The backwoods of clinical and research assessment of psychosis

Berendsen, S.

Publication date
2021

Chapter 10

An old but still burning problem: inter-rater reliability in clinical trials with antidepressant medication

Steven Berendsen, Loek M. A. Verdegaal, Mirjam J. van Tricht,
Matthijs Blankers, Henricus L. Van, Lieuwe de Haan

In: Journal of Affective Disorders, 276 (2020) 748-751

Abstract

Antidepressant trials are criticized due to potential methodological flaws. Root causes of failing methodology can be found in insufficient inter-rater reliability (IRR) and training practices, leading to higher placebo response and reduced study-power. However, it is unknown to what extent reliability estimates or training procedures are currently included in antidepressant reports. Therefore, we aimed to determine the proportion of publications concerning double-blind randomized controlled antidepressant trials that report IRR coefficients and training procedures. We extracted all double-blind randomized clinical trials (RCTs) from the meta-analysis of Cipriani et al. (2018) concerning the period from 2000 until January 2016. Further, we conducted a Medline-search for double-blind RCTs from January 2016 until January 2020 for additional reports. We identified IRR coefficients and training procedures in these publications. In total we identified 179 double-blind RCTs. Only 4.5% reported an IRR coefficient whereas 27.9% reported training procedures. We did not contact individual authors for additional information regarding implementation of training procedures or inter-rater reliability assessment. There is a substantial lack of reporting IRR coefficients and training procedures in RCTs with antidepressant medication. Considering the large implications of insufficient reliability, we urge researchers to conduct and report training procedures and reliability estimations.

Introduction

There is an ongoing fierce debate concerning the efficacy of antidepressant medication for major depressive disorder.[1] Grounds for this debate are negative results of antidepressant medication trials, potentially due to major shortcomings in study-methodology.[2] One of the noted limitations in methodology of antidepressant trails is failing training procedures resulting in poor inter-rater reliability (IRR).[3]

To achieve sufficient IRR in clinical trials intensive training procedures for assessors are necessary.[4] This implies knowledge of psychiatric disorders, training in instruments to assess severity of symptoms and feedback on performed assessments by experienced clinicians. Lacking training procedures leads to remarkable poor IRR estimates with large consequences on outcome.[5, 6] More specifically, power analysis is generally based on observed variance and does not include measurement error, i.e. reliability, in the calculation. Consequently, by disregarding measurement error the estimated variance used in the power-calculation is actually too low, causing unpowered sample sizes.[7] To illustrate this: an improvement of IRR from 0.5 to 0.8 leads to a reduction of approximately 40% of the required sample size. Secondly, training procedures for raters may reduce biases in clinical trials.[8] For instance, rater biases may cause inflated baseline scores enabling quick trial enrollment, while true symptom scores of these included patients are actually lower. Inflated baseline scores are associated with an increased placebo-response, making it more difficult to find significant differences between study-drug and placebo.[9]

Although training procedures and reliability estimates are profusely recognized for their importance, it is scarcely addressed in literature. The most recent estimate of training procedures or IRR in randomized controlled trials (RCTs) for depressive disorders goes back to 2002, reporting that merely 19% of the RCTs for depressive disorders reported training procedures or reliability estimates.[10] We think that it is particularly important to evaluate the inclusion of IRR coefficients and training procedures in more recent double-blind RCTs with antidepressant medication, in light of declining efficacy of antidepressant medication in major depressive disorders during the last two decades.[2]

Accordingly, the present study aims to determine the inclusion rate of IRR coefficients and training procedures in publications of double-blind randomized controlled trials with antidepressant medication during the past two decades. We hypothesize that the proportion of studies reporting IRR coefficients or training procedures will be higher in comparison to studies before 2000, due to increased awareness of the importance of IRR and improved study-methodology.

Methods

Included studies

We extracted all double-blind RCTs with antidepressant medication published between January 2000 and January 2016 from the influential network meta-analysis of Cipriani et al.[11] In addition, two authors (SB and LV) independently conducted Medline search for double-blind RCTs with antidepressant medication versus placebo from January 2016 till January 2020. Inclusion and exclusion criteria were identical as reported by Cipriani et al. Additionally, we included all Food and Drug Administration (FDA) approved antidepressant medication in our search. We reported the search strategy, terms and full list of included studies in the supplement. We used the full-text manuscript with supplement provided by Medline.

