BMJ Open

Exercise for patients with major depression: a systematic review with meta-analysis and Trial Sequential Analysis

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2016-014820
Article Type:	Research
Date Submitted by the Author:	19-Oct-2016
Complete List of Authors:	Krogh, Jesper; Kobenhavns Universitet Sundhedsvidenskabelige Fakultet, Mental Health Centre Copenhagen Speyer, Helene; University of Copenhagen Hjorthøj, Carsten; Copenhagen University Hospital, Mental Health Center Copenhagen Gluud, Christian; Copenhagen University Hospital Rigshospitalet, The Copenhagen Trial Unit, Centre for Clinical Intervention Research Nordentoft, Merete; University of Copenhagen, Mental Health Centre Copenhagen
Primary Subject Heading:	Mental health
Secondary Subject Heading:	Sports and exercise medicine, Evidence based practice
Keywords:	Exercise, MENTAL HEALTH, Systematic review, Meta-analysis, Evidence based medicine, Randomised clinical trials

Exercise for patients with major depression: a systematic review with meta-analysis and Trial Sequential Analysis

Jesper Krogh[1], Carsten Hjorthøj[1], Helene Speyer[1], Christian Gluud[1], Merete Nordentoft[1]

[1] Mental Health Centre Copenhagen. Faculty of Health Sciences. University of Copenhagen, Copenhagen, Denmark.

[2] Copenhagen Trial Unit, Centre for Clinical Intervention Research, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark.

Corrosponding author

Jesper Krogh, Jesper.krogh@dadlnet.dk

Mental Health Centre Copenhagen, Kildegårdsvej 28, Building 15. DK-2900, Hellerup, Denmark

Abstract

Objectives

To assess the benefits and harms of exercise in patients with depression.

Design

Systematic review

Data sources

Bibliographical databases were searched until the 17th of April, 2016.

Eligibility criteria and outcomes

Eligible were randomised clinical trials assessing the effect of exercise in patients diagnosed with depression. Primary outcomes were depression severity, lack of remission, and serious adverse events. Secondary outcomes were quality of life and adverse events, as well as assessment of depression severity and lack of remission during follow-up after the intervention.

Results

Thirty-one trials enrolling 2419 patients were included. The effect of exercise versus control on depression severity was -0.74 standardised mean difference (SMD) (95% CI -0.96 to -0.51; P < 0.001; GRADE: very low quality). Restricting this analysis to the four trials that seemed less affected of bias, the effect vanished to -0.11 SMD (-0.41 to 0.18; P = 0.45; GRADE: low quality). Exercise decreased the relative risk of no remission to 0.78 (0.68 to 0.90; P < 0.001; GRADE: very low quality). Restricting this analysis to the two trials that seemed less affected of bias, the effect vanished to 0.95 (0.74 to 1.23; P = 0.78). Trial Sequential Analysis excluded random error when all trials were analysed. Sub-group analyses found that trial size and intervention duration were inversely associated with effect size for both depression severity and lack of remission. There was no significant effect of exercise on secondary outcomes.

Conclusions

Trials with less risk of bias suggested no antidepressant effects of exercise and there were no significant effects of exercise on quality of life, depression severity, or lack of remission during follow-up. Data for serious adverse events and adverse events was scarce not allowing conclusions for these outcomes.

Systematic review registration

The protocol was published in the journal Systematic Reviews: 2015; **4**:40

DOI: 10.1186/s13643-015-0030-6.

Article Summary

Strengths and limitations of this study

- The protocol for this review has previously been published
- Using meta-regression analysis, trial sequential analysis and the GRADE system the conclusions form this review is based on a firm and transparent platform
- Based on an extensive literature search, this review included 31 one trials allocating more than 2000 participants to exercise or control interventions
- All included participants were diagnosed with depression according to a diagnostic system
- Effect estimates from included trials had considerable heterogeneity

Introduction

Depression is a common disease affecting up to 17% of the population during their lifetime.[1;2] Based on data from the World Health Organisation, depression is ranked as the second largest health-care problem globally, in terms of years lived with disability.[3] Depending on its severity, depression is often treated using psychotherapy, antidepressants, or a combination of both. However, the clinical benefits of antidepressants[4;5] and psychotherapy[6-8] has been challenged. Both treatments are costly in terms of time and money and may also have adverse effects. Compliance with antidepressant treatment is poor; the dropout rate in clinical trials is reported to be between 12% and 40% within the initial 6 to 8 weeks of treatment.[4;9]

The weakness of evidence for the beneficial effect of current interventions, along with problems related to low compliance and harms, has resulted in an interest in using alternative interventions. The use of exercise as an intervention has attracted considerable attention, and various forms of exercise varying in intensity have been assessed in a number of randomised clinical trials to test their effectiveness as a treatment for patients with depression. In 2011, we published a meta-analysis of randomised clinical trials examining the effect of exercise on depressive symptoms in patients with clinical depression.[10] The results suggested that referring patients with clinical depression to exercise programs was associated with a small to moderate effect on depressive symptoms. However, restricting the analysis to three trials with a low risk of bias, the effect estimate was non-significant. Since 2011, other reviews have been published on the effect of exercise on depressive symptoms,[11] in older people,[12] and in patients with chronic illnesses.[13] However, none of these reviews addressed the specific population of adults diagnosed with major depression according to valid diagnostic criteria, such as the International Classification of Diseases[14] or the Diagnostic and Statistical Manual of Mental Disorders.[15] The reviews contained a number of trials that included volunteers who were defined as being depressed on the basis of psychometric testing (for example, Beck

Depression Inventory[16]), as opposed to individuals with a clinical diagnosis of major depression. Furthermore, several randomised clinical trials investigating the effect of exercise in clinically depressed individuals have been published since our 2011 review.[10]

The objectives of the present systematic review are to investigate the beneficial and harmful effects of exercise, in terms of severity of depression, lack of remission, quality of life, and suicide versus controls with or without co-interventions in adults with a clinical diagnosis of major depression. The current systematic review differs from our previous review in a number of aspects.[10] We only considered trials including participants diagnosed with depression according to a validated diagnostic system. We also included trials including patients with somatic co-morbidity, e.g., cancer or diabetes. The harmful effects of exercise interventions are also addressed, the intervention effects being assessed according to the grading of recommendations assessment, development, and evaluation (GRADE) framework, and bibliographical searches have been extended to include a Chinese and a South-American database until 2016.

Methods/design

The protocol for this review has previously been published.[17]

Search strategy

The following bibliographical databases was searched until the 17th of April, 2016: CENTRAL, MEDLINE, EMBASE, Science Citation Index (Web of Science), LILACS, and Wanfang using medical subject headings (MeSH or similar) when possible or text word terms: depression, depressive disorder and exercise, aerobic, non-aerobic, physical activity, physical fitness, walking, jogging, running, bicycling, swimming, strength, or resistance.

Trial selection

One investigator (JK) examined titles and abstracts to remove obviously irrelevant reports. Two investigators (JK + HS) examined full text reports and abstracts determining compliance with inclusion criteria. A trial was considered eligible if it was a randomised clinical trials including patients diagnosed as having major depression according to a valid and recognised diagnostic system (that is, Research Diagnostic Criteria (RDC),[18] International Classification of Diseases (ICD),[14] or Diagnostic and Statistical Manual of Mental disorders (DSM)[15]) and included participants aged >17 years. Both abstracts and full text reports included.

Trials were excluded if they measured depression immediately after a single bout of exercise, compared one form of exercise versus another, or compared different exercise intensities without including a control group. The trials had to allocate participants to an exercise intervention versus a control group (that is, exercise versus a control group receiving no intervention or treatment as usual or an attention control using light exercise) or using exercise as an add-on-treatment (that is, exercise plus usual treatment in the experimental group versus usual treatment alone in the control group). Exercise intervention was defined as a systematic physical intervention with the intention to increase muscle strength and/or cardiovascular fitness, e.g., running, swimming or weight lifting. In case of attention control, it should specifically be mentioned by the authors that the intervention was intended as a control intervention.

Outcomes

The primary outcomes were 1) depressive symptoms measured on a continuous scale assessed at the end of the intervention; 2) lack of remission, that is, a binary outcome of the proportion of participants in each

intervention group of the trial who did not obtain remission at the end of the intervention according to the authors' own definition; and 3) serious adverse events defined according to ICH-GCP as any untoward medical occurrence that was life threatening, resulted in death or persistent or significant disability (ICH-GCP 1997).[19] Serious adverse events accordingly include suicide attempts as well as suicides. The secondary outcomes were quality of life, non-serious adverse events, as well as depressive symptoms and lack of remission assessed after the intervention.

Data extraction

Two authors (JK, HS) independently extracted data using a pre-piloted structured form. Any discrepancies in the data extraction or inclusion/exclusion of trials was resolved by referring to the original papers. CG or MN assisted as adjudicator in cases of disagreements. Data extraction included, in addition to outcomes, information regarding country of origin, number of randomised participants, number of participants included in efficacy analysis, mean age of participants, diagnostic system, baseline assessment of depression severity, type of intervention, frequency of intervention, and duration of intervention. JK and CH independently performed the assessment of bias domains. The authors JK, CG, and MN have previously published trial reports assessing the effect of exercise in patients with depression,[20;21] and to reduce the risk of academic bias two additional authors were included in the current systematic review (CH, HS).

Risk of bias assessment

Definitions in the assessment of bias risk of a trial was conducted according to the Cochrane Handbook for Systematic Reviews of Interventions[22] of the following domains: allocation sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessors, incomplete outcome data, selective outcome reporting, for-profit bias, and other bias. Trials assessed as having 'low risk of bias'

in all of the above specified domains were considered 'trials at low risk of bias'. Trials assessed as having 'uncertain risk of bias' or 'high risk of bias' in one or more of the above specified domains were considered trials with 'high risk of bias'. In line with our previous systematic review[10] and the latest Cochrane review on exercise for depression,[23] trials at low risk of bias in the allocation concealment domain, blinded outcome assessment domain, and the incomplete outcome data domain were characterised as 'trials potentially having less risk of bias than other trials at high risk of bias'. Trials assessing the effect of behavioural interventions are rarely able to mask the allocation, and participants and health care providers are therefore not blinded. Therefore, we will also report the number of trials at low risk of bias in the remaining domains.

Data synthesis and analysis

In order to be able to include all of the trials in our meta-analysis, estimates of standardised mean difference (SMD) for each individual trial was carried out. SMD is the mean difference in depression score between the exercise and control groups divided by the pooled standard deviation. The result is a unit free effect size. By convention, SMD effect sizes of 0.2, 0.5 and 0.8 are considered small, medium and large intervention effects. In case post-test scores as well as change from baseline was reported, post-test scores were preferred. For dichotomous variables, we calculated the risk ratio (RR) with a 95% confidence interval. It was expected that some trials would have several intervention groups. Data from the experimental groups was pooled and compared with the data from the control group. In case of discrepancies between the random-effects model analysis and the fixed-effect model analysis, both results are reported; otherwise, only results from the random-effects analysis is reported. The degree of heterogeneity was quantified using the I-squared statistic,[24] which can be interpreted as the percentage of variation observed between the trials attributable to between-trial differences, rather than sampling error (chance). Heterogeneity was explored by analyses of sub-groups (see below).

For the primary outcomes, Trial Sequential Analysis was performed.[25;26] In order to calculate the required information size and the cumulative Z-curve's eventual breach of relevant trial sequential monitoring boundaries, the required information size for the primary continuous outcome was based on type I error of 5%, a beta of 10%, the standard error of the meta-analysis, and a minimal difference of three points on the HAM-D_{17}.[17] In order to calculate the required information size and the cumulative Z-curve's eventual breach of relevant trial sequential monitoring boundaries, the required information size for lack of remission was based on type I error of 5%, a beta of 10%, the proportion of patients in the control group with the outcome, and a relative risk reduction of 15% and 30%.

Bayes factors were calculated for all primary outcomes.[27] Low P-values suggest that we can reject the null-hypothesis. But even a low *P*-value from a meta-analysis can be misleading if there is also a low probability that data are compatible with the anticipated intervention effect. In other words, the probability that the actual measured difference in effect of the compared interventions resulted from an a priori anticipated 'true' difference needs to be considered. For this purpose, it is helpful to calculate the Bayes factor, which is the ratio of the P-value probabilities of the meta-analysis result divided by the probability of the anticipated effect, or 'true' effect.[27] As suggested by Jakobsen et al.,[27] a Bayes factor lower than 0.1 together with a low P-value suggest, if bias can be ruled out, that the observed result is compatible with the a priori expected effect. If the Bayes factor is higher than 0.1 the result is not compatible with the a priori expected effect and the effect may be lower.

To assess the potential impact of missing data (incomplete outcome data bias) we did sensitivity analysis of missing data using the following strategy: a 'best-worst' case scenario was assessed, assuming that all participants lost to follow- up in the intervention group had a beneficial outcome (the group mean minus 1

standard deviation (SD)), and all those with missing outcomes in the placebo group have had a harmful outcome (the group mean plus 1 SD and 2 SD). In addition, the reverse 'worst-best-case' scenario analysis was also performed.[27] Missing data for the 'lack of remission' outcome was imputed in sensitivity analysis according to the following scenarios:[28] 1) poor outcome analysis: assuming that all of the drop-outs/participants lost from both the experimental and the control arms experienced the outcome, including all randomised participants in the denominator; 2) good outcome analysis: assuming that none of the drop-outs/participants lost from the experimental and the control arms experienced the outcome, including all randomised participants in the denominator; 3) extreme case analysis favouring the experimental intervention ('best-worse' case scenario): none of the drop-outs/participants lost from the experimental arm, but all of the drop-outs/participants lost from the control arm experienced the outcome, including all randomised participants in the denominator; and 4) extreme case analysis favouring the control ('worst-best' case scenario): all of the drop-outs/participants lost from the experimental arm, but none from the control arm experienced the outcome, including all randomised participants in the denominator.

Subgroup analyses

In subgroup analyses, the possible effects of variables on intervention effects on outcomes and heterogeneity were compared. Trials potentially having less risk of bias (i.e., trials with adequate allocation concealment, blinded outcome assessment, and intention to treat analysis) were compared to trials at high risk of bias. The effect of age was assessed by comparing trials including older participants (mean age >59 years) to trials including younger participants (mean age <60 years). The effect of type of exercise was assessed by comparing trials using group exercises compared to trials using individual exercise. The effect of duration of intervention was assessed by comparing trials with short duration of intervention to trials with long duration of intervention splitting by the median time of duration. The effect of type of control group was assessed by comparing trials using attention control to trials with waitlist controls and

comparing trials with exercise as add-on to medication to trials not using any medication. In addition, a within-study comparison of low-dose exercise versus high-dose exercise in trials using different exercise intensities was performed. The effect of co-morbid somatic disease was assessed by comparing the effect estimates from trials including patients with depression compared to trials including patients with depression in addition to a somatic disease. Publication bias was assessed by visual inspection of a funnel plot and by Egger's test and if publication bias plausible Duval's and Tweedie's trim and fill procedure was conducted.[29]

We assessed and graded the evidence according to the grading of recommendations assessment, development, and evaluation (GRADE) for high risk of bias, imprecision, indirectness, heterogeneity, and publication bias.[30] Based on this assessment, the intervention is graded accordingly: **'**high quality'**-** we are very confident that the true effect lies close to that of the estimate of the effect; **'**moderate quality'- we are moderately confident in the effect estimate. The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different; 'low quality**'**- our confidence in the effect estimate is limited: the true effect may be substantially different from the estimate of the effect; 'very low quality'- we have very little confidence in the effect estimate: the true effect is likely to be substantially different from the estimate of the effect.[31]

Deviations from our protocol

Post-hoc we included trials using the Chinese Classification of Mental Disorders (CCMD) as well as a few trials including patients classified as having 'minor depression'. The CCMD system closely adhere to the ICD and DSM systems and have been found highly compatible in field studies, so these studies were included.[32] A few trials included some patients classified as having 'minor depression' according to the trials chosen diagnostic system (e.g., DSM), and it is questionable if these patients have major depression. We therefore

decided to include these trials but also to conduct a sub-group analysis exclusively including patients with major depression. Post-hoc we also included a sub-group analysis according to trial size. Trials were divided into small or large trials using the median of total n included in the efficacy analysis. We did not conduct Trial Sequential Analysis based on a relative risk reduction of 30% of lack of remission as this was an implausible effect.

Patient involvement

Depressed patients were not involved in this study.

Results

Bibliographical search and trial characteristics

The main bibliographical search was conducted the 26th of August, 2015 and the final updates were conducted on the 17th of April, 2016. As illustrated in Figure S1, we identified 40 publications reporting the effect of exercise on depressive symptoms in 31 randomised clinical trials.[20;21;33-71] Four-teen trials were conducted in Europe,[20;21;38;47;50;51;53;59;63-66;72;73] seven in the U.S.A.,[36;37;41;43;58;62;74], six in Asia,[45;67-71] two in Australia,[52;56] and two in South-America.[54;61] A total of 2,419 patients were randomised and 2,331 were included in the efficacy analysis of benefit. 10 trials included inpatients[45;47;54;65;67-72] and five trials included participants with a mean age above 60 years.[50;52;56;58;59] No trials exclusively included patients with comorbid somatic disease. Please see Table 1 for trial characteristics.

Bias risk assessment

Sequence generation was adequate in 12/31 (39%), allocation concealment was adequate in 12/31 (39%) trials, blinding of participants and trial personnel was adequate in 0/31 (0%), blinded outcome assessment was performed in 16/31 (52%), low risk of bias in the 'incomplete outcome data' domain was found in 12/31 (39%) trials, selective outcome reporting domain was adequate in 27/31 (87%), for profit bias domain was adequate in 15/31 (48%) and 21/31 (68%) were free of other bias. All trials were at high risk of bias. Given the nature of the intervention, no trial had blinded participants or trial personnel, however, two trials had low risk of bias in all other bias domains.[21;52] Five trials (16%) were sponsored by for profit organisations: three trials were supported by pharmaceutical companies,[51;72;75] one trial by a company producing fitness machines,[43] and one trial by an insurance company.[20] According to our a priori defined criteria, 4/31 (13%) trials potentially had less risk of bias than the other trials at high risk of bias.[20;21;52;54] Please see Table 2 for details on assessment of risks of bias.

Primary outcomes

The effect of exercise on depression severity

All included trials provided a continuous outcome on depression severity for the assessment of the exercise intervention encompassing 2,331/2,419 randomised patients (96.4%). The effect of intervention versus control was a standardised mean difference (SMD) of -0.74 (95% CI -0.96 to -0.51; P<0.001) (Figure 1.). This corresponds to an effect on the HAM-D_{17} scale of -4.6 (95% CI -6.0 to -3.2) points.

Missing data

Missing outcome analysis for depression as a continuous outcome did not markedly change the effect estimates. The least favourable outcome for the exercise intervention was the worse/best outcome analysis using +2 SD resulting in an effect estimate of -0.61 SMD (95% CI -0.84 to -0.37; P<0.001) (Table S1).

Heterogeneity and subgroup analysis

The I^2 was 82% suggesting substantial heterogeneity. Subgroup analysis revealed that the effect estimates for trials potentially having less risk of bias was -0.11 SMD (95% CI. -0.41 to 0.18; P = 0.45; I^2 = 62%) compared to that of the trials at high risk of bias -0.85 SMD (95% CI. -1.10 to -0.60; P < 0.001; I^2 = 82%) (test of sub-group difference, P = 0.0002). In addition, trials including 52 patients or less had a pooled estimate of -1.30 SMD (95% CI -1.74 to -0.86; P < 0.001; I^2 = 77%) compared to that of larger trials of -0.40 SMD (95% CI -0.60 to -0.19; P < 0.001; I^2 = 76%) (test of sub-group difference, P < 0.001). Trials of short duration of intervention (less than 10 weeks) had a SMD of -0.93 (95% CI -1.11 to -0.88; P < 0.001; I^2 = 19%) compared to trials with longer duration of intervention, -0.58 SMD (95% CI -0.88 to -0.28; P < 0.001; I^2 = 86%) (test of sub-group difference, P = 0.05). Effect estimates from trials including patients with minor depression compared to trials exclusively including patient with major depression did not differ (test of sub-group difference, P = 0.67).

Four trials allocated 206 patients to different exercise intensities/doses.[43;56;71;76] Comparing the post-intervention depression scores for patients allocated to either high intensity/high dose or low intensity/low dose exercise showed a difference of -0.40 SMD (95% CI -0.67 to -0.12; P=0.005; I^2 = 0%) in favour of high intensity/high dose exercise. As shown in Table 3, no other trial characteristic significantly explained any of the observed heterogeneity. Please see Table S2 for trial characteristics used to explore heterogeneity.

Trial Sequential Analysis and diversity adjusted required information size

The diversity adjusted required information size for HAM-D_{17} as a continuous outcome was calculated based on our anticipated intervention effect of a minimal relevant difference of 3.0 HDRS points, a standard

deviation of 6.78 points, a risk of type I error of 0.05, a power of 90% and the observed diversity of 92% to 2610 participants. Only 14 trials reported results from HAM-D_{17}[20;21;36;37;41;42;50;51;53;54;56;66;68;76] with an accrued 1124 participants. As shown in Figure S2, the cumulative Z-curve just crossed the trial sequential monitoring boundary for benefit. With the aforementioned settings, the pooled estimate is therefore less likely to be a random finding due to lack of power or multiple testing if bias could be ignored.

Bayes factor

Fourteen trials reported effect estimates using the HAM-D_{17}.[20;21;36;37;41;43;50;51;53;61;66;68;76;77] Based on these trials, Bayes factor was calculated (δ = -3.37; SE_δ = 0.96; μ_a = -3.0) and was found to be 0.002, which is below the Bayes factor threshold for significance of 0.1, supporting the intervention effect if bias could be ignored.

Publication bias

Inspection of the funnel-plot (not shown) suggested that small trials with small or no effect of exercise were missing. Egger's test supported the suspicion of publication bias, P < 0.00001. Using the Duval and Tweedie's trim and fill procedure, the estimate was reduced to -0.28 SMD (95% CI -0.52 to -0.04). This corresponds to an effect on the HAM-D_{17} scale of -1.8 (95% CI -3.2 to -0.25).

The effect of exercise on depression – lack of remission

Nineteen trials, randomising 1825 patients and including 1639 patients (90%) in final analysis reported remission as an outcome.[20;21;36-38;41;43;45;47;51;52;54;58;59;63;66-68;70] The RR for lack of remission was 0.78 (95% CI

0.68 to 0.90; P=0.0008) in favour of the intervention using a random-effects analysis. The I^2 was 69% suggesting substantial heterogeneity. The forest plot for the intervention effect on lack of remission is illustrated in Figure S3.

Missing data

The scenario in least favour of the intervention was the 'poor' outcome analysis having an effect estimate of RR 0.88 (95% CI 0.83 to 0.94) P=0.0002; I^2 = 69%. As shown in Table S1, the remaining scenarios did not substantially differ from the main analysis.

Heterogeneity and subgroup analysis

I^2 was 69% for the outcome lack of remission suggesting substantial heterogeneity. For this outcome, only two trials[21;77] were considered as trials potentially having less risk of bias than the other trials at high risk of bias. The RR of these two trials was 0.95 (95% CI 0.74 to 1.23; P=0.78) compared to 0.77 (96% CI 0.64 to 0.92; P=0.003) for trials at high risk of bias, test of subgroup difference, P=0.19). Trials including 52 participants or less in their final analysis had a RR 0.62 (95% CI 0.50 to 0.76; P<0.001; I^2 = 45%) compared to 0.95 (95% CI 0.80 to 1.12; P=0.52; I^2 = 68%) for larger trials (test of sub-group difference, P=0.002). Also, trials with a duration of less than 10 weeks had a RR of 0.63 (95% CI 0.51 to 0.77; P<0.001; I^2 = 40%) compared to 0.93 (95% CI 0.78 to 1.10; P=0.39; I^2 = 69%) for trials of a longer duration (test of sub-group difference, P=0.004). As shown in Table S3, no other trial characteristic significantly explained any of the observed heterogeneity. Please see Table S2 for trial characteristics used to explore heterogeneity.

Trial Sequential Analysis and diversity adjusted required information size

The diversity adjusted required information size for lack of remission was calculated based on our observed diversity of 74%, a proportion in the control group with lack of remission of 66%, an anticipated intervention effect of 15% relative risk reduction, a risk of type I error of 0.05% and a power of 90%. As shown in Figure S4, the cumulative Z curve just crossed the trial sequential monitoring boundary for benefit. With the aforementioned settings, the pooled estimate is therefore less likely to be a random finding due to lack of power or multiple testing if bias could be ignored.

Bayes factor

Bayes factor was calculate based on the observed relative risk of remission, the associated standard error, and an anticipated intervention effect of relative increase in number of patients with remission by 15% (δ = -0.248; SE_δ= 0.08; μ_δ = -0.163). Bayes factor was 0.02, which is below the Bayes factor threshold for significance of 0.1.

Publication bias

Inspection of the funnel-plot (not shown) suggested that small trials with small or no effect of exercise were missing. Egger's test supported the suspicion of publication bias, P=0.002. Imputing theoretically missing studies by the Duval and Tweedie's trim and fill procedure, reduced the estimate of intervention effect to a relative risk reduction of 0.93 (95% CI 0.79 to 1.11).

The effect of exercise on serious adverse events

Serious adverse events (i.e., death or suicide attempts) were reported in only three trials.[20;21;56] In these trials, one suicide attempt[21] and one death by suicide[20] were recorded in the intervention groups. The RR for death or suicide in the two trials was 2.21 (95% CI 0.24 to 20.21; P=0.48; I^2 = 0%) as illustrated in Figure S5.

Missing data

Missing outcome analysis for 'serious adverse events' varied according to missing data scenario: poor outcome analysis relative risk, 0.92 (95% CI 0.37 to 2.30; P=0.86; I^2 = 60.0%), good outcome analysis, 2.19 (95% CI 0.23 to 20.76; P=0.50; I^2 = 0.0%), best/worst outcome analysis – 0.08 (95% CI 0.02 to 0.34; P=0.001; I^2 = 5.4%), worst/best outcome analysis 19.17 (95% CI 2.64 to 139.2; P=0.004; I^2 = 0.0%).

Trial Sequential Analysis and Bayes analysis

We decided not to conduct Trial Sequential Analysis or Bayes analysis due to too sparse data.

Publication bias

Only 3/31 trials reported on this outcome and no formal assessment for publication bias was made. However, the lack of reporting in the vast majority of trials suggest risk publication bias.

Secondary outcomes

The effect of exercise on quality of life

Eight trials randomising 901 participants reported on quality of life,[20;21;36;38;54;58;69;78] observing that patients allocated to exercise did not have significantly better quality of life (SMD 0.43; 95% CI -0.04 to 0.91; P=0.08). The I^2 was 89% showing substantial heterogeneity (Figure S6).

Non-serious adverse events

Non-serious adverse events were reported in only nine trials.[20;21;37;54;56;58;63;65;66] Five trials reported on musculoskeletal adverse events without conducting formal tests[56;58;63;65;66] and four trials reported on number of patients with high depression scores post-intervention compared to baseline assessment.[20;21;63;66] The RR for increased severity of depression post-intervention was 0.83 (95% CI 0.40 to 1.70; P=0.60; I^2 = 0.0%).

The effect of exercise on depression beyond the duration of the intervention

Assessment of depression beyond the intervention was conducted in seven trials,[20;36;38;50;58;61;79] with a median duration between end of intervention and assessment of depression of 6 months (range 5 to 23.5 months). The SMD between the intervention group and the control group using a random effects analysis was -0.10 (95% CI -0.28 to 0.09; P=0.31; I^2 = 19.5%). The I^2 for this estimate was 19.5% suggesting low heterogeneity (See Figure S7).

Remission beyond the intervention was assessed in five trials,[20;36-38;52] and the relative risk of lack of remission was 0.95 (95% CI 0.82 to 1.11; P=0.53) with an I^2 of 0.0% (See Figure S8).

GRADE assessments

The GRADE assessments are presented in Table 4, and quality of evidence for both primary and secondary outcomes was very low or low.

Discussion

Thirty-one clinical trials allocating more than 2400 participants diagnosed with depression according to validated diagnostic instruments were included in the present systematic review. Pooled estimates suggested moderate antidepressant effect assessed both as a continuous outcome and as lack of remission. Due to risk of bias, inconsistency of effect estimates, and publication bias we have, however, very little confidence in these effect estimates. Subgroup analyses exploring reasons for the heterogeneity found that trials potentially having less risk of bias than other trials at high risk of bias had no effect of exercise on depression. Furthermore, duration of intervention and trial size were inversely associated with effect estimates. Exercise did not improve quality of life or depression or remission after the intervention. Serious adverse event or adverse events were reported inconsistently and only by a few trials not permitting firm conclusions regarding these outcomes.

Strengths and limitations

The strengths of this systematic review are that it is based on the published protocol, a comprehensive search strategy, and the inclusion of patient centered outcomes such as quality of life as well as adverse events. Also, to avoid spurious finding from repeated testing, Trial Sequential Analysis and Bayes analysis were undertaken and these analyses did not suggest that the pooled estimates could be reduced to random errors for effect on depression severity or no remission. Neither Trial Sequential Analysis nor Bayes factor analysis are, however, able to wash of spurious effects induced by bias, or fraud or other reasons.[25;27;80-82] Had we restricted the Trial Sequential Analysis to trials of potentially lower risk of bias, the

number of trials and participants would be limited and we had seen evidence far from crossing boundaries for benefit, harms, or futility. The conclusions for serious adverse events and adverse events were associated with wide confidence intervals due to lack of data and firm conclusions for these outcomes are presently not available.

The number of trials with adequate allocation concealment was 39% in the current systematic review compared to only 15.1% in trials assessing non-drug interventions for depression.[83] Blinded outcome assessment was performed in 52% of the included trials compared to 44% in non-drug antidepressant trials in general.[83] The incomplete outcome bias domain was adequate in 48% of our included trials compared to 32.9% of antidepressant non-drug trials in general.[83] Compared to non-drug trials assessing interventions for patients with depression, the included exercise trials have more bias domains with low risk of bias. However, all our included trials were at high risk of bias. Two trials had low risk of bias for all bias domains except for blinding of participants and trial personnel, and four trials fulfilled our criteria for trials at potentially less risk of bias than the rest of the trials with at risk of bias. Despite a search strategy including bibliographical databases and trials from China and South-America, the vast majority of included trials were conducted in north America and western Europe, which is comparable to the geographical distribution of non-drug trials in general[83] limiting the applicability to other geographic regions.

The effect of exercise on depression

Our present results are similar to the latest Cochrane review by Cooney et al. (2013)[23] who found a moderate effect of exercise on depressive symptoms (-0.62 SMD) when including all trials and no effect when restricting the analysis to trials with less risk of bias (-0.18 SMD). The Cochrane review did find evidence of a small antidepressant effect beyond the intervention, which we could not confirm in our present systematic review. Bridle et al. (2012)[12] included 9 trials allocating old (> 60 years) patients with

depression to exercise interventions versus control interventions. Restricting the analysis to four trials at lower risk of bias they found small to moderate effect estimates (SMD -0.34) in favour of exercise. The studies by Cooney et al.[23] and Bridle et al.[12] both included trials allocating patient with depressive symptoms and not necessarily diagnosed using a validated diagnostic system, potentially explaining the differences in the effect sizes. However, in our present systematic review the estimate for four trials at potential less risk of bias than the remaining trials was -0.11 SMD and in the Cooney study the effect estimate for eight trials with lower risk of bias was -0.18 SMD[23] compared to -0.34 in the study by Bridle at al.[12] Meta-analysis of randomised clinical trials assessing the effects of exercise for depression consistently finds positive effects, however, when restricting the analysis to trials with less risk of bias the pooled effect sizes becomes very small or negligible. Meta-analysis examining the effect of exercise beyond the intervention also finds no or small effects of exercise. In the process of interpretation of effect estimates in the current research field, it is important to recognise that effect estimates from trials with non-blinded outcome assessment are at high risk of bias as reported by Savovic et al.[84] Thirteen of 31 trials in the current systematic review did not use blinded outcome assessment. In contradiction to the current systematic review, a recent meta-analysis by Schuch et al.[11] concluded that "exercise has a large and significant antidepressant effect in people with depression……Our data strongly support the claim that exercise is an evidence-based treatment for depression". This statement was based on a meta-analysis of 25 randomised clinical trials including patients with depression or depressive symptoms to exercise or control conditions and excluding trials using any form of active control group. Surprisingly, the authors found that adjusting for publication bias using the Trim and Fill procedure[29] the estimate *increased* from a SMD of 0.98 to 1.11. The effect in SMD in included studies ranged from -0.23 to 4.56 representing considerable heterogeneity.[11] The authors classified four trials as having lower risk of bias using the same criteria as in our systematic review and 21 trials as having high risk of bias. This illustrates some of the challenges in meta-analysis of exercise and depression: the large heterogeneity driven by small studies inflating the effects of random-effects analysis,[85] the misconception that we can restrict our analysis to

statistics and not consider the evident effect of bias.[22;84] We therefore recommend that future systematic reviews and meta-analysis a priori should have a primary outcome restricting effect analysis to larger trials with lower risk of bias and that any recommendations regarding exercise interventions for patients with depression should be assessed with the GRADE framework.

The I^2 of 82% and 71% for the primary outcomes indicate substantial evidence of heterogeneity of intervention effects that is variation in effect estimates beyond chance. Part of this heterogeneity was explained by bias and by trial size: trials at high risk of bias or small trials have very large effect estimates compared to trials potentially at less bias risk compared to the remaining trials at high risk of bias or larger trials. The funnel plots end Egger's test indicates publication bias, however, the association between trial size and effect estimates could suggest that the asymmetry in the funnel plots are due to small study bias rather than publication bias.[86] In addition, in line with our previous review we found duration of intervention inversely associated with effect size.[10] A number of studies compare exercise to control interventions rather than wait-list control to reduce the effect of non-specific effects, e.g., the DEMO trials and Mather et al.[20;21;50] Also, it could be speculated that the effect of exercise would be harder to detect if patients also received medical treatment in addition. The current systematic review could not confirm that the type of control condition explained heterogeneity. The discussion of control group is important in non-drug trials: choosing a waitlist control group the results potentially reflects non-specific effects, choosing an active control group (e.g., relaxation exercise) the trial is potentially a comparison between to active treatments. However, in the current systematic review we found no evidence that trials using an attention control group or exercise as add-on to pharmacotherapy had significantly different effect estimates compared to other trials.

