$, $p = 0.823$
Setting, number (%[b])				χ^2 (3) = 17.53, $p < 0.001$
Inpatient	32 (13)	25 (14)	7 (9)	
Outpatient	145 (58)	113 (65)	32 (43)	
In- and outpatient	27 (11)	13 (7)	14 (19)	
Unclear	45 (18)	24 (14)	21 (28)	
Dropouts, mean % of participants treated with DLX/VEN (SD)	24.03 (14.43)	24.73 (13.15)	22.21 (17.29)	$t(91) = 1.04$, $p = 0.299$
Methodological quality, mean (SD)[c]				
Reporting	0.94 (0.16)	0.97 (0.15)	0.89 (0.17)	$t(127) = 3.10$, $p = 0.002$
External validity	0.11 (0.19)	0.09 (0.18)	0.14 (0.21)	$t(118) = -1.64$, $p = 0.104$
Bias	0.69 (0.13)	0.74 (0.12)	0.59 (0.10)	$t(166) = 9.53$, $p < 0.001$
Confounding	0.62 (0.26)	0.72 (0.22)	0.40 (0.20)	$t(146) = 11.19$, $p < 0.001$

DLX, duloxetine; VEN, venlafaxine.

[a]Four RCTs provided data for duloxetine and venlafaxine analyses.

[b]Numbers may not sum up to 100% because of rounding.

[c]Downs and Black (1998)[27] subscales, higher values represent higher quality.

material S6.1, 6.3 and 6.4). Using the mean of reported pre-post correlations as imputation ($r = 0.35$ for HAMD and other scales, $r = 0.36$ for MADRS) did not suggest an influence on results (Supplement material S6.2).

Applying the model selection approach separately for venlafaxine and duloxetine revealed some differences for the drugs. For venlafaxine, 'study design' and 'dose' were identified as important and statistically significant predictors, while in the duloxetine data set only 'depression severity' reached the threshold of 0.8 and statistical significance. (Supplement material S6.5 and 6.6).

The meta-regression of the most important predictors including a quadratic polynomial term for 'dose' marginally improved the AICc from 588.22 (linear model) to 587.83 (quadratic polynomial model). However, in this model, both dose predictors (linear and quadratic) were no longer statistically significant. Therefore, this model is not supporting a u-shaped relationship (Supplement material S6.7 and 6.8).

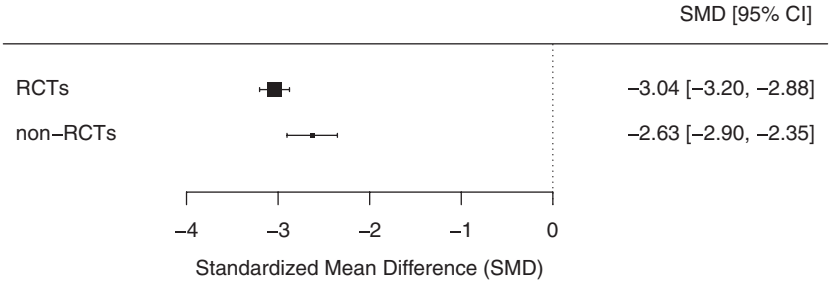

FIGURE 2 Forest plot of overall pre-post effect sizes of RCTs and non-RCTs; SMD, Standardized mean difference.

TABLE 2 Top meta-regression models[a] within two information criterion units of the best model

	Crossmodel predictors[b]	Model-specific predictors	AICc	AICc weight
1	Depression severity + dose + no. of participants + dropouts + study design + setting	Year	572.30	0.050
2		Year + medication	573.13	0.033
3		Year + published	573.15	0.033
4			573.52	0.027
5		Medication	573.53	0.027

AICc, corrected Akaike information criterion.

[a]Model = $y \sim 1 + \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p'} x_{ip'} + u_i$ with y as pre-post effect size of the depression score of the treatment with venlafaxine or duloxetine.

[b]Predictors which occur in all 5 models.