Data collection

The following data were extracted from the included studies: presence of an IRR coefficient with score, reported correlation coefficient between raters, percentage agreement and presence of rater training. Presence of an actual IRR coefficient was defined as an Intraclass Correlation Coefficient, Cohen's Kappa, Krippendorff's alpha or Agreement Coefficient-1. Presence of rater training was defined as any sort of reported training procedure for raters in the published manuscript. Examples of text that indicated training procedures: "All investigators involved in the trial took part in a 4-day practical pre-study check on the administration of the diagnostic and rating instruments.", or "Only those investigators who had actively participated in rater training sessions and who had received prior rater certification were allowed to rate patients", or merely stating that "ratings were performed by individuals that were trained".

Procedure

Two authors independently extracted data from the articles and supplements. To extract data from the included studies one author (SB or LV) read the method-section. Next, the author checked whether the search function of the PDF worked appropriately and searched for the following terms: *reliability, ICC, IRR, rater, training, consistency, agreement, Cohen, kappa, correlation, appendix and supplement*. If the search function of the PDF did not function properly, we converted the PDF to a Word document and used this document to search for the relevant terms. We searched for the term 'appendix' and 'supplement' to check if there was a corresponding appendix or supplement which we could explore for inclusion of inter-rater reliability and training procedures. In addition, we checked whether a supplement or appendix was provided on the website of the publisher. At the start of the trial and half-way during the data collection we assessed the IRR of the latter method among two authors SB and LV. To this end, authors SB and

LV applied the presented extraction-method to the first 25 manuscripts at the start of the trial, and another 20 manuscript half-way the data collection (after 107 manuscripts). Cohen's Kappa scores for presence of rater training was 0.915 [95% CI 0.719-1.000], and for the presence of an IRR coefficient 1.0. The second IRR assessment half-way the data-collection, resulted in Cohen's Kappa score for both the presence of an IRR coefficient and rater training of 1.0.

During the screening of the first 25 manuscripts there was one disagreement concerning the presence of rater training between L.V. and S.B. Both authors reviewed this manuscript again and reached consensus with regard to the presence of the variable.

All included studies applied commonly used observational instruments to assess severity of symptoms, one of the following instruments were always included: Hamilton Depression Rating Scale, Hamilton Anxiety Rating Scale, Montgomery Asberg Depression Rating Scale, and one study included the clinician-rated Inventory of Depressive Symptomatology (IDS-C-30).[12, 13] IRR assessment is deemed necessary for these instruments.

Statistical analysis

Chi-square tests were used to assess differences in reporting IRR and training procedures between an earlier study by Mulsant et al. and our results. Group differences and IRR analysis were performed with Statistical Package for the Social Sciences version 22.

Results

Study characteristics

In total we identified 179 double-blind RCTs with AD medication versus placebo or head-to-head for depressive disorders, the list of references can be found in the supplement section 1.3. We included 168 studies from the network meta-analysis of Cipriani et al. Additionally, our Medline search concerning the period from January 2016 to January 2020 resulted in 362 hits. Based on title and abstract we excluded 351 studies and we included 11 additional studies.

Reporting IRR coefficients and rater training

Of the 179 double-blind RCTs, 27.9% (N=50) reported training procedures, 4.5% (N=8) of all studies reported an IRR coefficient. We found that all reported ICC and Cohen's Kappa scores investigated inter-rater reliability of the outcome measures such as the HAMD or MADRS. Of these latter eight RCTs, six reported an ICC, two used Cohen's Kappa and one of the latter eight studies also reported a percentage agreement. Nine studies reported

that reliability of assessments was determined but did not include an IRR coefficient or score in their paper or supplemental material, and one other study reported the use of centralized raters. Compared to the study of Mulsant et al. 2002, the percentage of papers that reported IRR coefficients decreased significantly (X^2 (dF= 1, N=241) =13.408, P<0.001), while the percentage of papers that reported training procedures increased significantly (X^2 (dF= 1, N=241) =8.604, P=0.003).

Discussion

The current study aimed to determine the reporting practices of reliability estimates and training procedures in double-blind RCTs with antidepressant medication for major depressive disorder during the last two decades. Our study clearly demonstrated that IRR coefficients and training procedures were both strongly underreported.

