Author’s response to reviews

Title: Predictors of the effectiveness of an early medication change strategy in patients with Major Depressive Disorder

Authors:

Nadine Dreimüller (nadine.dreimueller@unimedizin-mainz.de)
Stefanie Wagner (stefwagner@uni-mainz.de)
Alice Engel (alice.engel@unimedizin-mainz.de)
Dieter Braus (Dieter.Braus@helios-gesundheit.de)
Sibylle Roll (Sibylle.C.Roll@vitos-rheingau.de)
Stefan Elsner (S.ELSNER@rmf.landeskrankenhaus.de)
André Tadic (tadic@uni-mainz.de)
Klaus Lieb (klaus.lieb@unimedizin-mainz.de)

Version: 3 Date: 24 Oct 2018

Author’s response to reviews:

October, 24th 2018

Manuscript ID: BPSY-D-18-00571R2

Dear Dr. Brettschneider,

please find attached our revised Ms. “Predictors of the effectiveness of an early medication change strategy in patients with Major Depressive Disorder” by Dreimüller et al. We are grateful to all reviewers for their helpful comments and recommendations. According to your suggestions we have improved and revised our manuscript according to the reviewers’ comments. In detail,
we have dealt as follows with the comments (changes in the revised MS are highlighted by using track changes text):

KELTIE MCDONALD (REVIEWER 1):

General comments: The document needs substantial editing to improve clarity and readability. There are many run on sentences, and corrections of grammar, use of punctuating and phrasing are necessary.

Reply/Action taken: Document was edited by an English native speaker in order to correct grammar, phrasing and punctuating (lines not shown).

Comment #1: The authors have used non-improvers, non-responders, and non-remitters seemingly interchangeably throughout the document, which can be confusing. I recommended that the authors select one (define it) and use it consistently throughout

Reply/Action taken:

We apologize that the usage of the different terms was somewhat confusing, and therefore reduced the terms into two: “non-improvement” (which we defined in the introduction section on page 5 line 101 as improvement of 20% of the HAMD-17 sum score after 2 weeks) and “remission” (which we defined on page 9 lines 192-193; as HAM-D17 score < 7 points after 8 weeks). However, we were not able to change the term “non-response” in the manuscript as we were adhering here to the study protocol, which states in case of non-response (≤50% reduction of HamD-17 sum score from baseline to week 4)(page 7 line 148) medication is changed from escitalopram to venlafaxine.

Comment #2: The endpoint, remission, has been defined as a HAM-D17 score < 7 points after 8 weeks. My understanding is that remission is typically considered a score of less than or equal to 7. Why was <7 chosen?

Reply: Several authors suggest to use < 7 as outcome parameter for HamD-17 (e.g. Spellmann et al. Validity of remission and recovery criteria for schizophrenia and major depression: comparison of the results of two one year follow up naturalistic studies. Eur Arch Psychiatry

Comment #3: The first analysis (Table 2) examined the effects of treatment group, predictor, and the interaction of treatment group and predictor on remission. None of the interactions were significant, so it seems unnecessary to further explore these associations using Chi-squared tests. This appears to be a repeat of the same analysis, only using a different measure of association and without the correction for multiple comparisons.

Reply/action taken: The reviewer is right that a secondary test of the association is normally only useful if there is a significant association in the regression analysis. Indeed, after correction for multiple testing no analysis remained significant. As we performed an exploratory analysis in a small sample size, which was not powered for this research question, we were interested in further investigating this association, because we were seeing an association (which was not significant after correction). In our opinion we did not repeat the same analysis as chi-squared tests are a different analysis, as it is not a modelling technique like the regression analyses, so there is no dependent variable. Furthermore, many readers have difficulties in interpreting the results of regression analyses and the chi square test is easy to interpret and the results are easy to present graphically. According to comment #4, we corrected the secondary analyses for multiple testing by Bonferroni-Holm correction (methods section page 10 line 220).

Comment #4: Why were the secondary analyses not corrected for multiple comparisons as well? Although the analysis was exploratory, issues with multiple comparisons still apply.

Reply/action taken: According to the reviewers suggestion we corrected the secondary exploratory analyses for multiple testing as well (methods section page 10 line 220, and in the results section page 12 lines 252-253, 265-266 and 268; page 14 lines 289-290; page 15 lines 293-294).