Our systematic review did not find indications of a positive effect on quality of life in patients with depression allocated to exercise interventions, which is in concordance with the review by Cooney et al.[23] Only 3/31 trials reported on serious adverse events, and found no significant risk of death or suicide attempt. No indication of increased severity of depression or other adverse events in participants allocated to exercise could be detected. However, data on adverse events was reported sporadically in a minority of trials and currently it is not possible to conclude on the risk of serious adverse events or adverse event from exercise interventions in patients with depression.

Conclusions

We have little confidence in the pooled effect estimates, especially because trials with less than high risk of bias produced significantly lower effect estimates, suggesting that exercise interventions only produce small or negligible antidepressant effects, depending on how much of the effect is caused by bias and how much is caused by the intervention. There was no effect of exercise on quality of life or depression beyond the intervention itself. There is currently no evidence in favour of exercise for patients with depression with a view to ameliorate depressive symptoms and at we do not recommend that exercise is prescribed to relieve depressive symptoms. Our systematic review did not evaluate possible beneficial effects of exercise on, e.g., metabolism or cardiovascular fitness,[21;87] and it is possible that exercise may have beneficial effects on these factors in patients diagnosed with depression.

Future perspectives

Despite the large number of published trials, further trials with more robust methodology seem still required to establish progress in this field. Also, additional trials from outside North-America and Europe may be required for results to be valid for patients in Asia, Africa, and South-America. To further elaborate

on the current findings, we recommend that future trials must include blinded outcome assessors and outcomes assessing quality of life, metabolic effects, and long-term effects beyond the intervention. It is also important that future trials systematically collect and report data on death, suicide events, musculoskeletal injuries and other potential adverse effects in both the intervention group as well as in the control group. Moreover, future trials ought to be designed according to the SPIRIT guidelines and reported according to the CONSORT guidelines[88;89] and transparently report deidentified individual patient data enabling individual patient data meta-analyses.[90]

Funding

No funding was obtained for this systematic review.

Competing interests

JK, CG, and MN have previously published two trials and a meta-analysis on this topic, which could introduce an academic bias in the current systematic review. We asked new authors to be involved in the preparation of the protocol, trial selection and bias assessment. No support from any organisation was received for the submitted work; no financial relationship with any organisations that might have an interest in the submitted work in the previous three years, and apart from the above no other relationship or activities that could appear to have influenced the submitted work.

Acknowledgement

We appreciate the help from Youling He for help with the Chinese Wanfang bibliographical database and translation of Chinese papers. We thank Janus C Jakobsen for assistance with the calculation of Bayes factor.

Contributors

JK conceived the project, collected data, did the statistical analysis, analysed the data, drafted and revised the manuscript. He is guarantor. CH collected the data, analysed the data and revised the manuscript. HS conceived the project, collected data, analysed the data, and revised the manuscript. CG conceived the project, analysed the data and revised the manuscript.MN conceived the project, analysed the data, and revised the manuscript.

References

(1) Lepine JP, Gastpar M, Mendelwicz J, Tylee A. Depression in the community: The first pan-european study DEPRES (Depression Research in European Society). *Int Clin Psychopharmocol* 1997; 12:19-29.

(2) Pedersen CB, Mors O, Bertelsen A, Waltoft BL, et al. A comprehensive nationwide study of the incidence rate and lifetime risk for treated mental disorders. *JAMA Psychiatry* 2014; 71(5):573-581.

(3) Ustun TB, Ayuso-Mateos JL, Chatterji S, Mathers C, Murray CJL. Global burden of depressive disorders in the year 2000. *Br J Psychiatry* 2004; 184:386-392.

(4) Kirsch I, Deacon B, Huedo-Medina T, Scoboria A, Moore T, Johnson B. Initial severity and antidepressant benefits: A meta-analysis of data submitted to the Food and Drug Administration. *PLoS Med* 2008; 5[2], e45. doi:10.1371/journal.pmed.0050045.

(5) Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy. *N Engl J Med* 2008; 358(3):252-260.

(6) Jakobsen JC, Hansen JL, Storebo OJ, Simonsen E, Gluud C. The effects of cognitive therapy versus 'no intervention' for major depressive disorder. *PLoS One* 2011; 6(12):e28299.

(7) Jakobsen JC, Hansen JL, Simonsen E, Gluud C. The effect of interpersonal psychotherapy and other psychodynamic therapies versus 'treatment as usual' in patients with major depressive disorder. *PLoS One* 2011; 6(4):e19044.

(8) Jakobsen JC, Hansen JL, Simonsen E, Gluud C. The effect of adding psychodynamic therapy to antidepressants in patients with major depressive disorder. A systematic review of randomized clinical trials with meta-analyses and trial sequential analyses. *J Affect Disord* 2012; 137(1-3):4-14.

(9) Cipriani A, Furukawa TA, Salanti G, Geddes JR, Higgins JP, Churchill R et al. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *Lancet* 2009; 373[9665]: 746-758.

(10) Krogh J, Nordentoft M, Sterne J, Lawlor D. The effect of exercise in clinically depressed adults: Systematic review and meta-analysis of randomized controlled trials. *J Clin Psychiatry* 2011; 72(4): 529-538.

(11) Schuch FB, Vancampfort D, Richards J, Rosenbaum S, Ward PB, Stubbs B. Exercise as a treatment for depression: A meta-analysis adjusting for publication bias. *J Psychiatr Res* 2016; 77:42-51.

(12) Bridle C, Spanjers K, Patel S, Atherton NM, Lamb SE. Effect of exercise on depression severity in older people: systematic review and meta-analysis of randomised controlled trials. *Br J Psychiatry* 2012; 201(3):180-185.

(13) Herring MP, Puetz TW, O'Connor PJ, Dishman RK. Effect of exercise training on depressive symptoms among patients with a chronic illness: a systematic review and meta-analysis of randomized controlled trials. *Arch Intern Med* 2012; 172(2):101-111.

(14) World Health Organisation. International Statistical Classification of Diseases, 10th Revision (ICD-10). Geneva, Switzerland, 1992.

(15) American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, 4th ed. Washington DC, 1994.

(16) Beck AT, Steer RA, Brown GK. BDI-II manual. 2 ed. New York: Psychological Corporation, 1996.

(17) Krogh J, Speyer H, Gluud C, Nordentoft M. Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential analysis. *Syst Rev* 2015; 4(1):40.

(18) Spitzer RL, Endicott J, Williams JB. Research diagnostic criteria. *Arch Gen Psychiatry* 1979; 36(12):1381-1383.

(19) ICH-secretariat. International Conference on Harmonisation (ICH) Harmonised Tripartite Guideline (Step 4). Guideline for Good Clinical Practice. Geneva, 1996.

(20) Krogh J, Saltin B, Gluud C, Nordentroft M. The DEMO trial: A randomized, parallel-group, observer-blinded clinical trial of strength versus aerobic versus relaxation training for patients with mild to moderate depression. *J Clin Psychiatry* 2009; 70[6]: 790-800.

(21) Krogh J, Videbech P, Thomsen C, Gluud C, Nordentoft M. DEMO-II trial.Aerobic Exercise Versus Stretching Exercise in Patients With Major Depression - A Randomised Clinical Trial. *PLoS One* 2012; 7(10):e48316. doi: 10.1371/journal.pone.0048316.

(22) Higgins JPT. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0. The Cochrane Collaboration. Available from www.cochrane-handbook.org The Cochrane Collaboration, 2009.

(23) Cooney GM, Dwan K, Greig CA, Lawlor DA, Rimer J, Waugh FR et al. Exercise for depression. *Cochrane Database Syst Rev* 2013; 9:CD004366.

(24) Higgins JPT, Thompson SG, Deeks JJ, Altmann DG. Measuring inconsistency in meta-analysis. *BMJ* 2003; 327:557-560.

(25) Wetterslev J, Thorlund K, Brok J, Gluud C. Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. *J Clin Epidemiol* 2008; 61(1):64-75.

(26) Brok J, Thorlund K, Gluud C, Wetterslev J. Trial sequential analysis reveals insufficient information size and potentially false positive results in many meta-analyses. *J Clin Epidemiol* 2008; 61(8):763-769.

(27) Jakobsen JC, Wetterslev J, Winkel P, Lange T, Gluud C. Thresholds for statistical and clinical significance in systematic reviews with meta-analytic methods. *BMC Med Res Methodol* 2014; 14:120.

(28) Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. *BMJ* 1999; 319(7211):670-674.

(29) Duval S, Tweedie R. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* 2000; 56(2):455-463.

(30) Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008; 336(7650):924-926.

(31) Andrews J, Guyatt G, Oxman AD, Alderson P, Dahm P, Falck-Ytter Y et al. GRADE guidelines: 14. Going from evidence to recommendations: the significance and presentation of recommendations. *J Clin Epidemiol* 2013; 66(7):719-725.

(32) Zheng YP, Lin KM, Zhao JP, Zhang MY, Yong D. Comparative study of diagnostic systems: Chinese Classification of Mental Disorders-Second Edition versus DSM-III-R. *Compr Psychiatry* 1994; 35(6):441-449.

(33) Babyak MA, Blumenthal JA, Herman S, Khatri P, Doraiswamy M, Moore K. Exercise treatment for major depression: Maintenance of therapeutic benefit at 10 months. *Psychosom Med* 2000; 62:633-638.

(34) Baxter H, Winder R, Chalder M, Wright C, Sherlock S, Haase A et al. Physical activity as a treatment for depression: the TREAD randomised trial protocol. *Trials* 2010; 11:105.

(35) Benson M, Blumenthal J, Babyak M, Smith P, Rogers S, Doraiswamy M et al. Exercise fails to improve neurocognition in depressed middle-aged and older adults. *Med Sci Sports Exerc* 2008; 40[7]: 1344-1352.

(36) Blumenthal J, Babyak MA, Moore KA, Craighead WE, Herman S, Khatri P. Effects of exercise training on older patients with major depression. *Arch Intern Med* 1999; 159:2349-2356.

(37) Blumenthal J, Babyak M, Doraiswamy P, Watkins L, Hoffman B, Barbour K et al. Exercise and pharmacotherapy in the treatment of major depressive disorder. *Psychosom Med* 2007; 69: 587-596.

(38) Chalder M, Wiles NJ, Campbell J, Hollinghurst SP, Haase AM, Taylor AH et al. Facilitated physical activity as a treatment for depressed adults: randomised controlled trial. *BMJ* 2012; 344:e2758.

(39) Chalder M, Wiles NJ, Campbell J, Hollinghurst SP, Searle A, Haase AM et al. A pragmatic randomised controlled trial to evaluate the cost-effectiveness of a physical activity intervention as a treatment for depression: the treating depression with physical activity (TREAD) trial. *Health Technol Assess* 2012; 16(10):1-iv.

(40) Combs K, Smith PJ, Sherwood A, Hoffman B, Carney RM, Freedland K et al. Impact of sleep complaints and depression outcomes among participants in the standard medical intervention and long-term exercise study of exercise and pharmacotherapy for depression. *J Nerv Ment Dis* 2014; 202: 167-171.

(41) Doyne EJ, Ossip-Klein DJ, Bowman ED, Osborn KM, McDougall-Wilson IB, Neimeyer RA. Running versus weight lifting in the treatment of depression. *J Consult Clin Psychol* 1987; 55(5):748-754.

(42) Dunn AL, Trivedi MH, Kampert JB, Clark CG, Chambliss HO. The DOSE study: a clinical trial to examine efficacy and dose response of exercise as treatment for depression. *Control Clin Trials* 2002; 23(5):584-603.

(43) Dunn AL, Trivedi MH, Kaplan K, Clark C, Chambliss H. Exercise treatment for depression. Efficacy and dose response. *Am J Prev Med* 2005; 28: 1-8.

(44) Herman S, Blumenthal JA, Babyak M, Khatri P, Craighead WE, Krishnan KR et al. Exercise therapy for depression in middle-aged and older adults: predictors of early dropout and treatment failure. *Health Psychol* 2002; 21:553-563.

(45) Ho CWH, Chan SC, Wong JS, Cheung WT, Chung DWS, Lau TFO. Effect of aerobic exercise training on chinese population with mild to moderate depression in Hong Kong. *Rehabil Res Pract* 2014;2014:627376. doi: 10.1155/2014/627376

(46) Hoffman BM, Babyak MA, Craighead WE, Sherwood A, Doraiswamy PM, Coons MJ et al. Exercise and pharmacotherapy in patients with major depression: one-year follow-up of the SMILE study. *Psychosom Med* 2011; 73(2):127-133.

(47) Kerling A, Tegtbur U, Gutzlaff E, Kuck M, Borchert L, Ates Z et al. Effects of adjunctive exercise on physiological and psychological parameters in depression: A randomized pilot trial. *J Affect Disord* 2015: 177:1-6.

(48) Khatri P, Blumenthal J, Babyak M, Craighead E, Herman S, Baldewicz T et al. Effects of exercise training on cognitive functioning among depressed older men and women. *J Aging Phys Act* 2001; 9: 43-57.

(49) Krogh J, Petersen L, Timmermann M, Saltin B, Nordentoft M. Design paper: The DEMO trial: A randomized, parallel-group, observer-blinded clinical trail of aerobic versus non-aerobic versus relaxation training for patients with light to moderate depression. *Contem Clin Trials* 2007; 28:79-89.

(50) Mather AS, Rodriguez C, Guthrie MF, McHarg AM, Reid IC, McMurdo ME. Effects of exercise on depressive symptoms in older adults with poorly responsive depressive disorder: randomised controlled trial. *Br J Psychiatry* 2002; 180:411-415.

(51) Mota-Pereira J, Silverio J, Carvalho S, Ribeiro JC, Fonte D, Ramos J. Moderate exercise improves depression parameters in treatment-resistant patients with major depressive disorder. *J Psychiatr Res* 2011; 45(8):1005-1011.

(52) Pfaff JJ, Alfonso H, Newton RU, Sim M, Flicker L, Almeida OP. ACTIVEDEP: a randomised, controlled trial of a home-based exercise intervention to alleviate depression in middle-aged and older adults. *Br J Sports Med* 2014; 48(3):226-232.

(53) Pilu A, Sorba M, Hardoy M, Floris A, Mannu F, Seruis M et al. Efficacy of physical activity in the adjunctive treatment of major depressive disorders: preliminary results. *Clin Pract Epidemiol Ment Health* 2007; 3(8).

(54) Schuch FB, Vasconcelos-Moreno MP, Borowsky C, Fleck MP. Exercise and severe depression: preliminary results of an add-on study. *J Affect Disord* 2011; 133(3):615-618.

(55) Singh NA, Clements KM, Fiatarone MA. A randomized controlled trial of the effect of exercise on sleep. *Sleep* 1997; 20(2):95-101.

(56) Singh NA, Stavrinos TM, Scarbek Y, Galambos G, Liber C, Fiatarone Singh MA. A randomized controlled trial of high versus low intensity weight training versus general practitioner care for clinical depression in older adults. *J Gerontol A Biol Sci Med Sci* 2005; 60(6):768-776.

(57) Singh NA, Clements KM, Singh MA. The efficacy of exercise as a long-term antidepressant in elderly subjects: a randomized, controlled trial. *J Gerontol A Biol Sci Med Sci* 2001; 56(8):M497-M504.

(58) Singh N, Clements K, Fiatarone M. A randomized controlled trial of progressive resistance training in depressed elders. *J Gerontol* 1997; 52:M27-M35.

(59) Toni G, Belvederi MM, Zanetidou S, Menchetti M, Ermini G, Tripi F et al. Physical exercise for late life Major Depression: The SEEDS study. *Eur J Prev Cardiol* 2014; 21:S107.

(60) Veale D, Le FK, Pantelis C, de S, V, Mann A, Sargeant A. Aerobic exercise in the adjunctive treatment of depression: a randomized controlled trial. *J R Soc Med* 1992; 85(9):541-544.

(61) Vieira JLL, Porcu M, da Rocha PGM. The regular physical exercise's practice as complementary therapy to the treatment of depressed women. *J Bras Psiquiatr* 2007; 56: 23-28.

(62) Epstein D. Reno, Dissertation: University of Nevada:Aerobic activity versus group cognitive therapy: An evaluative study of contrasting interventions for the alleviation of clinical depression. 1986.

(63) Danielsson L, Papoulias I, Petersson E, Carlsson J, Waern M. Exercise or basic body awareness therapy as add-on treatment for major depression: A controlled study. *J Affect Disord* 2014; 168:98-106.

(64) Carneiro L, Fonseca A, Vieira-Coelho M, Mota M, Vasconcelas-Raposo J. Effects of structured exercise and pharmacotherapy vs pharmactherapy for adults with depressive symptoms: A randomized clinical trial. *J Psych Res* 2015;(71):48-55.

(65) Legrand FD, Neff EM. Efficacy of exercise as an adjunct treatment for clinically depressed inpatients during the initial stages of antidepressant pharmacotherapy: An open randomized controlled trial. *J Affect Disord* 2016; 191:139-144.

(66) Doose M, Ziegenbein M, Hoos O, Reim D, Stengert W, Hoffer N et al. Self-selected intensity exercise in the treatment of major depression: A pragmatic RCT. *Int J Psychiatry Clin Pract* 2015; 19(4):266-275.

(67) Salehi I, Hosseini SM, Haghighi M, Jahangard L, Bajoghli H, Gerber M et al. Electroconvulsive therapy (ECT) and aerobic exercise training (AET) increased plasma BDNF and ameliorated depressive symptoms in patients suffering from major depressive disorder. *J Psychiatr Res* 2016; 76:1-8.

(68) Huipeng F, Xiaohui Z. Curative effect and observation of exercise on patients with depression. *Chin Med Pharm* 2013; 3:187-188.

(69) Jinchun W, Bo Y, Rong X, Chengfeng L, Yanan S. Effect of exercise therapy in improving social function and quality of life in patients with depression. *Nurs J Chin PLA* 2015; 32(21):24.

(70) Guifeng S. The role of setting - up exercise on rehabilitation in patients with depressive episode. *Chin J Health Psychol* 2015; 23:184-186.

(71) Fang D, Min R, Ai-qin W. The effect of aerobic exercise with different intensity and frequency on depressive symptoms. *Chin J Behav Med & Brain Sci* 2013; 22:830-832.

(72) Martinsen EW, Medhus A, Sandvik L. Effects of aerobic exercise on depression: A controlled study. *BMJ* 1985; 291:109.

(73) Veale D, Fevre K, Pantalis C, Souza V, Mann A, Sargeant A. Aerobic exercise in the adjunctive treatment of depression: A randomised controlled trial. *J R Soc Med* 1992; 85:541-544.

(74) Ossip-Klein DJ, Doyne EJ, Bowman ED, Osborn KM, Dougall-Wilson IB, Neimeyer RA. Effects of running or weight lifting on self-concept in clinically depressed women. *J Consult Clin Psychol* 1989; 57(1):158-161.

(75) Blumenthal JA, Babyak MA, Moore KA, Craighead WE, Herman S, Khatri P et al. Effects of exercise training on older patients with major depression. *Arch Intern Med* 1999; 159(19):2349-2356.

(76) Belvederi MM, Amore M, Menchetti M, Toni G, Neviani F, Cerri M et al. Physical exercise for late-life major depression. *Br J Psychiatry* 2015; 207(3):235-242.

(77) Schuch FB, Vasconcelos-Moreno MP, Borowsky C, Zimmermann AB, Rocha NS, Fleck MP. Exercise and severe major depression: effect on symptom severity and quality of life at discharge in an inpatient cohort. *J Psychiatr Res* 2015; 61:25-32.

(78) Carta MG, Hardoy MC, Pilu A, Sorba M, Floris AL, Mannu FA et al. Improving physical quality of life with group physical activity in the adjunctive treatment of major depressive disorder. *Clin Pract Epidemiol Ment Health* 2008; 4:1

(79) Klein MH, Greist JH, Gurman AS, Neimeyer R, Lesser D, Bushnell N et al. A comparative outcome study of group psychotherapy vs. exercise treatments for depression. *Int J Ment Health* 1985; 13:148-177.

(80) Roberts I, Ker K, Edwards P, Beecher D, Manno D, Sydenham E. The knowledge system underpinning healthcare is not fit for purpose and must change. *BMJ* 2015; 350:h2463.

(81) Gluud C, Jakobsen J, Imberger G, Lange T, Wetterslev J. Re: The knowledge system underpinning healthcare is not fit for purpose and must change - Responses to the opposing viewpoints of Roberts and colleagues and Tovey and colleagues. *BMJ* 2015; 350(h2463).

(82) Thorlund K, Engstrøm J, Wetterslev J, Brok J, Imberger G, Gluud C. User manual for Trial Sequential Analysis (TSA). Copenhagen Trial Unit ,Centre for Clinical Intervention Research, Copenhagen, Available from https://www.ctu.dk/tsa, 2011.

(83) Krogh J, Hjorthoj CR, Jakobsen JC, Lindschou J, Kessing LV, Nordentoft M et al. DEPERROR: Risks of systematic errors in drug and non-drug randomized clinical trials assessing intervention effects in patients with unipolar depression. *J Affect Disord* 2015; 179:121-127.

(84) Savovic J, Jones HE, Altman DG, Harris RJ, Juni P, Pildal J et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. *Ann Intern Med* 2012; 157(6):429-438.

(85) Poole C, Greenland S. Random-effects meta-analyses are not always conservative. *Am J Epidemiol* 1999; 150(5):469-475.

(86) Sterne JA, Sutton AJ, Ioannidis JP, Terrin N, Jones DR, Lau J et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ* 2011; 343:d4002.

(87) Stubbs B, Rosenbaum S, Vancampfort D, Ward PB, Schuch FB. Exercise improves cardiorespiratory fitness in people with depression: A meta-analysis of randomized control trials. *J Affect Disord* 2016; 190:249-253.

(88) Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. *Ann Intern Med* 2010; 152(11):726-732.

(89) Chan AW, Tetzlaff JM, Gotzsche PC, Altman DG, Mann H, Berlin JA et al. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. *BMJ* 2013; 346:e7586.

(90) Skoog M, Saarimäki J, Gluud C, Sheinin M, Erlandsson K, Aamdal S. Transparency and registration in clinical research in the Nordic countries. Nordic Trial Alliance, NordForsk; 2015: 1-108.

Figure 1. Effect of exercise on depression severity in patients diagnosed with depression

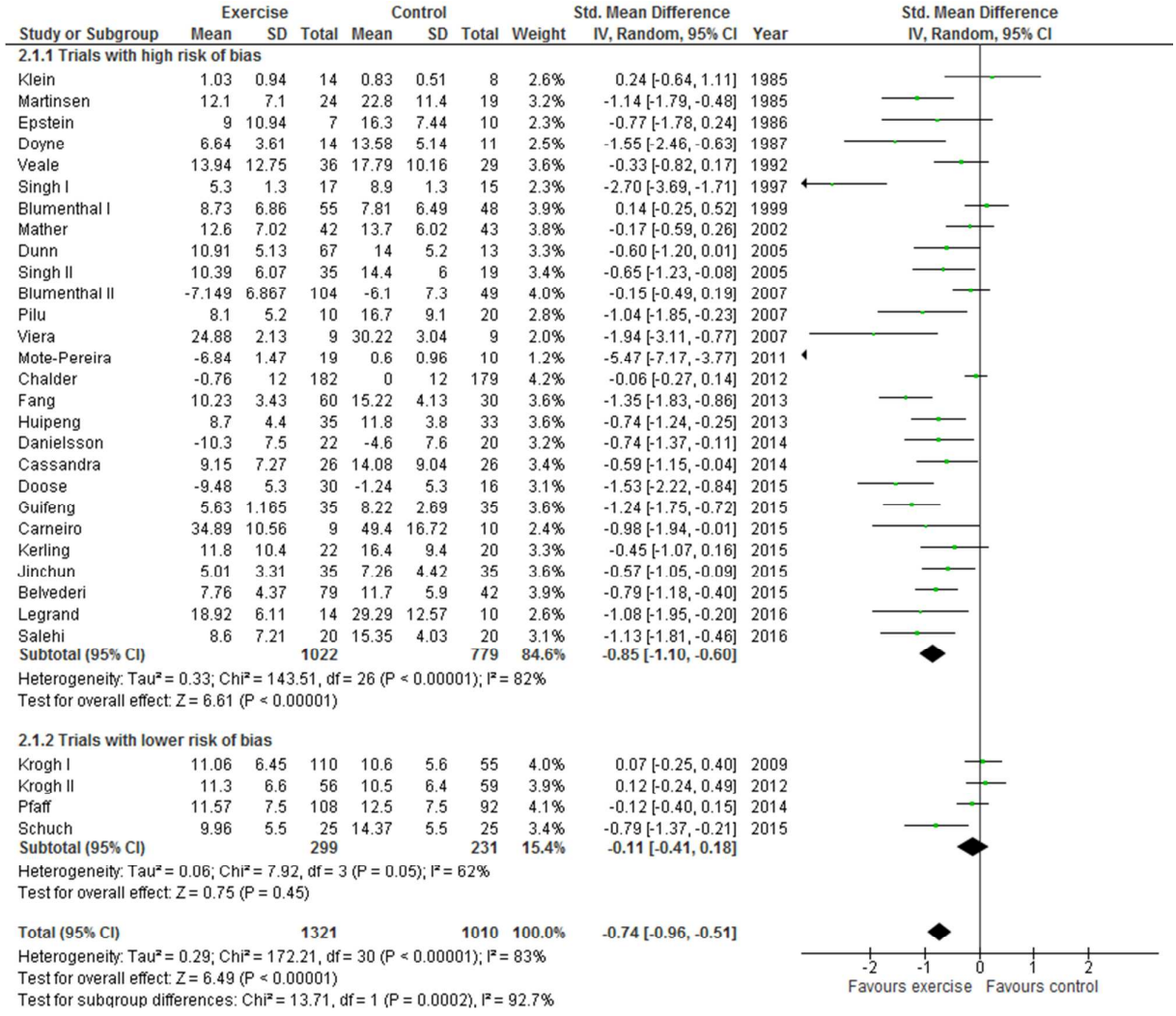


Table 1. Characteristics of trials assessing exercise for patients diagnosed with depression

Author, first Country of origin	Participants	Severity of depression at baseline	N at baseline (included in trial efficacy analysis)	Type of intervention	Frequency	Duration
Klein 1985 USA	Outpatients Mean age: 30 (SD 7) 72% female	SCL-D: 2.4 (SD 1)	50 (22)	*Aerobic exercise:* Supervised individual running. *Control group*: Supervised meditation in groups	2 sessions per week Control group: 1 session per week	12 weeks
Martinsen 1985 Norway	Inpatients Mean age: 40 (range 17-60 Distribution of sex not reported	BDI: 28.0 (SD 9)	49(43)	*Aerobic exercise*: Supervised group exercise. *Control group*: Occupational therapy.	3 sessions per week Control group: 3 sessions per week	9 weeks
Epstein 1986 USA	Outpatients Mean age: 39 (range 24 to 60) (NR) % female	BDI: 23.4 (SD 7)	21 (17)	*Aerobic exercise:* Supervised group exercise. *Control group*: Waitlist control.	3 sessions per week	8 weeks
Doyne 1987 USA	Outpatients Mean age: 29 (SD 4) 100 % female	HAM-D_{17}: 13.0 (SD 7)	52 (25)	*Aerobic exercise OR weightlifting*: Supervised individual exercise. *Control group*: Waiting list.	4 sessions per week	8 weeks
Veale 1992 UK	Outpatients Mean age: 35 (range 19-58) 64% female	BDI: 24.5 (SD 6)	83 (65)	*Aerobic exercise*: Supervised group exercise. *Control group*: Standard treatment from psychiatric services.	3 sessions per week	12 weeks
Singh 1997 USA	Outpatients Recruited from a register of volunteers Mean age: 71 (SD 1)	BDI: 19.9 (SD 2.3)	32 (32)	*Progressive resistance training*: Supervised group exercise. *Control group*: Attended seminars on health.	3 sessions per week Control group: 2 sessions per week	10 weeks
Blumenthal 1999 USA	Outpatients Mean age: 57 (SD 7) 71.8% female	HAM-D_{17}: Not reported	103 (103)	*Aerobic exercise*: Supervised exercise plus antidepressant medication (sertraline). *Control group*: Antidepressant medication (sertraline).	3 sessions per week	16 weeks
Mather 2002 UK	Outpatients Treatment resistant Mean age: 65 (range 53-91) 69% female	HAM-D_{17}: 17.1 (SD 6)	86 (85)	*Mixed aerobic and non-aerobic exercise*: Supervised group exercise. *Control group*: Attended health seminars.	2 sessions per week Control group: 2 seminars per week	10 weeks
Dunn 2005 USA	Outpatients Mean age: 36 (SD 6)	HAM-D_{17}: 19.4 (SD 2)	80 (80)	*Aerobic exercise*: Individually supervised	Group (1) and (2): 3 sessions	12 weeks

	75% female			exercise with (1) low energy expenditure (EE) OR (2) high EE OR (3) low EE OR (4) high EE. *Control group*: Flexibility exercise.	per week Group (3) and (4): 5 sessions per week Control group: 3 sessions per week	
Singh 2005 Australia	Outpatients Mean age: 69 (SD 6) 55% female	HAM-D_{17}: 18.9 (SD 4.2)	60 (54)	*Progressive resistance training (PRT)*: (1)Low intensity PRT OR (2) high intensity PRT. *Control group*: Standard GP care.	Group (1) and (2): 3 sessions per week	8 weeks
Pilu 2007 Italy	Outpatients Treatment resistant Age between 40 and 60 100% female	HAM-D_{17}: 19.7 (SD 6)	30 (30)	*Resistance exercise*: Supervised group sessions. *Control group*: Standard treatment.	2 sessions per week	32 weeks
Viera 2007 Brazil	Outpatients Mean age 43.66 (SD NR) 100% female	HAM-D_{21}: 31.9 (SD 3)	18 (18)	*Aerobic exercise:* Supervised water aerobics. *Control group*: Standard GP care.	2 sessions per week	12 weeks
Blumenthal 2007 USA	Outpatients Mean age: 52 (SD 8) 75.8% female	HAM-D_{17}: 16.7 (SD 4)	153 (153)	*Aerobic exercise*: (1) Supervised group exercise OR (2) home-based exercise. *Control group*: Placebo medication.	(1) and (2): 3 sessions per week	16 weeks
Krogh 2009 Denmark	Outpatients Mean age: 39 (SD 9) 74% female	HAM-D_{17}: 17.8 (SD 4)	165 (165)	*Exercise*: (1) Aerobic supervised group exercise OR (2) supervised group resistance training *Control group*: relaxation and stretching exercise.	(1)and (2): 2 sessions per week Control group: 2 sessions per week	16 weeks
Mota-Pereira 2011 Portugal	Outpatients Treatment resistant Mean age: 47.5 (SD 3) 65.5% female	HAM-D_{17}: 17.1 (SD 3)	33 (29)	*Aerobic exercise:* Homebased exercise + supervised. *Control group*: Attention control.	4 home-based sessions/week. 1 supervised session/week Control group: 1 supervised session/week	12 weeks
Krogh 2012 Denmark	Outpatients Mean age: 42 (SD 11) 67% female	HAM-D_{17}: 18.9 (SD 4)	115 (115)	*Aerobic exercise*: Supervised group exercise. *Control group*: Supervised stretching exercise in groups.	3 sessions per week Control group: 3 sessions per week	12 weeks
Chalder 2012 UK	Outpatients Mean age: 40 (SD 13) 66% female	BDI: 32.1 (SD 9)	361 (361)	*Exercise*: Participants received individually tailored support and encouragement to engage in physical activity. *Control group*: Standard GP care.	Individual	16 weeks

Fang 2013 China	Inpatients Mean age: 44 (SD 14) 66.9% female	HAM-D_{24}: 29.2 (SD 5)	90 (90)	*Aerobic exercise:* Group 1 and 2 had supervised group exercise, high intensity. *Control group:* 15 min stretching	Group 1 and 2 had 3 and 5 sessions per week, respectively Control group: 3 sessions per week	6 weeks
Huipeng 2013 China	Inpatients Mean age: 30 (SD 5) 100% female	HAM-D_{17}: 28 (SD 5)	68 (68)	*Aerobic exercise:* Jogging *Control group:* Standard treatment	5 sessions per week	6 weeks
Cassandra 2014 Honkong	Inpatients Mean age: 46 (SD 12) 67.3% female	MADRS: 19 (10)	52 (52)	*Aerobic exercise:* Supervised exercise. *Control group*: 10 min stretching.	5 sessions per week	3 weeks
Danielsson 2014 Sweden	Outpatients Mean age: 45 (SD 13) 76% female	MADRS: 24.0 (SD 5)	42 (42)	*Mixed aerobic and non-aerobic exercise*: First two weeks individual supervised exercise then supervised group exercise. *Control group*: One session with advice on physical activity.	2 sessions per week	10 weeks
Pfaff 2014 Australia	Outpatients Mean age: 61 (SD 8) 63% female	MADRS: 21.3 (SD NR)	200 (200)	*Resistance exercise*: Supervised home-based exercise *Control group*: Standard GP care	3 sessions per week	12 weeks
Guifeng 2015 China	Inpatients Mean age: 33 (SD 14) 70% female	HAM-D_{24}: 25.9 (SD 4)	70 (70)	*Aerobic exercise:* Supervised group exercise *Control group:* Standard treatment	5 sessions per week	8 weeks
Junchin 2015 China	Inpatients Mean age: 28 (SD 7) 61% female	HAM-D_{24}: 25.8 (SD 3)	70 (70)	*Aerobic exercise:* Supervised aerobic exercise of the patients own choice Control group: Standard treatment	5 sessions per week	8 weeks
Schuch 2015 Brazil	Inpatients Mean age: 40 (SD 11) 74% female	HAM-D_{17}: 26.7 (SD 2)	50 (50)	*Aerobic exercise:* Supervised individual exercise. *Control group*: Standard treatment.	3 sessions per week	2 weeks
Kerling 2015 Germany	Inpatients Mean age: 43 (SD 10)	MADRS: 24.0 (SD 9)	42 (42)	*Aerobic exercise:* Supervised exercise. *Control group*: Standard treatment.	3 sessions per week	6 weeks
Belvederi 2015 Italy	Outpatients Mean age: 75 (SD 6) 71% female	HAM-D_{17}: 20.1 (SD 3)	121 (121)	*Aerobic exercise:* (1) Sertraline + supervised non-progressive exercise OR (2) sertraline + supervised progressive aerobic	3 sessions per week	24 weeks

				exercise. *Control group*: Sertraline.		
Carneiro 2015 Portugal	Outpatients Mean age: 50.16 (SD 12) 100% female	BDI: 48.8 (SD 10)	26 (19)	*Aerobic exercise:* Supervised exercise *Control group:* Standard treatment	3 sessions per week	16 weeks
Doose 2015 Germany	Outpatients Mean age: 47.9 (SD 10.5) 63% female	HAM-D_{17}: 14.2 (SD 3)	46 (46)	*Aerobic exercise:* Supervised aerobic exercise *Control group*: Standard treatment	3 sessions per week	8 weeks
Salehi 2016 Iran	Inpatients Mean age: 30.0 (SD 6) 35% female	HAM-D_{21}: 43.4 (SD 8)	40 (40)	*Aerobic exercise + ECT:* Supervised aerobic exercise *Control group:* ECT	3 sessions per weeks Control group 3 ECTs per week	4 weeks
Legrand 2016 France	Inpatients Mean age: 46.9 (SD 13) 67% female	BDI: 36.0 (SD 6)	24 (24)	*Aerobic exercise:* Supervised aerobic exercise *Control group:* Standard treatment	10 sessions in 10 consecutive days	10 days