Earlier studies reported that only in 19% of the publications concerning RCTs for depressive disorders an IRR coefficient is mentioned and only 9% reported the use of training procedures.[10, 14] Our findings concerning the last two decades showed that only 4.5% of the reviewed papers reported IRR and only 27.9% reported on training practices. This implies that the proportion of RCTs reporting IRR coefficients even significantly declined over the last twenty years, while the percentage of training procedures significantly increased. Either way, IRR coefficients and training procedures are clearly neglected, this is an old problem but undoubtedly still alive and kicking.

The majority of RCTs that reported IRR used proper measures of IRR, such as the ICC or Cohen's Kappa. However, one RCT used percentage agreement as an IRR measure. As percentage agreement is not change-corrected, this kind of analysis is not suitable for IRR. We assume that reliability estimates and training procedures will be reported when they are performed. The question remains why reliability and training procedures in RCTs for antidepressant medication are neglected. We can only think of two highly disturbing reasons. Firstly, authors could be unaware of the relevance of training procedures and reliability estimates for their data. Illustrative in this respect is that current international guidelines such as CONSORT do not require the reporting of reliability or rater training procedures.[15] Secondly, designers of clinical trials could consider training procedures and reliability assessment as too labor-intensive and not feasible or necessary. Both reasons are disputable considering the high impact of lacking training procedures and reliability estimates on power, sample size and placebo-response.[7]

Limitations

Our study should be viewed in light of some limitations. Firstly, we did not contact individual authors of RCTs for additional information concerning training procedures or inter-rater reliability assessment. Secondly, the included manuscripts with antidepressant medication sparsely reported detailed information about training procedures. Therefore, we were not able to retrieve more specific data about for instance the number of raters, their professional background or the amount of provided training. Additionally, more information on the applied methods to achieve reliability such as videotaped or live interviews were generally lacking in the included studies. Thirdly, we only evaluated RCTs presenting the primary results of the outcome of antidepressants. It is therefore possible that we missed inter-rater reliability analysis or training procedures if they were reported in separate design manuscript concerning individual RCTs. However, we deemed this less likely since we think that most authors consider it a strength if IRR analysis or training procedures are performed and therefore will be inclined to mention these in their main paper.

The main strength of our study is the large number of included studies representing the most important studies on antidepressant medication during the last twenty years. Taken together, our study demonstrated that training procedures and reliability estimates were strongly underreported in double-blind RCTs with antidepressant medication, and that the situation concerning reliability estimates is even worse than two decades ago. In fact, the included trial reports in our review form the backbone of international antidepressant guidelines, making the identified flaws in reported methodology even more striking.

We feel that far-reaching recommendations to improve clinical trial assessments and outcome are needed. Firstly, clinical trials should report and conduct repeated training procedures followed by IRR assessment. This is important for outcome parameters but also for recruitment assessment or evaluation of adverse events. Training procedures may contain interview courses, videotaped as well as live interviews with patients with direct feedback (on assessment). Although this might be seen as problematic for multicenter trials, feasibility has been shown by an earlier study.[16] Secondly, we consider the use of centralized raters instead of site-raters the best option, who are reliable in assessment of instruments, highly trained and unrelated to study site, provided more accurate measurements leading to less bias and significantly decreased placebo response.[8, 17] Thirdly, clinical assessment may be recorded for reevaluation on reliability and rater drift. Thereafter, unreliable ratings can be adjusted and underperforming raters can receive additional training. Finally, and probably the most easy measure to implement, we think that the CONSORT guidelines and author instructions of scientific journals should include a statement concerning the relevance and consequences of rater-training and IRR.

Ethic approval

Since no patient data was used for this study we did not need approval of the Medical Ethics Committee.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgements

The authors would like to thank Jentien Vermeulen PhD for her help with the statistical analysis.

Conflict of interest statement

All authors declare not to have any conflicts of interest that might be interpreted as influencing the content of the manuscript.

Contributions

Authors SB and LMAV contributed to the study design and proposal, literature search, data collection, analysis and interpretation. Author SB drafted the manuscript and all other authors provided critical revisions. Authors MJT, MB, HLV and LH supervised statistical analysis, study design and writing of the manuscript. All authors contributed to and have approved the final manuscript.