Comment #5: Table 1 (page 9).

i. For items described in terms of n(%). The denominator for the percent appears to be inconsistent. For example, 55 females out of 97 EMC patients = 57%, but 78 melancholic MDD out of 97 EMC patients = 80% (versus 22% reported in the text). Please clarify.
ii. For items described in terms of (+/- SD). Is this mean +/- SD or median +/- SD or something else? Also, ensure consistency of reporting.

iii. It is unclear to me what is meant by 1st episode/recurrent MDD. Aren't these two different groups of patients, and wouldn't 100% of patients fit into these two categories? Or is this the proportion of 1st episode patients only?

Reply/action taken: i. According to the reviewers suggestion, we clarified the percentage figures. ii. For items described in Terms of “± SD”, it is “mean ± standard deviation” (SD). We added this information to the manuscript as well as to all tables (results section table 1 page 10-11).

Comment #6: Table 2 (pages 10-11). I don't think it is necessary to report both the regression coefficient and OR as they provide equivalent information. I suggest just including the OR for simplicity.

Reply/action taken: We are very thankful for this suggestion and deleted the regression coefficient for simplicity (results section: page 12 lines 247-248, 249-250, 261-62; pages 13-14 table 2, page 14 lines 285-286 and 287).

Comment #7: I could not recreate the Bonferroni-Holm adjusted p-values provided in Table 2. For example, if listed in ascending order, rank 1 would correspond to p=0.019 (Atypical MDD, treatment group). Given 36 comparisons, the adjusted p-value should be:

Adjusted p-value = (C-i+1)*p = (36-1+1)*0.019 = 0.684.

The corresponding adjusted p-value reported was 0.228. This p-value would correspond to only 12 comparisons. Were the adjustments made by groups?

Reply/action taken: The reviewer is right: The adjustments were made by groups, therefore we only corrected for 12 comparisons. However, according to reviewer two, we decided to change the whole “correction process” as the use of corrections for multiple comparisons is to decrease the "family-wise error rate" (FWER).

The inclusion of all 12 variables within one family of subtypes increases the risk of a Type 2 error by considering all variables/predictors as members of one "family" rather than separating them into the families/sets that were determined a-priori and listed under "Possible predictors of treatment outcome". Therefore we analyzed each of the four sets of predictors independent from
each other, correcting for the number of comparisons performed within each of the four sets rather than the grand total number of comparisons. According to the reviewers suggestion we corrected in the “family” for all calculations not only for the groups. This proceeding was now explained in more detail in the revised manuscript, to clarify the method of the correction for multiple testing used in our analyses (Methods section; pages 9-10; lines 210-215).

NICHOLAS ARA MISCHL, M.D., PH.D. (REVIEWER 2):

Comment #1 (General): There is an issue with the statistical analysis. There is an issue in the abstract and the full manuscript with inappropriate reporting of uncorrected p values as evidence, albeit labelled as "preliminary". It's my view, shared by many, that uncorrected p values are irrelevant at best and misleading at worst in cases of multiple comparisons. Any inferences drawn from these data are similarly irrelevant or misleading.

Reply/action taken: In principle, we agree with the reviewer's statement, but we do not agree that uncorrected p-values are totally irrelevant, as they can help to create new hypotheses to be tested in further studies with larger sample sizes and higher power (see also comment #1 of reviewer 3). Nevertheless, we followed the reviewer’s suggestion to report the corrected p-values. As we used the Bonferroni-Holm correction method and not the Bonferroni method, it was not possible only reporting the corrected p-values, but we deleted all misleading sentences reporting uncorrected p-values as evidence (Abstract: page 3 lines 51-52; 54-56; results section page 12 lines 249, 252-253, 262-263, 265-266 and 268; page 13 table 2; page 14-15 lines 287-294).

Comment #2 (Methods): The issue with the statistical analysis relates to the corrections for multiple comparisons. Is each predictor considered a discrete "subtype" in the analysis? More detail is required with respect to if corrections were made for the entire set of "subtypes", in this case corrected for 12 comparisons within the entire "family" of variables/predictors. The term family is used to refer to the statistical convention of calling a related set of variables, age/sex/race for example, a "family" and the use of corrections for multiple comparisons to decrease the "family-wise error rate" (FWER).