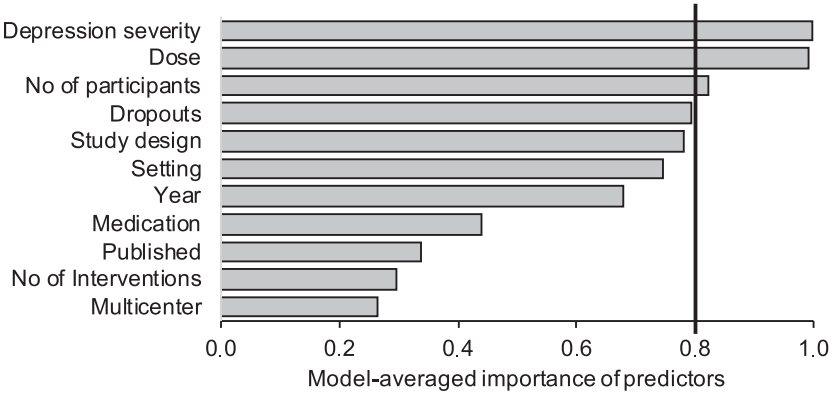

FIGURE 3 Model averaged importance of predictors. The vertical line shows the cut-off for most important variables.

3.6 | Publication bias

The visual inspection of the funnel plots is indicative of publication bias for the full data set and each subgroup (RCTs/non-RCTs). This is supported by the Egger's regression tests (each $p > 0.001$) (Supplement material S7).

4 | DISCUSSION

This study aimed at (1) testing whether RCTs and non-RCTs on the treatment of depressive disorders with duloxetine and venlafaxine differ in pre-post effect sizes for depression ratings, and (2) identifying predictors to explain differences in effect sizes.

The results confirmed the assumption of significantly larger effect sizes in RCTs compared to non-RCTs, which has been previously indicated in the literature.[6,18,19] Although we were able to show a gap in effect sizes between RCTs and non-RCTs, study characteristics of RCTs and non-RCTs were very similar. Methodological evaluations suggested that internal validity and reporting quality received significantly higher ratings in RCTs than in non-RCTs, but the ratings for external validity did not differ between the groups with generally very low external validity ratings for both groups. Therefore, our results question the common assumption that non-RCTs have the advantage of providing higher external validity and being closer to clinical routine conditions.

We were able to identify predictors to explain effect size differences, namely baseline 'depression severity', 'dose', 'number

TABLE 3 Inference of all a priori selected predictors across all models.

	Estimate	SE	z	Pr(>|z|)	Importance
Intercept	−32.657	27.936	−1.169	0.242	1.000
Depression severity	−0.039	0.010	−4.081	0.000	0.999
Dose	0.006	0.002	3.263	0.001	0.990
No. of participants	0.000	0.000	1.546	0.122	0.822
Dropouts	0.008	0.006	1.424	0.154	0.793
Study design[a]	0.292	0.210	1.386	0.166	0.781
Setting[b]	0.043	0.034	1.258	0.208	0.745
Publication year	0.015	0.014	1.074	0.283	0.679
Medication (Duloxetine)[c]	0.091	0.138	0.662	0.508	0.440
Published[d]	−0.074	0.146	−0.506	0.613	0.338
No. of interventions	−0.009	0.033	−0.289	0.772	0.295
Multicentre[e]	0.003	0.010	0.267	0.789	0.263

[a]1 = efficacy vs. 2 = effectiveness.

[b]1 = inpatients. 2 = outpatients. 3 = in- and outpatients. 9 = unclear.

[c]1 = yes (Duloxetine). 2 = no (Venlafaxine).

[d]0 = no. 1 = yes.

[e]0 = no. 1 = yes. 9 = unclear.

of participants', 'dropouts', and 'study design'. The variables 'number of participants', 'study design' and 'dropouts' did not reach statistical significance, but showed high importance scores and occurred in all top models. Following suggestions for the interpretation of results of information-theoretic approaches,[29,30] these predictors should also be considered.

Earlier research on the impact of 'depression severity' on effect sizes showed larger effects of antidepressant medication in more severely depressed patients.[52-54] In line with these studies, in our data, studies with more severely depressed patients showed larger effect sizes on the depression scale. Recent patient-level meta-analyses examining antidepressant-placebo differences found that the efficacy of this medication might be similar for different severity levels.[55,56] While our results suggest that the pre-post effect size is influenced by depression severity, comparative effect sizes may be unaffected.