References

1. Vinkers CH. [The fierce debate about antidepressant efficacy: will it ever end?]. *Nederlands tijdschrift voor geneeskunde* Jun 29 2018;162.
2. Gelenberg AJ, Thase ME, Meyer RE, et al. The history and current state of antidepressant clinical trial design: a call to action for proof-of-concept studies. *The Journal of clinical psychiatry* Oct 2008;69(10):1513-1528.
3. Marder SR, Laughren T, Romano SJ. Why Are Innovative Drugs Failing in Phase III? *Am J Psychiatry* Sep 1 2017;174(9):829-831.
4. Berendsen S, Kapitein P, Schirmbeck F, et al. Pre-training inter-rater reliability of clinical instruments in an international psychosis research project. . *Accepted for publication in Schizophrenia Research* 2020.
5. Loevdahl H, Friis S. Routine evaluation of mental health: reliable information or worthless "guesstimates'? *Acta psychiatrica Scandinavica* Feb 1996;93(2):125-128.
6. Muller MJ, Wetzel H. Improvement of inter-rater reliability of PANSS items and subscales by a standardized rater training. *Acta psychiatrica Scandinavica* Aug 1998;98(2):135-139.
7. Perkins DO, Wyatt RJ, Bartko JJ. Penny-wise and pound-foolish: the impact of measurement error on sample size requirements in clinical trials. *Biological psychiatry* Apr 15 2000;47(8):762-766.
8. Kobak KA, Leuchter A, DeBrota D, Engelhardt N, Williams JB, Cook IA, Leon AC, Alpert J. Site versus centralized raters in a clinical depression trial: impact on patient selection and placebo response. *Journal of clinical psychopharmacology* Apr 2010;30(2):193-197.
9. Rutherford BR, Pott E, Tandler JM, Wall MM, Roose SP, Lieberman JA. Placebo response in antipsychotic clinical trials: a meta-analysis. *JAMA psychiatry* Dec 1 2014;71(12):1409-1421.
10. Mulsant BH, Kastango KB, Rosen J, Stone RA, Mazumdar S, Pollock BG. Interrater reliability in clinical trials of depressive disorders. *Am J Psychiatry* Sep 2002;159(9):1598-1600.
11. 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* Apr 7 2018;391(10128):1357-1366.
12. Sharp R. The Hamilton Rating Scale for Depression. *Occupational medicine (Oxford, England)* Jun 2015;65(4):340.
13. Davidson J, Turnbull CD, Strickland R, Miller R, Graves K. The Montgomery-Asberg Depression Scale: reliability and validity. *Acta psychiatrica Scandinavica* May 1986;73(5):544-548.
14. Vacha-Haasem TN, C.; Nillson, J.; Reetz, D. Practices Regarding Reporting of Reliability Coefficients: A Review of Three Journals. *The Journal of Experimental Education* 1999;67:335-341.
15. http://www.consort-statement.org/.
16. Kobak KA, Lipsitz JD, Williams JB, Engelhardt N, Bellew KM. A new approach to rater training and certification in a multicenter clinical trial. *Journal of clinical psychopharmacology* Oct 2005;25(5):407-412.
17. Kobak KA, Engelhardt N, Williams JB, Lipsitz JD. Rater training in multicenter clinical trials: issues and recommendations. *Journal of clinical psychopharmacology* Apr 2004;24(2):113-117.

Supplement

Medline search
Search terms: (depress* OR dysthymi* OR “adjustment disorder” OR “affective disorder” OR “mood disorder” OR “affective symptoms”) AND (“agomelatine” OR “bupropion” OR “citalopram” OR “desvenlafaxine” OR “duloxetine” OR “escitalopram” OR “fluoxetine” OR “fluvoxamine” OR “levomilnacipran” OR “milnacipran” OR “mirtazapine” OR “paroxetine” OR “reboxetine” OR “sertraline” OR “venlafaxine” OR “vilazodone” OR “vortioxetine” OR “clomipramine” OR “nefazodone” OR “amitriptyline” OR “trazodone” OR “selegiline” OR “isocarboxazid” OR “phenelzine” OR “tranylcypromine” OR “amoxapine” OR “desipramine” OR “doxepin” OR “imipramine” OR “nortriptyline” OR “protriptyline” OR “trimipramine” OR “maprotiline” OR “esketamine” OR “brexanolone”).

