

1 **On the treatment effect heterogeneity of antidepressants in major depression.**

2 **A Bayesian meta-analysis**

3

4 Constantin Volkmann (MD)<sup>1</sup>, Alexander Volkmann (PhD)<sup>2,\*</sup>, Christian A. Müller (MD)<sup>1,\*</sup>

5

6 <sup>1</sup> *Department of Psychiatry and Psychotherapy, Charité - Universitätsmedizin Berlin, Campus*

7 *Charité Mitte, Charitéplatz 1, 10117 Berlin, Germany;*

8 <sup>2</sup> *Marienburger Strasse 21, 10405 Berlin, Germany;*

9 \*These authors contributed equally.

10

11

12

13 Corresponding author:

14 Dr. med. Constantin Volkmann

15 Department of Psychiatry and Psychotherapy

16 Charité – Universitätsmedizin Berlin, Charité Campus Mitte

17 Charitéplatz 1

18 10117 Berlin

19 Tel: +49 30 450 617406

20 Email: constantin.volkmann@charite.de

21

22

23

24 Tables: 1

25 Figures: 6

26 Word count: 5685

27 **Abstract**

28 **Background:** The average treatment effect of antidepressants in major depression was  
29 found to be about 2 points on the 17-item Hamilton Depression Rating Scale, which lies  
30 below clinical relevance. Here, we searched for evidence of a relevant treatment effect  
31 heterogeneity that could justify the usage of antidepressants despite their low average  
32 treatment effect.

33 **Methods:** Bayesian meta-analysis of 169 randomized, controlled trials including 58,687  
34 patients. We considered the effect sizes log variability ratio (lnVR) and log coefficient of  
35 variation ratio (lnCVR) to analyze the difference in variability of active and placebo response.  
36 We used Bayesian random-effects meta-analyses (REMA) for lnVR and lnCVR and fitted a  
37 random-effects meta-regression (REMR) model to estimate the treatment effect variability  
38 between antidepressants and placebo.

39 **Results:** The variability ratio was found to be very close to 1 in the best fitting models  
40 (REMR: 95% HPD [0.98, 1.02], REMA: 95% HPD [1.00, 1.02]). The between-study variance  
41  $\tau^2$  under the REMA was found to be low (95% HPD [0.00, 0.00]). Simulations showed that a  
42 large treatment effect heterogeneity is only compatible with the data if a strong correlation  
43 between placebo response and individual treatment effect is assumed.

44 **Conclusions:** The published data from RCTs on antidepressants for the treatment of major  
45 depression is compatible with a near-constant treatment effect. Although it is impossible to  
46 rule out a substantial treatment effect heterogeneity, its existence seems rather unlikely.  
47 Since the average treatment effect of antidepressants falls short of clinical relevance, the  
48 current prescribing practice should be re-evaluated.

49

50

51

52

53

54

55

56

57

58

59

60

61 Keywords: Major depressive disorder, Antidepressants, Individual treatment effect,  
62 Treatment effect heterogeneity, Average treatment effect.

63

64

## 65 **Introduction**

66

67 Depression is one of the most frequent psychiatric disorders and poses a major burden for  
68 individuals and society; it affects more than 300 million people worldwide and is ranked as  
69 the single largest contributor to disability [1]. The first-line treatment usually consists of  
70 psychotherapy and/or pharmacotherapy with antidepressant drugs [2, 3]. Within the last  
71 decades, the number of prescriptions of antidepressants has continuously increased in  
72 several regions of the world [4, 5]. However, whether antidepressants are effective in the  
73 treatment of major depression has been a highly controversial debate for many years [6-9].  
74 A recent meta-analysis by Cipriani et al. comprising 522 randomized, controlled trials (RCTs)  
75 of 21 antidepressants in 116 477 participants reported that all antidepressants were more  
76 effective than placebo in reducing depressive symptoms [10]. In contrast, the authors of a  
77 recent re-analysis criticised this meta-analysis for not taking into account several biases,  
78 such as publication bias [12]. They concluded that “the evidence does not support definitive  
79 conclusions regarding the efficacy of antidepressants for depression in adults, including  
80 whether they are more efficacious than placebo for depression”.

81 Albeit these contradictory conclusions, both analyses used the same dataset. The so-called  
82 average treatment effect, which measures the difference in mean outcomes between active  
83 and control group, was about 2 points on the 17-item Hamilton Depression Rating Scale  
84 (HAMD-17) [13] in this dataset [12]. According to Leucht et al. [14], a reduction of up to 3  
85 points on the HAMD corresponds to “no change” in the Clinical Global Impressions -  
86 Improvement Scale (CGI-I) [15] and the assumed threshold of clinical significance is 7 points  
87 [16]. Thus, a reduction of 2 points on the HAMD is not detectable by the treating physician  
88 and is presumably clinically irrelevant.

89 Crucially, Munkholm et al. [12] reported the average treatment effect as an outcome  
90 parameter, whereas Cipriani et al. [10] reported the odds ratio (OR) of “response rates”,  
91 signifying the fraction of patients crossing the rather arbitrary threshold of 50% in symptom  
92 reduction (“responders”). This approach translates into a number-needed-to-treat (NNT) of  
93 around 8-10 for “response” [17].

94

### 95 *Categorisation of continuous variables*

96 Categorising patients into “responders” and “non-responders” based on crossing an arbitrary  
97 threshold on a continuous scale has frequently been criticised by statisticians as it may lead  
98 to an artificial inflation of the measured treatment effect and to a loss of power [18, 19]. It  
99 may create the illusion of a subgroup of patients that benefit particularly well from a given

100 treatment where none exists. However, a NNT of 8 is compatible with every patient having  
101 the exact same treatment effect of 2 points on the HAMD-17 [12, 20]. Only if a substantial so-  
102 called treatment effect heterogeneity exists, meaning that there are true “responders” and  
103 “non-responders”, the calculation of response rates may be legitimate and the average  
104 treatment effect may not be an appropriate outcome measure. However, in the absence of  
105 clear evidence for a relevant treatment effect heterogeneity, the average treatment effect is  
106 the best predictor of the individual treatment effect [20, 21].