Contrary to expectations, effect sizes were smaller in studies which report higher maximum doses. Current evidence of the dose-response relationship of antidepressants is limited and inconsistent. While some studies indicate a steady increase in antidepressant response with higher doses,[57] other findings suggest a stronger increase at smaller doses and a weaker increase[58] or even a decrease in drug response at higher doses, that is an inverted u-shaped dose-response relationship.[59] Even though a non-linear relationship is plausible, this was not reflected in our data. Moreover, a relationship of antidepressant dose and of the number of dropouts, which was defined as index of acceptability of a drug,[59] is suggested in the literature.[59] In our study, the variable 'dropouts' achieved high importance scores. Higher dropout rates in the intervention arm were associated with lower effect sizes. Considering previous literature, an interaction effect with the variable 'dose' is possible and requires further consideration. However, meta-regressions including 'dose' on group level may be limited, thus making it necessary to examine the impact of dose on effect sizes and drug's acceptability in non-linear models with patient-level data.

In our data, studies with larger 'numbers of participants' were associated with smaller effect sizes. Previous research analysing placebo response rates found no influence of sample size on the effect sizes.[31,60] Smaller trials have been shown to produce more variable effect sizes, and it has been suggested that more recent (and larger) trials have more stable placebo effects.[61] Thus, controlled placebo trials may be different from our samples, which included some small uncontrolled studies with very large effect sizes.

The variable 'study' design occurred in all top models, but with regard to the overall importance and its changes in sensitivity analyses, the picture is unclear. In line with effect sizes differing between RCTs and non-RCTs, 'study design' appears to be relevant when analysing predictors of effect size. However, other predictors discussed here seem to be more relevant.

Even though the variable 'setting' occurred in all top models in the main analysis, the variable declined in importance values in most sensitivity analyses and no longer occurred in all top models. The importance score below the cut-off supports the assumption that this predictor is of minor importance in explaining effect sizes.

There are some other limitations in our study which have to be considered. Our analyses included only two antidepressants, duloxetine and venlafaxine, and may therefore not apply to other agents. Moreover, we analysed venlafaxine

and duloxetine data in one analysis. As 'medication' was not identified as an important predictor in our primary analysis, our results suggest that pre-post effect sizes depend less on medication than on other predictors. However, separate analyses of the data sets for each medication revealed some differences in predictor models. Considering the large number of predictors and the limited sample sizes for each of these data sets, we find these results not surprising. We assume that they reflect differences of the predictors' variances in duloxetine and venlafaxine data sets.

The information-theoretic approach used to examine predictors combines several advantages, but does not consider model validity and highly depends on an appropriate preselection of predictors.[30] Even though predictors were carefully selected, the set of potential predictors is somehow arbitrary; thus, it cannot be ruled out that a different selection might have led to different results.

Furthermore, the studies included were highly heterogeneous, even within the predictor analysis. The current study comprises a large number of trials with varying study designs; thus, substantial heterogeneity was expected.[62] Our analyses also suggest a significant publication bias for RCTs as well as for non-RCTs studies. Furthermore, the funnel plots suggest a stronger publication bias for non-RCTs than for RCTs and we were unable to identify any unpublished non-RCT.

External validity was rated low for the majority of studies in both groups. Thus, the subscale had very low variance and is therefore of limited value for the explanation of effect size differences. We used the Downs and Black (1998)[27] scale for our analyses, because there is a lack of evaluated instruments allowing the comparison of RCTs and non-RCTs. However, especially for the assessment of external trial validity, more sophisticated instruments to determine proximity to clinical reality should be developed.

The variable 'dose' was analysed as percentage of the maximum recommended dose. As mean doses are hardly reported (especially in non-RCTs), we assumed this operationalization to be the best approximation of drug dose. However, it is possible that, especially in non-RCTs or in studies with flexible dosing regimen, the reported possible maximum dose was not necessarily the dose taken by patients, as was shown recently.[63] As mentioned before, there is a need to replicate the influence of dose on effect sizes using individual patient data meta-analyses.

Moreover, the variable 'depression severity' was included as percentage of the scale maximum at baseline to allow comparisons across different scales. We included studies using the HAMD and MADRS, and one could argue that scales with different maxima have different cut-offs to distinguish mild, moderate and severe depression. However, cut-offs for depression severity have been shown to be very similar for HAMD and MADRS scores in comparison to the relevant scale maxima.[64]

In conclusion, our results show that RCTs and non-RCTs differ in effect sizes with RCTs showing larger treatment effects than non-RCTs. Furthermore, results reveal that external validity in non-RCTs was not superior to RCTs. In line with this, data show that different study designs (RCTs vs. non-RCTs) were very similar with respect to most study characteristics. Thus, other variables are likely to influence effect sizes. Predictor analyses revealed that the variables baseline 'depression severity', 'dose' and 'number of participants' are most important predictors of effect sizes, followed by 'dropouts', while the importance of the variable 'study design' remains unclear. In summary, we identified variables that predict antidepressant effect sizes but these variables were not necessarily descriptive attributes associated with 'study design' (RCT or non-RCT).