Criteria Medline
Selection criteria search: RCT, range from January 2016 till January 2020, humans. Results: 362 hits.

Selection of titles and abstracts on:
- Double-blind RCT, primary diagnosis of MDD, oral monotherapy, age >18 years, no post-hoc analysis.
- Primary or secondary outcome measurement of difference in depressive symptoms compared to baseline using a clinician-rated scale (HAM-D, MADRS or other clinician-rated observational instruments).

Studies reporting inter-rater reliability coefficients [1-8]

1. Mundt JC, Vogel AP, Feltner DE, Lenderking WR. Vocal acoustic biomarkers of depression severity and treatment response. *Biological psychiatry* Oct 1 2012;72(7):580-587.
2. Wang G, McIntyre A, Earley WR, Raines SR, Eriksson H. A randomized, double-blind study of the efficacy and tolerability of extended-release quetiapine fumarate (quetiapine XR) monotherapy in patients with major depressive disorder. *Neuropsychiatric disease and treatment* 2014;10:201-216.
3. Roose SP, Sackeim HA, Ranga Rama Krishnan K, Bruce Pollock CG, Alexopoulos G, Lavretsky H, Katz IR, Hakkarainen H. Article Antidepressant Pharmacotherapy in the Treatment of Depression in the Very Old: A Randomized, Placebo-Controlled Trial Old-Old Depression Study Group. *Am J Psychiatry* 2004;161:11.
4. Goldstein DJ, Mallinckrodt C, Lu Y, Demitrack MA. Duloxetine in the treatment of major depressive disorder: A double-blind clinical trial. *Journal of Clinical Psychiatry* 2002;63:225-231.
5. Dimidjian S, Hollon SD, Dobson KS, et al. Randomized trial of behavioral activation, cognitive therapy, and antidepressant medication in the acute treatment of adults with major depression. *Journal of Consulting and Clinical Psychology* 2006;74:658-670.
6. Barber JP, Barrett MS, Gallop R, Rynn MA, Rickels K. Short-term dynamic psychotherapy versus pharmacotherapy for major depressive disorder: A randomized, placebo-controlled trial. *Journal of Clinical Psychiatry* 2012;73:66-73.
7. Grunebaum MF, Ellis SP, Duan N, Burke AK, Oquendo MA, John Mann J. Pilot randomized clinical trial of an SSRI vs bupropion: effects on suicidal behavior, ideation, and mood in major depression. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology* Feb 2012;37(3):697-706.
8. Rossini D, Serretti A, Franchini L, Mandelli L, Smeraldi E, De Ronchi D, Zanardi R. Sertraline versus fluvoxamine in the treatment of elderly patients with major depression: a double-blind, randomized trial. *Journal of clinical psychopharmacology* Oct 2005;25(5):471-475.

Studies reporting training procedures [1-50]