107

#### 108 *Treatment effect heterogeneity*

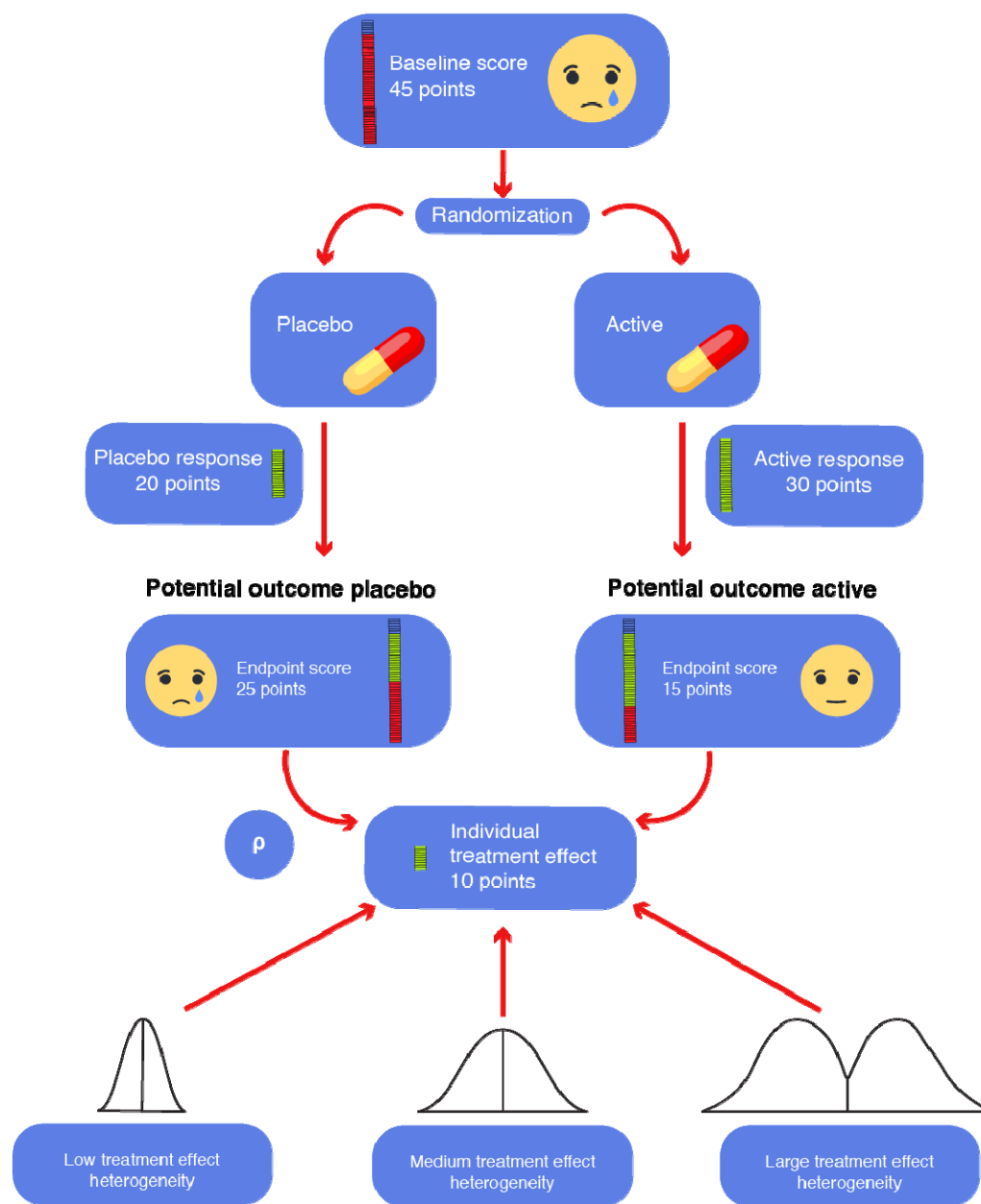
109 Treatment effect heterogeneity describes the extent to which a treatment might affect  
110 different individuals differentially. In other words, some patients may benefit a lot, others may  
111 be harmed by a given treatment, possibly resulting in a null finding when only considering the  
112 average treatment effect in clinical trials. However, the existence of a clinically relevant  
113 treatment effect heterogeneity, albeit widely believed and intuitively plausible, has not been  
114 shown yet.

115

116

117

118



119

120

121

122

123

124

125

126

127

128

129

130

**Figure 1:** Visualization of a hypothetical patient in a randomized placebo-controlled trial. The patient is randomized to either the placebo or the active arm, corresponding to two hypothetical “potential outcomes”. Only one of which can ever be observed, as a single patient cannot receive both placebo and the active intervention at the same time. The difference between the two potential outcomes corresponds to the “individual treatment effect” of the intervention (here, a clinically relevant difference of 10 HAMD-17 points). The individual treatment effect is unobservable and can be imaged to be drawn from hypothetical distributions of the treatment effect. The variance of this distribution corresponds to the treatment effect heterogeneity. The factor  $\rho$  is the correlation between the placebo response and the individual treatment effect. Here, we assume that a given patient has a fixed individual treatment effect.

131

132 *Investigation of treatment effect heterogeneity*

133 Simply labeling patients as “responders” and “non-responders” based on crossing an  
134 arbitrary threshold on a continuous outcome scale is not a valid way to investigate variation  
135 in individual treatment effect [22]. In order to assess treatment effect heterogeneity from the  
136 data of parallel group trials, the comparison of variances between the active and the control  
137 condition has been proposed [22, 23]. Here, an increase in variance in the active group might  
138 be a signal of a variation in the individual treatment effect [22].

139 Following a recent publication by Winkelbeiner et al. [21], analyzing differences in variances  
140 in 52 randomized, placebo-controlled antipsychotic drug trials, the present analysis aimed to  
141 assess the evidence for individual antidepressant drug response using the open dataset of  
142 the largest meta-analysis of the efficacy of antidepressants in major depressive disorder  
143 [10].

144 Here, we addressed the following research question: What is the evidence for a relevant  
145 treatment effect heterogeneity of antidepressants in the treatment of depression that justifies  
146 their usage despite the lack of a clinically relevant average treatment effect?

147

148 **Methods**

149

150 *Data Acquisition*

151 We obtained the dataset of the meta-analysis by Cipriani et al. [10] from the Mendeley  
152 database (<https://data.mendeley.com/datasets/83rthbp8ys/2>). This study included all RCTs  
153 comparing 21 antidepressants with placebo or another active antidepressant as oral  
154 monotherapy for the acute treatment of adults ( $\geq 18$  years old and of both sexes) with a  
155 primary diagnosis of major depressive disorder according to standard operationalized  
156 diagnostic criteria (Feighner criteria, Research Diagnostic Criteria, DSM-III, DSM-III-R, DSM-  
157 IV, DSM-5, and ICD-10). For further details on the inclusion criteria and study characteristics,  
158 see the original study [10].

159

160 *Data extraction and processing*

161 Of the total of 522 studies we kept the 304 that included a placebo arm. We excluded all  
162 studies for which the reported endpoint did not represent the change from baseline, leaving  
163 us with a total of 169 studies for the analysis (see PRISMA flow diagram, supplementary  
164 data, figure 1). We extracted both the mean and the standard deviation of pre- and post-  
165 treatment outcome difference scores (the “response”). The studies included in the data set  
166 comprised 8 different depression scales, namely HAMD-17, HAMD-21, HAMD-24, HAMD  
167 unspecified, HAMD-29, HAMD-31, Montgomery-Åsberg Depression Rating Scale (MADRS)

168 [24] and IDS-IVR-30 [25, 26]. Studies with different treatment arms were aggregated  
169 according to the recommendation of the Cochrane Collaboration [27]. In this manuscript, we  
170 define response as pre-post-difference of a given outcome scale.

171

## 172 *Statistical Analysis*

173 We considered two different effect size statistics as suggested by Nakagawa et al. [28] to  
174 analyze the difference in variability of active and placebo response.

175

### 176 1. The *log variability ratio*

$$177 \quad \ln VR = \ln \left( \frac{SD_a}{SD_p} \right) + \frac{1}{2(n_a-1)} - \frac{1}{2(n_p-1)}, \text{ where:}$$

$SD = \text{standard deviation, } a = \text{active, } p = \text{placebo}$

178

### 179 2. The *log coefficient of variation ratio*

$$180 \quad \ln CVR = \ln \left( \frac{CV_a}{CV_p} \right) + \frac{1}{2(n_a-1)} - \frac{1}{2(n_p-1)}, \text{ where: } CV = \frac{SD}{Mean}$$

181

182 These two effect sizes differ in the way they account for differences in means between the  
183 active and the placebo group. Whereas InVR assumes no correlation between concurrent  
184 changes in mean response and standard deviation of response, InCVR measures differences  
185 in variability between groups after accounting for differences in mean response. If the active  
186 and placebo arms have equal variance, a VR (or CVR) of 1 would be expected. A value  
187 greater than 1 indicates a larger variability in the active group.

188 A variability ratio that substantially differs from 1 implies a considerable treatment effect  
189 heterogeneity. Conversely, a VR of around 1 is compatible with a near-constant treatment  
190 effect but does not exclude the existence of treatment effect heterogeneity. It should be  
191 noted, that it is impossible to disprove the existence of a subgroup with a substantially  
192 greater than average effect. However, the magnitude of the treatment effect heterogeneity  
193 can be bounded by the distance of the variability ratio VR from the value 1.

194 All statistical analyses were carried out in the programming language Python (version 3.7)  
195 and the probabilistic programming language Stan (with pystan version 2.18.1.0 as a Python  
196 interface). We used a Bayesian approach to fit all our models using weakly informative  
197 priors. Firstly, we used a Bayesian random-effects meta-analysis (REMA) for the two effect  
198 statistics InVR and InCVR. Secondly, we used a Bayesian random-effects meta-regression  
199 (REMR) to fit the InVR effect statistic with the natural logarithm of the response ratio (lnRR)  
200 as a regressor [28], which is defined as:

201

202

203 3. The *log of the response ratio*

204  $\ln RR = \ln \left( \frac{Mean_a}{Mean_p} \right)$ , where “Mean” denotes the mean of the “Response” variable

205

206 An additional complexity in our analysis, as compared to recent analyses [21, 28, 29], came  
207 from the fact that our data set contained several different depression scales (several versions  
208 of the HAMD and the MADRS, see supplementary figure 2). For our analysis we made the  
209 assumption that these different scales are (locally) linearly transformable into each other.  
210 This assumption is well supported by the literature [30]. Fortunately, the InVR and InCVR  
211 effect statistics are invariant under linear transformations of the outcome scale.