Our results question the common assumption that non-RCTs are generally better suited to describe a drug's effectiveness in clinical practice than RCTs. In our study sample, non-RCTs did not show a significant advantage over RCTs in terms of external validity but were rated to be lower in internal validity and reporting quality. Thus, to achieve clinical relevance, a careful assessment of both internal and external study validity is required. The low external validity of RCTs as well as of non-RCTs limits the understanding of whether or not clinical studies offer a realistic assessment of effects in clinical practice. A better understanding would help us design trials that are better suited to inform clinical practice and, perhaps, improve clinical practise to achieve stronger treatment effects.

ACKNOWLEDGEMENTS

We would like to thank Ines Fiedler und Ann-Christin Holtrup for their support in the data management.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/acps.13293.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Carolin Schneider https://orcid.org/0000-0002-6197-1017
Johanna Breilmann https://orcid.org/0000-0002-0591-5969
Markus Kösters https://orcid.org/0000-0001-7018-6021

REFERENCES

1. Atkins MS, Rusch D, Mehta TG, Lakind D. Future directions for dissemination and implementation science: aligning ecological theory and public health to close the research to practice gap. J Clin Child Adolesc Psychol. 2016;45:215-226. https://doi.org/10.1080/15374416.2015.1050724

2. Paquet P-L. Probing the evidence: Can we bridge the theory-practice gap in language research? Book review. Emerg Trends Educat. 2019;2:102-105. https://doi.org/10.19136/etie.a2n3.3451.
3. Pullins EB, Timonen H, Kaski T, Holopainen M. An Investigation of the theory practice gap in professional sales. J Market Theor Pract. 2017;25:17-38. https://doi.org/10.1080/10696679.2016.1236665.
4. Herden J, Wittekind C, Weissbach L. Discrepancy between theory and practice of the clinical tumor-nodes-metastasis (TNM) classification for localized prostate cancer. Ann Transl Med. 2019;7:250. https://doi.org/10.21037/atm.2019.05.42.
5. Wolcott MD, Lobczowski NG, Lyons KL, McLaughlin JE. Design-based research: Connecting theory and practice in pharmacy educational intervention research. Curr Pharm Teach Learn. 2019;11:309-318. https://doi.org/10.1016/j.cptl.2018.12.002.
6. Nordon C, Karcher H, Groenwold RHH, et al. The "efficacy-effectiveness gap": historical background and current conceptualization. Value Health. 2016;19:75-81. https://doi.org/10.1016/j.jval.2015.09.2938.
7. Weiss AP, Guidi J, Fava M. Closing the efficacy-effectiveness gap: translating both the what and the how from randomized controlled trials to clinical practice. J Clin Psychiatry. 2009;70:446-449. https://doi.org/10.4088/JCP.08com04901.
8. Godwin M, Ruhland L, Casson I, et al. Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity. BMC Med Res Methodol. 2003;3:28. https://doi.org/10.1186/1471-2288-3-28.
9. Cipriani A, Furukawa TA, Salanti G, et al. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. Lancet. 2018;391:1357-1366. https://doi.org/10.1176/appi.focus.16407.
10. Schueler Y-B, Koesters M, Wieseler B, et al. A systematic review of duloxetine and venlafaxine in major depression, including unpublished data. Acta Psychiatr Scand. 2011;123:247-265. https://doi.org/10.1111/j.1600-0447.2010.01599.x.
11. Cipriani A, Furukawa TA, Salanti G, et al. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. Lancet. 2009;373:746-758. https://doi.org/10.1016/S0140-6736(09)60046-5.
12. Smith D, Dempster C, Glanville J, Freemantle N, Anderson I. Efficacy and tolerability of venlafaxine compared with selective serotonin reuptake inhibitors and other antidepressants: a meta-analysis. Br J Psychiatry. 2002;180:396-404. https://doi.org/10.1192/bjp.180.5.396.