1. Effect of Hypericum perforatum (St John's wort) in major depressive disorder: a randomized controlled trial. *Jama* Apr 10 2002;287(14):1807-1814.
2. Alvarez E, Perez V, Dragheim M, Loft H, Artigas F. A double-blind, randomized, placebo-controlled, active reference study of Lu AA21004 in patients with major depressive disorder. *International Journal of Neuropsychopharmacology* 2012;15:589-600.
3. Baldwin DS, Cooper JA, Huusom AK, Hindmarch I. A double-blind, randomized, parallel-group, flexible-dose study to evaluate the tolerability, efficacy and effects of treatment discontinuation with escitalopram and paroxetine in patients with major depressive disorder. *International clinical psychopharmacology* May 2006;21(3):159-169.
4. Baldwin DS, Loft H, Dragheim M. A randomised, double-blind, placebo controlled, duloxetine-referenced, fixed-dose study of three dosages of Lu AA21004 in acute treatment of major depressive disorder (MDD). *European Neuropsychopharmacology* 2012;22:482-491.
5. Benkert O, Szegedi A, Kohnen R. Mirtazapine compared with paroxetine in major depression. *The Journal of clinical psychiatry* Sep 2000;61(9):656-663.
6. Bjerkenstedt L, Edman GV, Alken RG, Mannel M. Hypericum extract LI 160 and fluoxetine in mild to moderate depression: A randomized, placebo-controlled multi-center study in outpatients. *European Archives of Psychiatry and Clinical Neuroscience* 2005;255:40-47.
7. Boulenger JP, Huusom AK, Florea I, Baekdal T, Sarchiapone M. A comparative study of the efficacy of long-term treatment with escitalopram and paroxetine in severely depressed patients. *Current medical research and opinion* Jul 2006;22(7):1331-1341.
8. Boulenger JP, 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. *International clinical psychopharmacology* 2014;29:138-149.
9. Cassano GB, Puca F, Scapicchio PL, Trabucchi M. Paroxetine and fluoxetine effects on mood and cognitive functions in depressed nondemented elderly patients. *The Journal of clinical psychiatry* May 2002;63(5):396-402.
10. Chang HH, Chou CH, Yang YK, Lee IH, Chen PS. Association between ABCB1 Polymorphisms and Antidepressant Treatment Response in Taiwanese Major Depressive Patients. *Clinical psychopharmacology and neuroscience : the official scientific journal of the Korean College of Neuropsychopharmacology* Dec 31 2015;13(3):250-255.
11. Colonna L, Andersen HF, Reines EH. A randomized, double-blind, 24-week study of escitalopram (10 mg/day) versus citalopram (20 mg/day) in primary care patients with major depressive disorder. *Current medical research and opinion* Oct 2005;21(10):1659-1668.
12. Croft HA, Pomara N, Gommoll C, Chen D, Nunez R, Mathews M. Efficacy and Safety of Vilazodone in Major Depressive Disorder. *The Journal of clinical psychiatry* 2014;75:e1291-e1298.
13. Cutler AJ. Extended release quetiapine fumarate monotherapy in major depressive disorder: A placebo- and duloxetine-controlled study (Journal Clinical of Psychiatry (2009) 70, 4, (526-539)). *Journal of Clinical Psychiatry* 2009;70:1729.
14. DeMartinis NA, Yeung PP, Entsuah R, Manley AL. A double-blind, placebo-controlled study of the efficacy and safety of desvenlafaxine succinate in the treatment of major depressive disorder. *Journal of Clinical Psychiatry* 2007;68:677-688.

15. Detke MJ, Lu Y, Goldstein DJ, McNamara RK, Demitrack MA. Duloxetine 60 mg once daily dosing versus placebo in the acute treatment of major depression. *Journal of psychiatric research* 2002;36:383-390.

16. Dimidjian S, Hollon SD, Dobson KS, et al. Randomized trial of behavioral activation, cognitive therapy, and antidepressant medication in the acute treatment of adults with major depression. *Journal of Consulting and Clinical Psychology* 2006;74:658-670.

17. Dubé S, Dellva MA, Jones M, Kielbasa W, Padich R, Saha A, Rao P. A study of the effects of LY2216684, a selective norepinephrine reuptake inhibitor, in the treatment of major depression. *Journal of psychiatric research* 2010;44:356-363.

18. Gastpar M, Singer A, Zeller K. Comparative efficacy and safety of a once-daily dosage of hypericum extract STW3-VI and citalopram in patients with moderate depression: A double-blind, randomised, multicentre, placebo-controlled study. *Pharmacopsychiatry* 2006;39:66-75.

19. Gentil V, Kerr-Correa F, Moreno R, et al. Double-blind comparison of venlafaxine and amitriptyline in outpatients with major depression with or without melancholia. *Journal of psychopharmacology (Oxford, England)* Mar 2000;14(1):61-66.

20. Grunebaum MF, Ellis SP, Duan N, Burke AK, Oquendo MA, John Mann J. Pilot randomized clinical trial of an SSRI vs bupropion: effects on suicidal behavior, ideation, and mood in major depression. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology* Feb 2012;37(3):697-706.

21. Hale A, Corral RM, Mencacci C, Ruiz JS, Severo CA, Gentil V. Superior antidepressant efficacy results of agomelatine versus fluoxetine in severe MDD patients: a randomized, double-blind study. *International clinical psychopharmacology* Nov 2010;25(6):305-314.