212

213 *Random-effects meta-analysis (REMA)*

214 We applied a Bayesian random-effects meta-analysis in order to estimate the effect sizes  
215  $ES = \ln VR, \ln CVR$ . For the REMA, the following model was applied, where  $\mu$  equals the “true”  
216 mean of the effect size. Finally,  $\eta$  represents the between-study-variance.

217

218  $ES_i \sim N(\mu + \eta_i, SD_i^2)$

219  $\eta_i \sim N(0, \tau^2)$

220

221 We specified the following weakly-informative hyper-priors:

222

$$\mu \sim \text{Cauchy}(0,1)$$

$$\tau \sim \text{Half - Cauchy}(0,1)$$

223

224 *Random-effects meta-regression (REMR)*

225 This approach is a “contrast-based” version of the “arm-based” meta-analysis in Nakagawa  
226 et al. [28] which models the log of the standard deviation of the outcome directly in a multi-  
227 level meta-regression. For the REMR, the following model was applied, where  $\mu$  equals the  
228 “true” mean of  $\ln VR$  over all studies and  $X$  the “true” value of  $\ln RR$ , if we account for  
229 measurement error. The variable  $\beta$  is the regression coefficient for  $X$  and thus signifies the  
230 degree of linear association between  $\ln VR$  and  $\ln RR$ . Finally,  $\eta$  represents the between-  
231 study-variance.

232

$$\ln VR_i \sim N(Y_i, s^2_{\ln VR_i})$$

$$Y_i = \mu + \beta * X_i + \eta_i$$

233  $\eta_i \sim N(0, \tau^2)$

$$X_i \sim N(0,100)$$



$$\ln RR_i \sim N(X_i, s^2_{\ln RR_i})$$

234

235 We specified the following weakly-informative hyper-priors:

236

$$\mu \sim \text{Cauchy}(0,1)$$

$$\beta \sim \text{Cauchy}(0,1)$$

$$\tau \sim \text{Half-Cauchy}(0,1)$$

237

### 238 *Simulation experiments*

239 For each simulation, the response under placebo and the response under treatment were  
240 simulated for 1000 patients. The response under placebo was drawn from a right skewed  
241 distribution with mean and standard deviation of 8.8 and 7.7 points on the HAMD-17 scale  
242 (based on Cipriani data [10]), respectively. For each patient, an outcome under treatment  
243 was computed from a mixed Gaussian distribution with a given  $SD_{TE}$ , where the outcome  
244 under placebo and the individual treatment effect were required to be correlated by the  
245 correlation coefficient  $\rho$ .

246 This yielded a potential outcome under placebo and a potential outcome under active  
247 treatment for every patient with a corresponding individual treatment effect (see figure 1 for  
248 illustration). Half of the patients were then randomly selected for treatment, the other half was  
249 assigned to placebo. Note that only one of these two outcomes can be observed in a real  
250 experiment.

251

## 252 **Results**

253

### 254 *Study selection*

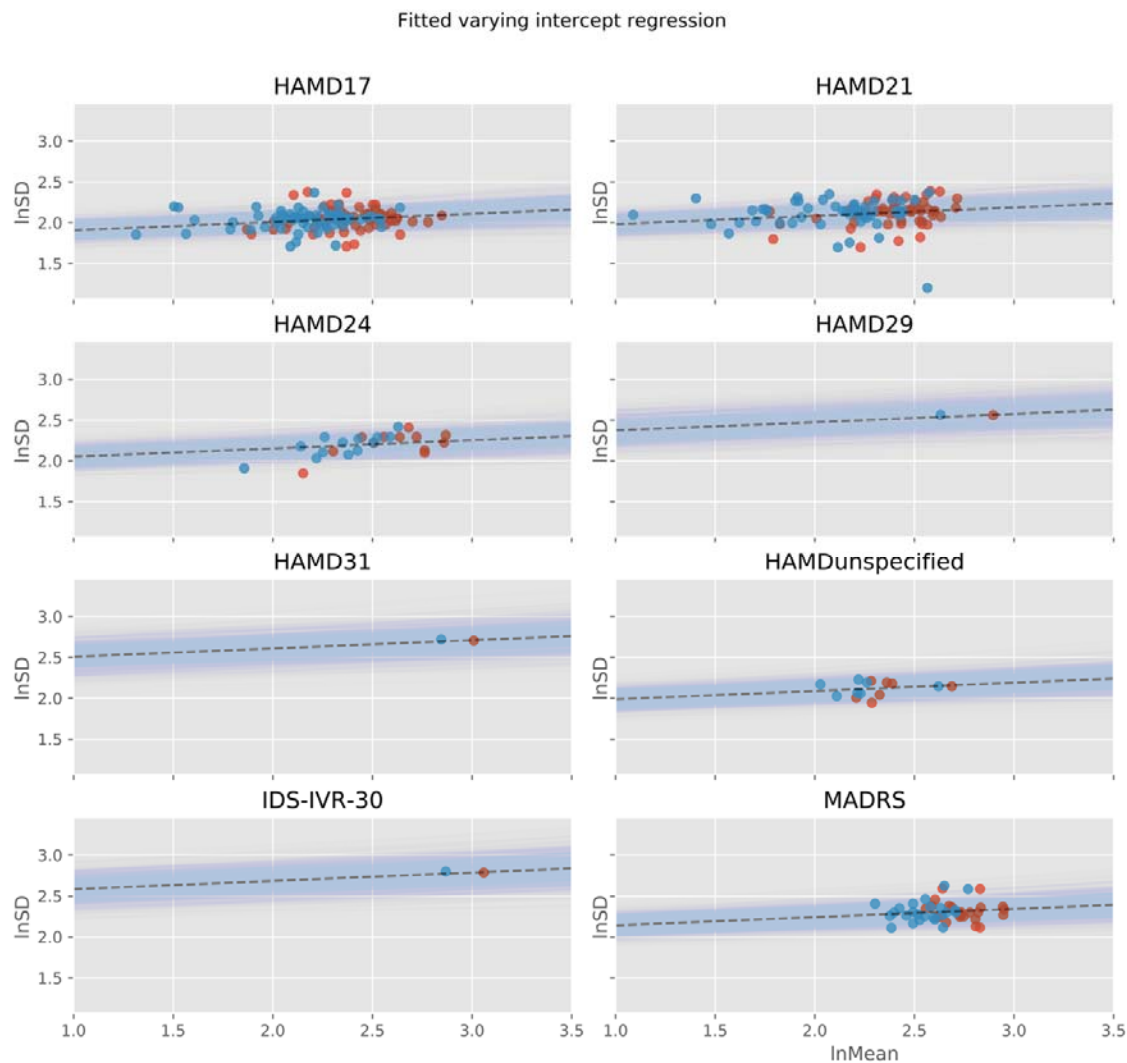
255 As mentioned above, we included 169 placebo-controlled studies that reported mean and  
256 standard deviation of change in depression scores. These studies included data on 58,687  
257 patients treated with 21 different antidepressants.

258

### 259 *Correlation between mean and standard deviation of depression scores*

260 In order to identify the more appropriate effect size (VR or CVR), we investigated the linear  
261 association between the logarithm of the mean response scores and the logarithm of their  
262 standard deviation using a varying intercept model, where the intercepts were allowed to  
263 vary between studies with different depression scales. Fitting a Bayesian varying intercept  
264 regression model with measurement error with  $\ln \text{Mean}$  as independent variable and  $\ln \text{SD}$  as  
265 dependent variable, we get a posterior mean for the slope coefficient of 0.10 with a 95%  
266 HPD (highest probability density) interval of [0.04, 0.16]. This can be interpreted as a weak

267 correlation between  $\ln\text{Mean}$  and  $\ln\text{SD}$ . We remark that simply computing the correlation of  
268 the two quantities without paying attention to the correct weighting and the different scales in  
269 the data would yield an overestimated slope coefficient of 0.25 (see supplementary table 1).  
270



271  
272 **Figure 2:** Linear association between  $\ln\text{Mean}$  and  $\ln\text{SD}$  using a varying intercept model,  
273 where the intercepts are allowed to vary between studies with different depression scales.  
274 Red dots represent active groups, blue dots represent placebo groups.

