**Online Supplementary Material**

Search strategy for Medline (Ovid)

1. exp depression/

2. exp depressive disorder/

3. 1 or 2

4. randomized controlled trial.pt.

5. controlled clinical trial.pt.

6. randomly.ab.

7. trial.ab.

8. groups.ab.

9. (control$ adj3 (trial$ or study or studies)).tw.

10. randomi#ed.ab.

11. placebo$.ab.

12. 11 or 10 or 9 or 8 or 7 or 6 or 5 or 4

13. exp Exercise/

14. exp Exercise Therapy/

15. exp Physical Fitness/

16. exp physical exertion/

17. exp walking/

18. exp running/

19. exp jogging/

20. exp yoga/

21. exp swimming/

22. exp bicycling/

23. exp resistance training/

24. exp weight lifting/

25. exp aerobic exercise/

26. exp physical activity/

27. (physical adj1 activ\*).mp.

28. (physical adj3 (education or training)).tw.

29. (exercise$ or exercising).tw.

30. exp physical education/ and train.mp.

31. 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30

32. 3 and 12 and 31

33. limit 32 to (english language and yr="1980 -Current")

**Table1:** Baseline Depression Severity Categories

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Scale** | **Depression Severity Category** | | | |
| **Sub-threshold** | **Threshold** | | |
| **Mild** | **Moderate** | **Severe** |
| BDI  (BDI & BDI-II)  (Beck *et al.* 1996; Beck *et al.* 1988) | 0 – 12 | 13 – 19 | 20 – 29 | ≥ 30 |
| CDI  (Friedberg R 2011; Kovacs 1985; Kovacs 1992; Matthey & Petrovski 2002) | 0 – 12 | ≥ 13 | Not specified | Not specified |
| HAMD  (Hamilton 1960) | 0 – 7 | 8 – 13 | 14 – 18 | ≥ 19 |
| CES-D  (Radloff 1977; Zick *et al.* 1990) | 0 – 15 | 16 - 26 | Not specified | Not specified |

**Figure 1:** Funnel Plot: trim and fill analysis.

Note: Trim and fill adjusted effect size estimate. White circles are included trials and black circles are imputed trials. The vertical line and black diamond indicate the adjusted effect size.

**GRADE** ratings assessing the quality of the evidence contributing to the effect estimate produced in the primary meta-analysis. (Balshem *et al.* 2011; Higgins *et al.* 2011; Schünemann *et al.* 2013)

RCT level evidence: HIGH ⊕⊕⊕⊕

**1. Limitation in study design (risk of bias):** Likely serious/very serious limitations

**Action:** Consider downgrading one/two levels ⊖ or ⊖⊖

**Reason:** High risk of bias across trials for at least one important domain and unclear for one or more domains. Specifically there was unclear risk of bias for randomization and allocation concealment (selection bias) for most trials. High risk of bias for outcome assessor blinding (ie self-report depression symptoms, detection bias) for most trials which likely increased effect size, particularly important as participants and personnel cannot be blinded to allocation and therefore performance bias is likely. Two thirds of trials were at high or unclear risk of bias for handling of incomplete outcome data.

**2. Inconsistency:** Unlikely

**Action:** do no downgrade

**Reasons:** All trial level 95% CIs overlapped, heterogeneity was low, Chi2 = 24.04, *df* = 15, *p* = 0.06, *I2* = 38% (low according to criteria set out in the Cochrane Handbook)

**3. Indirectness:** Unlikely

**Action:** Do not downgrade

**Reason:** All included trials were relevant to the review question, no indirect comparators were used, all trials recruited participants with at least threshold levels of depression symptoms, all reported depression symptoms as outcomes.

**4. Imprecision:** Unlikely

**Action:** Do not downgrade

**Reason:** Sample size 771, effect size is larger than 0.2 SD, 95% CI does not cross the line of no effect, 95%CI are relatively narrow and include the minimal important clinical difference when effect size converted to Beck Depression Inventory scale score.

**5. Publication Bias**: Suspected

**Action:** Consider downgrading one level ⊖

**Reason:** search strategy does not include unpublished literature, funnel plot suggest four missing trials from the side of no effect, imputation of these missing trials suggests a slight attenuation of effect size. However effect size remains moderate in size and significantly different from zero. Fail Safe *N* suggest over 400 trials with no effect would be required to make the effect no longer different from zero.

**Overall quality of evidence rating:** LOW ⊕⊕⊖⊖ to VERY LOW ⊕⊖⊖⊖

**Interpretation**

|  |  |  |
| --- | --- | --- |
| Low ⊕⊕⊖⊖ | Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect. |  |
| Very low ⊕⊖⊖⊖ | We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect. |  |

**References:**

**Balshem, H, Helfand, M, Schünemann, HJ, Oxman, AD, Kunz, R, Brozek, J, Vist, GE, Falck-Ytter, Y, Meerpohl, J, Norris, S, Guyatt, GH** (2011). GRADE guidelines: 3. Rating the quality of evidence. *Journal of Clinical Epidemiology* **64**, 401-406.

**Beck, AT, Steer, RA, Ball, R, Ranieri, W** (1996). Comparison of Beck Depression Inventories-IA and -II in Psychiatric Outpatients. *Journal of Personality Assessment* **67**, 588-589.

**Beck, AT, Steer, RA, Garbin, MG** (1988). Psychometric properties of the Beck Depression Inventory Twenty-five years of evaluation. *Clinicial Psychology Review* **8**, 77-100.

**Friedberg R, S, A. (2011). . , 1-5.** (2011). CDI Scores in Pediatric Psychiatric Inpatients: A Brief Retrospective Static Group Comparison. *Depression Research and Treatment* **2011**, 1-5.

**Hamilton, M** (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry* **23**, 56-62.

**Higgins, J, Green, S, (editors)** (2011). *Cochrane Handbook for Systematic Reviews of Interventions.* [V5.1 updated March 2011]*,* The Cochrane Collaboration:<http://www.cochrane-handbook.org>

**Kovacs, M** (1985). The children's depression, inventory (CDI). *Psychopharmacology Bulletin* **21**, 995–998.

**Kovacs, M** (1992). *The Children's Depression Inventory.* Mental Health Systems:North Tonawanda, New York

**Matthey, S, Petrovski, P** (2002). The children's depression inventory: error in cutoff scores for screening purposes. *Psychological Assessment* **14**, 146-149.

**Radloff, LS** (1977). The CES-D: A self report depression scale for research in the general population. *Applied Psychological Measurement* **1**, 385-401.

**Schünemann, H, Brożek, J, Guyatt, G, Oxman, A, (editors)** (2013). *Handbook for Grading the Quality of Evidence and the Strength of Recommendations Using the GRADE Approach.* The Grade Working Group:<http://gdt.guidelinedevelopment.org/app/handbook/handbook.html>

**Zick, JM, Attkisson, CC, Greenfield, TK** (1990). Screening for Depression in Primary Care Clinics: The CES-D and the BDI. *The International Journal of Psychiatry in Medicine* **20**, 259-277.