Comparing Apples and Oranges in Youth Depression Treatments? A Quantitative Critique of the Evidence Base and Guidelines

Major Revision: Letter to Reviewers

### 0.0.1 Reviewer 1

#### 0.0.1.1 Comment 1.7

I could not see the rational for excluding the all female trials in the sex analyses. [Comment 1.7]

#### 0.0.1.2 Response to 1.7

This is to not unduly bias the results. We now present the analyses with all studies included in the Supplemental Materials (see [Table 1](#tbl-all-studies-percfemale-results)).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Percentage female at baseline across medication and psychotherapy studies: Results for sample including all female studies   | **Subgroup** | **K** | **Mean** | **SE** | **Lower CI** | **Upper CI** | **T2** | **p-value** | | --- | --- | --- | --- | --- | --- | --- | --- | | **Overall** |  |  |  |  |  |  | **0.029** | | Psychotherapy | 45 | 61.05 | 2.43 | 56.16 | 65.94 | 264.99 |  | | Medication | 28 | 53.72 | 2.33 | 48.94 | 58.51 | 152.15 |  | | **Excluding subclinical** |  |  |  |  |  |  | **0.035** | | Psychotherapy | 38 | 61.4 | 2.8 | 55.72 | 67.07 | 297.93 |  | | Medication | 28 | 53.72 | 2.33 | 48.94 | 58.51 | 152.15 |  | | **Excluding waitlist** |  |  |  |  |  |  | **0.044** | | Psychotherapy | 37 | 61.01 | 2.77 | 55.39 | 66.63 | 284.33 |  | | Medication | 28 | 53.72 | 2.33 | 48.94 | 58.51 | 152.15 |  | |

### 0.0.2 Reviewer 3

#### 0.0.2.1 Comment 3.5a: Effect modification

1) Let’s first start with a pairwise meta-analysis (MA). In the MA, we are usually pooling RR (or OR) and not RD, simply because the former is more generalizable. For example,

|  |  |  |  |
| --- | --- | --- | --- |
|  | Control event rate (CER) | Experimental event rate (EER) | RR |
| Trial 1 | 5/100 | 10/100 | 2.0 |
| Trial 2 | 30/100 | 60/100 | 2.0 |

In this case, even though the CER (and probably patient characteristics) differ markedly between Trial1 and Trial2, the RR is the same between the two trials and there is no problem pooling these two (no heterogeneity).

Factors that change the CER are called prognostic factors, and factors that change the EER/CER are called effect modifiers. In this case there are apparently differences in PFs between the two trials but there are no differences in EMs. [Comment 3.5a]

#### 0.0.2.2 Response to 3.5a

We thank the reviewer for this example which perfectly illustrates our point and the rationale for our manuscript.

First, let’s start with a widely accepted definition of the effect modification as represented by the European Medicines Agency’s (EMA) group [European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENcePP) entry on effect modification](https://encepp.europa.eu/encepp-toolkit/methodological-guide/chapter-7-effect-modification-and-interaction_en) and the publications it recommends on this topic:

The clinical motivation behind the assessment of effect modification is to identify whether the effect of a treatment (or exposure) is different in groups of patients with different characteristics. If the effects are the same, the treatment (or exposure) effect is called homogeneous; if the effects are different, they are called heterogeneous (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5476432/> )

The ENcePP [also clearly states that](https://encepp.europa.eu/encepp-toolkit/methodological-guide/chapter-7-effect-modification-and-interaction_en):

The presence of effect modification depends on which measure is used in the study (**absolute or relative** [added emphasis by us]**)** and can be measured in two ways: on an additive scale (based on risk differences [RD]), or on a multiplicative scale (based on relative risks [RR]).

To make this clear, we reproduce the table above, only adding another two column, one is the Odds Ratio (OR) another relative/multiplicative measure of relative risk that the reviewer also refers to as “more generalizable”, and that of absolute risk, AR (i.e. the risk difference), an additive measure.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Control event rate (CER) | Experimental event rate (EER) | RR | OR | AR |
| Trial 1 | 5/100 | 10/100 | 2.0 | 2.1 | 0.05 |
| Trial 2 | 30/100 | 60/100 | 2.0 | 3.5 | 0.3 |

As can be seen the ORs differ quite substantially and the ARs differ by an order of magnitude. There is clearly an effect modification of some sort going on, when considering the OR or AR. Indeed, as the EMA and ENcePP state, whether effect modification is present (also) depends on the measure used. We therefore disagree with the reviewer that differences in the control event rate are ipso facto prognostic factors (PFs) if they do not affect the RR. Importantly, the RR is notorious for being potentially misleading (<https://pubmed.ncbi.nlm.nih.gov/15823376/>) and the necessity of reporting absolute risks is emphasised in standard clinical publications (e.g. <https://academic.oup.com/ndt/article/32/suppl_2/ii13/3056571> ).