275

276 *Log variability ratio ( $\ln\text{VR}$ ) and log coefficient of variation ( $\ln\text{CVR}$ ) models*

277 In order to estimate the difference in variability between antidepressant and placebo  
278 response, we modelled the  $\ln\text{VR}$  effect size using a Bayesian random effects model as  
279 heterogeneity between studies may be expected. The posterior mean estimate for the  
280 variability ratio was 1.01, with the 95 % highest posterior density (HPD) interval ranging from

281 1.00 to 1.02. The InVR effect size assumes no correlation between InMean and InSD and  
282 may give biased results if such a correlation exists. In the presence of a positive correlation  
283 between mean and standard deviation, Nakagawa et al. [28] suggest that the InCVR may be  
284 the more appropriate effect size to investigate the difference in variability between the active  
285 and control. The InCVR REMA showed a reduction in the coefficient of variation in the active  
286 versus the placebo group (posterior mean estimate for CVR: 0.82, 95% HPD [0.80,0.84]).

287

### 288 *Random-effects meta-regression*

289 Finally, we used a Bayesian random effects meta-regression (REMR). The advantage of this  
290 model over the InVR and InCVR meta-analyses is that we are not forced to make rigid  
291 assumptions about the association between the InMean and InSD, as the strength of this  
292 relationship is estimated directly from the data. Fitting this model, we obtained posterior  
293 statistics for the  $\mu$  and  $\beta$  coefficients. The posterior mean estimate for  $e^\mu$  was 1.00 (95% HPD  
294 [0.98,1.02]) and that for  $\beta$  0.04 (95% HPD [-0.03,0.12]), where we can (roughly, up to  
295 measurement error and random noise) interpret the coefficients as follows:

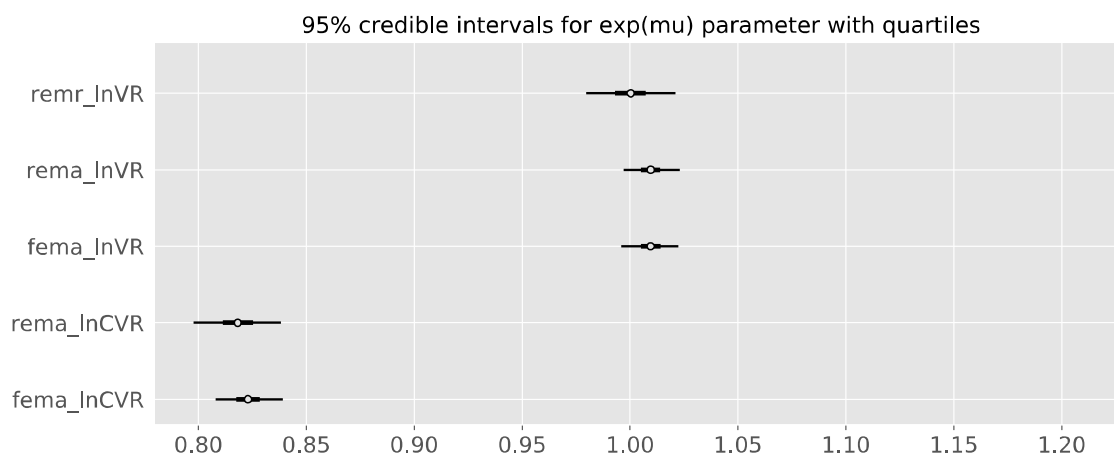
296

$$297 \quad 4. \quad VR \approx e^\mu * RR^\beta \quad (\ln VR \approx \mu + \beta * \ln RR)$$

298

299 Note that the InVR REMA corresponds to a InVR REMR with a  $\beta$  coefficient set to 0, whereas  
300 the InCVR REMA corresponds to a InVR REMR with a  $\beta$  coefficient set to 1. The REMR  
301 model learns the  $\beta$  coefficient and its posterior HDP interval is equal to 0.04 [-0.03, 0.12]  
302 suggesting that the InVR REMA is a more appropriate model than the InCVR REMA.

303



304

305 **Figure 3:** Posterior credible intervals for the  $\exp(\mu)$  parameter for the different models.  
306 REMA: random-effects meta-analysis. FEMA: fixed-effects meta-analysis. REMR: random-  
307 effects meta-regression. Note that the results are very similar for the REMR and the InVR  
308 meta-analyses.

309

310 *Between-study heterogeneity*

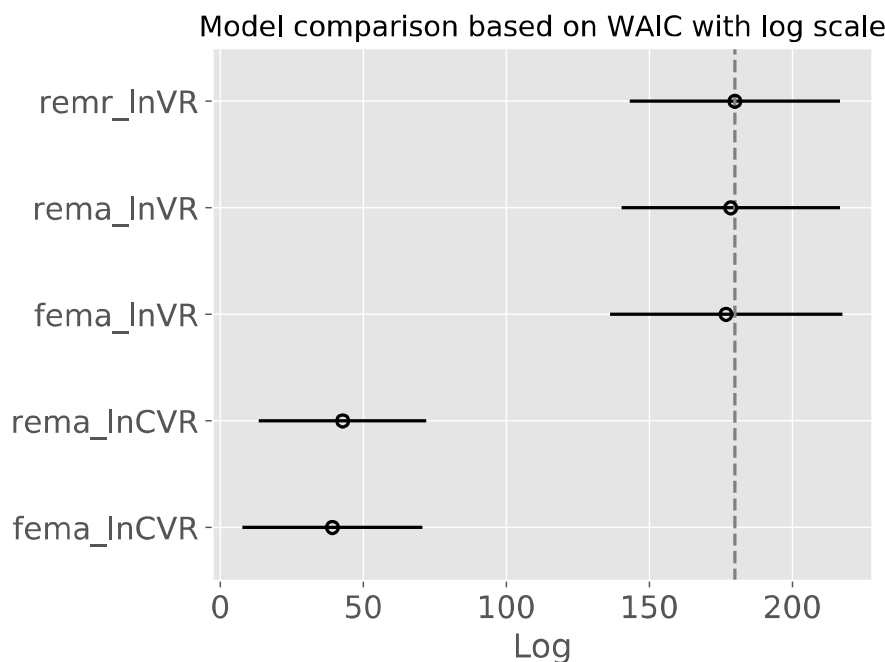
311 The between-study variance  $\tau^2$  under the REMA was found to be low for both InVR (95%  
312 HPD [0.00,0.00]) and InCVR (95% HPD for  $\tau^2$  [0.00,0.01]). Indeed, applying a fixed effects  
313 model instead of the REMA for the purpose of sensitivity analysis yielded similar results for  
314 the overall mean estimates of InVR and InCVR.

315

316 *Performance comparison of the different models*

317 In order to compare the performance of the different models applied, we used the so-called  
318 *widely applicable information criterion* (WAIC). This method estimates the pointwise  
319 prediction accuracy of fitted Bayesian models. Here, higher values of WAIC indicate a better  
320 out-of-sample predictive fit (“better” model). We refer to Vehtari et al. [31] for more details on  
321 WAIC. Figure 4 shows the logWAIC for the different models.

322



323

324 **Figure 4:** *Widely applicable information criterion* (WAIC) depicted on a logarithmic scale.  
325 Higher values signify a better predictive fit of the underlying model. Bars indicate standard  
326 errors. REMA: random-effects meta-analysis. FEMA: fixed-effects meta-analysis. REMR:  
327 random-effects meta-regression.

328

329 We observed that the InVR REMA and the InVR REMR outperformed the InCVR REMA with  
330 respect to the WAIC. The difference between the InVR REMA and the InVR REMR showed  
331 comparable performance with respect to the WAIC.