It appears that in this analysis, all variables were included within a family of 12 or so variables. It is possible that this approach unnecessarily increases the chance of a Type 2 error by considering all variables/predictors as members of one "family" rather than separating them into the families/sets that were determined a-priori and listed under "Possible predictors of treatment outcome". It is my opinion that the more appropriate approach would be to analyze each of the
four sets independent from each other, correcting for the number of comparisons performed within only each of the four sets rather than the grand total number of comparisons. It is possible that there is a signal in the MDD atypical subtype that is being missed since the analysis described here corrects for >10 comparisons rather than the four comparisons done within melancholic vs. anxious vs. atypical vs. suicidal set/family of predictors.

If I am mistaken regarding how the corrections for multiple comparisons were done, I apologize. In any case please be more descriptive in the manuscript or in a supplement regarding the how the number of comparisons was determined in post-hoc statistical correction procedures.

Reply/action taken: We are grateful for this recommendation and we explained the correction for multiple comparisons in more detail in the revised MS (Methods section; pages 9-10; lines 210-215). We corrected each set of predictors independently from the other sets, correcting for the number of comparisons performed within each of the four sets rather than the grand total number of comparisons. Additionally we described how the number of comparisons was determined in post-hoc statistical correction procedures (page10 lines 216-221).

Comment #3 (Results): Related to the above issues, the resulting analysis of the predictors is not listed according to the four priori determined sets that are listed under "Possible predictors of treatment outcome". Please reformat the tables to group these families/sets of predictors with one another and specify families/sets with sub-headings.

Reply/action taken: According to the reviewer’s suggestion, we changed the sequence of the predictors in table 2 and added subheadings with the names of each set of predictors (results section page 13-14 table 2).

Comment #3 (Results): As mentioned above, it is not appropriate to highlight or discuss any uncorrected p value in a multiple comparison procedure due to the problem of multiplicity making such values irrelevant and/or misleading.

Reply/action taken: According to the reviewers suggestion we changed the presentation of the results with only discussing the corrected p-values (see comment 1, reviewer 2 results section page 12 lines 249, 252-253, 262-263, 265-266 and 268; page 13 table 2; page 14-15 lines 293-297).

Comment #4 (Results): Any odds ratio reported in the abstract or text should include the 95% confidence interval in parentheses beside it and only corrected p values.
Reply / action taken: According to the reviewer’s suggestion we checked all confidence intervals and added the missing confidence interval (results section page 12, line 262). Concerning the reviewer’s suggestion to only report corrected p-values, it was not possible only reporting the corrected p-values as we used the Bonferroni-Holm correction method and not the Bonferroni method. However, we deleted all misleading sentences reporting uncorrected p-values as evidence (results section page 12 lines 254-259, page 15 lines 293-297).

Comment #5 (Results): It is not clear if or how the presence and severity of psychiatric somatic symptoms like irritable bowel, palpitations, headache, impacted the CIRS score relative to the presence and severity of non-psychiatric medical comorbidities like diabetes, heart disease, kidney disease, etc. If possible, report the presence and level of psychiatric somatic symptoms and any somatic symptom disorders in your cohort and account for this in the analysis.

Reply / action taken: According to the reviewer’s suggestion we now described the subitems of the CIRS scale (organ systems) with the mean value of all CIRS items and of the HAMD-item No. 11 of somatic symptoms. In the CIRS sum score we excluded the psychiatric subitem in order to separate psychiatric somatic symptoms from somatic symptoms (methods section pages 8-9 lines 185-19; results section page 10 line 229-232; table 1 page 11).

Additionally we assessed somatic symptoms of the depression by the HAMD interview and the raters were trained only to assess somatic comorbidities in the CIRS which are not a consequence of depression (page 9, lines 187-189).

Comment #6 (Discussion/Implications/Limitations/Conclusions): This will need to be re-written based on the feasibility and results of the analyses requested above.

Reply / action taken: The discussion/Implications/Limitations/Conclusions is rewritten based on the new analyses (pages 15-19, lines not shown).