To illustrate this point further in the context of trials, we point out that a key metric in intervention evaluation is the Numbers Needed to Treat (NNT) or its counterpart Numbers Needed to Harm (NNH). The NNT simply defined as the inverse of the AR, namely:

In the case above:

whereas:

The interpretation of NNT in the context of a placebo-controlled trial is simply: the average number of people who need to be treated with active treatment rather than placebo for one additional person to benefit. It is obvious from the above that, were such numbers to occur in a clinical trial, the clinical and public health implications of the difference between two treatments would be potentially vast and that it would be wrong to have considered RRs alone.

But we would go a step further and argue that all summary measures, including the AR, can conceal important clinical and public health facts, and that therefore reporting data that are close as possible to the raw numbers is key. We will illustrate this with an example.

Let there be two treatments A and B, with data on their efficacy coming comparisons that each has had with A\_control and B\_control, respectively. The efficacy of A is defined as:

and that of B as:

And therefore:

Since the AR is the same in each case, and one might conclude that we should be indifferent between the two treatments. However, it should be obvious that any rational agent, i.e. an agent who want so maximise benefit for themselves, would choose treatment A over B, if this were the data that they had to base their decision upon; indeed, nobody woudl choose a treatment with 40% probability recovery over one with 90% probability recovery? Unless of course there were a fundamental difference in the composition or circumstances of the trials.

All these considerations are very important and relevant to our paper and in the spirit of being very clear with the primary data, we have added, in addition to the  figure with the summary SMDs, this figure which presents the SMD of each study separately.

# 1. {r, warning=FALSE, message = FALSE, echo=FALSE, ft.align="left"} # # list\_df\_simulated[[1]] %>% # filter(arm\_effect\_size == "cohens\_d\_control") %>% # ggplot(aes(x = cohens\_d, y = 0, color = as.factor(psy\_or\_med), size = baseline\_n)) + # geom\_jitter(height = 0.2, alpha = 0.7) + # geom\_point(alpha = 0.7) + # labs( # title = "Pre-Post Standardised Mean Differences (SMD) of Control Arms", # x = "SMD", # y = "", # color = "Type of Study", # size = "Baseline n" # ) + # scale\_color\_discrete(name = "Type of Study", labels = c("Medication", "Psychological Therapy")) + # theme\_minimal() + # theme( # axis.text.y = element\_blank(), # axis.ticks.y = element\_blank(), # axis.title.y = element\_blank() # ) #

Given this finding, there must be two possibilities about such a difference (given what we have written above). First, the difference may be due to differences the method of control. It is easy to conceive of this as a difference in the quality of the controls, i.e. how well they were blinded, or what exactly participants in the trial were told.

Second, there may be differences in the members of the control group (and therefore, in the context of a successful randomisation of all trial members).

#### 1.0.0.1 Comment 3.8: Inaccurate statements

Given these theoretical considerations, here are examples of inaccurate or inadequate statements in the current manuscript. [Comment 3.8]

#### 1.0.0.2 Response to 3.8

Before we respond to the reviewer’s point one by one, we would like to suggest that a more useful—not least because more formally stated and generalizable—framework for theoretical considerations is the issue of external validity. We had included these formal treatments in previous versions of the manuscript, but decided to remove them becauseBriefly,

The average treatment effect is the measure of interest, defined as:

$$
ATE = E\[Y(1)\] - E\[Y(0)\] 
$$

And the conditional treatment effect:

$$
CATE(X = x) = E\[Y(1) - Y(0) \\mid X = x\]
$$

The CATE allows us to consider what may threaten external validity.

For example, if we consider W as a set of covariates affecting representativeness, then external validity would be violated if the following weren’t true.

E[Y \mid W = w] \approx E[Y \mid W = w\text{ in the population}]

Similarly, if we considered Z as a set of covariates related to trial conditions, then external validity would be violated if the following weren’t true.

E[Y \mid T = z] \approx E[Y \mid Z = z \

It is obvious, that these formalisms that apply to the relationship between a sample and a population, also apply when two samples are compared with each other, in other words, the apples vs oranges situation that we describe in our paper.

We have added these formalisms with an explanation to the Supplemental Materials of the paper.

#### 1.0.0.3 Comment 3.10: Use of baseline SD

Aside from this fundamental problem associated with pre-post SMDs, the current authors’ handling of pre-post SMDs has further problems.