332

333 *Upper bound on the treatment effect heterogeneity*

334 In order to investigate the compatibility of different assumptions regarding the treatment  
335 effect with the measured variability ratio, we use the following equation that was derived in  
336 [32]:

337

$$338 \quad 5. \quad VR^2 - 1 = \frac{(SD_{tx} + \rho * SD_p)^2}{SD_p^2} - \rho^2, \text{ where } tx = \text{individual treatment effect}$$

339

340 Note that the variable  $\rho$  signifies the degree of correlation between the treatment effect  $tx$   
341 and the response under placebo  $p$  (see figure 1) and is unobservable. Assuming that  $VR^2 - 1$   
342 is smaller than some number  $\varepsilon$ , the above equation implies that:

343

$$344 \quad 6. \quad SD_{tx} \leq SD_p * (\sqrt{(\varepsilon + \rho^2)} - \rho)$$

345

346 Using the Cipriani et al. dataset [10] we can estimate  $SD_p$  to be equal to around 7.66 on the  
347 HAMD-17 scale. From our meta-analysis of the  $\ln VR$  effect statistic, we have that the 95th  
348 percentile of the posterior distribution of  $e^{\mu}$  is 1.02. This implies the following inequality:

349

$$350 \quad 7. \quad SD_{tx} \leq 7.66 * (\sqrt{(0.04 + \rho^2)} - \rho)$$

351

352 This inequality tells us that if we assume any value  $\rho \in [-1, 1]$  (the correlation between the  
353 treatment effect and the response under placebo), we get an upper bound on the standard  
354 deviation of the treatment effect as above.

355

356 *Which distributions of the treatment effect are possible for a VR of nearly 1?*

357 Based on the above-mentioned formula 7, table 1 depicts different magnitudes of treatment  
358 effect heterogeneity compatible with a VR of 1.02, which is the 95th percentile of our VR  
359 estimate.

360 The left column (“any distribution”) of the table depicts the upper bound for the standard  
361 deviation of the treatment effect (the treatment effect heterogeneity) with a VR of 1.02. The  
362 upper bound for the treatment effect heterogeneity depends upon the population-level  
363 correlation  $\rho$  between the “response under placebo” and the “individual treatment effect” (see  
364 figure 1). The results can be interpreted as follows: assuming a correlation  $\rho$  between the  
365 individual treatment effect and the response under placebo of a given value in the table, we  
366 are 95% sure (under the assumptions of the meta-analytic model (REMA) of  $\ln VR$ ) that the  
367 standard deviation of the treatment effect variable is smaller than the corresponding value of  
368 the second column in the table. Furthermore, assuming a minimally clinically relevant effect

369 of 7 points on the HAMD-17 scale, the third column tells us the percentage of patients with a  
 370 medication effect at least as large for the largest possible standard deviation within the 95%  
 371 credible interval. Note that these results are independent of the distribution of the treatment  
 372 effect (normal, binormal, etc.), as these results were derived analytically from the above-  
 373 mentioned formula (see formula 7).

374

375 The right column assumes a dichotomous treatment effect (“responder”, “non-responder”).  
 376 Here, “non-responders” are assumed to have treatment effect of 0 (placebo response =  
 377 antidepressant response), whereas “responders” have a fixed treatment effect  $> 0$ . For a  
 378 given VR of 1.02, the percentage of “responders” and their respective “responder treatment  
 379 effect” depend upon the intra-individual correlation  $\rho$  between the potential outcome placebo  
 380 response and individual treatment effect.

381

382

$\rho$	any distribution		dichotomous response	
	SD(TE)	% > 7 HAMD	responder TE	% responder
-1.0	15.5	37.3	20.5	9.7
-0.8	12.4	34.4	24.3	8.2
-0.6	9.4	29.8	29.8	6.7
-0.4	6.5	22.1	38.1	5.2
-0.2	3.7	8.9	51.9	3.9
0.0	1.5	0.1	72.2	2.8
0.2	0.6	0.0	86.2	2.3
0.4	0.4	0.0	91.6	2.2
0.6	0.3	0.0	94.1	2.1
0.8	0.2	0.0	95.5	2.1
1.0	0.2	0.0	96.3	2.1

383

384 **Table 1:** Assuming a VR of 1.02, a  $SD_p$  of 7.66 (based on Cipriani [10]) and different  
 385 correlation coefficients  $\rho$  between the response under placebo and the treatment effect. Left  
 386 column (“any distribution”): Upper bounds for the standard deviation of the treatment effect.  
 387 Right column (“dichotomous response”): Patients are either “non-responders” with a  
 388 treatment effect of 0, or “responders” with the responder treatment effect. For a given  
 389 correlation coefficient  $\rho$ , there is one possible solution for this. TE: treatment effect.

390

391 These results show that, contrary to intuition, a variability ratio of 1.02 is (theoretically)  
392 compatible with a standard deviation of the treatment effect between 0 and 15.5. Conversely,  
393 a reduction in the variability in the treatment group is compatible with a substantial treatment  
394 effect heterogeneity if the response under placebo is correlated with the individual treatment  
395 effect (see simulations in supplementary file, figures 8 and 9).

396

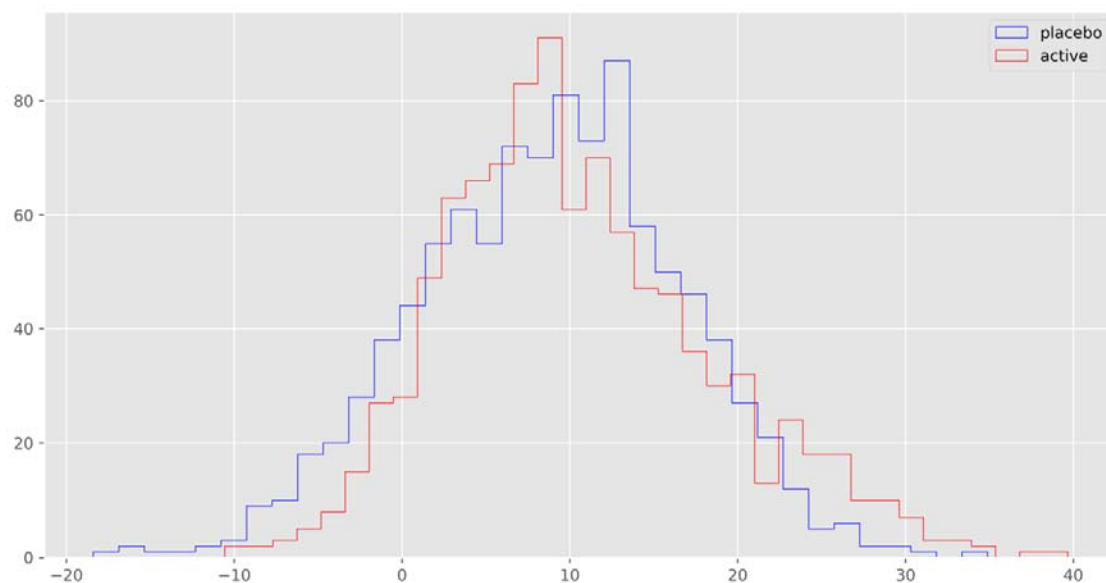
### 397 *Simulations*

398 We conducted simulation experiments in order to illustrate the compatibility of a VR of 1.02  
399 with different degrees of treatment effect heterogeneity. For a large treatment effect  
400 heterogeneity, the individual treatment effect was drawn from a distribution with an (arbitrarily  
401 chosen) standard deviation  $SD_{TE} = 6.5$  HAMD-17. For this  $SD_{TE}$ , the response under placebo  
402 and the individual treatment effect have to be correlated by the correlation factor  $\rho = -0.4$ , in  
403 order for the VR to credibly remain at or below 1.02 (see table 1). Figure 5 depicts the  
404 change scores of 1000 patients under placebo (blue) and under active treatment (red). Here,  
405 positive values denote an improvement of the depression severity.

406 Conversely, if the correlation factor  $\rho$  is equal to 0, a large treatment effect heterogeneity with  
407  $SD_{TE} = 6.5$  would yield a VR in the magnitude of 1.3. For the VR to be credibly lower than  
408 1.02 and the response under placebo and the individual treatment effect to be uncorrelated  
409 ( $\rho = 0$ ), the treatment effect heterogeneity has to be low. Supplementary figures 6 and 7  
410 depict the results of such a simulated experiment. Here, the treatment effect heterogeneity  
411 was imputed to be  $SD_{TE} = 1.5$  points on the HAMD-17 (derived from table 1). A VR closer to  
412 1 would yield an even smaller treatment effect heterogeneity.