SCL-D: Symptom Check List, depression subscale; HAM-D_{17}: Hamilton Depression Scale, 17 items; BDI: Beck's Depression Inventory; SD: Standard deviation; ECT: Electroconvulsive therapy

Table 2. Risk of bias in trials assessing exercise for patients diagnosed with depression

Author, Year of publication	Sequence generation	Allocation concealment	Blinding of participants and trial personnel assessors	Blinding of outcome assessors	Incomplete outcome data	Selective outcome reporting	For profit bias	Other bias	Comment on 'Other bias'
Klein 1985	Unclear	Unclear	High	High	High	Low	Low	Low	
Martinsen 1985	Unclear	Unclear	High	High	High	Low	High	Low	
Epstein 1986	Unclear	Unclear	High	High	High	Low	Unclear	High	Baseline difference
Doyne 1987	Unclear	Unclear	High	Low	High	Low	Unclear	High	Baseline difference
Veale 1992	Unclear	Unclear	High	High	High	Low	Low	High	Baseline difference
Singh 1997	Low	Unclear	High	Low	Low	Low	Low	High	Baseline difference
Blumenthal 1999	Unclear	Unclear	High	Low	High	Low	High	Low	
Mather 2002	Low	Low	High	Low	High	Low	Low	Low	
Dunn 2005	Low	Low	High	Low	High	High	High	Low	
Singh 2005	Low	Low	High	Low	High	Low	Unclear	Low	
Pilu 2007	Unclear	Unclear	High	Unclear	Low	Low	Unclear	Low	
Viera 2007	Unclear	Unclear	High	Unclear	Low	Low	Unclear	Low	
Blumenthal 2007	Low	Low	High	Low	High	High	Low	Low	
Krogh 2009	Low	Low	High	Low	Low[1]	High	High	High	Baseline difference
Mota-Pereira 2011	Unclear	Unclear	High	Low	High	Low	High	High	Baseline difference
Krogh 2012	Low	Low	High	Low	Low	Low	Low	Low	
Chalder 2012	Low	Low	High	High	Low	Low	Low	Low	
Fang 2013	Unclear	Unclear	High	Unclear	Unclear	High	Unclear	Low	
Huipeng 2013	Unclear	Unclear	High	Unclear	Low	Low	Unclear	Low	
Cassandra 2014	Low	Unclear	High	Low	High	Low	Low	Low	
Danielsson 2014	Unclear	Low	High	Low	High	Low	Low	Low	
Pfaff 2014	Low	Low	High	Low	Low[1]	Low	Low	High	Baseline difference
Guifeng 2015	Unclear	Unclear	High	Unclear	Low	Low	Unclear	Low	
Jinchun 2015	Unclear	Unclear	High	Unclear	Low	Low	Unclear	Low	
Schuch 2015	Unclear	Low	High	Low	Low	Low	Low	Low	
Kerling 2015	Unclear	Unclear	High	Unclear	Low	Low	Low	Low	
Belvederi 2015	Low	Low	High	Low	High	Low	Low	High	Post-hoc sample size
Carneiro 2015	Unclear	Low	High	High	Unclear	Low	Low	Low	
Doose 2015	Unclear	Unclear	High	High	High	Low	Low	High	No sample size calc.
Salehi 2016	High	High	High	Low	Unclear	Low	Low	High	Baseline difference
Legrand 2016	Low	High	High	High	High	Low	Unclear	Low	

Table 3. Heterogeneity of effect estimates for trials assessing the effect of exercise for patients diagnosed with depression explored by comparing sub-groups

Subgroups	Number of Trials (participants)	Random effects meta-analysis SMD (95% CI., p, I^2)	Subgroup explains heterogeneity P value
Risk of bias			
Less than high risk of bias[1]	4 (530)	-0.11 (-0.41 to 0.18; P=0.45; I^2 = 62%)	<0.001
High risk of bias	27 (1801)	-0.85 (-1.10 to -0.60; P<0.001; I^2 = 82%)	
Age			
Old (>59 years)	5 (492)	-0.77 (-1.34 to -0.19; P=0.009; I^2 = 87%)	0.99
Young (<59 years)	26 (1839)	-0.76 (-1.01 to -0.51; P<0.001; I^2 = 83%)	
Exercise context			
Group exercise	24 (1729)	-0.79 (-1.06 to -0.52; P<0.001; I^2 = 85%)	0.72
Individual exercise	7 (602)	-0.68 (-1.17 to -0.20; P=0.005; I^2 = 79%)	
Duration			
Less than 10 weeks	14 (691)	-0.93 (-1.11 to -0.88; P<0.001; I^2 = 19%)	0.05
10 weeks or more	17 (1640)	-0.58 (-0.88 to -0.28; P<0.001; I^2 = 86%)	
Attention control			
Attention control	7 (609)	-0.71 (-1.27 to -0.16; P=0.01; I^2 = 89%)	0.99
Waitlist	2 (47)	-0.67 (-2.48 to 1.13; P=0.47; I^2 = 88%)	
Pharmacotherapy			
Add-on	11 (734)	-0.92 (-1.38 to -0.46; P<0.001; I^2 = 86%)	0.82
No medication	6 (318)	-0.82 (-1.58 to -0.06; P=0.03; I^2 = 88%)	
Somatic comorbidity			
Somatic co-morbidity	0	N/A	
No co-morbidity	31 (2331)	N/A	
Minor depression			
Incl. minor depression	6 (350)	-0.90 (-1.65 to -0.15; P=0.02; I^2 = 86%)	0.67
No minor depression	25 (1981)	-0.73 (-0.97 to -0.49; P<0.001; I^2 = 88%)	
Patient setting			
Inpatients	10 (549)	-0.88 (-1.07 to -0.70; P<0.001; I^2 = 6%)	0.26
Outpatients	21 (1782)	-0.69 (-0.98 to -0.41; P<0.001; I^2 = 85%)	
Trial size			
Trials n ≤ 52	15 (479)	-1.30 (-1.74 to -0.86; P<0.001; I^2 = 77%)	<0.001
Trials n > 52	16 (1852)	-0.40 (-0.60 to -0.19; P<0.001; I^2 = 76%)	

Table 4. Summary of findings

Exercise compared to control or treatment as usual for depression

Patient or population: depression

Setting: In- or out-patients

Intervention: exercise

Comparison: control or treatment as usual

Outcomes	Anticipated absolute effects* (95% CI)		Relative effect (95% CI)	№ of participants (studies)	Quality of the evidence (GRADE)	Comments
	Risk with control or treatment as usual	Risk with exercise				
Severity of depression	-	**0.74 SMD lower** (0.51 lower to 0.96 lower)	-	2419 (31 RCTs)	⊕◯◯◯ VERY LOW [1]	Lower depression scores indicate improvement. SMD of 0.3 is considered clinically relevant.
Lack of remission	**Study population**		**RR 0.78** (0.68 to 0.90)	1639 (19 RCTs)	⊕◯◯◯ VERY LOW [2]	Remission is, with minor variations, defined as not full-filling the criteria for depression.
	646 per 1000	**504 per 1000** (426 to 594)				
Serious adverse events	**Study population**		**RR 2.21** (0.24 to 20.21)	335 (3 RCTs)	⊕⊕◯◯ LOW [3]	
	0 per 1000	**0 per 1000** (0 to 0)				
Quality of life	-	**0.43 SMD higher** (0.04 lower to 0.91 higher)	-	901 (8 RCTs)	⊕◯◯◯ VERY LOW [4]	Quality of life was assessed using a number of different methods. Higher score indicates improved quality of life. Seven of 24 trials reported on this outcome
Depression severity after the intervention	-	**0.06 SMD lower** (0.25 lower to 0.14 higher)	-	713 (7 RCTs)	⊕⊕◯◯ LOW [5]	Lower depression scores indicate improvement. SMD of 0.3 is considered clinically relevant.
Lack of remission after the intervention	**Study population**		**RR 0.95** (0.82 to 1.11)	777 (5 RCTs)	⊕⊕◯◯ LOW [6]	
	469 per 1000	**446 per 1000** (385 to 521)				
Depression severity. Restricted to trials with less than high risk of bias.	-	**0.11 SMD lower** (0.41 lower to 0.18 higher)	-	530 (4 RCTs)	⊕⊕◯◯ LOW [7]	Lower depression scores indicate improvement. SMD of 0.3 is considered clinically relevant.

***The risk in the intervention group** (and its 95% confidence interval) is based on the assumed risk in the comparison group and the **relative effect** of the intervention (and its 95% CI). **CI:** Confidence interval; **SMD:** Standardised mean difference; **RR:** Risk ratio

GRADE Working Group grades on evidence
High quality: We are very confident that the true effect lies close to that of the estimate of the effect
Moderate quality: We are moderately confident in the effect estimate. The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different
Low quality: 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 the effect

1. Downgraded by 3: risk of bias, inconsistency and publication bias
2. Downgraded by 3: risk of bias, inconsistency and publication bias
3. Downgraded by 2: imprecision and publication bias
4. Downgraded by 3: risk of bias, inconsistency and imprecision
5. Downgraded by 2: risk of bias and imprecision
6. Downgraded by 2: risk of bias and imprecision
7. Downgraded by 2: inconsistency and imprecision

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	5
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	6
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	1
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	7
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	6
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	6
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	6
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	8
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	7
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	8
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	9
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	9

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	8
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	11
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	Fig. 1
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Table 1
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	Table 2
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	Fig 2
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	Fig 3-fig8
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	Table 2
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	Table 3
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	21
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	21
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	22
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	26

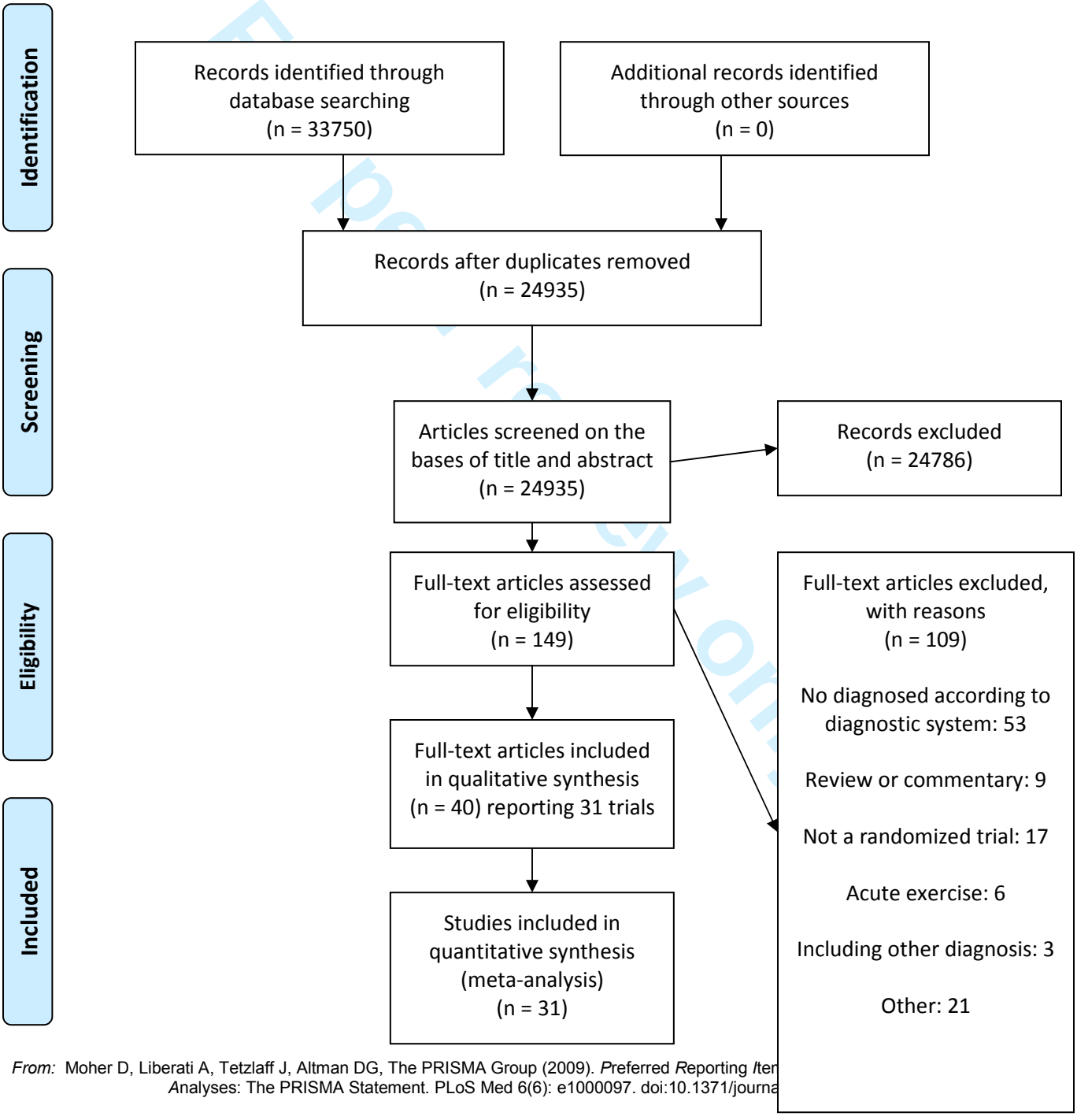
From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: www.prisma-statement.org.

Article:
Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplemental

Figure S1. Flow diagram for identification of trials assessing the effects of exercise for patients with depression.



From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). *P*referred *R*eporting *I*ter
*A*nalyses: The PRISMA Statement. PLoS Med 6(6): e1000097. doi:10.1371/journa

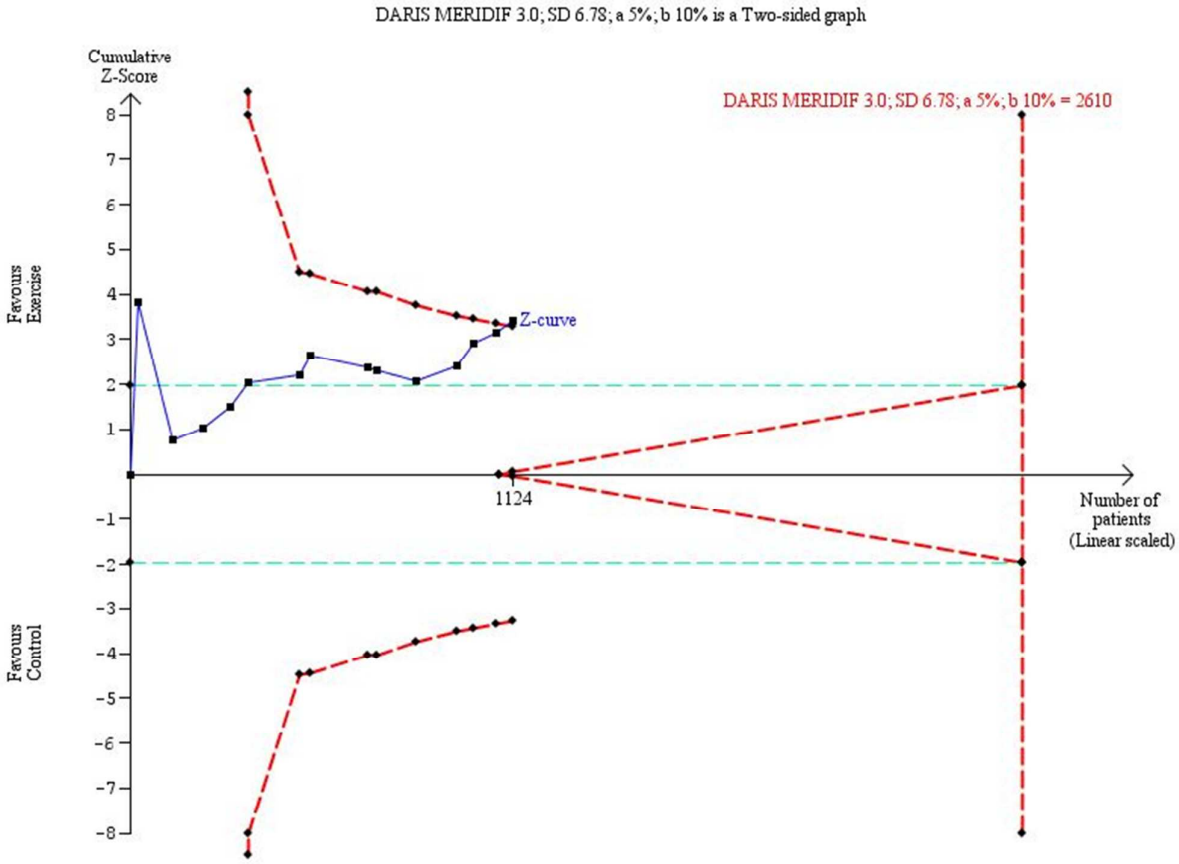
For more information, visit www.prisma-statement.org.

Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary Figure

Figure S2. Trial Sequential Analysis and required information size for the effect of exercise for depressive symptoms including twelve trials reporting on HAM-D_{17}.

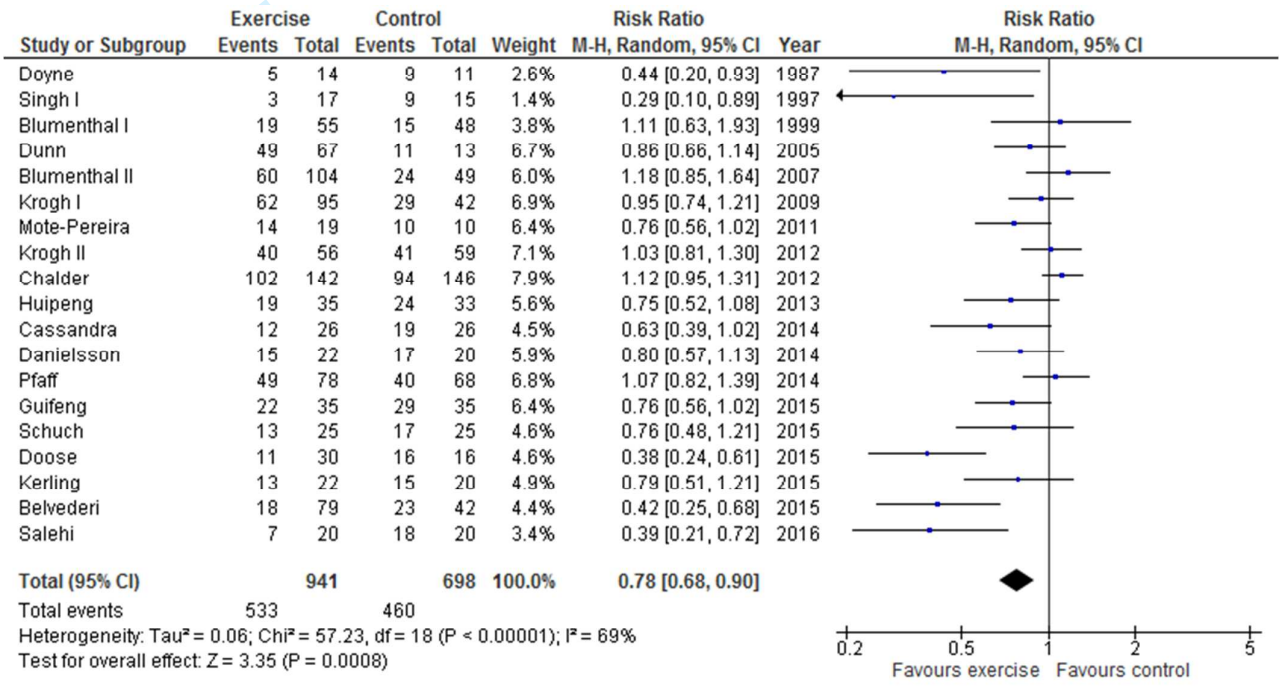


Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary Figure

Figure S3. Effect of exercise on lack of remission for patients diagnosed with depression

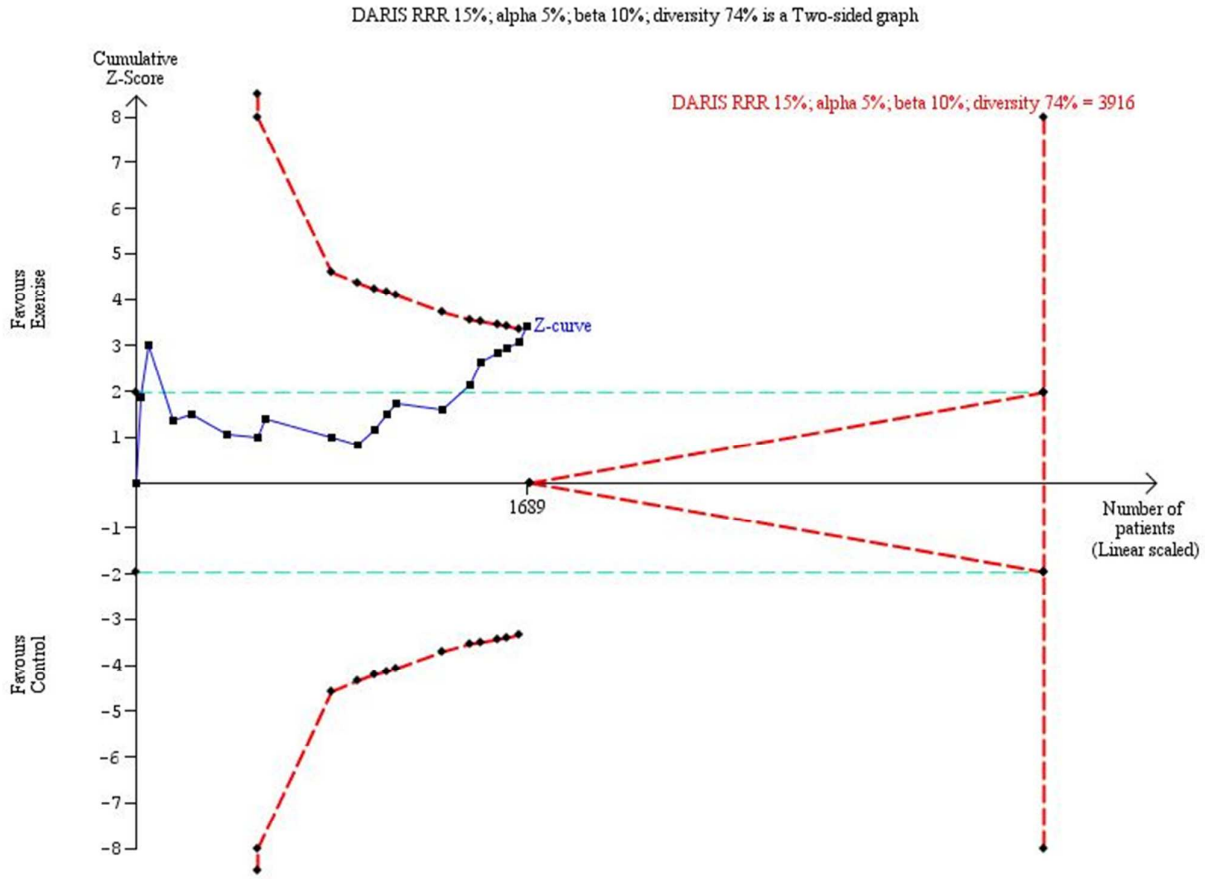


Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary Figure

Figure S3. Trial Sequential Analysis and required information size for the effect of exercise on lack of remission.

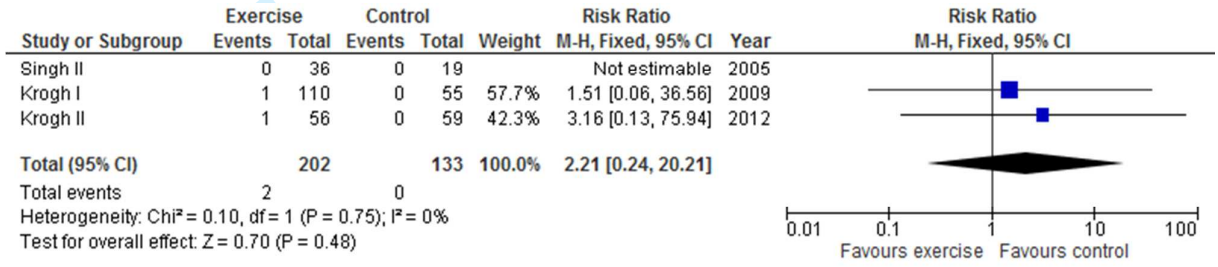


Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary Figure S5

Figure S5. Effect of exercise on risk of serious adverse events for patients diagnosed with depression

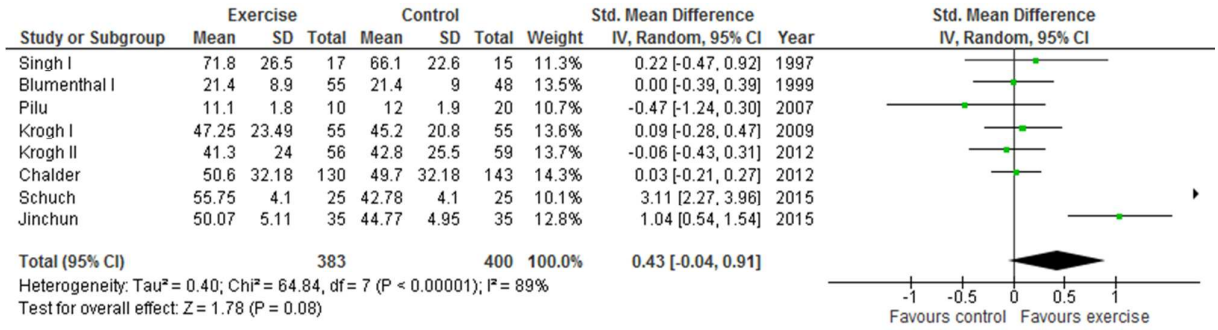


Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary Figure S6-S8

Figure S6. The effect of exercise on quality of life in patients diagnosed with depression



Quality of life was assessed using different scales: Singh I and Chalder used the SF-36, Blumenthal used Life Satisfaction Index, Pilu and Schuch used the WHOQOL, Krogh I and Krogh II used the WHO-Five Well-being Scale, and Jinchun used the GQOLI-74.

Figure S7. The effect of exercise on depression severity after the intervention in patients diagnosed with depression

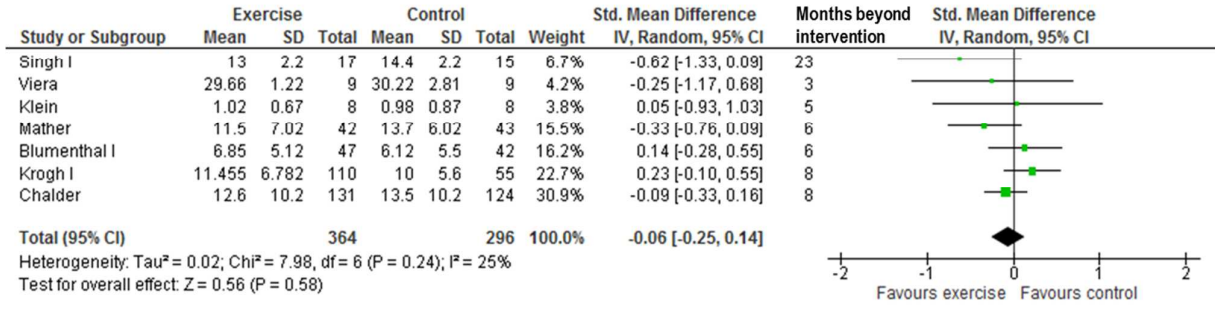
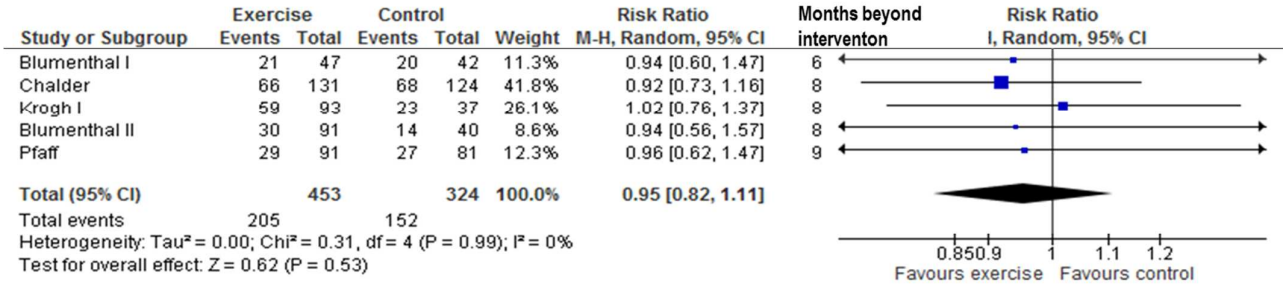


Figure S8. The effect of exercise on risk of lack of remission after the intervention in patients diagnosed with depression



Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary Table S1

Table S1. Imputation of missing data for trials assessing exercise for patients diagnosed with depression

Outcome	Result from review	Best/worse (1SD)	Best/worse (2SD)	Worse/best (1SD)	Worse/best (2SD)
Depression SMD (95% CI)	-0.74 (-0.96 to -0.51) p < 0.001; I^2 = 83%	-0.85 (-1.10 to -0.60) p < 0.001; I^2 = 87.2%	-0.85 (-1.11 to -0.60) p < 0.001; I^2 = 87.9%	-0.66 (-0.90 to -0.40) p < 0.001; I^2 = 85.4%	-0.61 (-0.84 to -0.38) p < 0.001; I^2 = 85.5%)
		Good Outcome	**Poor outcome**	**Good/poor outcome**	**Poor/good outcome**
Lack of remission (95% CL)	RR 0.78 (0.68 to 0.90) p < 0.001; I^2 = 69%	RR 0.75 (0.64 to 0.89) p = 0.0008; I^2 = 73%	RR 0.88 (0.83 to 0.94) p = 0.0002; I^2 = 69%	RR 0.71 (0.61 to 0.81) p < 0.001; I^2 = 68%	RR 0.86 (0.71 to 1.04) p = 0.12; I^2 = 83%
Serious adverse events (95% CL)	RR 2.21 (0.24 to 20.21) p = 0.48; I^2 = 0%	RR 2.19 (0.23 to 20.76) p = 0.50, I^2 = 50%	RR 0.92 (0.37 to 2.30) p = 0.86, I^2 = 60%	RR 0.08 (0.02 to 0.34) p = 0.001, I^2 = 5.4%	RR 19.17 (2.64 to 139.2) p = 0.004, I^2 = 0%

Imputation of missing data for continuous outcome: 'best-worst' - assumed that all participants lost to follow-up in the intervention group had a beneficial outcome (the group mean minus 1 or 2 SD), and all participants lost to follow-up in the placebo group have had a harmful outcome (the group mean plus 1 SD and 2 SD). The reverse 'worst-best-case' scenario is the reverse of the 'best-worst' scenario.

Missing data for the 'remission' outcome was imputed according to the following scenarios: 1) poor outcome analysis: none of the drop-outs/participants lost from both arms experienced the outcome; 2) good outcome analysis: all of the drop-outs/participants lost from both arms experienced the outcome; 3) extreme case analysis favouring the experimental intervention, all of the drop-outs/participants lost from the experimental arm, but none of the drop-outs/participants lost from the control arm experienced the outcome; and 4) extreme case analysis favouring the control: all drop-outs/participants lost from the experimental arm, but none from the control arm experienced the outcome. Missing data for 'serious adverse events' was calculated with the reverse assumptions.

Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary Table S2

Table S2. Trials characteristics for exploration of heterogeneity in trials assessing the effect of exercise in patients diagnosed with depression

Trial	Lower risk of bias	Age > 60	Group vs. individual	Duration	Attention control waitlist	Exercise as add on to drugs vs. exercise alone	Within-study dose exercise	Somatic disease vs. only MD	Trial Includes minor depression
Klein 1985	No	Young	Individual	12 weeks	Other	Exercise alone	No	No	Yes
Martinsen 1985	No	Young	Group	9 weeks	Attention control	Unclear	No	No	No
Epstein 1986	No	Young	Group	8 weeks	Waitlist	Unclear	No	No	Yes
Doyne 1987	No	Young	Individual	8 weeks	Waitlist	Exercise alone	No	No	Yes
Veale 1992	No	Young	Group	12 weeks	Other	Unclear	No	No	No
Singh 1997	No	Old	Group	10 weeks	Attention control	Exercise alone	No	No	Yes
Blumenthal 1999	No	Young	Group	16 weeks	Other	Add on	No	No	No
Mather 2002	No	Old	Group	10 weeks	Attention control	Add on	No	No	No
Dunn 2005	No	Young	Individual	12 weeks	Attention control	Exercise alone	Yes	No	No
Singh 2005	No	Old	Group	8 weeks	Other	Exercise alone	Yes	No	Yes
Pilu 2007	No	Young	Group	24 weeks	Other	Add on	No	No	No
Viera 2007	No	Young	Group	12 weeks	Other	Add on	No	No	No
Blumenthal 2007	No	Young	Group	16 weeks	Other	Add on	No	No	No
Krogh 2009	Yes	Young	Group	16 weeks	Attention control	No	No	No	No
Mota-Pereira 2011	No	Young	Group	12 weeks	Other	Add on	No	No	No
Krogh 2012	Yes	Young	Group	12 weeks	Attention control	Exercise alone	No	No	No
Chalder 2012	No	Young	Individual	32 weeks	Other	No	No	No	No
Fang 2013	No	Young	Group	6 weeks	Attention control	No	Yes	No	No
Huipeng 2013	No	Young	Group	6 weeks	Other	No	No	No	No
Cassandra 2014	No	Young	Group	3 weeks	Other	Add on	No	No	No
Danielsson 2014	No	Young	Group	10 weeks	Other	Add on	No	No	No
Pfaff 2014	Yes	Old	Group	12 weeks	Other	No	No	No	Yes
Guifeng 2015	No	Young	Group	8 weeks	Other	No	No	No	No
Jinchun 2015	No	Young	Group	8 weeks	Other	No	No	No	No
Schuch 2015	Yes	Young	Individual	2 weeks	Other	No	No	No	No
Kerling 2015	No	Young	Group	6 weeks	Other	No	No	No	No

Belvederi 2015	No	Old	Group	24 weeks	Other	Add on	Yes	No	No
Carneiro 2015	No	Young	Group	16 weeks	Other	Add on	No	No	No
Doose 2015	No	Young	Group	8 weeks	Other	No	No	No	No
Legrand 2016	No	Young	Individual	10 days	Other	No	No	No	No
Salehi 2016	No	Young	Individual	4 weeks	Other	Add on	No	No	No

Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary Table

Table S3. Heterogeneity of effect estimates for trials assessing the effect of exercise for patients diagnosed with depression on lack of remission.

Subgroups	Number of Trials (participants)	Random effects meta-analysis RR (95% CI., p, I^2)	Subgroup explains heterogeneity P value
Risk of bias			
Less than high risk of bias[1,2]	2 (165)	0.95 (0.74 to 1.23; p = 0.70; I^2 = 20%)	0.18
High risk of bias	17 (1474)	0.77 (0.64 to 0.92; p = 0.003; I^2 = 75%)	
Age			
Old (>59 years)	3 (299)	0.61 (0.21 to 1.02; p = 0.37; I^2 = 91%)	0.62
Young (<59 years)	16 (1340)	0.81 (0.70 to 0.93; p = 0.003; I^2 = 64%)	
Exercise context			
Group exercise	14 (1156)	0.80 (0.66 to 0.96; p = 0.02; I^2 = 72%)	0.69
Individual exercise	5 (483)	0.74 (0.52 to 1.04; p = 0.08; I^2 = 77%)	
Duration			
Less than 10 weeks	8 (393)	0.63 (0.51 to 0.77; p < 0.001; I^2 = 40%)	0.004
10 weeks or more	11 (1246)	0.93 (0.78 to 1.10; p = 0.39; I^2 = 69%)	
Attention control			
Attention control	4 (364)	0.91 (0.73 to 1.12; p = 0.38; I^2 = 42%)	0.07
Waitlist	1 (25)	0.44 (0.21 to 0.93; p = 0.03; I^2 = 0%)	
Pharmacotherapy			
Add-on	7 (540)	0.72 (0.54 to 0.96; p = 0.03; I^2 = 69%)	0.62
No medication	4 (252)	0.75 (0.52 to 1.09; p = 0.13; I^2 = 66%)	
Somatic comorbidity			
Somatic co-morbidity	0	N/A	
No co-morbidity	19 (1639)	N/A	
Minor depression			
Incl. minor depression	3 (203)	0.63 (0.21 to 1.89; p = 0.41; I^2 = 87%)	0.69
No minor depression	16 (1436)	0.79 (0.68 to 0.92; p = 0.002; I^2 = 69%)	
Patient setting			
Inpatients	6 (322)	0.71 (0.60 to 0.84; p < 0.001; I^2 = 0%)	0.21
Outpatients	13 (1317)	0.84 (0.69 to 1.01; p = 0.07; I^2 = 77%)	
Trial size			
Trials n ≤ 52	9 (358)	0.62 (0.50 to 0.76; p < 0.001; I^2 = 45%)	0.002
Trials n > 52	10 (1281)	0.95 (0.80 to 1.12; p = 0.52; I^2 = 68%)	

[1]Trials potentially having less bias than trials with high risk of bias.

BMJ Open

Exercise for patients with major depression: a systematic review with meta-analysis and Trial Sequential Analysis

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2016-014820.R1
Article Type:	Research
Date Submitted by the Author:	03-Apr-2017
Complete List of Authors:	Krogh, Jesper; Kobenhavns Universitet Sundhedsvidenskabelige Fakultet, Mental Health Centre Copenhagen Speyer, Helene; University of Copenhagen Hjorthøj, Carsten; Copenhagen University Hospital, Mental Health Center Copenhagen Gluud, Christian; Copenhagen University Hospital Rigshospitalet, The Copenhagen Trial Unit, Centre for Clinical Intervention Research Nordentoft, Merete; University of Copenhagen, Mental Health Centre Copenhagen
Primary Subject Heading:	Mental health
Secondary Subject Heading:	Sports and exercise medicine, Evidence based practice
Keywords:	Exercise, MENTAL HEALTH, Systematic review, Meta-analysis, Evidence based medicine, Randomised clinical trials

Exercise for patients with major depression: a systematic review with meta-analysis and Trial Sequential Analysis

Jesper Krogh[1], Carsten Hjorthøj[1], Helene Speyer[1], Christian Gluud[1], Merete Nordentoft[1]

[1] Mental Health Centre Copenhagen. Faculty of Health Sciences. University of Copenhagen, Copenhagen, Denmark.

[2] Copenhagen Trial Unit, Centre for Clinical Intervention Research, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark.

Corrosponding author

Jesper Krogh, Jesper.krogh@dadlnet.dk

Mental Health Centre Copenhagen, Kildegårdsvej 28, Building 15. DK-2900, Hellerup, Denmark

Abstract

Objectives

To assess the benefits and harms of exercise in patients with depression.

Design

Systematic review

Data sources

Bibliographical databases were searched until the 17th of April, 2016.

Eligibility criteria and outcomes

Eligible trials were randomised clinical trials assessing the effect of exercise in participants diagnosed with depression. Primary outcomes were depression severity, lack of remission, and serious adverse events (e.g. suicide) assessed at the end of the intervention. Secondary outcomes were quality of life and adverse events such as injuries, as well as assessment of depression severity and lack of remission during follow-up after the intervention.

Results

Thirty-one trials enrolling 2419 participants were included. The effect of exercise versus control on depression severity was -0.74 standardised mean difference (SMD) (95% CI -0.96 to -0.51; P < 0.001; GRADE: very low quality). Restricting this analysis to the four trials that seemed less affected of bias, the effect vanished to -0.11 SMD (-0.41 to 0.18; P = 0.45; GRADE: low quality). Exercise decreased the relative risk of no remission to 0.78 (0.68 to 0.90; P < 0.001; GRADE: very low quality). Restricting this analysis to the two trials that seemed less affected of bias, the effect vanished to 0.95 (0.74 to 1.23; P = 0.78). Trial Sequential Analysis excluded random error when all trials were analysed. Sub-group analyses found that

trial size and intervention duration were inversely associated with effect size for both depression severity and lack of remission. There was no significant effect of exercise on secondary outcomes.

Conclusions

Trials with less risk of bias suggested no antidepressant effects of exercise and there were no significant effects of exercise on quality of life, depression severity, or lack of remission during follow-up. Data for serious adverse events and adverse events was scarce not allowing conclusions for these outcomes.

Systematic review registration

The protocol was published in the journal Systematic Reviews: 2015; **4**:40

DOI: 10.1186/s13643-015-0030-6.

Article Summary

Strengths and limitations of this study

- The protocol for this review has previously been published
- Using meta-regression analysis, trial sequential analysis and the GRADE system the conclusions form this review is based on a firm and transparent platform
- Based on an extensive literature search, this review included 31 trials allocating more than 2000 participants to exercise or control interventions
- The effect estimates are largely based on trials at high risk of bias
- Effect estimates from included trials had considerable heterogeneity

Introduction

Depression is a common disorder affecting up to 17% of the population during their lifetime.[1;2] Based on data from the World Health Organisation, depression is ranked as the second largest health-care problem globally, in terms of years lived with disability.[3] Depending on its severity, depression is often treated using psychotherapy, antidepressants, or a combination of both. However, the clinical benefits of antidepressants[4;5] and psychotherapy[6-8] has been challenged. Both treatments are costly in terms of time and money and may also have adverse effects. Compliance with antidepressant treatment is poor; the dropout rate in clinical trials is reported to be between 12% and 40% within the initial 6 to 8 weeks of treatment.[4;9]

The weakness of evidence for the beneficial effect of current interventions, along with problems related to low compliance and harms, has resulted in an interest in using alternative interventions. The use of exercise as an intervention has attracted considerable attention, and various forms of exercise varying in intensity have been assessed in a number of randomised clinical trials to test their effectiveness as a treatment for patients with depression. In 2011, we published a meta-analysis of randomised clinical trials examining the effect of exercise on depressive symptoms in patients with clinical depression.[10] The results suggested that referring patients with clinical depression to exercise programs was associated with a small to moderate effect on depressive symptoms. However, restricting the analysis to three trials with a low risk of bias, the effect estimate was non-significant. Since 2011, other reviews have been published on the effect of exercise on depressive symptoms,[11] in older people,[12] and in patients with chronic illnesses.[13] However, none of these reviews addressed the specific population of adults diagnosed with major depression according to valid diagnostic criteria, such as the International Classification of Diseases[14] or the Diagnostic and Statistical Manual of Mental Disorders.[15] The reviews contained a number of trials that included volunteers who were defined as being depressed on the basis of psychometric testing (for example, Beck

Depression Inventory[16]), as opposed to individuals with a clinical diagnosis of major depression. Furthermore, several randomised clinical trials investigating the effect of exercise in clinically depressed individuals have been published since our 2011 review.[10]

The objectives of the present systematic review are to investigate the beneficial and harmful effects of exercise, in terms of severity of depression, lack of remission, quality of life, and suicide versus controls with or without co-interventions in adults with a clinical diagnosis of major depression. The current systematic review differs from our previous review in a number of aspects.[10] We only considered trials including participants diagnosed with depression according to a validated diagnostic system. We also included trials including participants with somatic co-morbidity, e.g., cancer or diabetes. The harmful effects of exercise interventions are also addressed, the intervention effects being assessed according to the grading of recommendations assessment, development, and evaluation (GRADE) framework, and bibliographical searches have been extended to include a Chinese and a South-American database until 2016.

Methods/design

The protocol for this review has previously been published.[17]

Search strategy

The following bibliographical databases was searched from April 2015 until the 17th of April, 2016: CENTRAL, MEDLINE, EMBASE, Science Citation Index (Web of Science), LILACS, and Wanfang using medical subject headings (MeSH or similar) when possible or text word terms: depression, depressive disorder and exercise, aerobic, non-aerobic, physical activity, physical fitness, walking, jogging, running, bicycling,

swimming, strength, or resistance. Please see supplementary material (S1) for an example of a bibliographical search.

Trial selection

One investigator (JK) examined titles and abstracts to remove obviously irrelevant reports. Two investigators (JK + HS) examined full text reports and abstracts determining compliance with inclusion criteria. A trial was considered eligible if it was a randomised clinical trials including participants diagnosed as having major depression according to a valid and recognised diagnostic system (that is, Research Diagnostic Criteria (RDC),[18] International Classification of Diseases (ICD),[14] or Diagnostic and Statistical Manual of Mental disorders (DSM)[15]) and included participants aged >17 years. Abstracts and full text reports were included.

Trials were excluded if they measured depression immediately after a single bout of exercise, compared one form of exercise versus another, or compared different exercise intensities without including a control group. The trials had to allocate participants to an exercise intervention versus a control group (that is, exercise versus a control group receiving no intervention or treatment as usual or an attention control using light exercise) or using exercise as an add-on-treatment (that is, exercise plus usual treatment in the experimental group versus usual treatment alone in the control group). Exercise intervention was defined as a systematic physical intervention with the intention to increase muscle strength and/or cardiovascular fitness, e.g., running, swimming or weight lifting. In case of attention control, it should specifically be mentioned by the authors of the trial report that the intervention was intended as a control intervention.

Outcomes

The primary outcomes were 1) depressive symptoms measured on a continuous scale assessed at the end of the intervention; 2) lack of remission, that is, a binary outcome of the proportion of participants in each intervention group of the trial who did not obtain remission at the end of the intervention according to the authors' own definition; and 3) serious adverse events defined according to ICH-GCP as any untoward medical occurrence that was life threatening, resulted in death or persistent or significant disability (ICH-GCP 1997).[19] Serious adverse events accordingly include suicide attempts as well as suicides. The secondary outcomes were quality of life, non-serious adverse events (e.g., muscle injuries) as well as depressive symptoms and lack of remission assessed after the intervention.

Data extraction

Two authors (JK, HS) independently extracted data using a pre-piloted structured form. Any discrepancies in the data extraction or inclusion/exclusion of trials was resolved by referring to the original papers. CG or MN assisted as adjudicator in cases of disagreements. Data extraction included, in addition to outcomes, information regarding country of origin, number of randomised participants, number of participants included in efficacy analysis, mean age of participants, diagnostic system, baseline assessment of depression severity, type of intervention, frequency of intervention, and duration of intervention. Continuous outcomes were preferred in the following order: post-intervention scores with corresponding standard deviations (SD), mean change from baseline with SD, mean difference between groups post-intervention and reported outcomes were preferred to figure's. JK and CH independently performed the assessment of bias domains. The authors JK, CG, and MN have previously published trial reports assessing the effect of exercise in participants with depression,[20;21] and to reduce the risk of academic bias two additional authors were included in the current systematic review (CH, HS).

Risk of bias assessment

Definitions in the assessment of bias risk of a trial was conducted according to the Cochrane Handbook for Systematic Reviews of Interventions[22] of the following domains: allocation sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessors, incomplete outcome data, selective outcome reporting, for-profit bias, and other bias. Trials assessed as having 'low risk of bias' in all of the above specified domains were considered 'trials at low risk of bias'. Trials assessed as having 'uncertain risk of bias' or 'high risk of bias' in one or more of the above specified domains were considered trials with 'high risk of bias'. In line with our previous systematic review[10] and the latest Cochrane review on exercise for depression,[23] trials at low risk of bias in the allocation concealment domain, blinded outcome assessment domain, and the incomplete outcome data domain were characterised as 'trials potentially having less risk of bias than other trials at high risk of bias'. Trials assessing the effect of behavioural interventions are rarely able to mask the allocation, and participants and health care providers are therefore not blinded. Therefore, we will also report the number of trials at low risk of bias in the remaining domains.

Data synthesis and analysis

In order to be able to include all of the trials in our meta-analysis, estimates of standardised mean difference (SMD) for each individual trial was carried out. SMD is the mean difference in depression score between the exercise and control groups divided by the pooled standard deviation. The result is a unit free effect size. By convention, SMD effect sizes of 0.2, 0.5 and 0.8 are considered small, medium and large intervention effects.[22] For dichotomous variables, we calculated the risk ratio (RR) with a 95% confidence interval. It was expected that some trials would have several intervention groups. Data from the experimental groups was pooled and compared with the data from the control group. In case of discrepancies between the random-effects model analysis and the fixed-effect model analysis, both results are reported; otherwise, only results from the random-effects analysis is reported. The degree of

heterogeneity was quantified using the I-squared statistic,[24] which can be interpreted as the percentage of variation observed between the trials attributable to between-trial differences, rather than sampling error (chance). Heterogeneity was explored by analyses of sub-groups (see below).

For the primary outcomes, Trial Sequential Analysis was performed.[25;26] In order to calculate the required information size and the cumulative Z-curve's eventual breach of relevant trial sequential monitoring boundaries, the required information size for the primary continuous outcome was based on type I error of 5%, a beta of 10%, the standard error of the meta-analysis, and a minimal difference of three points on the HAM-D_{17}.[17] Post-hoc we calculated the required information size including all trials. This was done by converting effect estimates from trials reporting other outcome scales into the HAM-D_{17} scale as described by Thorlund et al.[27] In order to calculate the required information size and the cumulative Z-curve's eventual breach of relevant trial sequential monitoring boundaries, the required information size for lack of remission was based on type I error of 5%, a beta of 10%, the proportion of participants in the control group with the outcome, and a relative risk reduction of 15% and 30%.

Bayes factors were calculated for all primary outcomes.[28] Low P-values suggest that we can reject the null-hypothesis. But even a low *P*-value from a meta-analysis can be misleading if there is also a low probability that data are compatible with the anticipated intervention effect. In other words, the probability that the actual measured difference in effect of the compared interventions resulted from an a priori anticipated 'true' difference needs to be considered. For this purpose, it is helpful to calculate the Bayes factor, which is the ratio of the P-value probabilities of the meta-analysis result divided by the probability of the anticipated effect, or 'true' effect.[28] As suggested by Jakobsen et al.,[28] a Bayes factor lower than 0.1 together with a low P-value suggest, if bias can be ruled out, that the observed result is compatible with the

a priori expected effect. If the Bayes factor is higher than 0.1 the result is not compatible with the a priori expected effect and the effect may be lower.

To assess the potential impact of missing data (incomplete outcome data bias) we did sensitivity analysis of missing data using the following strategy: a 'best-worst' case scenario was assessed, assuming that all participants lost to follow- up in the intervention group had a beneficial outcome (the group mean minus 1 standard deviation (SD)), and all those with missing outcomes in the control group have had a harmful outcome (the group mean plus 1 SD and 2 SD). In addition, the reverse 'worst-best-case' scenario analysis was also performed.[28] Missing data for the 'lack of remission' outcome was imputed in sensitivity analysis according to the following scenarios:[29] 1) poor outcome analysis: assuming that all of the drop-outs/participants lost from both the experimental and the control arms experienced the outcome, including all randomised participants in the denominator; 2) good outcome analysis: assuming that none of the drop-outs/participants lost from the experimental and the control arms experienced the outcome, including all randomised participants in the denominator; 3) extreme case analysis favouring the experimental intervention ('best-worse' case scenario): none of the drop-outs/participants lost from the experimental arm, but all of the drop-outs/participants lost from the control arm experienced the outcome, including all randomised participants in the denominator; and 4) extreme case analysis favouring the control ('worst-best' case scenario): all of the drop-outs/participants lost from the experimental arm, but none from the control arm experienced the outcome, including all randomised participants in the denominator.

Subgroup analyses

In subgroup analyses, the possible effects of variables on intervention effects on outcomes and heterogeneity were compared. Trials potentially having less risk of bias (i.e., trials with adequate allocation concealment, blinded outcome assessment, and intention to treat analysis) were compared to trials at high

risk of bias. The effect of age was assessed by comparing trials including older participants (mean age >59 years) to trials including younger participants (mean age <60 years). The effect of type of exercise was assessed by comparing trials using group exercises compared to trials using individual exercise. The effect of duration of intervention was assessed by comparing trials with short duration of intervention to trials with long duration of intervention splitting by the median time of duration. The effect of type of control group was assessed by comparing trials using attention control to trials with waitlist controls and comparing trials with exercise as add-on to medication to trials not using any medication. In addition, a within-study comparison of low-dose exercise versus high-dose exercise in trials using different exercise intensities was performed. The effect of co-morbid somatic disease was assessed by comparing the effect estimates from trials including participants with depression compared to trials including participants with depression in addition to a somatic disease. Publication bias was assessed by visual inspection of a funnel plot and by Egger's test and if publication bias plausible Duval's and Tweedie's trim and fill procedure was conducted.[30]

We assessed and graded the evidence according to the grading of recommendations assessment, development, and evaluation (GRADE) for high risk of bias, imprecision, indirectness, heterogeneity, and publication bias.[31] Based on this assessment, the intervention is graded accordingly: **'**high quality'**-** we are very confident that the true effect lies close to that of the estimate of the effect; **'**moderate quality'- we are moderately confident in the effect estimate. The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different; 'low quality**'**- our confidence in the effect estimate is limited: the true effect may be substantially different from the estimate of the effect; 'very low quality'- we have very little confidence in the effect estimate: the true effect is likely to be substantially different from the estimate of the effect.[32]

Deviations from our protocol

Post-hoc we included trials using the Chinese Classification of Mental Disorders (CCMD) as well as a few trials including participants classified as having 'minor depression'. The CCMD system closely adhere to the ICD and DSM systems and have been found highly compatible in field studies, so these studies were included.[33] A few trials included some participants classified as having 'minor depression' according to the trials chosen diagnostic system (e.g., DSM), and it is questionable if these participants have major depression. We therefore decided to include these trials but also to conduct a sub-group analysis exclusively including participants with major depression. To further explore heterogeneity, we post-hoc included sub-group analysis comparing intervention effects in inpatients and outpatients as well as an analysis according to trial size. Trials were divided into small or large trials using the median of total n included in the efficacy analysis. The effect of exercise capacity was post-hoc assessed by comparing trials with a high increase in maximal oxygen uptake (VO2max) with studies with lower increase in maximal oxygen uptake. Assessment of exercise capacity was based on the increase of VO2max in the intervention groups and trials were stratified to either high or low increase in exercise capacity by median. We did not conduct Trial Sequential Analysis based on a relative risk reduction of 30% of lack of remission as this was an implausible effect.

Participant involvement

Depressed participants were not involved in this study.

Results

Bibliographical search and trial characteristics

The main bibliographical search was conducted the 26th of August, 2015 and the final updates were conducted on the 17th of April, 2016. As illustrated in Figure S1, we identified 40 publications reporting the effect of exercise on depressive symptoms in 31 randomised clinical trials.[20;21;34-72] Four-teen trials were conducted in Europe,[20;21;39;48;51;52;54;60;64-67;73;74] seven in the U.S.A.,[37;38;42;44;59;63;75], six in Asia,[46;68-72] two in Australia,[53;57] and two in South-America.[55;62] A total of 2,419 participants were randomised and 2,331 were included in the efficacy analysis of benefit. 10 trials included inpatients[46;48;55;66;68-73] and five trials included participants with a mean age above 60 years.[51;53;57;59;60] No trials exclusively included participants with comorbid somatic disease. Four trials reported the continuous outcome as mean change from baseline in each group with a corresponding SD, [38;52;64;67] and one trial presented data as mean difference between groups post-intervention.[39]. The remaining trials reported post-scores in each group with corresponding SD. Please see Table 1 for trial characteristics.

Bias risk assessment

Sequence generation was adequate in 12/31 (39%), allocation concealment was adequate in 12/31 (39%) trials, blinding of participants and trial personnel was adequate in 0/31 (0%), blinded outcome assessment was performed in 16/31 (52%), low risk of bias in the 'incomplete outcome data' domain was found in 12/31 (39%) trials, selective outcome reporting domain was adequate in 27/31 (87%), for profit bias domain was adequate in 15/31 (48%) and 21/31 (68%) were free of other bias. All trials were at high risk of bias. Given the nature of the intervention, no trial had blinded participants or trial personnel, however, two trials had low risk of bias in all other bias domains.[21;53] Five trials (16%) were sponsored by for profit organisations: three trials were supported by pharmaceutical companies,[52;73;76] one trial by a company producing fitness machines,[44] and one trial by an insurance company.[20] According to our a priori defined criteria, 4/31 (13%) trials potentially had less risk of bias than the other trials at high risk of bias.[20;21;53;55] Please see Table 2 for details on assessment of risks of bias.

Primary outcomes

The effect of exercise on depression severity

All included trials provided a continuous outcome on depression severity for the assessment of the exercise intervention encompassing 2,331/2,419 randomised participants (96.4%). The effect of intervention versus control was a standardised mean difference (SMD) of -0.74 (95% CI -0.96 to -0.51; P<0.001) (Figure 1.). This corresponds to an effect on the HAM-D_{17} scale of -4.6 (95% CI -6.0 to -3.2) points.

Missing data

Missing outcome analysis for depression as a continuous outcome did not markedly change the effect estimates. The least favourable outcome for the exercise intervention was the worse/best outcome analysis using +2 SD resulting in an effect estimate of -0.61 SMD (95% CI -0.84 to -0.37; P<0.001) (Table S1).

Heterogeneity and subgroup analysis

The I^2 was 82% suggesting substantial heterogeneity. Subgroup analysis revealed that the effect estimates for trials potentially having less risk of bias was -0.11 SMD (95% CI. -0.41 to 0.18; P = 0.45; I^2 = 62%) compared to that of the trials at high risk of bias -0.85 SMD (95% CI. -1.10 to -0.60; P < 0.001; I^2 = 82%) (test of sub-group difference, P = 0.0002). In addition, trials including 52 participants or less had a pooled estimate of -1.30 SMD (95% CI -1.74 to -0.86; P < 0.001; I^2 = 77%) compared to that of larger trials of -0.40 SMD (95% CI -0.60 to -0.19; P < 0.001; I^2 = 76%) (test of sub-group difference, P < 0.001). Trials of short duration of intervention (less than 10 weeks) had a SMD of -0.93 (95% CI -1.11 to -0.88; P < 0.001; I^2 = 19%) compared to trials with longer duration of intervention, -0.58 SMD (95% CI -0.88 to -0.28; P < 0.001; I^2 = 86%) (test of sub-group difference, P = 0.05). Effect estimates from trials including participants with minor

depression compared to trials exclusively including participants with major depression did not differ (test of sub-group difference, P = 0.67).

Four trials allocated 206 participants to different exercise intensities/doses.[44;57;72;77] Comparing the post-intervention depression scores for participants allocated to either high intensity/high dose or low intensity/low dose exercise showed a difference of -0.40 SMD (95% CI -0.67 to -0.12; P=0.005; I^2 = 0%) in favour of high intensity/high dose exercise. As shown in Table 3, no other trial characteristic significantly explained any of the observed heterogeneity. Please see Table S2 for trial characteristics used to explore heterogeneity.

Trial Sequential Analysis and diversity adjusted required information size

The diversity adjusted required information size for HAM-D_{17} as a continuous outcome was calculated based on our anticipated intervention effect of a minimal relevant difference of 3.0 HDRS points, a standard deviation of 6.78 points, a risk of type I error of 0.05, a power of 90% and the observed diversity of 92% to 2610 participants. Only 14 trials reported results from HAM-D_{17}[20;21;37;38;42;43;51;52;54;55;57;67;69;77] with an accrued 1124 participants. As shown in Figure S2, the cumulative Z-curve just crossed the trial sequential monitoring boundary for benefit. With the aforementioned settings, the pooled estimate is therefore less likely to be a random finding due to lack of power or multiple testing if bias could be ignored. Post-hoc we calculated the adjusted required information size for HAM-D_{17} including all trials as shown in Figure S3. As with the original analysis the Z-curve crossed the trial sequential monitoring boundary for benefit supporting that the pooled estimate is less likely to represent a Type 1 error if bias could be ignored.

Bayes factor

Fourteen trials reported effect estimates using the HAM-D_{17}.[20;21;37;38;42;44;51;52;54;62;67;69;77;78] Based on these trials, Bayes factor was calculated (δ = -3.37; SE_δ = 0.96; μ_a = -3.0) and was found to be 0.002, which is below the Bayes factor threshold for significance of 0.1, supporting the intervention effect if bias could be ignored.

Publication bias

Inspection of the funnel-plot (not shown) suggested that small trials with small or no effect of exercise were missing (Figure S4). Egger's test supported the suspicion of publication bias, P < 0.00001. Using the Duval and Tweedie's trim and fill procedure, the estimate was reduced to -0.28 SMD (95% CI -0.52 to -0.04). This corresponds to an effect on the HAM-D_{17} scale of -1.8 (95% CI -3.2 to -0.25).

The effect of exercise on depression – lack of remission

Nineteen trials, randomising 1825 participants and including 1639 participants (90%) in final analysis reported remission as an outcome.[20;21;37-39;42;44;46;48;52;53;55;59;60;64;67-69;71] Remission post-intervention was defined in various ways: A post-intervention score on the HAM-D_{17} less than 8 points,[43;52;55;68;69] not fulfilling the DSM criteria for depression *and* a HAM-D_{17} less than 8 points,[20;21;38] not fulfilling the DSM criteria for depression,[37;53;59] a BDI score less than 9 points,[42] a BDI score less than 10 points,[39] a HAM-D_{17} score less than 10 points,[77] a MADRS score less than 10 points,[46] a MADRS score less than 10 points *and* a 50% reduction in symptom score,[64] a 75% reduction in HAM-D_{24},[71] a HAM-D_{17} score less than 11.28 points *and* a reduction in HAM-D_{17} scores > 7.74 points,[67] and one study used MADRS not specifying the cut-off for remission.[48] The RR for lack of remission was 0.78 (95% CI 0.68 to 0.90; P=0.0008) in favour of the

intervention using a random-effects analysis. The I^2 was 69% suggesting substantial heterogeneity. The forest plot for the intervention effect on lack of remission is illustrated in Figure S5.

Missing data

The scenario in least favour of the intervention was the 'poor' outcome analysis having an effect estimate of RR 0.88 (95% CI 0.83 to 0.94) P=0.0002; I^2 = 69%. As shown in Table S1, the remaining scenarios did not substantially differ from the main analysis.

Heterogeneity and subgroup analysis

I^2 was 69% for the outcome lack of remission suggesting substantial heterogeneity. For this outcome, only two trials[21;78] were considered as trials potentially having less risk of bias than the other trials at high risk of bias. The RR of these two trials was 0.95 (95% CI 0.74 to 1.23; P=0.78) compared to 0.77 (96% CI 0.64 to 0.92; P=0.003) for trials at high risk of bias, test of subgroup difference, P=0.19). Trials including 52 participants or less in their final analysis had a RR 0.62 (95% CI 0.50 to 0.76; P<0.001; I^2 = 45%) compared to 0.95 (95% CI 0.80 to 1.12; P=0.52; I^2 = 68%) for larger trials (test of sub-group difference, P=0.002). Also, trials with a duration of less than 10 weeks had a RR of 0.63 (95% CI 0.51 to 0.77; P<0.001; I^2 = 40%) compared to 0.93 (95% CI 0.78 to 1.10; P=0.39; I^2 = 69%) for trials of a longer duration (test of sub-group difference, P=0.004). As shown in Table S3, no other trial characteristic significantly explained any of the observed heterogeneity. Please see Table S2 for trial characteristics used to explore heterogeneity.

Trial Sequential Analysis and diversity adjusted required information size

The diversity adjusted required information size for lack of remission was calculated based on our observed diversity of 74%, a proportion in the control group with lack of remission of 66%, an anticipated

intervention effect of 15% relative risk reduction, a risk of type I error of 0.05% and a power of 90%. As shown in Figure S6, the cumulative Z curve just crossed the trial sequential monitoring boundary for benefit. With the aforementioned settings, the pooled estimate is therefore less likely to be a random finding due to lack of power or multiple testing if bias could be ignored.

Bayes factor

Bayes factor was calculate based on the observed relative risk of remission, the associated standard error, and an anticipated intervention effect of relative increase in number of participants with remission by 15% (δ = -0.248; SE_δ= 0.08; μ_δ = -0.163). Bayes factor was 0.02, which is below the Bayes factor threshold for significance of 0.1.

Publication bias

Inspection of the funnel-plot (not shown) suggested that small trials with small or no effect of exercise were missing. Egger's test supported the suspicion of publication bias, P=0.002. Imputing theoretically missing studies by the Duval and Tweedie's trim and fill procedure, reduced the estimate of intervention effect to a relative risk reduction of 0.93 (95% CI 0.79 to 1.11).

The effect of exercise on serious adverse events

Serious adverse events (i.e., death or suicide attempts) were reported in only three trials.[20;21;57] In these trials, one suicide attempt[21] and one death by suicide[20] were recorded in the intervention groups. The RR

for death or suicide in the two trials was 2.21 (95% CI 0.24 to 20.21; P=0.48; I^2 = 0%) as illustrated in Figure S7.

Missing data

Missing outcome analysis for 'serious adverse events' varied according to missing data scenario: poor outcome analysis relative risk, 0.92 (95% CI 0.37 to 2.30; P=0.86; I^2 = 60.0%), good outcome analysis, 2.19 (95% CI 0.23 to 20.76; P=0.50; I^2 = 0.0%), best/worst outcome analysis – 0.08 (95% CI 0.02 to 0.34; P=0.001; I^2 = 5.4%), worst/best outcome analysis 19.17 (95% CI 2.64 to 139.2; P=0.004; I^2 = 0.0%).

Trial Sequential Analysis and Bayes analysis

We decided not to conduct Trial Sequential Analysis or Bayes analysis due to too sparse data.

Publication bias

Only 3/31 trials reported on this outcome and no formal assessment for publication bias was made. However, the lack of reporting in the vast majority of trials suggest risk publication bias.

Secondary outcomes

The effect of exercise on quality of life

Eight trials randomising 901 participants reported on quality of life,[20;21;37;39;55;59;70;79] observing that participants allocated to exercise did not have significantly better quality of life (SMD 0.43; 95% CI -0.04 to 0.91; P=0.08). The I^2 was 89% showing substantial heterogeneity (Figure S8).

Non-serious adverse events

Non-serious adverse events were reported in only nine trials.[20;21;38;55;57;59;64;66;67] Five trials reported on musculoskeletal adverse events without conducting formal tests[57;59;64;66;67] and four trials reported on number of participants with high depression scores post-intervention compared to baseline assessment.[20;21;64;67] The RR for increased severity of depression post-intervention was 0.83 (95% CI 0.40 to 1.70; P=0.60; I^2 = 0.0%).

The effect of exercise on depression beyond the duration of the intervention

Assessment of depression beyond the intervention was conducted in seven trials,[20;37;39;51;59;62;80] with a median duration between end of intervention and assessment of depression of 6 months (range 5 to 23.5 months). The SMD between the intervention group and the control group using a random effects analysis was -0.10 (95% CI -0.28 to 0.09; P=0.31; I^2 = 19.5%). The I^2 for this estimate was 19.5% suggesting low heterogeneity (See Figure S9).