22. Hsu JW, Su TP, Huang CY, Chen YS, Chou YH. Faster onset of antidepressant effects of citalopram compared with sertraline in drug-naive first-episode major depressive disorder in a Chinese population: a 6-week double-blind, randomized comparative study. *Journal of clinical psychopharmacology* Oct 2011;31(5):577-581.

23. Jefferson JW, Rush AJ, Nelson JC, VanMeter SA, Krishen A, Hampton KD, Wightman DS, Modell JG. Extended-release bupropion for patients with major depressive disorder presenting with symptoms of reduced energy, pleasure, and interest: Findings from a randomized, double-blind, placebo-controlled study. *Journal of Clinical Psychiatry* 2006;67:865-873.

24. Kasper S, Ebert B, Larsen K, Tonnoir B. Combining escitalopram with gaboxadol provides no additional benefit in the treatment of patients with severe major depressive disorder. *International Journal of Neuropsychopharmacology* 2012;15:715-725.

25. Katona C, Hansen T, Olsen CK. A randomized, double-blind, placebo-controlled, duloxetine-referenced, fixed-dose study comparing the efficacy and safety of Lu AA21004 in elderly patients with major depressive disorder. *International clinical psychopharmacology* 2012;27:215-223.

26. Keller MB, Trivedi MH, Thase ME, et al. The Prevention of Recurrent Episodes of Depression with Venlafaxine for Two Years (PREVENT) Study: Outcomes from the 2-year and combined maintenance phases. *The Journal of clinical psychiatry* Aug 2007;68(8):1246-1256.

27. Koshino Y, Bahk WM, Sakai H, Kobayashi T. The efficacy and safety of bupropion sustained-release formulation for the treatment of major depressive disorder: a multi-center, randomized, double-blind, placebo-controlled study in Asian patients. *Neuropsychiatric disease and treatment* 2013;9:1273-1280.

28. Mahableshwarkar AR, Jacobsen PL, Serenko M, Chen Y, Trivedi MH. A randomized, double-blind, placebo-controlled study of the efficacy and safety of 2 doses of vortioxetine in adults with major depressive disorder. *Journal of Clinical Psychiatry* 2015;76:583-591.

29. Moller HJ, Glaser K, Leverkus F, Gobel C. Double-blind, multicenter comparative study of sertraline versus amitriptyline in outpatients with major depression. *Pharmacopsychiatry* Nov 2000;33(6):206-212.

30. Montgomery SA, Huusom AK, Bothmer J. A randomised study comparing escitalopram with venlafaxine XR in primary care patients with major depressive disorder. *Neuropsychobiology* 2004;50(1):57-64.

31. Montgomery SA, Kennedy SH, Burrows GD, Lejoyeux M, Hindmarch I. Absence of discontinuation symptoms with agomelatine and occurrence of discontinuation symptoms with paroxetine: a randomized, double-blind, placebo-controlled discontinuation study. *International clinical psychopharmacology* Sep 2004;19(5):271-280.

32. Moore N, Verdoux H, Fantino B. Prospective, multicentre, randomized, double-blind study of the efficacy of escitalopram versus citalopram in outpatient treatment of major depressive disorder. *International clinical psychopharmacology* May 2005;20(3):131-137.

33. Moreno RA, Teng CT, Almeida KMD. Hypericum perforatum versus fluoxetine in the treatment of mild to moderate depression : a randomized double-blind trial in a Brazilian sample Hypericum perforatum versus fluoxetina no tratamento da depressão leve a moderada : estudo duplo-cego randomizad. *Depression* 2005;28:29-32.

34. Ou JJ, Xun GL, Wu RR, et al. Efficacy and safety of escitalopram versus citalopram in major depressive disorder: a 6-week, multicenter, randomized, double-blind, flexible-dose study. *Psychopharmacology* Feb 2011;213(2-3):639-646.

35. Quera-Salva MA, Hajak G, Philip P, Montplaisir J, Keufer-Le Gall S, Laredo J, Guilleminault C. Comparison of agomelatine and escitalopram on nighttime sleep and daytime condition and efficacy in major depressive disorder patients. *International clinical psychopharmacology* Sep 2011;26(5):252-262.