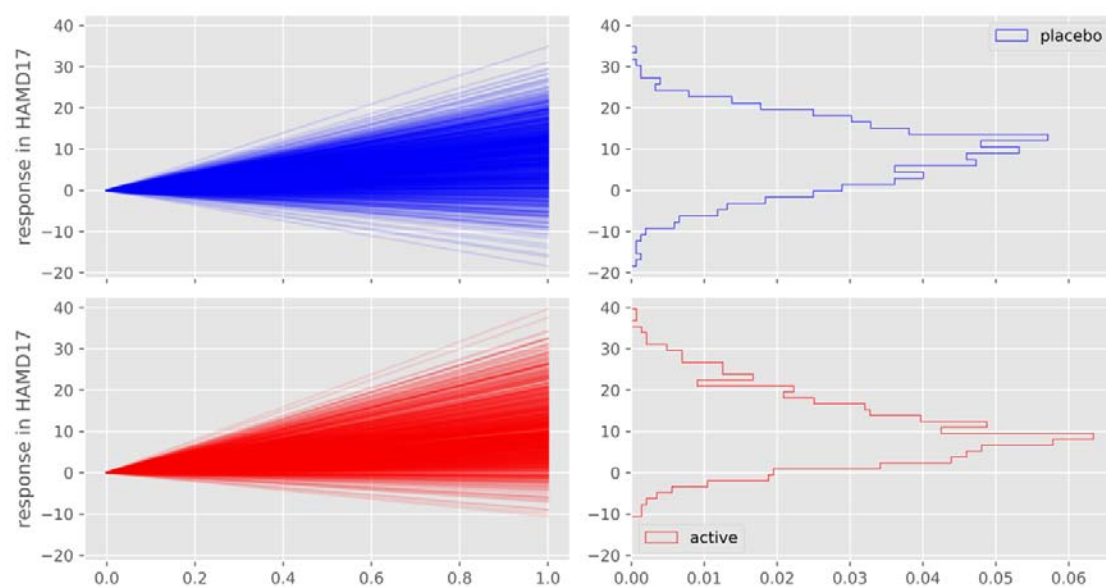
413

Histogram of potential outcome response under placebo and active treatment



414

Potential outcome responses with baseline gauged to 0



415

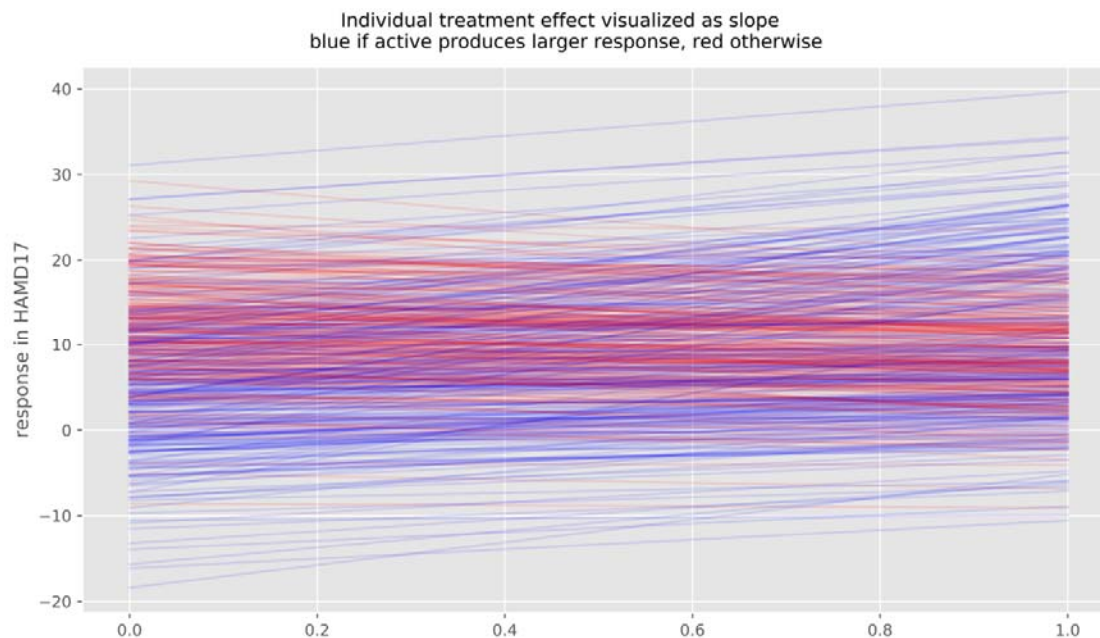
416 **Figure 5:** Change score of 1000 simulated patients under placebo (blue) and under active  
417 treatment (red) for  $\rho = -0.4$ ,  $SD_{TE} = 6.35$  HAMD-17 points and  $VR = 1.02$ . Note, that in this  
418 particular simulation, the  $SD_{TE}$  is not exactly equal to 6.5, as all simulations contain random  
419 processes.

420

421 Figure 6 shows the magnitude of the individual treatment effect of 100 individuals of this  
422 simulation experiment. The values on the left ( $x = 0.0$ ) denote the response under placebo of  
423 all 100 patients, the values on the right ( $x = 1.0$ ) represent the response under active



424 treatment and the difference between the two values corresponds to the individual treatment  
425 effect of a patient. As can be seen in the figure, for a large treatment effect heterogeneity to  
426 be compatible with a VR of 1.02, the response under placebo and the individual treatment  
427 effect have to be correlated. Specifically, patients that would remain unchanged under  
428 placebo (response close to 0 at  $x = 0.0$ ) have a larger benefit from active medication (higher  
429 density of blue slopes) than patients that would have improved under placebo (response  
430 above 0 at  $x = 0.0$ ). For some patients to benefit substantially, however, other patients have  
431 to be harmed by the medication (red slopes).



432  
433 **Figure 6:** Potential outcomes and individual treatment effect of 100 simulated patients. Value  
434 at  $x = 0$  depicts response under placebo, value at  $x = 1$  response under active treatment.  
435 Slopes represent individual treatment effect, which varies substantially in this simulation.  
436 Blue lines indicated improvement under active treatment, red lines deterioration.

437  
438

### 439 Discussion

440

441 The efficacy of antidepressants in the treatment of major depressive disorder has been the  
442 topic of an ongoing debate for years in the psychiatric community and the public [6-9]. In a  
443 recent re-analysis [12] of a network meta-analysis [10], the average treatment effect of  
444 antidepressants was found to be about 2 points on the HAM-D-17 scale, which is almost  
445 undetectable by clinicians [14] and clearly lies below the assumed minimally clinically  
446 relevant effect of 7 points [16]. In addition, it should be noted that relevant biases may have

447 led to an overestimation of the drugs' efficacy [12, 33]. Jakobsen et al. recently concluded  
448 that, based on current evidence, antidepressants should not be used in adult patients with  
449 major depressive disorder [34].

450

#### 451 *Evidence for treatment effect heterogeneity*

452 On the basis of these facts, the question arises as to why these compounds are considered  
453 to be effective, nevertheless. The main reason for this might be the assumption of a  
454 substantial treatment effect heterogeneity, meaning that subpopulations of patients exist that  
455 benefit substantially more than average from the medication. If the treatment effect  
456 heterogeneity is low, no patient would have a clinically relevant benefit. In the face of the well  
457 documented harms and side effects, antidepressants may then not be considered useful in  
458 the treatment of major depression.

459 Albeit widely believed and putatively observed in clinical routine, substantial differences in  
460 the individual treatment effect of antidepressants have not been shown to exist yet. A  
461 "responder", usually defined as someone crossing an arbitrary threshold of symptom  
462 severity, is a person who was observed to improve and not necessarily caused by the  
463 medication to get better. Even constant treatment effects lead to differences in observed  
464 response rates, creating the illusion of a differential treatment effect, where none exists [20].  
465 Therefore, the frequently performed calculation of response rates is misleading and seems  
466 inappropriate to answer the question of treatment effect heterogeneity.