13. Lisinski A, Hieronymus F, Näslund J, Nilsson S, Eriksson E. Item-based analysis of the effects of duloxetine in depression: a patient-level post hoc study. Neuropsychopharmacol. 2020;45:553-560. https://doi.org/10.1038/s41386-019-0523-4.
14. Depp C, Lebowitz BD. Clinical trials: bridging the gap between efficacy and effectiveness. Int Rev Psychiatry. 2007;19:531-539. https://doi.org/10.1080/09540260701563320.
15. Glasgow RE, Lichtenstein E, Marcus AC. Why don't we see more translation of health promotion research to practice? Rethinking the efficacy-to-effectiveness transition. Am J Public Health. 2003;93:1261-1267. https://doi.org/10.2105/AJPH.93.8.1261.
16. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. JAMA. 2003;290:1624-1632. https://doi.org/10.1001/jama.290.12.1624.
17. Mulder RT, Frampton C, Joyce PR, Porter R. Randomized controlled trials in psychiatry. Part II: their relationship to clinical practice. Aust N Z J Psychiatry. 2003;37:265-269. https://doi.org/10.1046/j.1440-1614.2003.01176.x.
18. Naudet F, Maria AS, Falissard B. Antidepressant response in major depressive disorder: a meta-regression comparison of randomized controlled trials and observational studies. PLoS One. 2011;6(6):e20811. https://doi.org/10.1371/journal.pone.0020811
19. van der Lem R, van der Wee NJA, van Veen T, Zitman FG. Efficacy versus effectiveness: a direct comparison of the outcome of treatment for mild to moderate depression in randomized controlled trials and daily practice. Psychother Psychosom. 2012;81:226-234. https://doi.org/10.1159/000330890.
20. Koesters M, Holtrup A-C, Fiedler I, Becker T. Systematic Evaluation of the "Efficacy-Effectiveness Gap" in the Treatment of Depression with Venlafaxine and Duloxetine (Protocol); 2013. Available from https://oparu.Uni-Ulm.De/xmlui/bitstream/123456789/2867/1/vts_8406_12363.Pdf [Accessed 7 April 2020].
21. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med. 2009;151:264-269. https://doi.org/10.7326/0003-4819-151-4-200908180-00135.
22. Zwarenstein M, Treweek S, Gagnier JJ, et al. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. BMJ. 2008;337(nov11 2):a2390. https://doi.org/10.1136/bmj.a2390
23. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. Int J Surg. 2014;12:1495-1499. https://doi.org/10.1016/j.ijsu.2014.07.013.
24. Sanderson S, Tatt ID, Higgins JPT. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. Int J Epidemiol. 2007;36:666-676. https://doi.org/10.1093/ije/dym018.
25. Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry. 1960;23:56-62. https://doi.org/10.1136/jnnp.23.1.56.
26. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. Br J Psychiatry. 1979;134:382-389. https://doi.org/10.1192/bjp.134.4.382.
27. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J Epidemiol Community Health. 1998;52:377-384. https://doi.org/10.1136/jech.52.6.377.
28. Furukawa TA, Barbui C, Cipriani A, Brambilla P, Watanabe N. Imputing missing standard deviations in meta-analyses can provide accurate results. J Clin Epidemiol. 2006;59:7-10. https://doi.org/10.1016/j.jclinepi.2005.06.006.
29. Burnham KP, Anderson DR. Model Selection and Multimodel Inference*:* A Practical Information-Theoretic Approach, 2nd edn. New York: Springer; 2002.
30. Burnham KP, Anderson DR, Huyvaert KP. AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. Behav Ecol Sociobiol. 2011;65:23-35. https://doi.org/10.1007/s00265-010-1029-6.
31. Breilmann J, Furukawa TA, Becker T, Koesters M. Differences in the placebo response in duloxetine and venlafaxine trials. Acta Psychiatr Scand. 2018;137:472-480. https://doi.org/10.1111/acps.12881.