Remission beyond the intervention was assessed in five trials,[20;37-39;53] and the relative risk of lack of remission was 0.95 (95% CI 0.82 to 1.11; P=0.53) with an I^2 of 0.0% (See Figure S10).

GRADE assessments

The GRADE assessments are presented in Table 4, and quality of evidence for both primary and secondary outcomes was very low or low.

Additional analysis

Four studies reported change in scores from baseline with corresponding SD's, and one study reported mean difference between groups post-intervention. Comparing the effect size of these five studies with the remaining did not explain part of the heterogeneity (p = 0.23).

Discussion

Thirty-one clinical trials allocating more than 2400 participants diagnosed with depression according to validated diagnostic instruments were included in the present systematic review. Pooled estimates suggested moderate antidepressant effect assessed both as a continuous outcome and as lack of remission. Due to risk of bias, inconsistency of effect estimates, and publication bias we have, however, very little confidence in these effect estimates. Subgroup analyses exploring reasons for the heterogeneity found that trials potentially having less risk of bias than other trials at high risk of bias had no effect of exercise on depression. Furthermore, duration of intervention and trial size were inversely associated with effect estimates. Exercise did not improve quality of life or depression or remission after the intervention. Serious adverse event or adverse events were reported inconsistently and only by a few trials not permitting firm conclusions regarding these outcomes.

Strengths and limitations

The strengths of this systematic review are that it is based on the published protocol, a comprehensive search strategy, and the inclusion of patient centered outcomes such as quality of life as well as adverse events. Also, to avoid spurious finding from repeated testing, Trial Sequential Analysis and Bayes analysis were undertaken and these analyses did not suggest that the pooled estimates could be reduced to random errors for effect on depression severity or no remission. Neither Trial Sequential Analysis nor Bayes

factor analysis are, however, able to wash of spurious effects induced by bias, or fraud or other reasons.[25;28;81-83] Had we restricted the Trial Sequential Analysis to trials of potentially lower risk of bias, the number of trials and participants would be limited and we had seen evidence far from crossing boundaries for benefit, harms, or futility. The conclusions for serious adverse events and adverse events were associated with wide confidence intervals due to lack of data and firm conclusions for these outcomes are presently not available.

The number of trials with adequate allocation concealment was 39% in the current systematic review compared to only 15.1% in trials assessing non-drug interventions for depression.[84] Blinded outcome assessment was performed in 52% of the included trials compared to 44% in non-drug antidepressant trials in general.[84] The incomplete outcome bias domain was adequate in 48% of our included trials compared to 32.9% of antidepressant non-drug trials in general.[84] Compared to non-drug trials assessing interventions for participants with depression, the included exercise trials have more bias domains with low risk of bias. However, all our included trials were at high risk of bias. Two trials had low risk of bias for all bias domains except for blinding of participants and trial personnel, and four trials fulfilled our criteria for trials at potentially less risk of bias than the rest of the trials with at risk of bias. Despite a search strategy including bibliographical databases and trials from China and South-America, the vast majority of included trials were conducted in north America and western Europe, which is comparable to the geographical distribution of non-drug trials in general[84] limiting the applicability to other geographic regions.

All outcomes for the primary analysis reflect depression severity, however, the different psychometrics may represent different aspects of depression not reflected in the pooled estimate. An in-depth discussion of the included assessment scales is beyond the scope of this review, but in the current systematic review we

found no significant differences of effect estimates from trials using HAM-D_{17} compared to trials using other assessment scales (data not shown).

The effect of exercise on depression

Our present results are similar to the latest Cochrane review by Cooney et al. (2013)[23] who found a moderate effect of exercise on depressive symptoms (-0.62 SMD) when including all trials and no effect when restricting the analysis to trials with less risk of bias (-0.18 SMD). The Cochrane review did find evidence of a small antidepressant effect beyond the intervention, which we could not confirm in our present systematic review. Bridle et al. (2012)[12] included 9 trials allocating old (> 60 years) participants with depression to exercise interventions versus control interventions. Restricting the analysis to four trials at lower risk of bias they found small to moderate effect estimates (SMD -0.34) in favour of exercise. The studies by Cooney et al.[23] and Bridle et al.[12] both included trials allocating participants with depressive symptoms and not necessarily diagnosed using a validated diagnostic system, potentially explaining the differences in the effect sizes. However, in our present systematic review the estimate for four trials at potential less risk of bias than the remaining trials was -0.11 SMD and in the Cooney study the effect estimate for eight trials with lower risk of bias was -0.18 SMD[23] compared to -0.34 in the study by Bridle at al.[12] Meta-analysis of randomised clinical trials assessing the effects of exercise for depression consistently finds positive effects, however, when restricting the analysis to trials with less risk of bias the pooled effect sizes becomes very small or negligible. Meta-analysis examining the effect of exercise beyond the intervention also finds no or small effects of exercise. In the process of interpretation of effect estimates in the current research field, it is important to recognise that effect estimates from trials with non-blinded outcome assessment are at high risk of bias as reported by Savovic et al.[85] Thirteen of 31 trials in the current systematic review did not use blinded outcome assessment. In contradiction to the current systematic review, a recent meta-analysis by Schuch et al.[11] concluded that “exercise has a large and

significant antidepressant effect in people with depression……Our data strongly support the claim that exercise is an evidence-based treatment for depression". This statement was based on a meta-analysis of 25 randomised clinical trials including participants with depression or depressive symptoms to exercise or control conditions and excluding trials using any form of active control group. Surprisingly, the authors found that adjusting for publication bias using the Trim and Fill procedure[30] the estimate *increased* from a SMD of 0.98 to 1.11. The effect in SMD in included studies ranged from -0.23 to 4.56 representing considerable heterogeneity.[11] The authors classified four trials as having lower risk of bias using the same criteria as in our systematic review and 21 trials as having high risk of bias. This illustrates some of the challenges in meta-analysis of exercise and depression: the large heterogeneity driven by small studies inflating the effects of random-effects analysis,[86] the misconception that we can restrict our analysis to statistics and not consider the evident effect of bias.[22;85] Compared to our previous review,[10] we now included 31 trials including 2419 participants versus previously 13 trials and 687 participants. It may seem as a paradox that this large increase in data has not provided us with a similar increase in certainty of conclusions reflected by heterogeneity of trial results as well as our conclusions from the systematic reviews. The increase in available data is, however, primarily provided by small trials at high risk of bias introducing exaggerated effect estimates. In the current systematic review, we included four trials with 530 participants at lower risk of bias compared to three trials with 239 participants in our previous review, reflecting that only a small part of the additional data comes from trials at lower risk of bias. The continuous increase in data associated with high risk of bias will not provide patients, clinicians or policymakers with adequate information and represents an unethical enrollment of trial participants and waste of resources.[87-93] We therefore recommend that future systematic reviews and meta-analysis a priori should have a primary outcome restricting effect analysis to larger trials with lower risk of bias and that any recommendations regarding exercise interventions for participants with depression should be assessed with the GRADE framework.

The I^2 of 82% and 71% for the primary outcomes indicate substantial evidence of heterogeneity of intervention effects that is variation in effect estimates beyond chance. Part of this heterogeneity was explained by bias and by trial size: trials at high risk of bias or small trials have very large effect estimates compared to trials potentially at less bias risk compared to the remaining trials at high risk of bias or larger trials. The funnel plots end Egger's test indicates publication bias, however, the association between trial size and effect estimates could suggest that the asymmetry in the funnel plots are due to small study bias rather than publication bias.[94] It could be argued that both the delivery of exercise as well as the actual increase in fitness are fundamental to the assessment of the antidepressant effects of exercise, and in line with our previous review we found duration of intervention inversely associated with effect size.[10] Comparing different exercise intensities, we did find a small effect of high intensity exercise compared to lower intensity exercise. However, assessing delivered exercise expressed as increase in maximal oxygen uptake we could not reproduce this finding. Future trials need to pay more attention to the dose of the intervention as well as compliance with intervention.[95] We suggest using maximal oxygen uptake or 1 repetition maximum as the gold standard to assess the received exercise. Several studies compare exercise to control interventions rather than wait-list control to reduce the effect of non-specific effects, e.g., the DEMO trials and Mather et al.[20;21;51] Also, it could be speculated that the effect of exercise would be harder to detect if participants also received medical treatment in addition. The current systematic review could not confirm that the type of control condition explained heterogeneity. The discussion of control group is important in non-drug trials: choosing a waitlist control group the results potentially reflects non-specific effects, choosing an active control group (e.g., relaxation exercise) the trial is potentially a comparison between to active treatments. However, in the current systematic review we found no evidence that trials using an attention control group or exercise as add-on to pharmacotherapy had significantly different effect estimates compared to other trials.

Our systematic review did not find indications of a positive effect on quality of life in participants with depression allocated to exercise interventions, which is in concordance with the review by Cooney et al.[23] Only 3/31 trials reported on serious adverse events, and found no significant risk of death or suicide attempt. No indication of increased severity of depression or other adverse events in participants allocated to exercise could be detected. However, data on adverse events was reported sporadically in a minority of trials and currently it is not possible to conclude on the risk of serious adverse events or adverse event from exercise interventions in participants with depression.

Conclusions

We have little confidence in the pooled effect estimates, especially because trials with less than high risk of bias produced significantly lower effect estimates, suggesting that exercise interventions only produce small or negligible antidepressant effects, depending on how much of the effect is caused by bias and how much is caused by the intervention. There was no effect of exercise on quality of life or depression beyond the intervention itself. There is currently no evidence in favour of exercise for patients with depression with a view to ameliorate depressive symptoms and at we do not recommend that exercise is prescribed to relieve depressive symptoms. Our systematic review did not evaluate possible beneficial effects of exercise on, e.g., metabolism or cardiovascular fitness,[21;96] and it is possible that exercise may have beneficial effects on these factors in patients diagnosed with depression.

Future perspectives

Despite the large number of published trials, further trials with more robust methodology seem still required to establish progress in this field. Also, additional trials from outside North-America and Europe may be required for results to be valid for patients in Asia, Africa, and South-America. To further elaborate

on the current findings, we recommend that future trials must include blinded outcome assessors and outcomes assessing quality of life, metabolic effects, and long-term effects beyond the intervention. It is also important that future trials systematically collect and report data on death, suicide events, musculoskeletal injuries and other potential adverse effects in both the intervention group as well as in the control group. Moreover, future trials ought to be designed according to the SPIRIT guidelines and reported according to the CONSORT guidelines[97;98] and transparently report deidentified individual participant data enabling individual participant data meta-analyses.[99]

Funding

No funding was obtained for this systematic review.

Competing interests

JK, CG, and MN have previously published two trials and a meta-analysis on this topic, which could introduce an academic bias in the current systematic review. We asked new authors (HS and CH) to be involved in the preparation of the protocol, trial selection and bias assessment. No support from any organisation was received for the submitted work; no financial relationship with any organisations that might have an interest in the submitted work in the previous three years, and apart from the above no other relationship or activities that could appear to have influenced the submitted work.

Acknowledgement

We appreciate the help from Youling He for help with the Chinese Wanfang bibliographical database and translation of Chinese papers. We thank Janus C Jakobsen for assistance with the calculation of Bayes factor.

Contributors

JK conceived the project, collected data, did the statistical analysis, analysed the data, drafted and revised the manuscript. He is guarantor. CH collected the data, analysed the data and revised the

manuscript. HS conceived the project, collected data, analysed the data, and revised the manuscript. CG conceived the project, analysed the data and revised the manuscript.MN conceived the project, analysed the data, and revised the manuscript.

Data sharing statement

All data used in this study are available in Figures and Fables. No other data were used.

Reference List

(1) Lepine JP, Gastpar M, Mendelwicz J, Tylee A. Depression in the community: The first pan-european study DEPRES (Depression Research in European Society). *Int Clin Psychopharmocol* 1997;12:19-29.

(2) Pedersen CB, Mors O, Bertelsen A, *et al.* A comprehensive nationwide study of the incidence rate and lifetime risk for treated mental disorders. *JAMA Psychiatry* 2014;71(5):573-81.

(3) Ustun TB, Ayuso-Mateos JL, Chatterji S, *et al.* Global burden of depressive disorders in the year 2000. *Br J Psychiatry* 2004;184:386-92.

(4) Kirsch I, Deacon B, Huedo-Medina T, *et al.* Initial severity and antidepressant benefits: A meta-analysis of data submitted to the Food and Drug Administration. *PLoS Med* 2008;5[2], e45. doi:10.1371/journal.pmed.0050045.

(5) Turner EH, Matthews AM, Linardatos E, *et al*. Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy. *N Engl J Med* 2008; 358(3):252-60.

(6) Jakobsen JC, Hansen JL, Storebo OJ, *et al.* The effects of cognitive therapy versus 'no intervention' for major depressive disorder. *PLoS One* 2011;6(12):e28299.

(7) Jakobsen JC, Hansen JL, Simonsen E, Gluud C. The effect of interpersonal psychotherapy and other psychodynamic therapies versus 'treatment as usual' in patients with major depressive disorder. *PLoS One* 2011; 6(4):e19044.

(8) Jakobsen JC, Hansen JL, Simonsen E, Gluud C. The effect of adding psychodynamic therapy to antidepressants in patients with major depressive disorder. A systematic review of randomized clinical trials with meta-analyses and trial sequential analyses. *J Affect Disord* 2012;137(1-3):4-14.

(9) Cipriani A, Furukawa TA, Salanti G, *et al.* Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *Lancet* 2009;373[9665]:746-58.

(10) Krogh J, Nordentoft M, Sterne J, Lawlor D. The effect of exercise in clinically depressed adults: Systematic review and meta-analysis of randomized controlled trials. *J Clin Psychiatry* 2011;72(4):529-38.

(11) Schuch FB, Vancampfort D, Richards J, *et al.* Exercise as a treatment for depression: A meta-analysis adjusting for publication bias. *J Psychiatr Res* 2016;77:42-51.

(12) Bridle C, Spanjers K, Patel S, *et al.* Effect of exercise on depression severity in older people: systematic review and meta-analysis of randomised controlled trials. *Br J Psychiatry* 2012; 201(3):180-185.

(13) Herring MP, Puetz TW, O'Connor PJ, Dishman RK. Effect of exercise training on depressive symptoms among patients with a chronic illness: a systematic review and meta-analysis of randomized controlled trials. *Arch Intern Med* 2012; 172(2):101-111.

(14) World Health Organisation. International Statistical Classification of Diseases, 10th Revision (ICD-10). *Geneva, Switzerland* 1992.

(15) American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, 4th ed. *Washington DC* 1994.

(16) Beck AT, Steer RA, Brown GK. BDI-II manual. 2 ed. New York: Psychological Corporation, 1996.

(17) Krogh J, Speyer H, Gluud C, Nordentoft M. Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential analysis. *Syst Rev* 2015; 4(1):40.

(18) Spitzer RL, Endicott J, Williams JB. Research diagnostic criteria. *Arch Gen Psychiatry* 1979; 36(12):1381-1383.

(19) ICH-secretariat. International Conference on Harmonisation (ICH) Harmonised Tripartite Guideline (Step 4). Guideline for Good Clinical Practice. *Geneva* 1996.

(20) Krogh J, Saltin B, Gluud C, Nordentroft M. The DEMO trial: A randomized, parallel-group, observer-blinded clinical trial of strength versus aerobic versus relaxation training for patients with mild to moderate depression. *J Clin Psychiatry* 2009;70[6]: 790-00.

(21) Krogh J, Videbech P, Thomsen C, *et al.* DEMO-II trial.Aerobic Exercise Versus Stretching Exercise in Patients With Major Depression - A Randomized Clinical Trial. *PLoS One* 2012;7(10): e48316. doi:10.1371/journal.pone.0048316

(22) Higgins JPT. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0. The Cochrane Collaboration. Available from www.cochrane-handbook.org The Cochrane Collaboration, 2009.

(23) Cooney GM, Dwan K, Greig CA, *et al.* Exercise for depression. *Cochrane Database Syst Rev* 2013; 9:CD004366.

(24) Higgins JPT, Thompson SG, Deeks JJ, Altmann DG. Measuring inconsistency in meta-analysis. *BMJ* 2003; 327:557-60.

(25) Wetterslev J, Thorlund K, Brok J, Gluud C. Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. *J Clin Epidemiol* 2008; 61(1):64-75.

(26) Brok J, Thorlund K, Gluud C, Wetterslev J. Trial sequential analysis reveals insufficient information size and potentially false positive results in many meta-analyses. *J Clin Epidemiol* 2008; 61(8):763-69.

(27) Thorlund K, Walter SD, Johnston BC, *et al.* Pooling health-related quality of life outcomes in meta-analysis-a tutorial and review of methods for enhancing interpretability. *Res Synth Methods* 2011; 2(3):188-03.

(28) Jakobsen JC, Wetterslev J, Winkel P, *et al.* Thresholds for statistical and clinical significance in systematic reviews with meta-analytic methods. *BMC Med Res Methodol* 2014; 14:120.

(29) Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. *BMJ* 1999; 319(7211):670-74.

(30) Duval S, Tweedie R. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* 2000; 56(2):455-63.

(31) Guyatt GH, Oxman AD, Vist GE, *et al.* GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008; 336(7650):924-26.

(32) Andrews J, Guyatt G, Oxman AD, *et al.* GRADE guidelines: 14. Going from evidence to recommendations: the significance and presentation of recommendations. *J Clin Epidemiol* 2013; 66(7):719-25.

(33) Zheng YP, Lin KM, Zhao JP, *et al.* Comparative study of diagnostic systems: Chinese Classification of Mental Disorders-Second Edition versus DSM-III-R. *Compr Psychiatry* 1994; 35(6):441-49.

(34) Babyak MA, Blumenthal JA, Herman S, *et al.* Exercise treatment for major depression: Maintenance of therapeutic benefit at 10 months. *Psychosom Med* 2000; 62:633-38.

(35) Baxter H, Winder R, Chalder M, *et al.* Physical activity as a treatment for depression: the TREAD randomised trial protocol. *Trials* 2010; 11:105.

(36) Benson M, Blumenthal J, Babyak M, *et al.* Exercise fails to improve neurocognition in depressed middle-aged and older adults. *Med Sci Sports Exerc* 2008;40[7]:1344-352.

(37) Blumenthal J, Babyak MA, Moore KA, *et al.* Effects of exercise training on older patients with major depression. *Arch Int Med* 1999;159:2349-356.

(38) Blumenthal J, Babyak M, Doraiswamy P, *et al.* Exercise and pharmacotherapy in the treatment of major depressive disorder. *Psychosom Med* 2007;69:587-96.

(39) Chalder M, Wiles NJ, Campbell J, *et al.* Facilitated physical activity as a treatment for depressed adults: randomised controlled trial. *BMJ* 2012; 344:e2758.

(40) Chalder M, Wiles NJ, Campbell J, *et al.* A pragmatic randomised controlled trial to evaluate the cost-effectiveness of a physical activity intervention as a treatment for depression: the treating depression with physical activity (TREAD) trial. *Health Technol Assess* 2012; 16(10):1-iv.

(41) Combs K, Smith PJ, Sherwood A, *et al.* Impact of sleep complaints and depression outcomes among participants in the standard medical intervention and long-term exercise study of exercise and pharmacotherapy for depression. *J Nerv Ment Dis* 2014;202(2):167-71.

(42) Doyne EJ, Ossip-Klein DJ, Bowman ED, *et al.* Running versus weight lifting in the treatment of depression. *J Consult Clin Psychol* 1987;55(5):748-54.

(43) Dunn AL, Trivedi MH, Kampert JB, *et al.* The DOSE study: a clinical trial to examine efficacy and dose response of exercise as treatment for depression. *Control Clin Trials* 2002;23(5):584-03.

(44) Dunn AL, Trivedi MH, Kaplan K, *et al.* Exercise treatment for depression. Efficacy and dose response. *Am J Prev Med* 2005;28:1-8.

(45) Herman S, Blumenthal JA, Babyak M, *et al.* Exercise therapy for depression in middle-aged and older adults: predictors of early dropout and treatment failure. *Health psychology* 2002;21:553-63.

(46) Ho CWH, Chan SC, Wong JS, *et al.* Effect of aerobic exercise training on chinese population with mild to moderate depression in Hong Kong. *Rehabil Res Pract* 2014;2014:627376

(47) Hoffman BM, Babyak MA, Craighead WE, *et al*. Exercise and pharmacotherapy in patients with major depression: one-year follow-up of the SMILE study. *Psychosom Med* 2011;73(2):127-33.

(48) Kerling A, Tegtbur U, Gutzlaff E, *et al.* Effects of adjunctive exercise on physiological and psychological parameters in depression: A randomized pilot trial. *J Affect Disord* 2015;177:1-6.

(49) Khatri P, Blumenthal J, Babyak M, *et al.* Effects of exercise training on cognitive functioning among depressed older men and women. *J Aging Phys Act* 2001;9:43-57

(50) Krogh J, Petersen L, Timmermann M, *et al.* Design paper: The DEMO trial: A randomized, parallel-group, observer-blinded clinical trail of aerobic versus non-aerobic versus relaxation training for patients with light to moderate depression. *Contemp Clin Trials* 2007;28:79-89.

(51) Mather AS, Rodriguez C, Guthrie MF, *et al.*. Effects of exercise on depressive symptoms in older adults with poorly responsive depressive disorder: randomised controlled trial. *Br J Psychiatry* 2002;180:411-15.

(52) Mota-Pereira J, Silverio J, Carvalho S, et al. Moderate exercise improves depression parameters in treatment-resistant patients with major depressive disorder. *J Psychiatr Res* 2011; 45(8):1005-011.

(53) Pfaff JJ, Alfonso H, Newton RU, *et al.* ACTIVEDEP: a randomised, controlled trial of a home-based exercise intervention to alleviate depression in middle-aged and older adults. *Br J Sports Med* 2014;48(3):226-32.

(54) Pilu A, Sorba M, Hardoy M, *et al.* Efficacy of physical activity in the adjunctive treatment of major depressive disorders: preliminary results. *Clin Pract Epidemio Ment Health* 2007;3:8.

(55) Schuch FB, Vasconcelos-Moreno MP, Borowsky C, Fleck MP. Exercise and severe depression: preliminary results of an add-on study. *J Affect Disord* 2011;133(3):615-18.

(56) Singh NA, Clements KM, Fiatarone MA. A randomized controlled trial of the effect of exercise on sleep. *Sleep* 1997; 20(2):95-01.

(57) Singh NA, Stavrinos TM, Scarbek Y, *et al.* A randomized controlled trial of high versus low intensity weight training versus general practitioner care for clinical depression in older adults. *J Gerontol A Biol Sci Med Sci* 2005; 60(6):768-776.

(58) Singh NA, Clements KM, Singh MA. The efficacy of exercise as a long-term antidepressant in elderly subjects: a randomized, controlled trial. *J Gerontol A Biol Sci Med Sci* 2001; 56(8):M497-M504.

(59) Singh N, Clements K, Fiatarone M. A randomized controlled trial of progressive resistance training in depressed elders. *J Gerontol* 1997; 52:M27-M35.

(60) Toni G, Belvederi MM, Zanetidou S, *et al.* Physical exercise for late life Major Depression: The SEEDS study. *Eur J Prev Cardiol* 2014; 21:S107.

(61) Veale D, Le FK, Pantelis C, *et al*. Aerobic exercise in the adjunctive treatment of depression: a randomized controlled trial. *J R Soc Med* 1992; 85(9):541-544.

(62) Vieira JLL, Porcu M, da Rocha PGM. The regular physical exercise's practice as complementary therapy to the treatment of depressed women. *J Bras Psiquiatr* 2007;56(1):23-8.

(63) Epstein D. Reno: University of Nevada:Aerobic activity versus group cognitive therapy: An evaluative study of contrasting interventions for the alleviation of clinical depression. *(dissertation)* 1986.

(64) Danielsson L, Papoulias I, Petersson E, *et al.* Exercise or basic body awareness therapy as add-on treatment for major depression: A controlled study. *J Affect Disord* 2014;168:98-06.

(65) Carneiro L, Fonseca A, Vieira-Coelho M, *et al*. Effects of structured exercise and pharmacotherapy vs pharmactherapy for adults with depressive symptoms: A randomized clinical trial. *J Psychitr Res* 2015;(71):48-55.

(66) Legrand FD, Neff EM. Efficacy of exercise as an adjunct treatment for clinically depressed inpatients during the initial stages of antidepressant pharmacotherapy: An open randomized controlled trial. *J Affect Disord* 2016;191:139-44.

(67) Doose M, Ziegenbein M, Hoos O, *et al.* Self-selected intensity exercise in the treatment of major depression: A pragmatic RCT. *Int J Psychiatry Clin Pract* 2015;19(4):266-75.

(68) Salehi I, Hosseini SM, Haghighi M, *et al.* Electroconvulsive therapy (ECT) and aerobic exercise training (AET) increased plasma BDNF and ameliorated depressive symptoms in patients suffering from major depressive disorder. *J Psychiatr Res* 2016;76:1-8.

(69) Huipeng F, Xiaohui Z. Curative effect and observation of exercise on patients with depression. *Chin Med Pharmacy* 2013;3:187-88.

(70) Jinchun W, Bo Y, Rong X, *et al.* Effect of exercise therapy in improving social function and quality of life in patients with depression. *Nurs J Chin PLA* 2015; 32(21):24.

(71) Guifeng S. The role of setting - up exercise on rehabilitation in patients with depressive episode. *Chin J Health Psychol* 2015;23:184-186.

(72) Fang D, Min R, Ai-qin W. The effect of aerobic exercise with different intensity and frequency on depressive symptoms. *Chin J Behav Med & Brain Sci* 2013;22:830-32.

(73) Martinsen EW, Medhus A, Sandvik L. Effects of aerobic exercise on depression: A controlled study. *BMJ* 1985; 291:109.

(74) Veale D, Fevre K, Pantalis C, *et al.* Aerobic exercise in the adjunctive treatment of depression: A randomised controlled trial. *J R Soc Med* 1992;85:541-44.

(75) Ossip-Klein DJ, Doyne EJ, Bowman ED, *et al.* Effects of running or weight lifting on self-concept in clinically depressed women. *J Consult Clin Psychol* 1989;57(1):158-61.

(76) Blumenthal JA, Babyak MA, Moore KA, *et al.* Effects of exercise training on older patients with major depression. *Arch Intern Med* 1999; 159(19):2349-356.

(77) Belvederi MM, Amore M, Menchetti M, *et al.* Physical exercise for late-life major depression. *Br J Psychiatry* 2015; 207(3):235-42.

(78) Schuch FB, Vasconcelos-Moreno MP, Borowsky C, *et al.* Exercise and severe major depression: effect on symptom severity and quality of life at discharge in an inpatient cohort. *J Psychiatr Res* 2015; 61:25-32.

(79) Carta MG, Hardoy MC, Pilu A, *et al.* Improving physical quality of life with group physical activity in the adjunctive treatment of major depressive disorder. *Clin Pract Epidemiol Ment Health* 2008; 4:1.doi: 10.1186/1745-0179-4-1.

(80) Klein MH, Greist JH, Gurman AS, *et al.* A comparative outcome study of group psychotherapy vs. exercise treatments for depression. *Int J Ment Health* 1985;13:148-77.

(81) Roberts I, Ker K, Edwards P, et al. The knowledge system underpinning healthcare is not fit for purpose and must change. *BMJ* 2015; 350:h2463.

(82) Gluud C, Jakobsen J, Imberger G, *et al.* Re: The knowledge system underpinning healthcare is not fit for purpose and must change - Responses to the opposing viewpoints of Roberts and colleagues and Tovey and colleagues . *BMJ* 2015; 350(h2463).

(83) Thorlund K, Engstrøm J, Wetterslev J, Brok J, Imberger G, Gluud C. User manual for Trial Sequential Analysis (TSA). *Copenhagen Trial Unit ,Centre for Clinical Intervention Research, Copenhagen, Denmark P1-115, Available from https://www ctu dk/tsa* 2011.

(84) Krogh J, Hjorthoj CR, Jakobsen JC, *et al.* DEPERROR: Risks of systematic errors in drug and non-drug randomized clinical trials assessing intervention effects in patients with unipolar depression. *J Affect Disord* 2015; 179:121-27.

(85) Savovic J, Jones HE, Altman DG, *et al*. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. *Ann Intern Med* 2012; 157(6):429-38.

(86) Poole C, Greenland S. Random-effects meta-analyses are not always conservative. *Am J Epidemiol* 1999; 150(5):469-75.

(87) Ioannidis JP, Greenland S, Hlatky MA, *et al.* Increasing value and reducing waste in research design, conduct, and analysis. *Lancet* 2014; 383(9912):166-75.

(88) Chan AW, Song F, Vickers A, *et al.* Increasing value and reducing waste: addressing inaccessible research. *Lancet* 2014; 383(9913):257-66.

(89) Moher D, Glasziou P, Chalmers I, *et al.* Increasing value and reducing waste in biomedical research: who's listening? *Lancet* 2016; 387(10027):1573-586.

(90) Glasziou P, Altman DG, Bossuyt P, *et al.* Reducing waste from incomplete or unusable reports of biomedical research. *Lancet* 2014; 383(9913):267-76.

(91) Chalmers I, Bracken MB, Djulbegovic B, Garattini S, Grant J, Gulmezoglu AM et al. How to increase value and reduce waste when research priorities are set. *Lancet* 2014; 383(9912):156-65.

(92) Al-Shahi SR, Beller E, Kagan J, *et al.* Increasing value and reducing waste in biomedical research regulation and management. *Lancet* 2014;383(9912):176-85.

(93) Garattini S, Jakobsen JC, Wetterslev J, *et al.* Evidence-based clinical practice: Overview of threats to the validity of evidence and how to minimise them. *Eur J Intern Med* 2016;32:13-21.

(94) Sterne JA, Sutton AJ, Ioannidis JP, *et al.* Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ* 2011; 343:d4002.

(95) Slade SC, Dionne CE, Underwood M, *et al.* Consensus on Exercise Reporting Template (CERT): Explanation and Elaboration Statement. *Br J Sports Med* 2016.

(96) Stubbs B, Rosenbaum S, Vancampfort D, *et al.* Exercise improves cardiorespiratory fitness in people with depression: A meta-analysis of randomized control trials. *J Affect Disord* 2016; 190:249-53.

(97) Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. *Ann Intern Med* 2010; 152(11):726-32.

(98) Chan AW, Tetzlaff JM, Gotzsche PC, *et al.* SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. *BMJ* 2013; 346:e7586.

(99) Skoog M, Saarimäki J, Gluud C, *et al.* Transparency and registration in clinical research in the Nordic countries. *Nordic Trial Alliance, NordForsk*; 2015 1-108.