1. The authors used the mean of the baseline SD and the endpoint SD to calculate the pre-post SMD. The use of baseline SD is problematic in depression trials, many of which use thresholds for baseline eligibility 2. In other words, the baseline SD is unnaturally small and, moreover, how small it is is partially determined by the threshold of the study, which can be different from study to study. [Comment 3.10]

#### 1.0.0.4 Response to 3.10

We thank the reviewer for pointing this out. The key question is whether the baseline SDs are systematically different between medication and psychological therapy trials, i.e. whether this introduces a bias. If the differences are random, then this only contributes to noise. We are doing two things to address this.

First, we plot the mean differences (endpoint – baseline) of the min-max transformed scores. As can be seen these are substantial, indicating that the difference is unlikely to stem simply from a difference in standard deviation calculation.

# 2. {r, warning=FALSE, message = FALSE, echo=FALSE} # # mean\_diff\_min\_max\_scores <- df\_long\_for\_metan %>% # filter(arm\_effect\_size == "cohens\_d\_control") %>% # mutate(mean\_dif = perc\_post\_mean - perc\_baseline\_mean) %>% # ggplot(aes(y = mean\_dif, x = as.factor(psy\_or\_med), size = baseline\_n)) + # geom\_jitter(width = 0.2) + # Add jitter to points # ggtitle("Post-Pre Differences in Mean Min-Max Scores") + # ylab("Post-Pre Difference") + # xlab("Type of Study") + # scale\_x\_discrete(labels = c("0" = "Medication", "1" = "Psychological Therapy")) + # Customize x-axis labels # theme\_minimal() # # print(mean\_diff\_min\_max\_scores) # # png("mean\_diff\_min\_max\_scores.png") # mean\_diff\_min\_max\_scores # dev.off() # #

Second, we plot below the baseline variances (i.e. squared sds), which largely overlap, making it unlikely that differences in baseline SD is the cause of the difference in the outcome.

# 3. {r, warning=FALSE, message = FALSE, echo=FALSE} # # create baseline variances # df\_long\_for\_metan$variances <- (df\_long\_for\_metan$perc\_baseline\_sd)^2 # # plot\_baseline\_var <- df\_long\_for\_metan %>% # filter(arm\_effect\_size == "cohens\_d\_control") %>% # mutate(mean\_dif = perc\_post\_mean - perc\_baseline\_mean) %>% # ggplot(aes(y = variances, x = as.factor(psy\_or\_med), size = baseline\_n)) + # geom\_jitter(width = 0.2) + # Add jitter to points # ggtitle("Post-Pre Differences in Study Baseline Variances") + # ylab("Baseline Variance") + # xlab("Type of Study") + # scale\_x\_discrete(labels = c("0" = "Medication", "1" = "Psychological Therapy")) + # Customize x-axis labels # theme\_minimal() # # print(plot\_baseline\_var) # # #png("plot\_baseline\_var.png") # # plot\_baseline\_var # # dev.off() #

Third, we conducted an addition analysis, restricting to studies with variances below 0.02 (see graph above).

# 4. {r, warning=F} # # #| label: fig-plot-means-var2 # #| fig-cap: "Meta-analytic estimates of within-group changes: studies with variance <0.02" # #| fig-cap-location: top # # coef\_and\_se\_means\_var2\_study <- coef\_and\_se\_means\_var2\_study %>% # mutate(condition = case\_when( # condition == "medication\_control" ~ "Medication Control", # condition == "medication\_active" ~ "Medication Active", # condition == "psychotherapy\_control" ~ "Psychotherapy Control", # condition == "psychotherapy\_active" ~ "Psychotherapy Active", # TRUE ~ as.character(condition))) # # # subtitle\_text\_hamd <- "Metanalytic estimates of within-group changes: HAM-D studies only" # plot\_means\_multi\_var2 <- plot\_means\_function(coef\_and\_se\_means\_var2\_study) # print(plot\_means\_multi\_var2) # # # png("plot\_means\_multi\_var2.png") # # plot\_means\_multi\_var2 # # dev.off() # #

#### 4.0.0.1 Comment 3.11: Studies with post SD

1. I would predict that few studies, if any, reported both the baseline and endpoint SD and the authors have apparently used various imputations based on so many assumptions to impute the pre-post SMD. [Comment 3.11]

#### 4.0.0.2 Response to 3.11

Actually, we have endpoint SDs in 61 (69.3% of) trials. To address the reviewer’s point, we have now reanalysed our data restricted to just those studies and as can be seen below, the same inferences hold as before. We include this figure now as Fig XXX in the Supplement.

|  |
| --- |
| Figure 1: Meta-analytic estimates of within-group changes: studies with post SD |

Coefficients across simulations (not sure in which response to add this):