32. Baujat B, Mahé C, Pignon J-P, Hill C. A graphical method for exploring heterogeneity in meta-analyses: application to a meta-analysis of 65 trials. Stat Med. 2002;21:2641-2652. https://doi.org/10.1002/sim.1221.
33. Calcagno V, Mazancourt CD. glmulti: An R package for easy automated model selection with (Generalized) linear models. J Stat Soft. 2010;34:1-29. https://doi.org/10.18637/jss.v034.i12.
34. Viechtbauer W. Conducting meta-analyses in R with the metafor package. J Stat Soft. 2010;36:1-48.
35. Bhagwagar Z, Torbeyns A, Hennicken D, et al. Assessment of the efficacy and safety of BMS-820836 in patients with treatment-resistant major depression: results from 2 randomized double-blind studies. J Clin Psychopharmacol. 2015;35:454-459. https://doi.org/10.1097/JCP.0000000000000335.
36. Dichter GS, Tomarken AJ, Freid CM, Addington S, Shelton RC. Do venlafaxine XR and paroxetine equally influence negative and positive affect? J Affect Disord. 2005;85:333-339. https://doi.org/10.1016/j.jad.2004.10.007.
37. F1J-MC-HMCA. Duloxetine versus placebo in the treatment of fibromyalgia patients with or without major depressive disorder.
38. Kok RM, Nolen WA, Heeren TJ. Venlafaxine versus nortriptyline in the treatment of elderly depressed inpatients: a randomised, double-blind, controlled trial. Int J Geriatr Psychiatry. 2007;22:1247-1254. https://doi.org/10.1002/gps.1823.
39. Güzel Özdemir P, Boysan M, Smolensky MH, Selvi Y, Aydin A, Yilmaz E. Comparison of venlafaxine alone versus venlafaxine plus bright light therapy combination for severe major depressive disorder. J Clin Psychiatry. 2015;76:e645-e654. https://doi.org/10.4088/JCP.14m09376.
40. Mulvahill JS, Nicol GE, Dixon D, et al. Effect of metabolic syndrome on late-life depression: associations with disease severity and treatment resistance. J Am Geriatr Soc. 2017;65:2651-2658. https://doi.org/10.1111/jgs.15129.
41. Di Nasso E, Chiesa A, Serretti A, de Ronchi D, Mencacci C. Clinical and demographic predictors of improvement during duloxetine treatment in patients with major depression: an open-label study. Clin Drug Investig. 2011;31:385-405. https://doi.org/10.2165/11588800-000000000-00000.
42. Hung C-I, Liu C-Y, Yang C-H, Wang S-J. Negative impact of migraine on quality of life after 4 weeks of treatment in patients with major depressive disorder. Psychiatry Clin Neurosci. 2012;66:8-16. https://doi.org/10.1111/j.1440-1819.2011.02286.x.
43. Joffe H, Soares CN, Petrillo LF, et al. Treatment of depression and menopause-related symptoms with the serotonin-norepinephrine reuptake inhibitor duloxetine. J Clin Psychiatry. 2007;68:943-950. https://doi.org/10.4088/jcp.v68n0619.
44. Badyal DK, Khosla PP, Deswal RS, Matreja PS. Safety and efficacy of duloxetine versus venlafaxine in major depression in Indian patients. JK Science. 2006;8:195-199.
45. Araya AV, Rojas P, Fritsch R, et al. Early response to venlafaxine antidepressant correlates with lower ACTH levels prior to pharmacological treatment. Endocrine. 2006;30:289-298.
46. Boulenger J-P, Loft H, Olsen CK. Efficacy and safety of vortioxetine (Lu AA21004), 15 and 20 mg/day: A randomized, double-blind, placebo-controlled, duloxetine-referenced study in the acute treatment of adult patients with major depressive disorder. Int Clin Psychopharmacol. 2014;29(3):138-149. https://doi.org/10.1097/YIC.0000000000000018.
47. Sauer H, Huppertz-Helmhold S, Dierkes W. Efficacy and safety of venlafaxine ER vs. amitriptyline ER in patients with major depression of moderate severity. Pharmacopsychiatry. 2003;36:169-175. https://doi.org/10.1055/s-2003-43052.
48. Tourian KA, Padmanabhan SK, Groark J, Brisard C, Farrington D. Desvenlafaxine 50 and 100 mg/d in the treatment of major depressive disorder: An 8-week, phase III, multicenter, randomized, double-blind, placebo-controlled, parallel-group trial and a post hoc pooled analysis of three studies. Clin Ther. 2009;31:1405-1423.