Figure 1. Effect of exercise on depression severity in patients diagnosed with depression

Table 1. Characteristics of trials assessing exercise for patients diagnosed with depression

Author, first Country of origin	Participants	Severity of depression at baseline	N at baseline (included in trial efficacy analysis)	Type of intervention	Frequency	Duration
Klein 1985 USA	Outpatients Mean age: 30 (SD 7) 72% female	SCL-D: 2.4 (SD 1)	50 (22)	*Aerobic exercise:* Supervised individual running. *Control group*: Supervised meditation in groups	2 sessions per week Control group: 1 session per week	12 weeks
Martinsen 1985 Norway	Inpatients Mean age: 40 (range 17-60 Distribution of sex not reported	BDI: 28.0 (SD 9)	49(43)	*Aerobic exercise*: Supervised group exercise. *Control group*: Occupational therapy.	3 sessions per week Control group: 3 sessions per week	9 weeks
Epstein 1986 USA	Outpatients Mean age: 39 (range 24 to 60) (NR) % female	BDI: 23.4 (SD 7)	21 (17)	*Aerobic exercise:* Supervised group exercise. *Control group*: Waitlist control.	3 sessions per week	8 weeks
Doyne 1987 USA	Outpatients Mean age: 29 (SD 4) 100 % female	HAM-D_{17}: 13.0 (SD 7)	52 (25)	*Aerobic exercise OR weightlifting*: Supervised individual exercise. *Control group*: Waiting list.	4 sessions per week	8 weeks
Veale 1992 UK	Outpatients Mean age: 35 (range 19-58) 64% female	BDI: 24.5 (SD 6)	83 (65)	*Aerobic exercise*: Supervised group exercise. *Control group*: Standard treatment from psychiatric services.	3 sessions per week	12 weeks
Singh 1997 USA	Outpatients Recruited from a register of volunteers Mean age: 71 (SD 1)	BDI: 19.9 (SD 2.3)	32 (32)	*Progressive resistance training*: Supervised group exercise. *Control group*: Attended seminars on health.	3 sessions per week Control group: 2 sessions per week	10 weeks
Blumenthal 1999 USA	Outpatients Mean age: 57 (SD 7) 71.8% female	HAM-D_{17}: Not reported	103 (103)	*Aerobic exercise*: Supervised exercise plus antidepressant medication (sertraline). *Control group*: Antidepressant medication (sertraline).	3 sessions per week	16 weeks
Mather 2002 UK	Outpatients Treatment resistant Mean age: 65 (range 53-91) 69% female	HAM-D_{17}: 17.1 (SD 6)	86 (85)	*Mixed aerobic and non-aerobic exercise*: Supervised group exercise. *Control group*: Attended health seminars.	2 sessions per week Control group: 2 seminars per week	10 weeks
Dunn 2005 USA	Outpatients Mean age: 36 (SD 6)	HAM-D_{17}: 19.4 (SD 2)	80 (80)	*Aerobic exercise*: Individually supervised	Group (1) and (2): 3 sessions	12 weeks

	75% female			exercise with (1) low energy expenditure (EE) OR (2) high EE OR (3) low EE OR (4) high EE. *Control group*: Flexibility exercise.	per week Group (3) and (4): 5 sessions per week Control group: 3 sessions per week	
Singh 2005 Australia	Outpatients Mean age: 69 (SD 6) 55% female	HAM-D_{17}: 18.9 (SD 4.2)	60 (54)	*Progressive resistance training (PRT)*: (1)Low intensity PRT OR (2) high intensity PRT. *Control group*: Standard GP care.	Group (1) and (2): 3 sessions per week	8 weeks
Pilu 2007 Italy	Outpatients Treatment resistant Age between 40 and 60 100% female	HAM-D_{17}: 19.7 (SD 6)	30 (30)	*Resistance exercise*: Supervised group sessions. *Control group*: Standard treatment.	2 sessions per week	32 weeks
Viera 2007 Brazil	Outpatients Mean age 43.66 (SD NR) 100% female	HAM-D_{21}: 31.9 (SD 3)	18 (18)	*Aerobic exercise:* Supervised water aerobics. *Control group*: Standard GP care.	2 sessions per week	12 weeks
Blumenthal 2007 USA	Outpatients Mean age: 52 (SD 8) 75.8% female	HAM-D_{17}: 16.7 (SD 4)	153 (153)	*Aerobic exercise*: (1) Supervised group exercise OR (2) home-based exercise. *Control group*: Placebo medication.	(1) and (2): 3 sessions per week	16 weeks
Krogh 2009 Denmark	Outpatients Mean age: 39 (SD 9) 74% female	HAM-D_{17}: 17.8 (SD 4)	165 (165)	*Exercise*: (1) Aerobic supervised group exercise OR (2) supervised group resistance training *Control group*: relaxation and stretching exercise.	(1)and (2): 2 sessions per week Control group: 2 sessions per week	16 weeks
Mota-Pereira 2011 Portugal	Outpatients Treatment resistant Mean age: 47.5 (SD 3) 65.5% female	HAM-D_{17}: 17.1 (SD 3)	33 (29)	*Aerobic exercise:* Homebased exercise + supervised. *Control group*: Attention control.	4 home-based sessions/week. 1 supervised session/week Control group: 1 supervised session/week	12 weeks
Krogh 2012 Denmark	Outpatients Mean age: 42 (SD 11) 67% female	HAM-D_{17}: 18.9 (SD 4)	115 (115)	*Aerobic exercise*: Supervised group exercise. *Control group*: Supervised stretching exercise in groups.	3 sessions per week Control group: 3 sessions per week	12 weeks
Chalder 2012 UK	Outpatients Mean age: 40 (SD 13) 66% female	BDI: 32.1 (SD 9)	361 (361)	*Exercise*: Participants received individually tailored support and encouragement to engage in physical activity. *Control group*: Standard GP care.	Individual	16 weeks

Fang 2013 China	Inpatients Mean age: 44 (SD 14) 66.9% female	HAM-D_{24}: 29.2 (SD 5)	90 (90)	*Aerobic exercise:* Group 1 and 2 had supervised group exercise, high intensity. *Control group:* 15 min stretching	Group 1 and 2 had 3 and 5 sessions per week, respectively Control group: 3 sessions per week	6 weeks
Huipeng 2013 China	Inpatients Mean age: 30 (SD 5) 100% female	HAM-D_{17}: 28 (SD 5)	68 (68)	*Aerobic exercise:* Jogging *Control group:* Standard treatment	5 sessions per week	6 weeks
Cassandra 2014 Honkong	Inpatients Mean age: 46 (SD 12) 67.3% female	MADRS: 19 (10)	52 (52)	*Aerobic exercise:* Supervised exercise. *Control group*: 10 min stretching.	5 sessions per week	3 weeks
Danielsson 2014 Sweden	Outpatients Mean age: 45 (SD 13) 76% female	MADRS: 24.0 (SD 5)	42 (42)	*Mixed aerobic and non-aerobic exercise*: First two weeks individual supervised exercise then supervised group exercise. *Control group*: One session with advice on physical activity.	2 sessions per week	10 weeks
Pfaff 2014 Australia	Outpatients Mean age: 61 (SD 8) 63% female	MADRS: 21.3 (SD NR)	200 (200)	*Resistance exercise*: Supervised home-based exercise *Control group*: Standard GP care	3 sessions per week	12 weeks
Guifeng 2015 China	Inpatients Mean age: 33 (SD 14) 70% female	HAM-D_{24}: 25.9 (SD 4)	70 (70)	*Aerobic exercise:* Supervised group exercise *Control group:* Standard treatment	5 sessions per week	8 weeks
Junchin 2015 China	Inpatients Mean age: 28 (SD 7) 61% female	HAM-D_{24}: 25.8 (SD 3)	70 (70)	*Aerobic exercise:* Supervised aerobic exercise of the patients own choice Control group: Standard treatment	5 sessions per week	8 weeks
Schuch 2015 Brazil	Inpatients Mean age: 40 (SD 11) 74% female	HAM-D_{17}: 26.7 (SD 2)	50 (50)	*Aerobic exercise:* Supervised individual exercise. *Control group*: Standard treatment.	3 sessions per week	2 weeks
Kerling 2015 Germany	Inpatients Mean age: 43 (SD 10)	MADRS: 24.0 (SD 9)	42 (42)	*Aerobic exercise:* Supervised exercise. *Control group*: Standard treatment.	3 sessions per week	6 weeks
Belvederi 2015 Italy	Outpatients Mean age: 75 (SD 6) 71% female	HAM-D_{17}: 20.1 (SD 3)	121 (121)	*Aerobic exercise:* (1) Sertraline + supervised non-progressive exercise OR (2) sertraline + supervised progressive aerobic	3 sessions per week	24 weeks

				exercise. *Control group*: Sertraline.		
Carneiro 2015 Portugal	Outpatients Mean age: 50.16 (SD 12) 100% female	BDI: 48.8 (SD 10)	26 (19)	*Aerobic exercise:* Supervised exercise *Control group:* Standard treatment	3 sessions per week	16 weeks
Doose 2015 Germany	Outpatients Mean age: 47.9 (SD 10.5) 63% female	HAM-D_{17}: 14.2 (SD 3)	46 (46)	*Aerobic exercise:* Supervised aerobic exercise *Control group*: Standard treatment	3 sessions per week	8 weeks
Salehi 2016 Iran	Inpatients Mean age: 30.0 (SD 6) 35% female	HAM-D_{21}: 43.4 (SD 8)	40 (40)	*Aerobic exercise + ECT:* Supervised aerobic exercise *Control group:* ECT	3 sessions per weeks Control group 3 ECTs per week	4 weeks
Legrand 2016 France	Inpatients Mean age: 46.9 (SD 13) 67% female	BDI: 36.0 (SD 6)	24 (24)	*Aerobic exercise:* Supervised aerobic exercise *Control group:* Standard treatment	10 sessions in 10 consecutive days	10 days

SCL-D: Symptom Check List, depression subscale; HAM-D_{17}: Hamilton Depression Scale, 17 items; BDI: Beck's Depression Inventory; SD: Standard deviation; ECT: Electroconvulsive therapy

Table 2. Risk of bias in trials assessing exercise for patients diagnosed with depression

Author, Year of publication	Sequence generation	Allocation concealment	Blinding of participants and trial personnel assessors	Blinding of outcome assessors	Incomplete outcome data	Selective outcome reporting	For profit bias	Other bias	Comment on 'Other bias'
Klein 1985	Unclear	Unclear	High	High	High	Low	Low	Low	
Martinsen 1985	Unclear	Unclear	High	High	High	Low	High	Low	
Epstein 1986	Unclear	Unclear	High	High	High	Low	Unclear	High	Baseline difference
Doyne 1987	Unclear	Unclear	High	Low	High	Low	Unclear	High	Baseline difference
Veale 1992	Unclear	Unclear	High	High	High	Low	Low	High	Baseline difference
Singh 1997	Low	Unclear	High	Low	Low	Low	Low	High	Baseline difference
Blumenthal 1999	Unclear	Unclear	High	Low	High	Low	High	Low	
Mather 2002	Low	Low	High	Low	High	Low	Low	Low	
Dunn 2005	Low	Low	High	Low	High	High	High	Low	
Singh 2005	Low	Low	High	Low	High	Low	Unclear	Low	
Pilu 2007	Unclear	Unclear	High	Unclear	Low	Low	Unclear	Low	
Viera 2007	Unclear	Unclear	High	Unclear	Low	Low	Unclear	Low	
Blumenthal 2007	Low	Low	High	Low	High	High	Low	Low	
Krogh 2009	Low	Low	High	Low	Low[1]	High	High	High	Baseline difference
Mota-Pereira 2011	Unclear	Unclear	High	Low	High	Low	High	High	Baseline difference
Krogh 2012	Low	Low	High	Low	Low	Low	Low	Low	
Chalder 2012	Low	Low	High	High	Low	Low	Low	Low	
Fang 2013	Unclear	Unclear	High	Unclear	Unclear	High	Unclear	Low	
Huipeng 2013	Unclear	Unclear	High	Unclear	Low	Low	Unclear	Low	
Cassandra 2014	Low	Unclear	High	Low	High	Low	Low	Low	
Danielsson 2014	Unclear	Low	High	Low	High	Low	Low	Low	
Pfaff 2014	Low	Low	High	Low	Low[1]	Low	Low	High	Baseline difference
Guifeng 2015	Unclear	Unclear	High	Unclear	Low	Low	Unclear	Low	
Jinchun 2015	Unclear	Unclear	High	Unclear	Low	Low	Unclear	Low	
Schuch 2015	Unclear	Low	High	Low	Low	Low	Low	Low	
Kerling 2015	Unclear	Unclear	High	Unclear	Low	Low	Low	Low	
Belvederi 2015	Low	Low	High	Low	High	Low	Low	High	Post-hoc sample size
Carneiro 2015	Unclear	Low	High	High	Unclear	Low	Low	Low	
Doose 2015	Unclear	Unclear	High	High	High	Low	Low	High	No sample size calc.
Salehi 2016	High	High	High	Low	Unclear	Low	Low	High	Baseline difference
Legrand 2016	Low	High	High	High	High	Low	Unclear	Low	

[1]For the outcome 'lack of remission' this bias domain was high risk of bias

Table 3. Heterogeneity of effect estimates for trials assessing the effect of exercise for patients diagnosed with depression explored by comparing sub-groups

Subgroups	Number of Trials (participants)	Random effects meta-analysis SMD (95% CI., p, I^2)	Subgroup explains heterogeneity P value
Risk of bias			
Less than high risk of bias[1]	4 (530)	-0.11 (-0.41 to 0.18; p = 0.45; I^2 = 62%)	<0.001
High risk of bias	27 (1801)	-0.85 (-1.10 to -0.60; p < 0.001; I^2 = 82%)	
Age			
Old (>59 years)	5 (492)	-0.77 (-1.34 to -0.19; p = 0.009; I^2 = 87%)	0.99
Young (<59 years)	26 (1839)	-0.76 (-1.01 to -0.51; p < 0.001; I^2 = 83%)	
Exercise context			
Group exercise	24 (1729)	-0.79 (-1.06 to -0.52; p < 0.001; I^2 = 85%)	0.72
Individual exercise	7 (602)	-0.68 (-1.17 to -0.20; p = 0.005; I^2 = 79%)	
Duration			
Less than 10 weeks	14 (691)	-0.93 (-1.11 to -0.88; p < 0.001; I^2 = 19%)	0.05
10 weeks or more	17 (1640)	-0.58 (-0.88 to -0.28; p < 0.001; I^2 = 86%)	
Attention control			
Attention control	7 (609)	-0.71 (-1.27 to -0.16; p = 0.01; I^2 = 89%)	0.99
Waitlist	2 (47)	-0.67 (-2.48 to 1.13; p = 0.47; I^2 = 88%)	
Pharmacotherapy			
Add-on	11 (734)	-0.92 (-1.38 to -0.46; p < 0.001; I^2 = 86%)	0.82
No medication	6 (318)	-0.82 (-1.58 to -0.06; p = 0.03; I^2 = 88%)	
Somatic comorbidity			
Somatic co-morbidity	0	N/A	
No co-morbidity	31 (2331)	N/A	
Minor depression			
Incl. minor depression	6 (350)	-0.90 (-1.65 to -0.15; p = 0.02; I^2 = 86%)	0.67
No minor depression	25 (1981)	-0.73 (-0.97 to -0.49; p < 0.001; I^2 = 88%)	
Patient setting			
Inpatients	10 (549)	-0.88 (-1.07 to -0.70; p < 0.001; I^2 = 6%)	0.26
Outpatients	21 (1782)	-0.69 (-0.98 to -0.41; p < 0.001; I^2 = 85%)	
Trial size			
Trials n ≤ 52	15 (479)	-1.30 (-1.74 to -0.86; p < 0.001; I^2 = 77%)	<0.001
Trials n > 52	16 (1852)	-0.40 (-0.60 to -0.19; p < 0.001; I^2 = 76%)	
Increase in exercise capacity			
VO2max > 2.6 ml/kg/min	5 (356)	-0.55 (-0.65 to 0.07; p = 0.08; I^2 = 86%)	0.49
VO2max ≤ 2.6 ml/kg/min	5 (601)	-0.30 (-0.63 to 0.03; p = 0.07; I^2 = 73%)	

[1]Trials potentially having less bias than trials with high risk of bias.

Table 4. Summary of findings

Exercise compared to control or treatment as usual for depression

Patient or population: depression

Setting: In- or out-patients

Intervention: exercise

Comparison: control or treatment as usual

Outcomes	Anticipated absolute effects* (95% CI)		Relative effect (95% CI)	№ of participants (studies)	Quality of the evidence (GRADE)	Comments
	Risk with control or treatment as usual	Risk with exercise				
Severity of depression	-	**0.74 SMD lower** (0.51 lower to 0.96 lower)	-	2419 (31 RCTs)	⊕◯◯◯ VERY LOW [1]	Lower depression scores indicate improvement. SMD of 0.3 is considered clinically relevant.
Lack of remission	**Study population**		**RR 0.78** (0.68 to 0.90)	1639 (19 RCTs)	⊕◯◯◯ VERY LOW [2]	Remission is, with minor variations, defined as not full-filling the criteria for depression.
	646 per 1000	**504 per 1000** (426 to 594)				
Serious adverse events	**Study population**		**RR 2.21** (0.24 to 20.21)	335 (3 RCTs)	⊕⊕◯◯ LOW [3]	
	0 per 1000	**0 per 1000** (0 to 0)				
Quality of life	-	**0.43 SMD higher** (0.04 lower to 0.91 higher)	-	901 (8 RCTs)	⊕◯◯◯ VERY LOW [4]	Quality of life was assessed using a number of different methods. Higher score indicates improved quality of life. Seven of 24 trials reported on this outcome
Depression severity after the intervention	-	**0.06 SMD lower** (0.25 lower to 0.14 higher)	-	713 (7 RCTs)	⊕⊕◯◯ LOW [5]	Lower depression scores indicate improvement. SMD of 0.3 is considered clinically relevant.
Lack of remission after the intervention	**Study population**		**RR 0.95** (0.82 to 1.11)	777 (5 RCTs)	⊕⊕◯◯ LOW [6]	
	469 per 1000	**446 per 1000** (385 to 521)				
Depression severity. Restricted to trials with less than high risk of bias.	-	**0.11 SMD lower** (0.41 lower to 0.18 higher)	-	530 (4 RCTs)	⊕⊕◯◯ LOW [7]	Lower depression scores indicate improvement. SMD of 0.3 is considered clinically relevant.

***The risk in the intervention group** (and its 95% confidence interval) is based on the assumed risk in the comparison group and the **relative effect** of the intervention (and its 95% CI). **CI:** Confidence interval; **SMD:** Standardised mean difference; **RR:** Risk ratio

GRADE Working Group grades on evidence
High quality: We are very confident that the true effect lies close to that of the estimate of the effect
Moderate quality: We are moderately confident in the effect estimate. The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different
Low quality: 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 the effect

1. Downgraded by 3: risk of bias, inconsistency and publication bias
2. Downgraded by 3: risk of bias, inconsistency and publication bias
3. Downgraded by 2: imprecision and publication bias
4. Downgraded by 3: risk of bias, inconsistency and imprecision

5. Downgraded by 2: risk of bias and imprecision
6. Downgraded by 2: risk of bias and imprecision
7. Downgraded by 2: inconsistency and imprecision

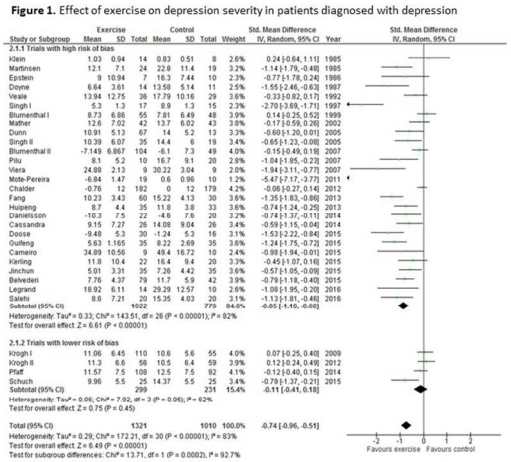


Figure 1

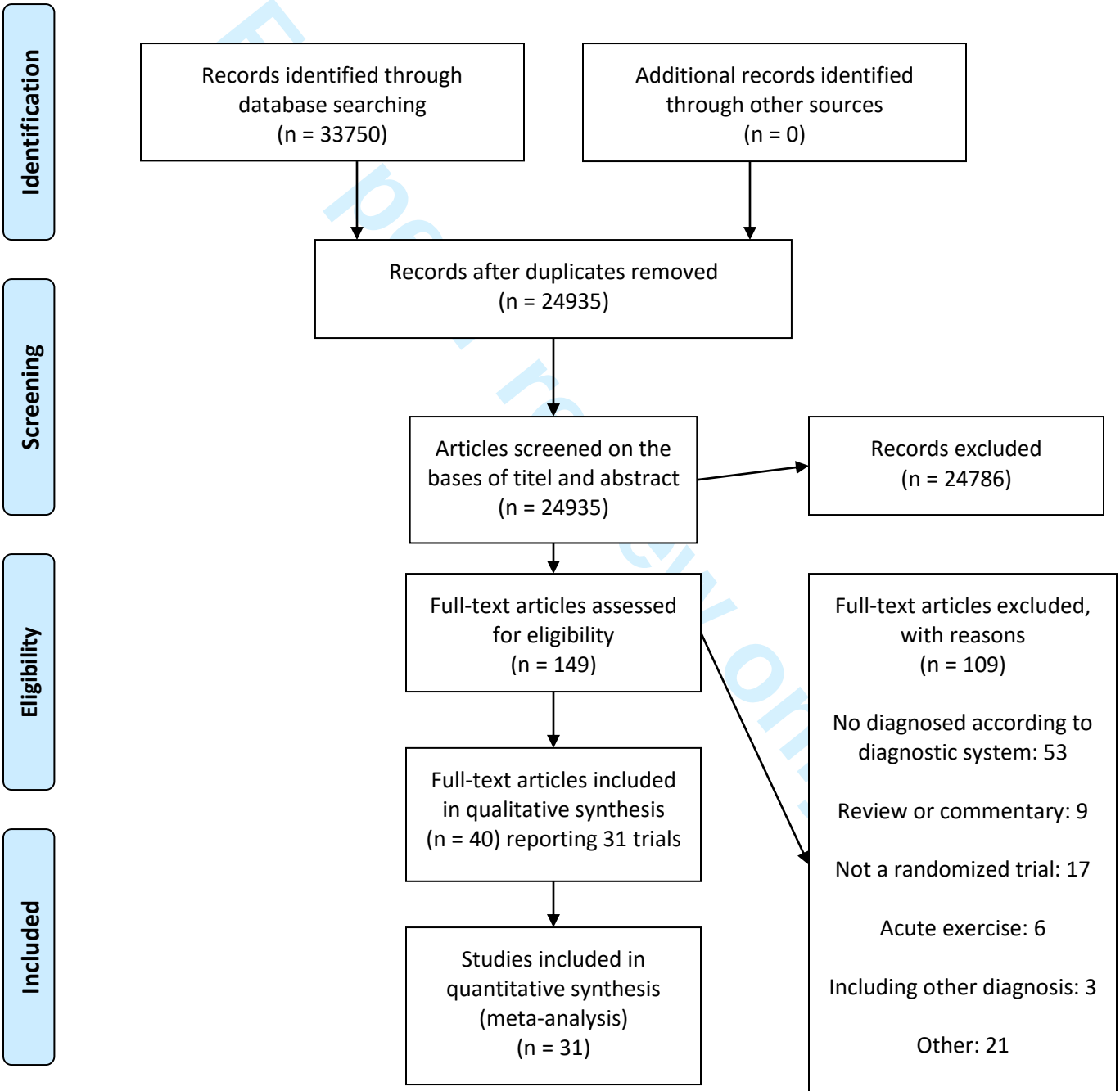
338x190mm (96 x 96 DPI)

Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplemental

Figure S1. Flow diagram for identification of trials assessing the effects of exercise for patients with depression.



From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). *P*referred *R*eporting *I*ter
*A*nalyses: The PRISMA Statement. PLoS Med 6(6): e1000097. doi:10.1371/journa

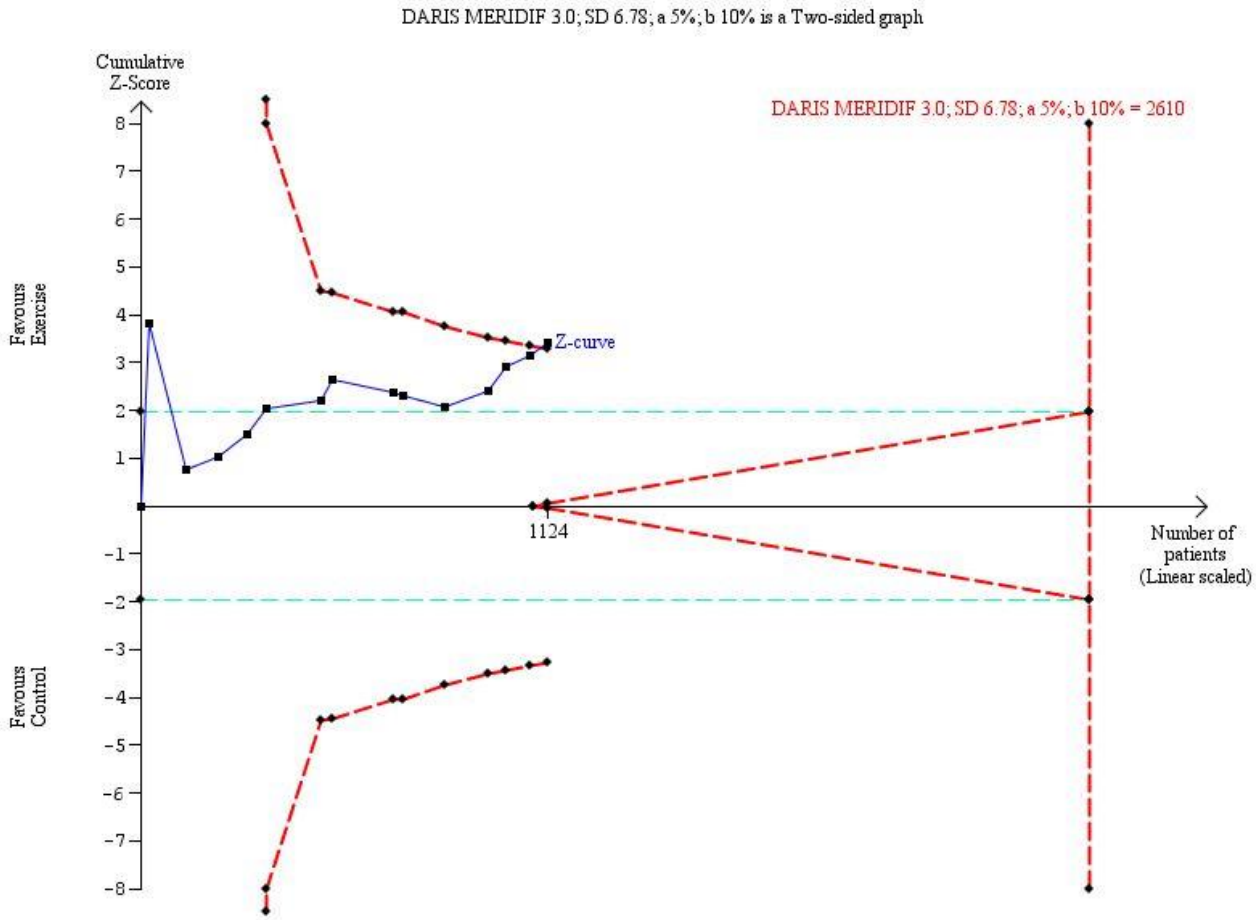
For more information, visit www.prisma-statement.org.

Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary Figure

Figure S2. Trial Sequential Analysis and required information size for the effect of exercise for depressive symptoms including twelve trials reporting on HAM-D_{17}.

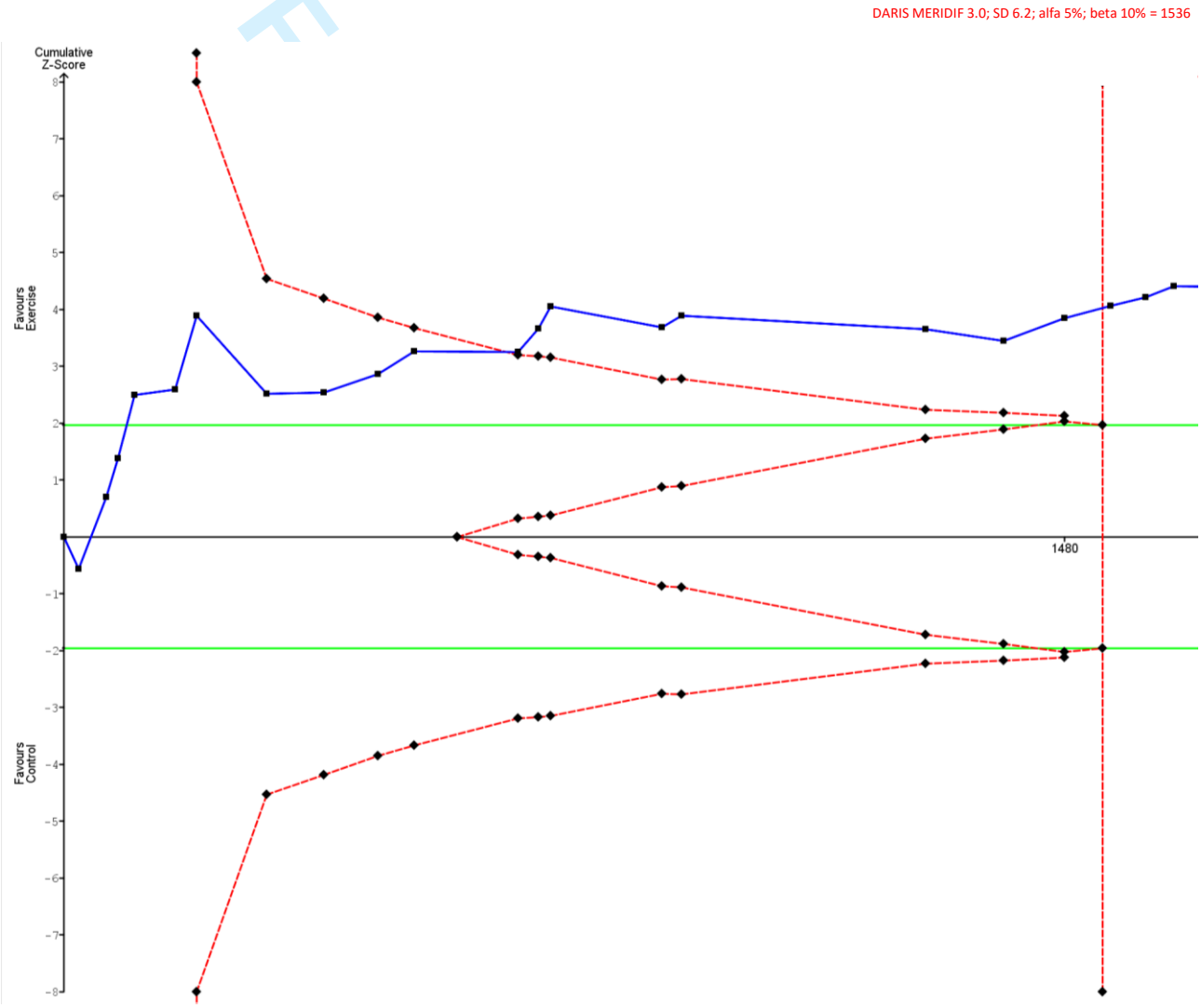


Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary Figure

Figure S3. Trial Sequential Analysis and required information size for the effect of exercise for depressive symptoms including 31 trials 'converted' to a HAM-D_{17} scale.



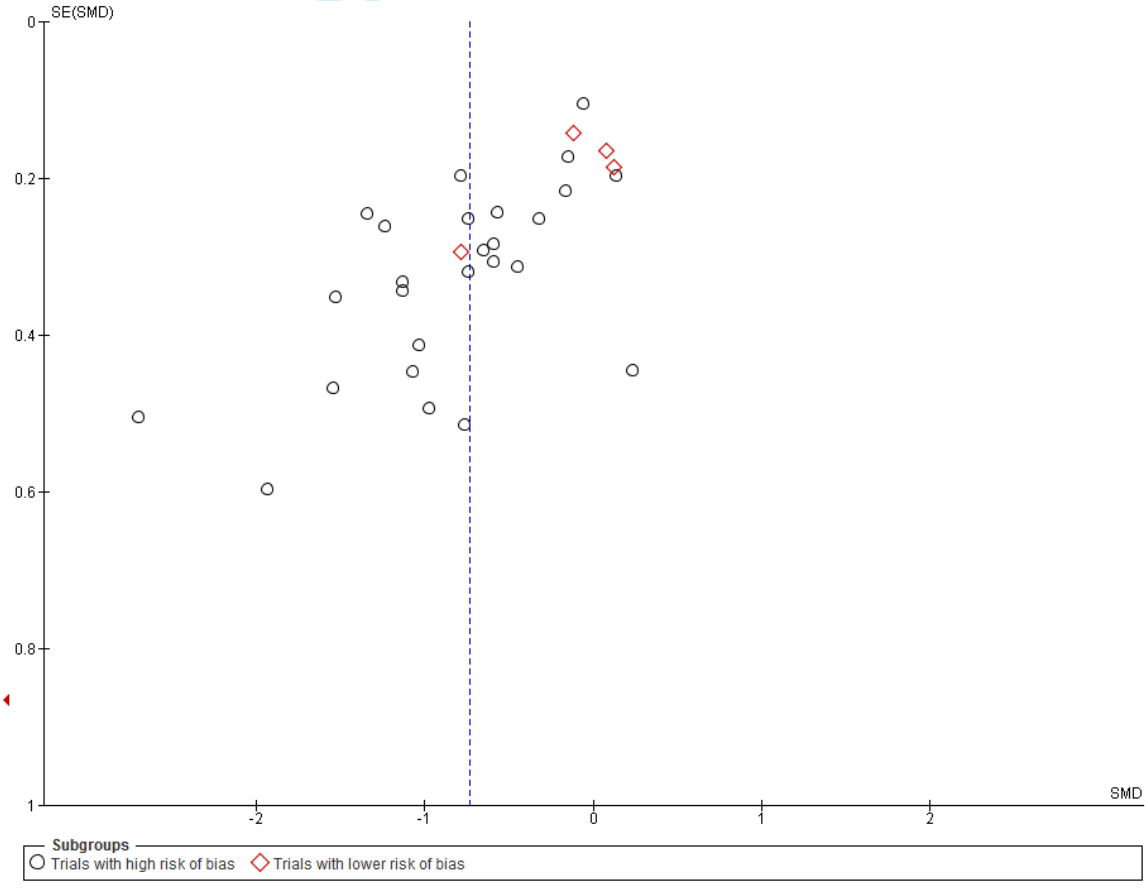
Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary Figure

Figure S4.