467 This work aimed to estimate the treatment effect heterogeneity of antidepressants in the  
468 treatment of major depressive disorder using a large dataset of a recent network meta-  
469 analysis [10]. To this end, we applied the effect size statistics  $\ln VR$  and  $\ln CVR$  suggested by  
470 Nakagawa et al. [28], using a Bayesian random-effects meta-analytical approach (REMA)  
471 and fitted a multi-level meta-regression (REMR) model to estimate the treatment effect  
472 variability between antidepressants and placebo. Both the  $\ln VR$  REMR and the  $\ln VR$  REMA,  
473 which were found to outperform the  $\ln CVR$  REMA, showed that the variability ratio was very  
474 close to 1 (REMR: 95% HPD [0.98, 1.02], REMA: 95% HPD [1.00, 1.02]), perfectly  
475 compatible with a near-constant effect of antidepressants on depression severity. These  
476 findings are in line with those of a recently published meta-analysis of antidepressants using  
477 the same dataset [35].

478

#### 479 *Methodological aspects*

480 In order to determine the variability of the treatment response, the correlation between the  
481 mean and standard deviation of the underlying measuring scale has to be taken into account.  
482 The  $\ln VR$  and the  $\ln CVR$  effect sizes naively assume a slope coefficient of 0 and 1,  
483 respectively. In other words, how much of the (logarithmic) difference in variances is

484 explained by the difference in (logarithmic) means. Both scales may thus give biased results,  
485 if the true slope coefficient differs from the one assumed.

486 By applying a varying intercept model, taking into account the occurrence of different  
487 depression scales, we could show that the correlation between (logarithmic) mean and  
488 (logarithmic) standard deviation is of a small magnitude (slope coefficient = 0.10), indicating  
489 that the lnVR is a more appropriate measure as opposed to the lnCVR. A regression over all  
490 depression scales yields a slope coefficient of 0.25, which is 2.5 x as large as our estimate.

491 When simply conducting a significance test for the existence of such correlation, the lnCVR  
492 effect size would appear to be the appropriate measure, leading to the incorrect conclusion  
493 of a substantially reduced variability in the active arm. It is important to note that a VR (or  
494 CVR) sufficiently smaller than 1 is in fact evidence of relevant treatment effect heterogeneity  
495 (see supplementary figure 8 and 9). Therefore, considering the lnCVR as the main outcome  
496 would lead to the opposite conclusion of substantial treatment effect heterogeneity [36].

497 Our work adds accuracy to the existing literature, as we developed a generalized model  
498 (REMR) that incorporates the slope coefficient for the correlation between mean and  
499 standard deviation directly from the data. This approach yielded a mean estimate for the VR  
500 of 1.00 (95% HPD [0.98,1.02]), again compatible with a near-constant effect.

501 We applied the WAIC in order to estimate the predictive power of our models. This effect  
502 measure showed that the models using the lnVR effect size had a better out of sample  
503 predictive power than the models using the lnCVR effect size.

504

#### 505 *Upper bound for the treatment effect heterogeneity*

506 As a relevant treatment effect heterogeneity cannot be ruled out even with a variability ratio  
507 near 1, we aimed at estimating the upper bound in treatment effect variation compatible with  
508 our results. We were able to analytically derive an inequality that provides an upper bound  
509 for the treatment effect heterogeneity, taking into account a possible correlation between the  
510 placebo response and the individual treatment effect. We could show that a VR of 1.02 (the  
511 upper bound of the 95% HPD interval of the REMR) is theoretically compatible with a  
512 standard deviation of the treatment effect between 0 and 15.5 points on the HAMD-17 scale,  
513 translating into a maximum of 37% of patients with an individual treatment effect of more  
514 than 7 points on the HAMD-17 scale.

515 However, such a large standard deviation and hence treatment effect heterogeneity would  
516 require the treatment effect and the response under placebo of a patient to be strongly and  
517 negatively correlated in order to be compatible with a VR of 1.02. If no such correlation  
518 exists, the treatment effect heterogeneity would be negligibly small ( $SD_{TE} = 1.5$  HAMD-17  
519 points, 0.1% of patients benefitting more than 7 points on the HAMD-17 scale).

520 As only one outcome per patient (either under placebo or under active treatment; figure 1)  
521 can be measured in a real experiment, the true correlation  $\rho$  between the response under  
522 placebo and the treatment effect cannot be derived from RCT data.

523

524 *How should these results be interpreted?*

525 The VR is a measure that can potentially detect evidence for subgroups that benefit  
526 (substantially) more than average from an intervention. A VR that differs substantially from 1  
527 is evidence of such subgroups (of large treatment effect heterogeneity), while a VR near 1  
528 is compatible with both a small and a large treatment effect heterogeneity. A VR of exactly 1  
529 (which is the mean-estimate of our REMR model) would be proof of a constant treatment  
530 effect. It is, however, impossible to ever prove identity, as we can never reach an uncertainty  
531 of 0 (credible interval with width of 0). Furthermore, an exactly constant treatment effect  
532 seems impossible also from a theoretical point of view. So how should a VR of 1 (95% HPD  
533 [0.98,1.02]) be interpreted? For this, consider the following illustration:

534

535 *Hypothesis 1 (H1):* The treatment effect heterogeneity is close to 0 (e.g. 99% of patients  
536 have an individual treatment effect of 1 to 3 HAMD points).

537 *Hypothesis 2 (H2):* The treatment effect heterogeneity is greater than in H1.

538

539 There are now three possibilities:

- 540 1. H1 is true and  $VR \approx 1$  (very close to 1, e.g. 0.98 to 1.02)
- 541 2. H2 is true and  $VR \approx 1$
- 542 3. H2 is true and  $VR \neq 1$  (not very close to 1)

543

544 Our results indicate that  $VR \approx 1$ . We can thus rule out one of the three possibilities, namely a  
545 large treatment heterogeneity combined with a  $VR \neq 1$ . From a Bayesian perspective, the  
546 probability of H1 being true increases, while that of H2 being true decreases. How we now  
547 regard the probability of H1 or H2 being true depends on how plausible we considered these  
548 scenarios to begin with (the prior probabilities).

549 In order for H2 to be true and the VR being close to 1, strong assumptions regarding  
550 the correlation between the placebo response and the individual treatment effect of  
551 antidepressants are necessary. Specifically, those patients whose depression severity would  
552 remain unchanged under placebo would need to have the strongest antidepressant  
553 medication effect. If this were the case, we might expect patients with certain features (such  
554 as chronic depression) to benefit substantially more than average from  
555 antidepressants. Since no such subpopulations have been identified to date, such a

556 correlation seems unlikely. If no such correlation is assumed, a VR of 1.02 indicates a low  
557 degree of treatment effect heterogeneity.

558

### 559 *Conclusion*

560 By applying a multiple level Bayesian regression model and simulations, this work could  
561 show that the published data on antidepressants in the treatment of major depression is  
562 compatible with a near-constant treatment effect, which is also the simplest explanation for  
563 the observed data. Although it is not possible to rule out a substantial treatment effect  
564 heterogeneity using summary data from RCTs, we could show that a substantial treatment  
565 effect heterogeneity is only compatible with the published data under strong assumptions  
566 that seem rather unlikely. Until the existence of benefiting subgroups has been demonstrated  
567 prospectively, the average treatment effect is the best estimator for the individual treatment  
568 effect. Since the average treatment effect of antidepressants probably falls short of clinical  
569 relevance, the current prescribing practice in the treatment of major depression should be  
570 critically re-evaluated.

571

572

### 573 *Python code*

574 <https://github.com/volkale/advr>

575

### 576 *Statement of Ethics*

577 The authors have no ethical conflicts to disclose.

578

### 579 *Disclosure Statement*

580 CAM received consulting fees from Silence Therapeutics, outside the submitted work. The  
581 other authors declared no competing interest. All authors declare no other relationships or  
582 activities that could appear to have influenced the submitted work. No funder had any role in:  
583 the design and conduct of the study; collection, management, analysis, and interpretation of  
584 the data; preparation, review, or approval of the manuscript; and decision to submit the  
585 manuscript for publication.