49. Cristancho P, O'Connor B, Lenze EJ, et al. Treatment emergent suicidal ideation in depressed older adults. Int J Geriatr Psychiatry. 2017;32:596-604.
50. Gaynor PJ, Gopal M, Zheng W, Martinez JM, Robinson MJ, Marangell LB. A randomized placebo-controlled trial of duloxetine in patients with major depressive disorder and associated painful physical symptoms. Curr Med Res Opin. 2011;27:1849-1858.
51. Davies J, Lloyd KR, Jones IK, Barnes A, Pilowsky LS. Changes in regional cerebral blood flow with venlafaxine in the treatment of major depression. Am J Psychiatry. 2003;160:374-376. https://doi.org/10.1176/appi.ajp.160.2.374.
52. Fournier JC, DeRubeis RJ, Hollon SD, et al. Antidepressant drug effects and depression severity: a patient-level meta-analysis. JAMA. 2010;303:47-53. https://doi.org/10.1001/jama.2009.1943.
53. Schmitt AB, Bauer M, Volz H-P, et al. Differential effects of venlafaxine in the treatment of major depressive disorder according to baseline severity. Eur Arch Psychiatry Clin Neurosci. 2009;259:329-339. https://doi.org/10.1007/s00406-009-0003-7.
54. Shelton RC, Andorn AC, Mallinckrodt CH, et al. Evidence for the efficacy of duloxetine in treating mild, moderate, and severe depression. Int Clin Psychopharmacol. 2007;22:348-355. https://doi.org/10.1097/YIC.0b013e32821c6189.
55. Furukawa TA, Maruo K, Noma H, et al. Initial severity of major depression and efficacy of new generation antidepressants: individual participant data meta-analysis. Acta Psychiatr Scand. 2018;137:450-458. https://doi.org/10.1111/acps.12886.
56. Hieronymus F, Lisinski A, Nilsson S, Eriksson E. Influence of baseline severity on the effects of SSRIs in depression: an item-based, patient-level post-hoc analysis. Lancet Psychiatry. 2019;6:745-752. https://doi.org/10.1016/S2215-0366(19)30216-0.
57. Jakubovski E, Varigonda AL, Freemantle N, Taylor MJ, Bloch MH. Systematic review and meta-analysis: dose-response relationship of selective serotonin reuptake inhibitors in major depressive disorder. Am J Psychiatry. 2016;173:174-183. https://doi.org/10.1176/appi.ajp.2015.15030331.
58. Bollini P, Pampallona S, Tibaldi G, Kupelnick B, Munizza C. Effectiveness of antidepressants: meta-analysis of dose-effect relationships in randomised clinical trials. Br J Psychiatry. 1999;174:297-303.
59. Furukawa TA, Cipriani A, Cowen PJ, Leucht S, Egger M, Salanti G. Optimal dose of selective serotonin reuptake inhibitors, venlafaxine, and mirtazapine in major depression: a systematic review and dose-response meta-analysis. Lancet Psychiatry. 2019;6:601-609. https://doi.org/10.1016/S2215-0366(19)30217-2.
60. Furukawa TA, Cipriani A, Atkinson L, et al. Revisiting placebo response rates in antidepressant trials: a systematic review of published and unpublished double-blind studies. Lancet Psychiatry. 2016;3:1059-1066. https://doi.org/10.1016/S2215-0366(16)30307-8.
61. Khan A, Mar KF, Brown WA. The conundrum of depression clinical trials: one size does not fit all. Int Clin Psychopharmacol. 2018;33:239-248. https://doi.org/10.1097/YIC.0000000000000229.
62. Higgins JPT. Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. Int J Epidemiol. 2008;37:1158-1160. https://doi.org/10.1093/ije/dyn204.

63. Hieronymus F, Eriksson E. Inclusion of flexible-dose trials in the meta-analysis of SSRI dose-dependency. Am J Psychiatry. 2016;173:836. https://doi.org/10.1176/appi.ajp.2016.16030304.
64. Müller M. Differentiating moderate and severe depression using the Montgomery-Åsberg depression rating scale (MADRS). J Affect Disord. 2003;77:255-260. https://doi.org/10.1016/S0165-0327(02)00120-9.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Schneider C, Breilmann J, Reuter B, Becker T, Kösters M. Systematic evaluation of the 'efficacy-effectiveness gap' in the treatment of depression with venlafaxine and duloxetine. *Acta Psychiatr Scand*. 2021;144:113–124. https://doi.org/10.1111/acps.13293