Funnel plot of 31 trials assessing the antidepressant effect of exercise as a continuous outcome

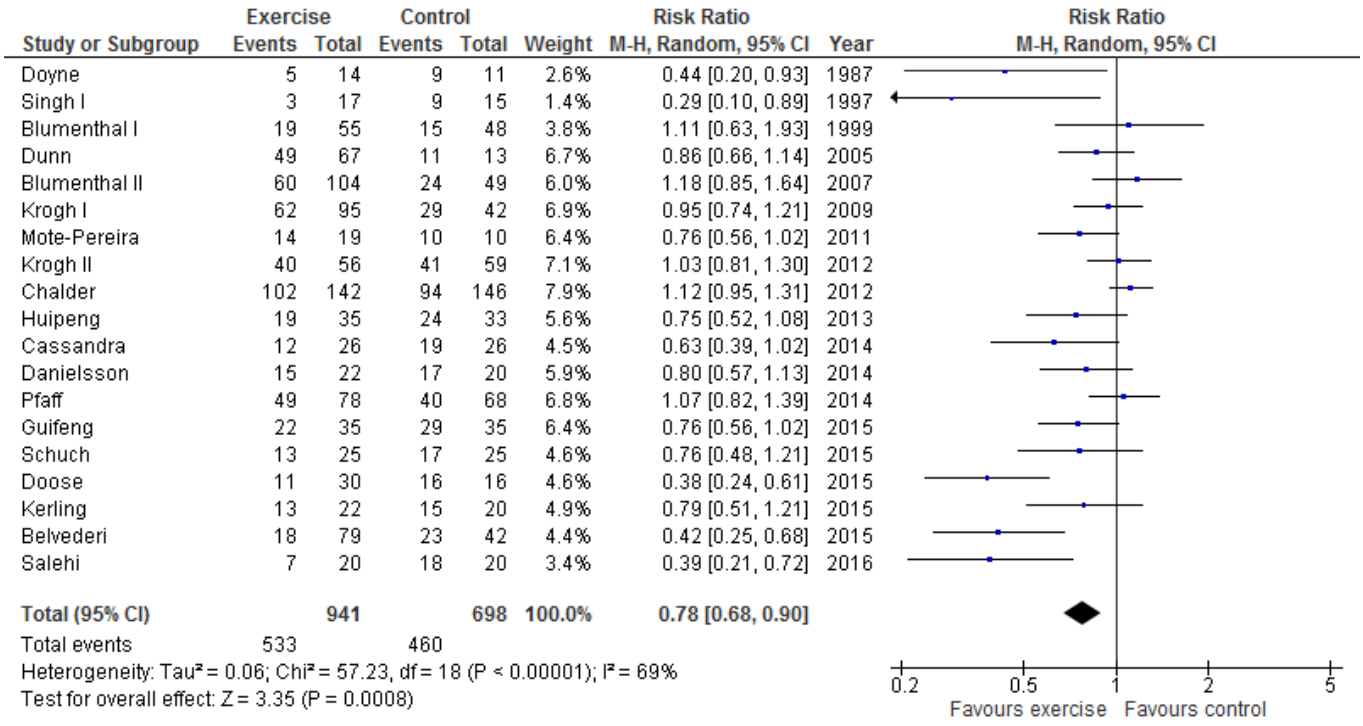


Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary Figure

Figure S5. Effect of exercise on lack of remission for patients diagnosed with depression

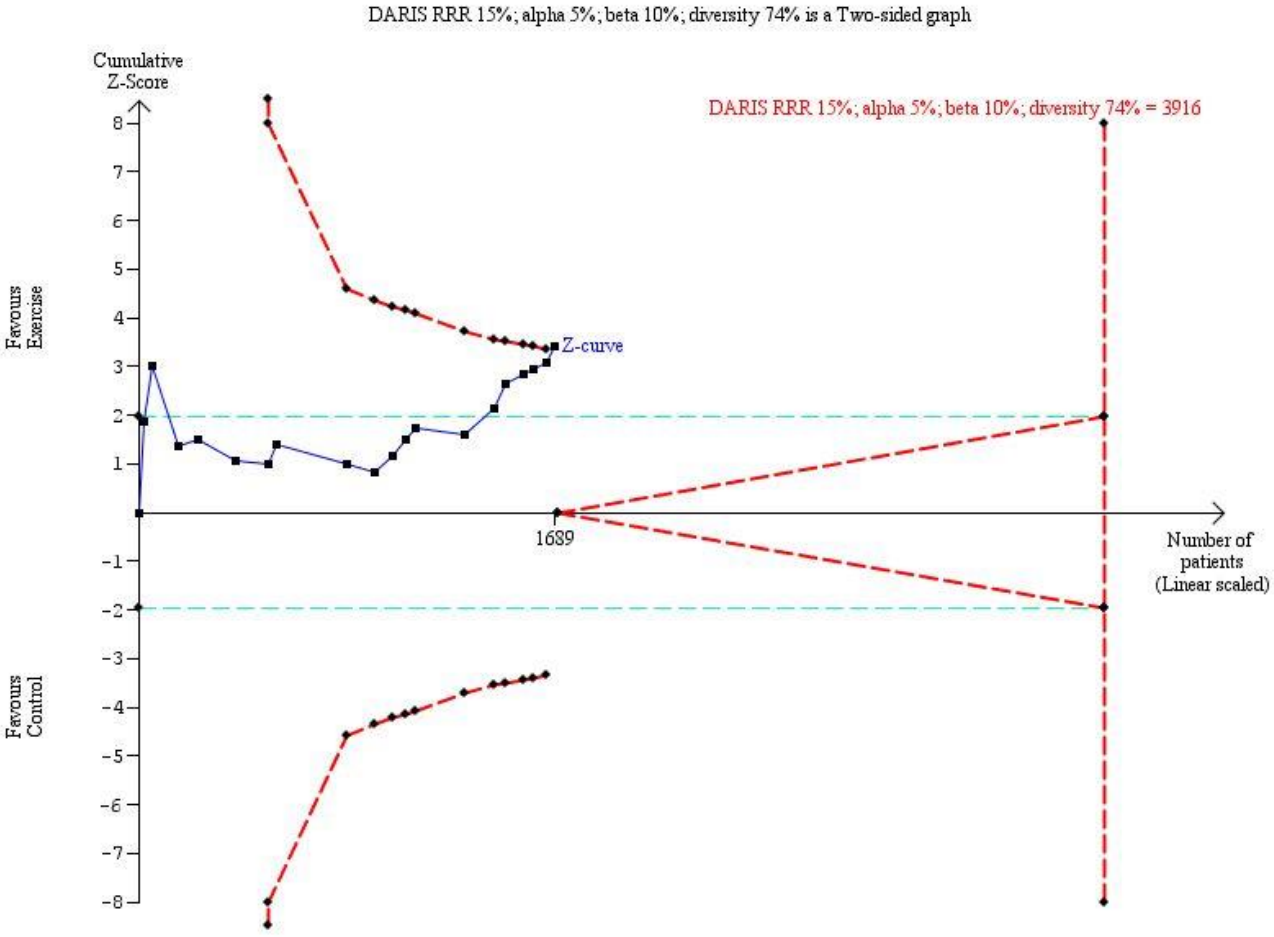


Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary Figure

Figure S6. Trial Sequential Analysis and required information size for the effect of exercise on lack of remission.



Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary Figure S7

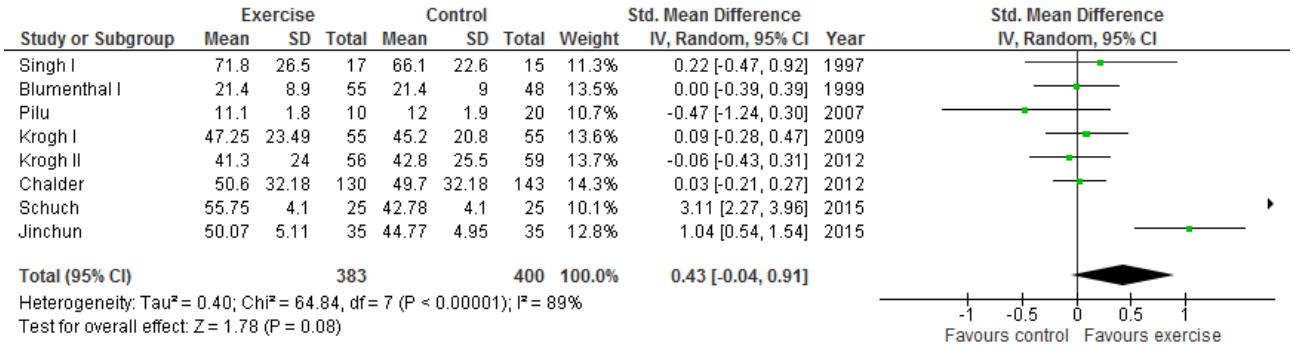
Figure S7. Effect of exercise on risk of serious adverse events for patients diagnosed with depression

Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary Figure S8-S10

Figure S8. The effect of exercise on quality of life in patients diagnosed with depression



Quality of life was assessed using different scales: Singh I and Chalder used the SF-36, Blumenthal used Life Satisfaction Index, Pilu and Schuch used the WHOQOL, Krogh I and Krogh II used the WHO-Five Well-being Scale, and Jinchun used the GQOLI-74.

Figure S9. The effect of exercise on depression severity after the intervention in patients diagnosed with depression

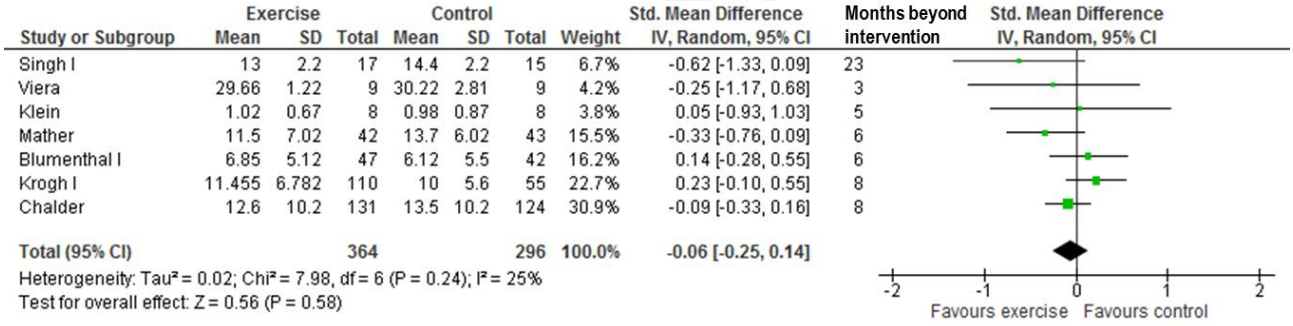
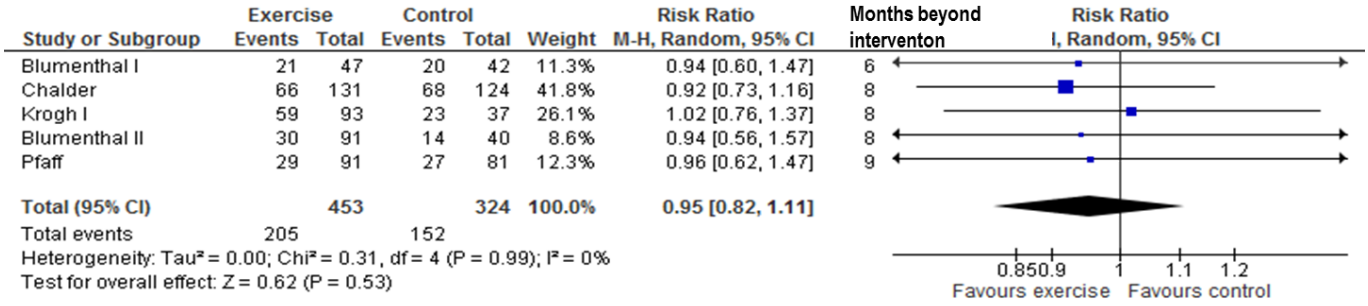


Figure S10. The effect of exercise on risk of lack of remission after the intervention in patients diagnosed with depression



Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary Table S1

Table S1. Imputation of missing data for trials assessing exercise for patients diagnosed with depression

Outcome	Result from review	Best/worse (1SD)	Best/worse (2SD)	Worse/best (1SD)	Worse/best (2SD)
Depression SMD (95% CI)	-0.74 (-0.96 to -0.51) p < 0.001; I^2 = 83%	-0.85 (-1.10 to -0.60) p < 0.001; I^2 = 87.2%	-0.85 (-1.11 to -0.60) p < 0.001; I^2 = 87.9%	-0.66 (-0.90 to -0.40) p < 0.001; I^2 = 85.4%	-0.61 (-0.84 to -0.38) p < 0.001; I^2 = 85.5%)
		Good Outcome	**Poor outcome**	**Good/poor outcome**	**Poor/good outcome**
Lack of remission (95% CL)	RR 0.78 (0.68 to 0.90) p < 0.001; I^2 = 69%	RR 0.75 (0.64 to 0.89) p = 0.0008; I^2 = 73%	RR 0.88 (0.83 to 0.94) p = 0.0002; I^2 = 69%	RR 0.71 (0.61 to 0.81) p < 0.001; I^2 = 68%	RR 0.86 (0.71 to 1.04) p = 0.12; I^2 = 83%
Serious adverse events (95% CL)	RR 2.21 (0.24 to 20.21) p = 0.48; I^2 = 0%	RR 2.19 (0.23 to 20.76) p = 0.50, I^2 = 50%	RR 0.92 (0.37 to 2.30) p = 0.86, I^2 = 60%	RR 0.08 (0.02 to 0.34) p = 0.001, I^2 = 5.4%	RR 19.17 (2.64 to 139.2) p = 0.004, I^2 = 0%

Imputation of missing data for continuous outcome: 'best-worst' - assumed that all participants lost to follow-up in the intervention group had a beneficial outcome (the group mean minus 1 or 2 SD), and all participants lost to follow-up in the placebo group have had a harmful outcome (the group mean plus 1 SD and 2 SD). The reverse 'worst-best-case' scenario is the reverse of the 'best-worst' scenario.

Missing data for the 'remission' outcome was imputed according to the following scenarios: 1) poor outcome analysis: none of the drop-outs/participants lost from both arms experienced the outcome; 2) good outcome analysis: all of the drop-outs/participants lost from both arms experienced the outcome; 3) extreme case analysis favouring the experimental intervention, all of the drop-outs/participants lost from the experimental arm, but none of the drop-outs/participants lost from the control arm experienced the outcome; and 4) extreme case analysis favouring the control: all drop-outs/participants lost from the experimental arm, but none from the control arm experienced the outcome. Missing data for 'serious adverse events' was calculated with the reverse assumptions.

Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary Table S2

Table S2. Trials characteristics for exploration of heterogeneity in trials assessing the effect of exercise in patients diagnosed with depression

Trial	Lower risk of bias	Age > 60	Group vs. individual	Duration	Attention control waitlist	Exercise as add on to drugs vs. exercise alone	Within-study dose exercise	Increase in VO2max [1]	Somatic disease vs. only MD	Trial Includes minor depression
Klein 1985	No	Young	Individual	12 weeks	Other	Exercise alone	No	No	No	Yes
Martinsen 1985	No	Young	Group	9 weeks	Attention control	Unclear	No	11[a]	No	No
Epstein 1986	No	Young	Group	8 weeks	Waitlist	Unclear	No	No	No	Yes
Doyne 1987	No	Young	Individual	8 weeks	Waitlist	Exercise alone	No	No	No	Yes
Veale 1992	No	Young	Group	12 weeks	Other	Unclear	No	No	No	No
Singh 1997	No	Old	Group	10 weeks	Attention control	Exercise alone	No	N/A	No	Yes
Blumenthal 1999	No	Young	Group	16 weeks	Other	Add on	No	2.3	No	No
Mather 2002	No	Old	Group	10 weeks	Attention control	Add on	No	No	No	No
Dunn 2005	No	Young	Individual	12 weeks	Attention control	Exercise alone	Yes	No	No	No
Singh 2005	No	Old	Group	8 weeks	Other	Exercise alone	Yes	N/A	No	Yes
Pilu 2007	No	Young	Group	24 weeks	Other	Add on	No	No	No	No
Viera 2007	No	Young	Group	12 weeks	Other	Add on	No	No	No	No
Blumenthal 2007	No	Young	Group	16 weeks	Other	Add on	No	2.0[a]	No	No
Krogh 2009	Yes	Young	Group	16 weeks	Attention control	No	No	2.9	No	No
Mota-Pereira 2011	No	Young	Group	12 weeks	Other	Add on	No	No	No	No
Krogh 2012	Yes	Young	Group	12 weeks	Attention control	Exercise alone	No	3.4	No	No
Chalder 2012	No	Young	Individual	32 weeks	Other	No	No	No	No	No
Fang 2013	No	Young	Group	6 weeks	Attention control	No	Yes	No	No	No
Huipeng 2013	No	Young	Group	6 weeks	Other	No	No	No	No	No
Cassandra 2014	No	Young	Group	3 weeks	Other	Add on	No	No	No	No
Danielsson 2014	No	Young	Group	10 weeks	Other	Add on	No	2.4	No	No
Pfaff 2014	Yes	Old	Group	12 weeks	Other	No	No	1.5	No	Yes
Guifeng 2015	No	Young	Group	8 weeks	Other	No	No	No	No	No
Jinchun 2015	No	Young	Group	8 weeks	Other	No	No	No	No	No
Schuch 2015	Yes	Young	Individual	2 weeks	Other	No	No	No	No	No
Kerling 2015	No	Young	Group	6 weeks	Other	No	No	2.8	No	No

Belvederi 2015	No	Old	Group	24 weeks	Other	Add on	Yes	0.3[a]	No	No
Carneiro 2015	No	Young	Group	16 weeks	Other	Add on	No	No	No	No
Doose 2015	No	Young	Group	8 weeks	Other	No	No	3.2	No	No
Legrand 2016	No	Young	Individual	10 days	Other	No	No	No	No	No
Salehi 2016	No	Young	Individual	4 weeks	Other	Add on	No	No	No	No

[1]Increase in VO2max is based on increase in intervention group, if [a] then value is based on an estimate from text or figures.

Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary Table

Table S3. Heterogeneity of effect estimates for trials assessing the effect of exercise for patients diagnosed with depression on lack of remission.

Subgroups	Number of Trials (participants)	Random effects meta-analysis RR (95% CI., p, I^2)	Subgroup explains heterogeneity P value
Risk of bias			
Less than high risk of bias[1,2]	2 (165)	0.95 (0.74 to 1.23; p = 0.70; I^2 = 20%)	0.18
High risk of bias	17 (1474)	0.77 (0.64 to 0.92; p = 0.003; I^2 = 75%)	
Age			
Old (>59 years)	3 (299)	0.61 (0.21 to 1.02; p = 0.37; I^2 = 91%)	0.62
Young (<59 years)	16 (1340)	0.81 (0.70 to 0.93; p = 0.003; I^2 = 64%)	
Exercise context			
Group exercise	14 (1156)	0.80 (0.66 to 0.96; p = 0.02; I^2 = 72%)	0.69
Individual exercise	5 (483)	0.74 (0.52 to 1.04; p = 0.08; I^2 = 77%)	
Duration			
Less than 10 weeks	8 (393)	0.63 (0.51 to 0.77; p < 0.001; I^2 = 40%)	0.004
10 weeks or more	11 (1246)	0.93 (0.78 to 1.10; p = 0.39; I^2 = 69%)	
Attention control			
Attention control	4 (364)	0.91 (0.73 to 1.12; p = 0.38; I^2 = 42%)	0.07
Waitlist	1 (25)	0.44 (0.21 to 0.93; p = 0.03; I^2 = 0%)	
Pharmacotherapy			
Add-on	7 (540)	0.72 (0.54 to 0.96; p = 0.03; I^2 = 69%)	0.62
No medication	4 (252)	0.75 (0.52 to 1.09; p = 0.13; I^2 = 66%)	
Somatic comorbidity			
Somatic co-morbidity	0	N/A	
No co-morbidity	19 (1639)	N/A	
Minor depression			
Incl. minor depression	3 (203)	0.63 (0.21 to 1.89; p = 0.41; I^2 = 87%)	0.69
No minor depression	16 (1436)	0.79 (0.68 to 0.92; p = 0.002; I^2 = 69%)	
Patient setting			
Inpatients	6 (322)	0.71 (0.60 to 0.84; p < 0.001; I^2 = 0%)	0.21
Outpatients	13 (1317)	0.84 (0.69 to 1.01; p = 0.07; I^2 = 77%)	
Trial size			
Trials n ≤ 52	9 (358)	0.62 (0.50 to 0.76; p < 0.001; I^2 = 45%)	0.002
Trials n > 52	10 (1281)	0.95 (0.80 to 1.12; p = 0.52; I^2 = 68%)	

[1]Trials potentially having less bias than trials with high risk of bias.

Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary material (S1)

An example of bibliographical search for PubMEd

#1 Depression [MeSH]

#2 Depresive disorder [MeSH]

#3 Exercise [Text Word]

#4 Aerobic [Text Word]

#5 Non-aerobic [Text Word]

#6 Physical activity [Text Word]

#7 Physical fitness [Text Word]

#8 Walking [MeSH]

#9 Jogging [MeSH]

#10 Running [MeSH]

#11 Bicycling [MeSH]

#12 Swimming [MeSH]

#13 Strength [Text Word]

#14 Resistance [Text Word]

#15 #1 OR #2

#16 #3 OR #4 OR #5 OR #6 OR #7 OR #8 OR #9 OR #10 OR #11 OR #12 OR #13 OR #14

#17 #15 AND #16

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	5
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	6
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	1
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	7
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	6
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	6
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	6
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	8
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	7
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	8
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	9
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	9

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	8
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	11
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	Fig. 1
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Table 1
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	Table 2
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	Fig 2
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	Fig 3-fig8
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	Table 2
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	Table 3
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	21
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	21
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	22
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	26

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: www.prisma-statement.org.

BMJ Open

Exercise for patients with major depression: a systematic review with meta-analysis and Trial Sequential Analysis

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2016-014820.R2
Article Type:	Research
Date Submitted by the Author:	26-Jun-2017
Complete List of Authors:	Krogh, Jesper; Kobenhavns Universitet Sundhedsvidenskabelige Fakultet, Mental Health Centre Copenhagen Hjorthøj, Carsten; Copenhagen University Hospital, Mental Health Center Copenhagen Speyer, Helene; University of Copenhagen Gluud, Christian; Copenhagen University Hospital Rigshospitalet, The Copenhagen Trial Unit, Centre for Clinical Intervention Research Nordentoft, Merete; University of Copenhagen, Mental Health Centre Copenhagen
Primary Subject Heading:	Mental health
Secondary Subject Heading:	Sports and exercise medicine, Evidence based practice
Keywords:	Exercise, Systematic review, Meta-analysis, Evidence based medicine, Randomised clinical trials, Depression & mood disorders < PSYCHIATRY

Exercise for patients with major depression: a systematic review with meta-analysis and Trial Sequential Analysis

Jesper Krogh[1], Carsten Hjorthøj[1], Helene Speyer[1], Christian Gluud[1], Merete Nordentoft[1]

[1] Mental Health Centre Copenhagen. Faculty of Health Sciences. University of Copenhagen, Copenhagen, Denmark.

[2] Copenhagen Trial Unit, Centre for Clinical Intervention Research, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark.

Corrosponding author

Jesper Krogh, Jesper.krogh@dadlnet.dk

Mental Health Centre Copenhagen, Kildegårdsvej 28, Building 15. DK-2900, Hellerup, Denmark

Abstract

Objectives

To assess the benefits and harms of exercise in patients with depression.

Design

Systematic review

Data sources

Bibliographical databases were searched until the 20th of June, 2017.

Eligibility criteria and outcomes

Eligible trials were randomised clinical trials assessing the effect of exercise in participants diagnosed with depression. Primary outcomes were depression severity, lack of remission, and serious adverse events (e.g. suicide) assessed at the end of the intervention. Secondary outcomes were quality of life and adverse events such as injuries, as well as assessment of depression severity and lack of remission during follow-up after the intervention.

Results

Thirty-five trials enrolling 2498 participants were included. The effect of exercise versus control on depression severity was -0.66 standardised mean difference (SMD) (95% CI -0.86 to -0.46; P < 0.001; GRADE: very low quality). Restricting this analysis to the four trials that seemed less affected of bias, the effect vanished into -0.11 SMD (-0.41 to 0.18; P = 0.45; GRADE: low quality). Exercise decreased the relative risk of no remission to 0.78 (0.68 to 0.90; P < 0.001; GRADE: very low quality). Restricting this analysis to the two trials that seemed less affected of bias, the effect vanished into 0.95 (0.74 to 1.23; P = 0.78). Trial Sequential Analysis excluded random error when all trials were analysed, but not if focusing on trials less affected of bias. Sub-group analyses found that trial size and intervention duration were inversely

associated with effect size for both depression severity and lack of remission. There was no significant effect of exercise on secondary outcomes.

Conclusions

Trials with less risk of bias suggested no antidepressant effects of exercise and there were no significant effects of exercise on quality of life, depression severity, or lack of remission during follow-up. Data for serious adverse events and adverse events was scarce not allowing conclusions for these outcomes.

Systematic review registration

The protocol was published in the journal Systematic Reviews: 2015; **4**:40.

DOI: 10.1186/s13643-015-0030-6.

Article Summary

Strengths and limitations of this study

- The protocol for this review has previously been published
- Using meta-regression analysis, Trial Sequential Analysis and the GRADE system the conclusions from this review is based on a firm and transparent platform
- Based on an extensive literature search, this review included 35 trials allocating almost 2500 participants diagnosed with depression to exercise or control interventions than could be analysed
- The effect estimates are largely based on trials at high risk of bias
- Effect estimates from included trials had considerable heterogeneity

Introduction

Depression is a common disorder affecting up to 17% of the population during their lifetime.[1;2] Based on data from the World Health Organisation, depression is ranked as the second largest health-care problem globally, in terms of years lived with disability.[3] Depending on its severity, depression is often treated using psychotherapy, antidepressants, or a combination of both. However, the clinical benefits of antidepressants[4-6] and psychotherapy[7-9] has been challenged. Both treatments are costly in terms of time and money and may also have adverse effects. Compliance with antidepressant treatment is poor; the dropout rate in clinical trials is reported to be between 12% and 40% within the initial 6 to 8 weeks of treatment.[4;10]

The weakness of evidence for the beneficial effect of current interventions, along with problems related to low compliance and harms, has resulted in an interest in using alternative interventions. The use of exercise as an intervention has attracted considerable attention, and various forms of exercise varying in intensity have been assessed in a number of randomised clinical trials to test their effectiveness as a treatment for patients with depression. In 2011, we published a meta-analysis of randomised clinical trials examining the effect of exercise on depressive symptoms in patients with clinical depression.[11] The results suggested that referring patients with clinical depression to exercise programs was associated with a small to moderate effect on depressive symptoms. However, restricting the analysis to three trials at low risk of bias, the effect estimate was non-significant. Since 2011, other reviews have been published on the effect of exercise on depressive symptoms,[12] in older people,[13] and in patients with chronic illnesses.[14] However, none of these reviews addressed the specific population of adults diagnosed with major depression according to valid diagnostic criteria, such as the International Classification of Diseases[15] or the Diagnostic and Statistical Manual of Mental Disorders.[16] The reviews contained a number of trials that included volunteers who were defined as being depressed on the basis of psychometric testing (for example, Beck

Depression Inventory[17]), as opposed to individuals with a clinical diagnosis of major depression. Furthermore, several randomised clinical trials investigating the effect of exercise in clinically depressed individuals have been published since our 2011 review.[11]

The objectives of the present systematic review are to investigate the beneficial and harmful effects of exercise, in terms of severity of depression, lack of remission, quality of life, and suicide versus controls with or without co-interventions in adults with a clinical diagnosis of major depression. The current systematic review differs from our previous review in a number of aspects.[11] We only considered trials including participants diagnosed with depression according to a validated diagnostic system. We also included trials including participants with somatic co-morbidity, e.g. cancer or diabetes. The harmful effects of exercise interventions are also addressed, the intervention effects being assessed according to the grading of recommendations assessment, development, and evaluation (GRADE) framework, and bibliographical searches have been extended to include a Chinese and a South-American database until 2016.

Methods/design

The protocol for this review has previously been published.[18]

Search strategy

The following bibliographical databases was searched: CENTRAL, MEDLINE, EMBASE, Science Citation Index (Web of Science), LILACS, and Wanfang using medical subject headings (MeSH or similar) when possible or text word terms: depression, depressive disorder and exercise, aerobic, non-aerobic, physical activity, physical fitness, walking, jogging, running, bicycling, swimming, strength, or resistance. Please see

supplementary material (S1) for an example of a bibliographical search. The main search was conducted in August 2015, and the latest search was conducted on 20th of June, 2017.

Trial selection

One investigator (JK) examined titles and abstracts to remove obviously irrelevant reports. Two investigators (JK + HS) examined full text reports and abstracts determining compliance with inclusion criteria. A trial was considered eligible if it was a randomised clinical trials including participants diagnosed as having major depression according to a valid and recognised diagnostic system (that is, Research Diagnostic Criteria (RDC),[19] International Classification of Diseases (ICD)[15] or Diagnostic and Statistical Manual of Mental disorders (DSM)[16]) and included participants aged >17 years. Abstracts and full text reports were included.

Trials were excluded if they measured depression immediately after a single bout of exercise, compared one form of exercise versus another, or compared different exercise intensities without including a control group. The trials had to allocate participants to an exercise intervention versus a control group (that is, exercise versus a control group receiving no intervention or treatment as usual or an attention control using light exercise) or using exercise as an add-on-treatment (that is, exercise plus usual treatment in the experimental group versus usual treatment alone in the control group). Exercise intervention was defined as a systematic physical intervention with the intention to increase muscle strength and/or cardiovascular fitness, e.g. running, swimming or weight lifting. In case of attention control, it should specifically be mentioned by the authors of the trial report that the intervention was intended as a control intervention.

Outcomes

The primary outcomes were 1) depressive symptoms measured on a continuous scale assessed at the end of the intervention; 2) lack of remission, that is, a binary outcome of the proportion of participants in each intervention group of the trial who did not obtain remission at the end of the intervention according to the authors' own definition; and 3) serious adverse events defined according to ICH-GCP as any untoward medical occurrence that was life threatening, resulted in death or persistent or significant disability (ICH-GCP 1997).[20] Serious adverse events accordingly include suicide attempts as well as suicides. The secondary outcomes were quality of life, non-serious adverse events (e.g. muscle injuries) as well as depressive symptoms and lack of remission assessed after the intervention.

Data extraction

Two authors (JK, HS) independently extracted data using a pre-piloted structured form. Any discrepancies in the data extraction or inclusion/exclusion of trials was resolved by referring to the original papers. CG or MN assisted as adjudicator in cases of disagreements. Data extraction included, in addition to outcomes, information regarding country of origin, number of randomised participants, number of participants included in efficacy analysis, mean age of participants, diagnostic system, baseline assessment of depression severity, type of intervention, frequency of intervention, and duration of intervention. Continuous outcomes were preferred in the following order: post-intervention scores with corresponding standard deviations (SD), mean change from baseline with SD, mean difference between groups post-intervention and reported outcomes were preferred to figures. JK and CH independently performed the assessment of bias domains. The authors JK, CG, and MN have previously published trial reports assessing the effect of exercise in participants with depression,[21;22] and to reduce the risk of academic bias two additional authors were included in the current systematic review (CH, HS).

Risk of bias assessment

Definitions in the assessment of bias risk of a trial was conducted according to the Cochrane Handbook for Systematic Reviews of Interventions[23] of the following domains: allocation sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessors, incomplete outcome data, selective outcome reporting, for-profit bias, and other bias. Trials assessed as having 'low risk of bias' in all of the above specified domains were considered 'trials at low risk of bias'. Trials assessed as having 'uncertain risk of bias' or 'high risk of bias' in one or more of the above specified domains were considered trials at 'high risk of bias'. In line with our previous systematic review[11] and the latest Cochrane review on exercise for depression,[24] trials at low risk of bias in the allocation concealment domain, blinded outcome assessment domain, and the incomplete outcome data domain were characterised as 'trials potentially having less risk of bias than other trials at high risk of bias'. Trials assessing the effect of behavioural interventions are rarely able to mask the allocation, and participants and health care providers are therefore not blinded. Therefore, we will also report the number of trials at low risk of bias in the remaining domains.

Data synthesis and analysis

In order to be able to include all of the trials in our meta-analysis, estimates of standardised mean difference (SMD) for each individual trial was carried out. SMD is the mean difference in depression score between the exercise and control groups divided by the pooled standard deviation at follow-up. The result is a unit free effect size. By convention, SMD effect sizes of 0.2, 0.5 and 0.8 are considered small, medium and large intervention effects.[23] For dichotomous variables, we calculated the risk ratio (RR) with a 95% confidence interval. It was expected that some trials would have several intervention groups. Data from the experimental groups was pooled and compared with the data from the control group. In case of discrepancies between the random-effects model analysis and the fixed-effect model analysis, both results are reported; otherwise, only results from the random-effects analysis is reported. The degree of

heterogeneity was quantified using the I-squared statistic,[25] which can be interpreted as the percentage of variation observed between the trials attributable to between-trial differences, rather than sampling error (chance). Heterogeneity was explored by analyses of sub-groups (see below).

For the primary outcomes, Trial Sequential Analysis was performed.[26;27] In order to calculate the required information size and the cumulative Z-curve's eventual breach of relevant trial sequential monitoring boundaries, the required information size for the primary continuous outcome was based on type I error of 5%, a beta of 10%, the standard error of the meta-analysis, and a minimal difference of three points on the HAM-D_{17}.[18] Post-hoc we calculated the required information size including all trials. This was done by converting effect estimates from trials reporting other outcome scales into the HAM-D_{17} scale as described by Thorlund et al.[28] In order to calculate the required information size and the cumulative Z-curve's eventual breach of relevant trial sequential monitoring boundaries, the required information size for lack of remission was based on type I error of 5%, a beta of 10%, the proportion of participants in the control group with the outcome, and a relative risk reduction of 15% and 30%.

Bayes factors were calculated for all primary outcomes.[29] Low P-values suggest that we can reject the null-hypothesis. But even a low *P*-value from a meta-analysis can be misleading if there is also a low probability that data are compatible with the anticipated intervention effect. In other words, the probability that the actual measured difference in effect of the compared interventions resulted from an a priori anticipated 'true' difference needs to be considered. For this purpose, it is helpful to calculate the Bayes factor, which is the ratio of the P-value probabilities of the meta-analysis result divided by the probability of the anticipated effect, or 'true' effect.[29] As suggested by Jakobsen et al.,[29] a Bayes factor lower than 0.1 together with a low P-value suggest, if bias can be ruled out, that the observed result is compatible with the

a priori expected effect. If the Bayes factor is higher than 0.1 the result is not compatible with the a priori expected effect and the effect may be lower.

To assess the potential impact of missing data (incomplete outcome data bias) we did sensitivity analysis of missing data using the following strategy: a 'best-worst' case scenario was assessed, assuming that all participants lost to follow- up in the intervention group had a beneficial outcome (the group mean minus 1 standard deviation (SD)), and all those with missing outcomes in the control group have had a harmful outcome (the group mean plus 1 SD and 2 SD). In addition, the reverse 'worst-best-case' scenario analysis was also performed.[29] Missing data for the 'lack of remission' outcome was imputed in sensitivity analysis according to the following scenarios:[30] 1) poor outcome analysis: assuming that all of the drop-outs/participants lost from both the experimental and the control arms experienced the outcome, including all randomised participants in the denominator; 2) good outcome analysis: assuming that none of the drop-outs/participants lost from the experimental and the control arms experienced the outcome, including all randomised participants in the denominator; 3) extreme case analysis favouring the experimental intervention ('best-worse' case scenario): none of the drop-outs/participants lost from the experimental arm, but all of the drop-outs/participants lost from the control arm experienced the outcome, including all randomised participants in the denominator; and 4) extreme case analysis favouring the control ('worst-best' case scenario): all of the drop-outs/participants lost from the experimental arm, but none from the control arm experienced the outcome, including all randomised participants in the denominator.