586

### 587 *Funding Sources*

588 The authors received no specific funding for this work.

589

### 590 *Author Contributions*

591 Study idea and design: CV and CAM. Data extraction, statistical analyses, visualizations and  
592 simulation experiments: AV and CV. Mathematical modelling, code implementation and

593 derivation of formula: AV. All authors contributed to drafting the manuscript. All authors  
594 provided a critical review and approved the final paper.  
595



596 **References**

- 597 1. WHO. *Depression and Other Common Mental Disorders: Global Health*  
598 *Estimates*. 2017; Available from:  
599 [https://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-](https://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf)  
600 2017.2-eng.pdf.
- 601 2. DGPPN, *S3-Leitlinie und Nationale VersorgungsLeitlinie (NVL) Unipolare*  
602 *Depression, 2. Auflage*. 2015.
- 603 3. NICE, *Depression in adults: recognition and management*. 2009. [28.08.2019]  
604 Available from: [https://www.nice.org.uk/guidance/cg90/chapter/1-](https://www.nice.org.uk/guidance/cg90/chapter/1-Guidance#treatment-choice-based-on-depression-subtypes-and-personal-characteristics)  
605 [Guidance#treatment-choice-based-on-depression-subtypes-and-personal-](https://www.nice.org.uk/guidance/cg90/chapter/1-Guidance#treatment-choice-based-on-depression-subtypes-and-personal-characteristics)  
606 [characteristics](https://www.nice.org.uk/guidance/cg90/chapter/1-Guidance#treatment-choice-based-on-depression-subtypes-and-personal-characteristics)
- 607 4. BMJ, *NHS prescribed record number of antidepressants last year*, in *BMJ*.  
608 2019. l1508.
- 609 5. Kantor, E.D., et al., *Trends in Prescription Drug Use Among Adults in the*  
610 *United States From 1999-2012*. *JAMA*, 2015. 314(17): p. 1818-31.
- 611 6. Moncrieff, J. and I. Kirsch, *Efficacy of antidepressants in adults*. *BMJ*, 2005.  
612 331(7509): p. 155-7.
- 613 7. Fountoulakis, K.N. and H.J. Moller, *Efficacy of antidepressants: a re-analysis*  
614 *and re-interpretation of the Kirsch data*. *Int J Neuropsychopharmacol*, 2011.  
615 14(3): p. 405-12.
- 616 8. Davis, J.M., et al., *Should we treat depression with drugs or psychological*  
617 *interventions? A reply to Ioannidis*. *Philos Ethics Humanit Med*, 2011. 6: p. 8.
- 618 9. Gotzsche, P.C., *Why I think antidepressants cause more harm than good*.  
619 *Lancet Psychiatry*, 2014. 1(2): p. 104-6.
- 620 10. Cipriani, A., et al., *Comparative efficacy and acceptability of 21 antidepressant*  
621 *drugs for the acute treatment of adults with major depressive disorder: a*  
622 *systematic review and network meta-analysis*. *Lancet*, 2018. 391(10128): p.  
623 1357-1366.
- 624 11. *BMJ, Large meta-analysis ends doubts about efficacy of antidepressants*.  
625 2018. k847.
- 626 12. Munkholm, K., A.S. Paludan-Muller, and K. Boesen, *Considering the*  
627 *methodological limitations in the evidence base of antidepressants for*  
628 *depression: a reanalysis of a network meta-analysis*. *BMJ Open*, 2019. 9(6): p.  
629 e024886.
- 630 13. Hamilton, M., *Development of a rating scale for primary depressive illness*. *Br*  
631 *J Soc Clin Psychol*, 1967. 6(4): p. 278-96.
- 632 14. Leucht, S., et al., *What does the HAM-D mean?* *J Affect Disord*, 2013. 148(2-  
633 3): p. 243-8.
- 634 15. Guy, W., *Clinical Global Impressions ECDEU Assessment Manual for*  
635 *Psychopharmacology, Revised (DHEW Publ. No. ADM 76-338)*. 1976,  
636 National Institute of Mental Health: Rockville, MD. p. 218–222.
- 637 16. Moncrieff, J. and I. Kirsch, *Empirically derived criteria cast doubt on the clinical*  
638 *significance of antidepressant-placebo differences*. *Contemp Clin Trials*, 2015.  
639 43: p. 60-2.
- 640 17. *BMJ, Effectiveness of antidepressants*. 2018. k1073.
- 641 18. *BMJ, The cost of dichotomising continuous variables*. 2006. 1080.
- 642 19. Austin, P.C. and L.J. Brunner, *Inflation of the type I error rate when a*  
643 *continuous confounding variable is categorized in logistic regression analyses*.  
644 *Stat Med*, 2004. 23(7): p. 1159-78.

- 645 20. Senn, S., *Statistical pitfalls of personalized medicine*. Nature, 2018.  
646 563(7733): p. 619-621.
- 647 21. Winkelbeiner, S., et al., *Evaluation of Differences in Individual Treatment*  
648 *Response in Schizophrenia Spectrum Disorders: A Meta-analysis*. JAMA  
649 Psychiatry, 2019.
- 650 22. Senn, S., *Mastering variation: variance components and personalised*  
651 *medicine*. Stat Med, 2016. 35(7): p. 966-77.
- 652 23. Fisher, R.A. and others, *Statistical inference and analysis: Selected*  
653 *correspondence of ra fisher, edited by jh bennett*. 1990: Oxford: Clarendon  
654 Press.
- 655 24. Montgomery, S.A. and M. Asberg, *A new depression scale designed to be*  
656 *sensitive to change*. Br J Psychiatry, 1979. 134: p. 382-9.
- 657 25. Rush, A.J., et al., *Self-reported depressive symptom measures: sensitivity to*  
658 *detecting change in a randomized, controlled trial of chronically depressed,*  
659 *nonpsychotic outpatients*. Neuropsychopharmacology, 2005. 30(2): p. 405-16.
- 660 26. Jefferson, J.W., et al., *Extended-release bupropion for patients with major*  
661 *depressive disorder presenting with symptoms of reduced energy, pleasure,*  
662 *and interest: findings from a randomized, double-blind, placebo-controlled*  
663 *study*. J Clin Psychiatry, 2006. 67(6): p. 865-73.
- 664 27. The Cochrane Collaboration, *Cochrane Handbook for Systematic Reviews of*  
665 *Interventions Version 5.1.0*. 2011 [02.09.2019]; Available from:  
666 [www.handbook.cochrane.org](http://www.handbook.cochrane.org).
- 667 28. Nakagawa, S., et al., *Meta-analysis of variation: ecological and evolutionary*  
668 *applications and beyond*. Methods in Ecology and Evolution, 2015. 6(2): p.  
669 143-152.
- 670 29. McCutcheon, R.A., et al., *The efficacy and heterogeneity of antipsychotic*  
671 *response in schizophrenia: A meta-analysis*. Mol Psychiatry, 2019. doi:  
672 10.1038/s41380-019-0502-5.
- 673 30. Leucht, S., et al., *Translating the HAM-D into the MADRS and vice versa with*  
674 *equipercntile linking*. J Affect Disord, 2018. 226: p. 326-331.
- 675 31. Vehtari, A., A. Gelman, and J. Gabry. *Practical Bayesian model evaluation*  
676 *using leave-one-out cross-validation and WAIC*. 2016. doi: 10.1007/s11222-  
677 016-9696-4.
- 678 32. Volkman, A. "A bound on the Treatment Effect Heterogeneity using estimates  
679 on the Variability Ratio", in preparation.
- 680 33. Hengartner, M.P., *Methodological Flaws, Conflicts of Interest, and Scientific*  
681 *Fallacies: Implications for the Evaluation of Antidepressants' Efficacy and*  
682 *Harm*. Front Psychiatry, 2017. 8: p. 275.
- 683 34. Jakobsen, J.C., C. Gluud, and I. Kirsch, *Should antidepressants be used for*  
684 *major depressive disorder?* BMJ Evid Based Med, 2019. doi:  
685 10.1136/bmjebm-2019-111238.
- 686 35. Plöderl, M. and M.P. Hengartner, *Can we expect that some patients respond*  
687 *better to antidepressants? A secondary variance-ratio meta-analysis*. 2019.  
688 [osf.io/98kex](https://osf.io/98kex).
- 689 36. Maslej, M.M., et al., *Individual Differences in Response to Antidepressants: A Meta-*  
690 *analysis of Placebo-Controlled Randomized Clinical Trials*. JAMA Psychiatry, 2020.  
691 DOI: 10.1001/jamapsychiatry.2019.4815  
692