Waldemar Greil, MD (Reviewer 3):

Comment #1 (General): The paper can stimulate to carry out analogous studies - if possible with a larger number of cases and thus higher statistical power.
Then the described findings - for example in a meta-analysis of several studies - could be statistically significant (even after Bonferroni correction).

This could be emphasized even more in the discussion and in the conclusions.

"The study yielded statistically negative results" - although true - puts the important results into perspective.

Reply/ action taken: We are grateful for this important comment. We emphasized that statistically not significant results does not mean that the results are clinically irrelevant or that the identified predictors might not be proofed to be relevant in further studies with larger samples sizes. We added this point to the discussion and in the conclusions (page 19 lines 397-401 409-415 and page 20 line 434-436.

Comment #2 (Results): The presentation of the results (p. 11 and p. 12) is difficult to read and partly probably not correct.

S.12 line 242 should read: remission rates were 22.6% and 20% (not 27.3%, please check).

It would be easier to control the data if the exact numbers in the figures 1a to 1c were given - in the figure itself or in the legend.

Reply/ action taken: We checked the results and rephrased misleading sentences and checked all remission rates and changed them accordingly to the reviewers suggestion (results section: page 12 lines 250-269, page 15 293-297). We added the exact number in the figures 1a-1c (legend section page 26-27 lines 667-671, 676-678, 683-686).

Comment #3 (Results): On page 11, on line 229, it says "versus 7.6" on line 232, but "7.5%", meaning the same value (better uniform).

Reply/action taken: According to the reviewers suggestion we changed the value for equalization (page 12 line 267).

Comment #3: P 10: Table 2 "Suicidality" better than "suicidality" (capital letter) Termination point missed at p 18 line 412 between "Mainz." and Monika Seifert and p 19 line 422 to "Andernach."

Reply/action taken: We changed the term into Suicidality and added the missing Termination points (results section, table 2, page 14 and Acknowledgements page 22, lines 489 and 499).
Heather A. MacPherson (Reviewer 4):

Comment #1 (general): This study examined predictors of the effectiveness of an early medication change strategy in patients with major depressive disorder (MDD). Primary outcomes from this trial demonstrated that remission rates in an early medication change (EMC) group versus treatment as usual (TAU) were not significantly different. The current paper similarly reports statistically negative results after correction for multiple comparisons. However, recurrent MDD, prior medication at study entry, and lack of atypical depression showed promise as potential predictors of higher remission rates in EMC vs. TAU. Though this is certainly an important and understudied topic, the publication of a second paper with null findings from the same sample offers limited value above and beyond the initial publication, and therefore limits the conclusions that can be drawn, and hampers my overall enthusiasm for the paper.

Reply: Nearly all publications in medicine report “positive” findings (Fanelli 2010) with psychiatry and psychology having the highest proportion of studies reporting support for the tested hypothesis. 95% of all published data in psychiatry and psychotherapy yielded positive results, in contrast to 22% in astrophysics. 36% of study replications yielded significant findings (p value below 0.05) compared to 97% of the original studies that had significant effects. The mean effect size in the replications was approximately half the magnitude of the effects reported in the original studies. (Collaboration, Open Science (2015-08-28). "Estimating the reproducibility of psychological science". Science. 349 (6251): aac4716. doi:10.1126/science.aac4716. 0036-8075. PMID 2631544) To our point of view, the "success" of study results should not be evaluated on the basis of whether they are significant or not, but on the importance of the clinical research question and the potential the study has to stimulate further research. We would like to cite reviewer 3: “The paper can stimulate to carry out analogous studies - if possible with a larger number of cases and thus higher statistical power. Then the described findings - for example in a meta-analysis of several studies - could be statistically significant (even after Bonferroni correction). “

Comment #2 (general): Also, on a minor note, the writing is a bit choppy and there are some grammatical errors and awkward wording/transitions throughout, which should be addressed.

Reply/action taken: Document was edited by an English native speaker in order to correct grammar, wording and punctuating. (lines not shown).
We again wish to thank the editor and the reviewers for their efforts and very helpful comments and we hope that you will find our revised manuscript now fully suitable for publication.

Nadine Dreimüller, MD

(Corresponding author)