36. Rapaport MH, Schneider LS, Dunner DL, Davies JT, Pitts CD. Efficacy of controlled-release paroxetine in the treatment of late-life depression. *Journal of Clinical Psychiatry* 2003;64:1065-1074.

37. Rickels K, Athanasiou M, Robinson DS, Gibertini M, Whalen H, Reed CR. Evidence for efficacy and tolerability of vilazodone in the treatment of major depressive disorder: A randomized, double-blind, placebo-controlled trial. *Journal of Clinical Psychiatry* 2009;70:326-333.

38. Rossini D, Serretti A, Franchini L, Mandelli L, Smeraldi E, De Ronchi D, Zanardi R. Sertraline versus fluvoxamine in the treatment of elderly patients with major depression: a double-blind, randomized trial. *Journal of clinical psychopharmacology* Oct 2005;25(5):471-475.

39. Schatzberg AF, Kremer C, Rodrigues HE, Murphy GM, Jr. Double-blind, randomized comparison of mirtazapine and paroxetine in elderly depressed patients. *The American journal of geriatric psychiatry : official journal of the American Association for Geriatric Psychiatry* Sep-Oct 2002;10(5):541-550.

40. Schneider LS, Nelson JC, Clary CM, Newhouse P, Krishnan KRR, Shiovitz T, Weihs K. An 8-week multicenter, parallel-group, double-blind, placebo-controlled study of sertraline in elderly outpatients with major depression. *American Journal of Psychiatry* 2003;160:1277-1285.

41. Shu L, Sulaiman AH, Huang YS, Fones Soon Leng C, Crutel VS, Kim YS. Comparable efficacy and safety of 8 weeks treatment with agomelatine 25-50mg or fluoxetine 20-40mg in Asian out-patients with major depressive disorder. *Asian journal of psychiatry* Apr 2014;8:26-32.

42. Versiani M, Amin M, Chouinard G. Double-blind, placebo-controlled study with reboxetine in inpatients with severe major depressive disorder. *Journal of clinical psychopharmacology* Feb 2000;20(1):28-34.

43. Versiani M, Moreno R, Ramakers-van Moorsel CJ, Schutte AJ. Comparison of the effects of mirtazapine and fluoxetine in severely depressed patients. *CNS drugs* 2005;19(2):137-146.
44. Wade A, Crawford GM, Angus M, Wilson R, Hamilton L. A randomized, double-blind, 24-week study comparing the efficacy and tolerability of mirtazapine and paroxetine in depressed patients in primary care. *International clinical psychopharmacology* May 2003;18(3):133-141.
45. Wade A, Gembert K, Florea I. A comparative study of the efficacy of acute and continuation treatment with escitalopram versus duloxetine in patients with major depressive disorder. *Current medical research and opinion* Jul 2007;23(7):1605-1614.
46. Wang G, Gislum M, Filippov G, Montgomery S. Comparison of vortioxetine versus venlafaxine XR in adults in Asia with major depressive disorder: a randomized, double-blind study. *Current medical research and opinion* Apr 2015;31(4):785-794.
47. Wang G, McIntyre A, Earley WR, Raines SR, Eriksson H. A randomized, double-blind study of the efficacy and tolerability of extended-release quetiapine fumarate (quetiapine XR) monotherapy in patients with major depressive disorder. *Neuropsychiatric disease and treatment* 2014;10:201-216.
48. Yevtushenko VY, Belous AI, Yevtushenko YG, Gusinin SE, Buzik OJ, Agibalova TV. Efficacy and tolerability of escitalopram versus citalopram in major depressive disorder: a 6-week, multicenter, prospective, randomized, double-blind, active-controlled study in adult outpatients. *Clinical therapeutics* Nov 2007;29(11):2319-2332.
49. Yu YM, Gao KR, Yu H, Shen YF, Li HF. Efficacy and Safety of Agomelatine vs Paroxetine Hydrochloride in Chinese Han Patients with Major Depressive Disorder: A Multicentre, Double-Blind, Noninferiority, Randomized Controlled Trial. *Journal of clinical psychopharmacology* Jun 2018;38(3):226-233.
50. Zhang L, Xie WW, Li LH, et al. Efficacy and safety of prolonged-release trazodone in major depressive disorder: a multicenter, randomized, double-blind, flexible-dose trial. *Pharmacology* 2014;94:199-206.