Subgroup analyses

In subgroup analyses, the possible effects of variables on intervention effects on outcomes and heterogeneity were compared. Trials potentially having less risk of bias (i.e., trials with adequate allocation concealment, blinded outcome assessment, and intention to treat analysis) were compared to trials at high

risk of bias. The effect of age was assessed by comparing trials including older participants (mean age >59 years) to trials including younger participants (mean age <60 years). The effect of type of exercise was assessed by comparing trials using group exercises compared to trials using individual exercise. The effect of duration of intervention was assessed by comparing trials with short duration of intervention to trials with long duration of intervention splitting by the median time of duration. The effect of type of control group was assessed by comparing trials using attention control to trials with waitlist controls and comparing trials with exercise as add-on to medication to trials not using any medication. In addition, a within-study comparison of low-dose exercise versus high-dose exercise in trials using different exercise intensities was performed. The effect of co-morbid somatic disease was assessed by comparing the effect estimates from trials including participants with depression compared to trials including participants with depression in addition to a somatic disease. Publication bias was assessed by visual inspection of a funnel plot and by Egger's test and if publication bias plausible Duval's and Tweedie's trim and fill procedure was conducted.[31]

We assessed and graded the evidence according to the grading of recommendations assessment, development, and evaluation (GRADE) for high risk of bias, imprecision, indirectness, heterogeneity, and publication bias.[32] Based on this assessment, the intervention was graded accordingly: **'**high quality'**-** we are very confident that the true effect lies close to that of the estimate of the effect; **'**moderate quality'- we are moderately confident in the effect estimate. The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different; 'low quality**'**- our confidence in the effect estimate is limited: the true effect may be substantially different from the estimate of the effect; 'very low quality'- we have very little confidence in the effect estimate: the true effect is likely to be substantially different from the estimate of the effect.[33]

Deviations from our protocol

Post-hoc we included trials using the Chinese Classification of Mental Disorders (CCMD) as well as a few trials including participants classified as having 'minor depression'. The CCMD system closely adhere to the ICD and DSM systems and have been found highly compatible in field studies, so these studies were included.[34] A few trials included some participants classified as having 'minor depression' according to the trials chosen diagnostic system (e.g. DSM), and it is questionable if these participants have major depression. We therefore decided to include these trials but also to conduct a sub-group analysis exclusively including participants with major depression. To further explore heterogeneity, we post-hoc included sub-group analysis comparing intervention effects in inpatients and outpatients as well as an analysis according to trial size. Trials were divided into small or large trials using the median of total n included in the efficacy analysis. The effect of exercise capacity was post-hoc assessed by comparing trials with a high increase in maximal oxygen uptake (VO2max) with studies with lower increase in maximal oxygen uptake. Assessment of exercise capacity was based on the increase of VO2max in the intervention groups and trials were stratified to either high or low increase in exercise capacity by median. We did not conduct Trial Sequential Analysis based on a relative risk reduction of 30% of lack of remission as this was an implausible effect.

Participant involvement

Depressed participants were not involved in this study.

Results

Bibliographical search and trial characteristics

The main bibliographical search was conducted the 26th of August, 2015 and the final updates were conducted on the 20th of June, 2017. As illustrated in Figure S1, we identified 45 publications reporting the effect of exercise on depressive symptoms in 35 randomised clinical trials.[21;22;35-78] Seven-teen trials were conducted in Europe,[21;22;40;49;52;53;55;61;65-68;74;75;77;79;80] eight in the U.S.A.,[38;39;43;45;60;64;76;81], six in Asia,[47;69-73] two in Australia,[54;58] and two in South-America.[56;63] A total of 2,630 participants were randomised and 2,498 were included in the efficacy analysis of benefit. 10 trials included inpatients[47;49;56;67;69-73;79] and five trials included participants with a mean age above 60 years.[52;54;58;60;61] No trials exclusively included participants with comorbid somatic disease. Four trials reported the continuous outcome as mean change from baseline in each group with a corresponding SD, [39;53;65;68] and one trial presented data as mean difference between groups post-intervention.[40]. The remaining trials reported post-scores in each group with corresponding SD. Please see Table 1 for trial characteristics.

Bias risk assessment

Sequence generation was adequate in 15/35(43%), allocation concealment was adequate in 13/35 (37%) trials, blinding of participants and trial personnel was adequate in 0/35 (0%), blinded outcome assessment was performed in 16/35 (46%), low risk of bias in the 'incomplete outcome data' domain was found in 12/35 (34%) trials, selective outcome reporting domain was adequate in 31/35 (89%), for profit bias domain was adequate in 19/35 (54%) and 25/35 (71%) were free of other bias. Accordingly, all trials were at high risk of bias. Given the nature of the intervention, no trial had blinded participants or trial personnel, however, two trials had low risk of bias in all other bias domains.[22;54] Five trials (16%) were sponsored by for profit organisations: three trials were supported by pharmaceutical companies,[53;79;82] one trial by a company producing fitness machines,[45] and one trial by an insurance company.[21] According to our a priori defined criteria, 4/35 (11%) trials potentially had less risk of bias than the other trials at high risk of bias.[21;22;54;56] Please see Table 2 for details on assessment of risks of bias.

Primary outcomes

The effect of exercise on depression severity

All included trials provided a continuous outcome on depression severity for the assessment of the exercise intervention encompassing 2,498/2,630 randomised participants (95%). The effect of intervention versus control was a standardised mean difference (SMD) of -0.66 (95% CI -0.86 to -0.46; P<0.001) (Figure 1.). This corresponds to an effect on the HAM-D_{17} scale of -4.1 (95% CI -5.3 to -2.9) points.

Missing data

Missing outcome analysis for depression as a continuous outcome did not markedly change the effect estimates. The least favourable outcome for the exercise intervention was the worse/best outcome analysis using +2 SD resulting in an effect estimate of -0.57 SMD (95% CI -0.78 to -0.36; P<0.001) (Table S1).

Heterogeneity and subgroup analysis

The I^2 was 81% suggesting substantial heterogeneity. Subgroup analysis revealed that the effect estimates for trials potentially having less risk of bias was -0.11 SMD (95% CI. -0.41 to 0.18; P = 0.45; I^2 = 62%) compared to that of the trials at high risk of bias -0.75 SMD (-0.98 to -0.52; p < 0.001; I^2 = 81%) (test of sub-group difference, P < 0.001). In addition, trials including 50 participants or less had a pooled estimate of -1.11 (-1.52 to -0.72; p < 0.001; I^2 = 78%) compared to that of larger trials of -0.37 (-0.57 to -0.18; p < 0.001; I^2 = 75%) (test of sub-group difference, P = 0.001). Trials of short duration of intervention (less than 10 weeks) had a SMD of -0.92 (-1.09 to -0.74; p < 0.001; I^2 = 14%) compared to trials with longer duration of intervention, -0.49 (-0.75 to -0.23; p < 0.001; I^2 = 83%) (test of sub-group difference, P = 0.007). Effect

estimates from trials including participants with minor depression compared to trials exclusively including participants with major depression did not differ (test of sub-group difference, P = 0.53).

Four trials allocated 206 participants to different exercise intensities/doses.[45;58;73;83] Comparing the post-intervention depression scores for participants allocated to either high intensity/high dose versus low intensity/low dose exercise showed a difference of -0.40 SMD (95% CI -0.67 to -0.12; P=0.005; I^2 = 0%) in favour of high intensity/high dose exercise. As shown in Table 3, no other trial characteristic significantly explained any of the observed heterogeneity. Please see Table S2 for trial characteristics used to explore heterogeneity.

Trial Sequential Analysis and diversity adjusted required information size

The diversity adjusted required information size for HAM-D_{17} as a continuous outcome was calculated based on our anticipated intervention effect of a minimal relevant difference of 3.0 HDRS points, a standard deviation of 6.78 points, a risk of type I error of 0.05, a power of 90% and the observed diversity of 92% to 2610 participants. Only 14 trials reported results from HAM-D_{17}[21;22;38;39;43;44;52;53;55;56;58;68;70;83] with an accrued 1124 participants. As shown in Figure S2, the cumulative Z-curve just crossed the trial sequential monitoring boundary for benefit. With the aforementioned settings, the pooled estimate is therefore less likely to be a random finding due to lack of power or multiple testing if bias could be ignored. Post-hoc we calculated the adjusted required information size for HAM-D_{17} including all trials as shown in Figure S3. As with the original analysis the Z-curve crossed the trial sequential monitoring boundary for benefit supporting that the pooled estimate is less likely to represent a Type 1 error if bias could be ignored.

Bayes factor

Fourteen trials reported effect estimates using the HAM-D_{17}.[21;22;38;39;43;45;52;53;55;63;68;70;83;84] Based on these trials, Bayes factor was calculated (δ = -3.37; SE_δ = 0.96; μ_a = -3.0) and was found to be 0.002, which is below the Bayes factor threshold for significance of 0.1, supporting the intervention effect if bias could be ignored.

Publication bias

Inspection of the funnel-plot (not shown) suggested that small trials with small or no effect of exercise were missing (Figure S4). Egger's test supported the suspicion of publication bias, P < 0.00001. Using the Duval and Tweedie's trim and fill procedure, the estimate was reduced into -0.27 SMD (95% CI -0.50 to -0.05). This corresponds to an effect on the HAM-D_{17} scale of -1.7 (95% CI -3.1 to -0.31) points.

The effect of exercise on depression – lack of remission

Nineteen trials, randomising 1825 participants and including 1639 participants (90%) in final analysis reported remission as an outcome.[21;22;38-40;43;45;47;49;53;54;56;60;61;65;68-70;72] Remission post-intervention was defined in various ways: a post-intervention score on the HAM-D_{17} less than 8 points,[44;53;56;69;70] not fulfilling the DSM criteria for depression *and* a HAM-D_{17} less than 8 points,[21;22;39] not fulfilling the DSM criteria for depression,[38;54;60] a BDI score less than 9 points,[43] a BDI score less than 10 points,[40] a HAM-D_{17} score less than 10 points,[83] a MADRS score less than 10 points,[47] a MADRS score less than 10 points *and* a 50% reduction in symptom score,[65] a 75% reduction in HAM-D_{24},[72] a HAM-D_{17} score less than 11.28 points *and* a reduction in HAM-D_{17} scores > 7.74 points,[68] and one study used MADRS not specifying the cut-off for remission.[49] The RR for lack of remission was 0.78 (95% CI 0.68 to 0.90; P=0.0008) in favour of the

intervention using a random-effects analysis. The I^2 was 69% suggesting substantial heterogeneity. The forest plot for the intervention effect on lack of remission is illustrated in Figure S5.

Missing data

The scenario in least favour of the intervention was the 'poor' outcome analysis having an effect estimate of RR 0.88 (95% CI 0.83 to 0.94) P=0.0002; I^2 = 69%. As shown in Table S1, the remaining scenarios did not substantially differ from the main analysis.

Heterogeneity and subgroup analysis

I^2 was 69% for the outcome lack of remission suggesting substantial heterogeneity. For this outcome, only two trials[22;84] were considered as trials potentially having less risk of bias than the other trials at high risk of bias. The RR of these two trials was 0.95 (95% CI 0.74 to 1.23; P=0.78) compared to 0.77 (96% CI 0.64 to 0.92; P=0.003) for trials at high risk of bias, test of subgroup difference, P=0.19). Trials including 52 participants or less in their final analysis had a RR 0.62 (95% CI 0.50 to 0.76; P<0.001; I^2 = 45%) compared to 0.95 (95% CI 0.80 to 1.12; P=0.52; I^2 = 68%) for larger trials (test of sub-group difference, P=0.002). Also, trials with a duration of less than 10 weeks had a RR of 0.63 (95% CI 0.51 to 0.77; P<0.001; I^2 = 40%) compared to 0.93 (95% CI 0.78 to 1.10; P=0.39; I^2 = 69%) for trials of a longer duration (test of sub-group difference, P=0.004). As shown in Table S3, no other trial characteristic significantly explained any of the observed heterogeneity. Please see Table S2 for trial characteristics used to explore heterogeneity.

Trial Sequential Analysis and diversity adjusted required information size

The diversity adjusted required information size for lack of remission was calculated based on our observed diversity of 74%, a proportion in the control group with lack of remission of 66%, an anticipated

intervention effect of 15% relative risk reduction, a risk of type I error of 0.05% and a power of 90%. As shown in Figure S6, the cumulative Z curve just crossed the trial sequential monitoring boundary for benefit. With the aforementioned settings, the pooled estimate is therefore less likely to be a random finding due to lack of power or multiple testing if bias could be ignored.

Bayes factor

Bayes factor was calculate based on the observed relative risk of remission, the associated standard error, and an anticipated intervention effect of relative increase in number of participants with remission by 15% (δ = -0.248; SE_δ= 0.08; μ_δ = -0.163). Bayes factor was 0.02, which is below the Bayes factor threshold for significance of 0.1.

Publication bias

Inspection of the funnel-plot (not shown) suggested that small trials with small or no effect of exercise were missing. Egger's test supported the suspicion of publication bias, P=0.002. Imputing theoretically missing studies by the Duval and Tweedie's trim and fill procedure, reduced the estimate of intervention effect into a relative risk reduction of 0.93 (95% CI 0.79 to 1.11).

The effect of exercise on serious adverse events

Serious adverse events (i.e., death or suicide attempts) were reported in only three trials.[21;22;58] In these trials, one suicide attempt[22] and one death by suicide[21] were recorded in the intervention groups. The RR

for death or suicide in the two trials was 2.21 (95% CI 0.24 to 20.21; P=0.48; I^2 = 0%) as illustrated in Figure S7.

Missing data

Missing outcome analysis for 'serious adverse events' varied according to missing data scenario: poor outcome analysis relative risk, 0.92 (95% CI 0.37 to 2.30; P=0.86; I^2 = 60.0%), good outcome analysis, 2.19 (95% CI 0.23 to 20.76; P=0.50; I^2 = 0.0%), best/worst outcome analysis – 0.08 (95% CI 0.02 to 0.34; P=0.001; I^2 = 5.4%), worst/best outcome analysis 19.17 (95% CI 2.64 to 139.2; P=0.004; I^2 = 0.0%).

Trial Sequential Analysis and Bayes analysis

We decided not to conduct Trial Sequential Analysis or Bayes analysis due to too sparse data.

Publication bias

Only 3/35 trials reported on this outcome and no formal assessment for publication bias was made. However, the lack of reporting in the vast majority of trials suggest risk publication bias.

Secondary outcomes

The effect of exercise on quality of life

Nine trials randomising 827 participants reported on quality of life,[21;22;38;40;56;60;71;76;85] observing that participants allocated to exercise did not have significantly better quality of life (SMD 0.40; 95% CI -0.03 to 0.83; P=0.07). The I^2 was 88% showing substantial heterogeneity (Figure S8).

Non-serious adverse events

Non-serious adverse events were reported in only ten trials.[21;22;39;56;58;60;65;67;68;75] Five trials reported on musculoskeletal adverse events without conducting formal tests[58;60;65;67;68] and four trials reported on number of participants with high depression scores post-intervention compared to baseline assessment.[21;22;65;68] The RR for increased severity of depression in patients allocated to exercise post-intervention was 0.83 (95% CI 0.40 to 1.70; P=0.60; I^2 = 0.0%).

The effect of exercise on depression beyond the duration of the intervention

Assessment of depression beyond the intervention was conducted in seven trials,[21;38;40;52;60;63;86] with a median duration between end of intervention and assessment of depression of 6 months (range 5 to 23.5 months). The SMD between the intervention group and the control group using a random effects analysis was -0.10 (95% CI -0.28 to 0.09; P=0.31; I^2 = 19.5%). The I^2 for this estimate was 19.5% suggesting low heterogeneity (See Figure S9).

Remission beyond the intervention was assessed in five trials,[21;38-40;54] and the relative risk of lack of remission was 0.95 (95% CI 0.82 to 1.11; P=0.53) with an I^2 of 0.0% (See Figure S10).

GRADE assessments

The GRADE assessments are presented in Table 4, and quality of evidence for both primary and secondary outcomes was very low or low.

Additional analysis

Four studies reported change in scores from baseline with corresponding SD's, and one study reported mean difference between groups post-intervention. Comparing the effect size of these five studies with the remaining did not seem to explain part of the heterogeneity (p = 0.23).

Discussion

Thirty-five clinical trials allocating more than 2498 participants diagnosed with depression according to validated diagnostic instruments were included in the present systematic review. Pooled estimates suggested moderate antidepressant effect assessed both as a continuous outcome and as lack of remission. Due to risk of bias, inconsistency of effect estimates, and publication bias we have, however, very little confidence in these effect estimates. Subgroup analyses exploring reasons for the heterogeneity found that trials potentially having less risk of bias than other trials at high risk of bias had no effect of exercise on depression. Furthermore, duration of intervention and trial size were inversely associated with effect estimates. Exercise did not improve quality of life or depression or remission after the intervention. Serious adverse event or adverse events were reported inconsistently and only by a few trials not permitting firm conclusions regarding these outcomes.

Strengths and limitations

The strengths of this systematic review are that it is based on the published protocol, a comprehensive search strategy, and the inclusion of patient centered outcomes such as quality of life as well as adverse events. Also, to avoid spurious finding from repeated testing, Trial Sequential Analysis and Bayes analysis were undertaken and these analyses did not suggest that the pooled estimates could be reduced to random errors for effect on depression severity or no remission. Neither Trial Sequential Analysis nor Bayes

factor analysis are, however, able to wash of spurious effects induced by bias, fraud or other reasons.[26;29;87-89] Had we restricted the Trial Sequential Analysis to trials of potentially lower risk of bias, the number of trials and participants would be limited and we had seen evidence far from crossing any boundaries for benefit, harms, or futility. The conclusions for serious adverse events and adverse events were associated with wide confidence intervals due to lack of data and firm conclusions for these outcomes are presently not available.

The number of trials with adequate allocation concealment was 37% in the current systematic review compared to only 15.1% in trials assessing non-drug interventions for depression.[90] Blinded outcome assessment was performed in 46% of the included trials compared to 44% in non-drug antidepressant trials in general.[90] The incomplete outcome bias domain was adequate in 34% of our included trials compared to 32.9% of antidepressant non-drug trials in general.[90] Compared to non-drug trials assessing interventions for participants with depression, the included exercise trials have more bias domains with low risk of bias. However, all our included trials were at high risk of bias. Two trials had low risk of bias for all bias domains except for blinding of participants and trial personnel, and four trials fulfilled our criteria for trials at potentially less risk of bias than the rest of the trials with at risk of bias. Despite a search strategy including bibliographical databases and trials from China and South-America, the vast majority of included trials were conducted in north America and western Europe, which is comparable to the geographical distribution of non-drug trials in general[90] limiting the applicability to other geographic regions.

All outcomes for the primary analysis reflect depression severity, however, the different psychometrics may represent different aspects of depression not reflected in the pooled estimate. An in-depth discussion of the included assessment scales is beyond the scope of this review, but in the current systematic review we

found no significant differences of effect estimates from trials using HAM-D_{17} compared to trials using other assessment scales (data not shown).

The effect of exercise on depression

Our present results are similar to the latest Cochrane review by Cooney et al. (2013)[24] who found a moderate effect of exercise on depressive symptoms (-0.62 SMD) when including all trials and no effect when restricting the analysis to trials with less risk of bias (-0.18 SMD). The Cochrane review did find evidence of a small antidepressant effect beyond the intervention, which we could not confirm in our present systematic review. Bridle et al. (2012)[13] included 9 trials allocating old (> 60 years) participants with depression to exercise interventions versus control interventions. Restricting the analysis to four trials at lower risk of bias they found small to moderate effect estimates (SMD -0.34) in favour of exercise. The studies by Cooney et al.[24] and Bridle et al.[13] both included trials allocating participants with depressive symptoms and not necessarily diagnosed using a validated diagnostic system, potentially explaining the differences in the effect sizes. However, in our present systematic review the estimate for four trials at potential less risk of bias than the remaining trials was -0.11 SMD and in the Cooney study the effect estimate for eight trials with lower risk of bias was -0.18 SMD[24] compared to -0.34 in the study by Bridle at al.[13] Meta-analysis of randomised clinical trials assessing the effects of exercise for depression consistently finds positive effects, however, when restricting the analysis to trials with less risk of bias the pooled effect sizes becomes very small or negligible. Meta-analysis examining the effect of exercise beyond the intervention also finds no or small effects of exercise. In the process of interpretation of effect estimates in the current research field, it is important to recognise that effect estimates from trials with non-blinded outcome assessment are at high risk of bias as reported by Savovic et al.[91] Sixteen of 35 trials in the current systematic review did not use blinded outcome assessment. In contradiction to the current systematic review, a recent meta-analysis by Schuch et al.[12] concluded that “exercise has a large and significant

antidepressant effect in people with depression……Our data strongly support the claim that exercise is an evidence-based treatment for depression". This statement was based on a meta-analysis of 25 randomised clinical trials including participants with depression or depressive symptoms to exercise or control conditions and excluding trials using any form of active control group. Surprisingly, the authors found that adjusting for publication bias using the Trim and Fill procedure[31] the estimate *increased* from a SMD of 0.98 to 1.11. The effect in SMD in included studies ranged from -0.23 to 4.56 representing considerable heterogeneity.[12] The authors classified four trials as having lower risk of bias using the same criteria as in our systematic review and 21 trials as having high risk of bias. This illustrates some of the challenges in meta-analysis of exercise and depression: the large heterogeneity driven by small studies inflating the effects of random-effects analysis,[92] the misconception that we can restrict our analysis to statistics and not consider the evident effect of bias.[23;91] Compared to our previous review,[10] we now included 35 trials including 2498 participants versus previously 13 trials and 687 participants. It may seem as a paradox that this large increase in data has not provided us with a similar increase in certainty of conclusions reflected by heterogeneity of trial results as well as our conclusions from the systematic reviews. The increase in available data is, however, primarily provided by small trials at high risk of bias introducing exaggerated effect estimates. In the current systematic review, we included four trials with 530 participants at lower risk of bias compared to three trials with 239 participants in our previous review, reflecting that only a small part of the additional data comes from trials at lower risk of bias. The continuous increase in data associated with high risk of bias will not provide patients, clinicians or policymakers with adequate information and represents an unethical enrollment of trial participants and waste of resources.[93-99] We therefore recommend that future systematic reviews and meta-analysis a priori should have a primary outcome restricting effect analysis to larger trials with lower risk of bias and that any recommendations regarding exercise interventions for participants with depression should be assessed with the GRADE framework.

The I^2 of 81% and 69% for the primary outcomes indicate substantial evidence of heterogeneity of intervention effects that is variation in effect estimates beyond chance. Part of this heterogeneity was explained by bias and by trial size: trials at high risk of bias or small trials have very large effect estimates compared to trials potentially at less risk of bias or larger trials. The funnel plots and Egger's test indicates publication bias, however, the association between trial size and effect estimates could suggest that the asymmetry in the funnel plots are due to small study bias rather than publication bias.[100] It could be argued that both the delivery of exercise as well as the actual increase in fitness are fundamental to the assessment of the antidepressant effects of exercise, and in line with our previous review we found duration of intervention inversely associated with effect size.[11] Comparing different exercise intensities, we did find a small effect of high intensity exercise compared to lower intensity exercise. However, assessing delivered exercise expressed as increase in maximal oxygen uptake we could not reproduce this finding. Future trials need to pay more attention to the dose of the intervention as well as compliance with intervention.[101] We suggest using maximal oxygen uptake or 1 repetition maximum as the gold standards to assess the received exercise. Several studies compare exercise to control interventions rather than wait-list control to reduce the effect of non-specific effects, e.g. the DEMO trials and Mather et al.[21;22;52] Also, it could be speculated that the effect of exercise would be harder to detect if participants also received medical treatment in addition. The current systematic review could not confirm that the type of control condition explained heterogeneity. The discussion of control group is important in non-drug trials: choosing a waitlist control group the results potentially reflects non-specific effects, choosing an active control group (e.g., relaxation exercise) the trial is potentially a comparison between two active treatments. However, in the current systematic review we found no evidence that trials using an attention control group or exercise as add-on to pharmacotherapy had significantly different effect estimates compared to other trials.

Our systematic review did not find indications of a positive effect on quality of life in participants with depression allocated to exercise interventions, which is in concordance with the review by Cooney et al.[24] Only 3/35 trials reported on serious adverse events, and we found no significant effects of exercise on risk of death or suicide attempt. No indication of increased severity of depression or other adverse events in participants allocated to exercise could be detected. However, data on adverse events was reported sporadically in a minority of trials and currently it is not possible to conclude on the risk of serious adverse events or adverse event from exercise interventions in participants with depression.

Conclusions

We have little confidence in the pooled effect estimates, especially because trials with less than high risk of bias produced significantly lower effect estimates, suggesting that exercise interventions only produce small or negligible antidepressant effects, depending on how much of the effect is caused by bias and how much is caused by the intervention. There was no effect of exercise on depression beyond the intervention itself. We found no effect on quality of life. There is currently no evidence in favour of exercise for patients with depression with a view to ameliorate depressive symptoms. Our systematic review did not evaluate possible beneficial effects of exercise on, e.g., metabolism or cardiovascular fitness,[22;102] and it is possible that exercise may have beneficial effects on these factors in patients diagnosed with depression.

Future perspectives

Despite the large number of published trials, further trials with more robust methodology seem still required to establish progress in this field. Also, additional trials from outside North-America and Europe may be required for results to be valid for patients in Asia, Africa and South-America. To further elaborate on the current findings, we recommend that future trials must include blinded outcome assessors and

outcomes assessing quality of life, metabolic effects, and long-term effects beyond the intervention. It is also important that future trials systematically collect and report data on death, suicide events, musculoskeletal injuries and other potential adverse effects in both the intervention group as well as in the control group. Moreover, future trials ought to be designed according to the SPIRIT guidelines and reported according to the CONSORT guidelines[103;104] and transparently report deidentified individual participant data enabling individual participant data meta-analyses.[105]

Funding

No funding was obtained for this systematic review.

Competing interests

JK, CG, and MN have previously published two trials and a meta-analysis on this topic, which could introduce an academic bias in the current systematic review. We asked new authors (HS and CH) to be involved in the preparation of the protocol, trial selection and bias assessment. No support from any organisation was received for the submitted work; no financial relationship with any organisations that might have an interest in the submitted work in the previous three years; and apart from the above no other relationship or activities that could appear to have influenced the submitted work.

Acknowledgement

We appreciate the help from Youling He for help with the Chinese Wanfang bibliographical database and translation of Chinese papers. We thank Janus C Jakobsen for assistance with the calculation of Bayes factor.

Contributors

JK conceived the project, collected data, did the statistical analysis, analysed the data, drafted and revised the manuscript. He is guarantor. CH collected the data, analysed the data and revised the

manuscript. HS conceived the project, collected data, analysed the data, and revised the manuscript. CG conceived the project, analysed the data and revised the manuscript. MN conceived the project, analysed the data, and revised the manuscript.

Data sharing statement

All data used in this study are available in Figures and Fables. No other data were used.

Reference List

(1) Lepine JP, Gastpar M, Mendelwicz J, Tylee A. Depression in the community: The first pan-european study DEPRES (Depression Research in European Society). *Int Clin Psychopharmocol* 1997; 12:19-29.

(2) Pedersen CB, Mors O, Bertelsen A, *et al.* A comprehensive nationwide study of the incidence rate and lifetime risk for treated mental disorders. *JAMA Psychiatry* 2014; 71(5):573-581.

(3) Ustun TB, Ayuso-Mateos JL, Chatterji S, *et al.* Global burden of depressive disorders in the year 2000. *Br J Psychiatry* 2004; 184:386-392.

(4) Kirsch I, Deacon B, Huedo-Medina T, *et al.* Initial severity and antidepressant benefits: A meta-analysis of data submitted to the Food and Drug Administration. *PLoS Med* 2008; 5[2], e45. doi:10.1371/journal.pmed.0050045.

(5) Turner EH, Matthews AM, Linardatos E, *et al.* Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy. *N Engl J Med* 2008; 358(3):252-260.

(6) Jakobsen JC, Katakam KK, Schou A, *et al.* Hellmuth SG, Stallknecht SE, Leth-Moller K et al. Selective serotonin reuptake inhibitors versus placebo in patients with major depressive disorder. A systematic review with meta-analysis and Trial Sequential Analysis. *BMC Psychiatry* 2017; 17(1):58.

(7) Jakobsen JC, Hansen JL, Storebo OJ, *et al.* The effects of cognitive therapy versus 'no intervention' for major depressive disorder. *PLoS One* 2011; 6(12):e28299.

(8) Jakobsen JC, Hansen JL, Simonsen E, Gluud C. The effect of interpersonal psychotherapy and other psychodynamic therapies versus 'treatment as usual' in patients with major depressive disorder. *PLoS One* 2011; 6(4):e19044.

(9) Jakobsen JC, Hansen JL, Simonsen E, Gluud C. The effect of adding psychodynamic therapy to antidepressants in patients with major depressive disorder. A systematic review of randomized clinical trials with meta-analyses and trial sequential analyses. *J Affect Disord* 2012; 137(1-3):4-14.

(10) Cipriani A, Furukawa TA, Salanti G, *et al.* Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *Lancet* 2009; 373: 746-758.

(11) Krogh J, Nordentoft M, Sterne J, Lawlor D. The effect of exercise in clinically depressed adults: Systematic review and meta-analysis of randomized controlled trials. *J Clin Psychiatry* 2011; 72:529-538.

(12) Schuch FB, Vancampfort D, Richards J, *et al.* Exercise as a treatment for depression: A meta-analysis adjusting for publication bias. *J Psychiatr Res* 2016; 77:42-51.

(13) Bridle C, Spanjers K, Patel S, *et al.* Effect of exercise on depression severity in older people: systematic review and meta-analysis of randomised controlled trials. *Br J Psychiatry* 2012; 201(3):180-185.

(14) Herring MP, Puetz TW, O'Connor PJ, Dishman RK. Effect of exercise training on depressive symptoms among patients with a chronic illness: a systematic review and meta-analysis of randomized controlled trials. *Arch Intern Med* 2012; 172(2):101-111.

(15) World Health Organisation. International Statistical Classification of Diseases, 10th Revision (ICD-10). *Geneva, Switzerland* 1992.

(16) American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, 4th ed. *Washington DC* 1994.

(17) Beck AT, Steer RA, Brown GK. BDI-II manual. 2 ed. New York: Psychological Corporation, 1996.

(18) Krogh J, Speyer H, Gluud C, Nordentoft M. Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential analysis. *Syst Rev* 2015; 4(1):40.

(19) Spitzer RL, Endicott J, Williams JB. Research diagnostic criteria. *Arch Gen Psychiatry* 1979; 36(12):1381-1383.

(20) ICH-secretariat. International Conference on Harmonisation (ICH) Harmonised Tripartite Guideline (Step 4). Guideline for Good Clinical Practice. *Geneva* 1996.

(21) Krogh J, Saltin B, Gluud C, Nordentroft M. The DEMO trial: A randomized, parallel-group, observer-blinded clinical trial of strength versus aerobic versus relaxation training for patients with mild to moderate depression. *J Clin Psychiatry* 2009; 70[6]: 790-800.

(22) Krogh J, Videbech P, Thomsen C, *et al.* DEMO-II trial.Aerobic Exercise Versus Stretching Exercise in Patients With Major Depression - A Randomised Clinical Trial. *PLoS One* 2012; 7(10): e48316. doi:10.1371/journal.pone.0048316.

(23) Higgins JPT GSe. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0. The Cochrane Collaboration. Available from www.cochrane-handbook.org The Cochrane Collaboration, 2009.

(24) Cooney GM, Dwan K, Greig CA, *et al.* Exercise for depression. *Cochrane Database Syst Rev* 2013; 9:CD004366.

(25) Higgins JPT, Thompson SG, Deeks JJ, Altmann DG. Measuring inconsistency in meta-analysis. *BMJ* 2003; 327:557-560.

(26) Wetterslev J, Thorlund K, Brok J, Gluud C. Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. *J Clin Epidemiol* 2008; 61(1):64-75.

(27) Brok J, Thorlund K, Gluud C, Wetterslev J. Trial sequential analysis reveals insufficient information size and potentially false positive results in many meta-analyses. *J Clin Epidemiol* 2008; 61(8):763-769.

(28) Thorlund K, Walter SD, Johnston BC, *et al.* Pooling health-related quality of life outcomes in meta-analysis-a tutorial and review of methods for enhancing interpretability. *Res Synth Methods* 2011; 2(3):188-203.

(29) Jakobsen JC, Wetterslev J, Winkel P, *et al.* Thresholds for statistical and clinical significance in systematic reviews with meta-analytic methods. *BMC Med Res Methodol* 2014; 14:120.

(30) Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. *BMJ* 1999; 319(7211):670-674.

(31) Duval S, Tweedie R. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* 2000; 56(2):455-463.

(32) Guyatt GH, Oxman AD, Vist GE, *et al.* GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008; 336(7650):924-926.

(33) Andrews J, Guyatt G, Oxman AD, *et al.* GRADE guidelines: 14. Going from evidence to recommendations: the significance and presentation of recommendations. *J Clin Epidemiol* 2013; 66(7):719-725.

(34) Zheng YP, Lin KM, Zhao JP, *et al.* Comparative study of diagnostic systems: Chinese Classification of Mental Disorders-Second Edition versus DSM-III-R. *Compr Psychiatry* 1994; 35(6):441-449.

(35) Babyak MA, Blumenthal JA, Herman S, *et al.* Exercise treatment for major depression: Maintenance of therapeutic benefit at 10 months. *Psychosom Med* 2000; 62:633-638.

(36) Baxter H, Winder R, Chalder M, *et al.* Physical activity as a treatment for depression: the TREAD randomised trial protocol. *Trials* 2010; 11:105.

(37) Benson M, Blumenthal J, Babyak M, *et al.* Exercise fails to improve neurocognition in depressed middle-aged and older adults. *Med Sci Sports Exerc* 2008; 40[7]: 1344-1352.

(38) Blumenthal J, Babyak MA, Moore KA, *et al.* Effects of exercise training on older patients with major depression. *Arch Int Med* 1999; 159:2349-2356.

(39) Blumenthal J, Babyak M, Doraiswamy P, *et al.* Exercise and pharmacotherapy in the treatment of major depressive disorder. *Psychosom Med* 2007; 69: 587-596.

(40) Chalder M, Wiles NJ, Campbell J, et al. Facilitated physical activity as a treatment for depressed adults: randomised controlled trial. *BMJ* 2012; 344:e2758.

(41) Chalder M, Wiles NJ, Campbell J, *et al.* A pragmatic randomised controlled trial to evaluate the cost-effectiveness of a physical activity intervention as a treatment for depression: the treating depression with physical activity (TREAD) trial. *Health Technol Assess* 2012; 16(10):1-iv.

(42) Combs K, Smith PJ, Sherwood A, *et al.* Impact of sleep complaints and depression outcomes among participants in the standard medical intervention and long-term exercise study of exercise and pharmacotherapy for depression. *J Nerv Ment Dis* 2014;202(2):167-71

(43) Doyne EJ, Ossip-Klein DJ, Bowman ED, *et al.* Running versus weight lifting in the treatment of depression. *J Consult Clin Psychol* 1987; 55(5):748-754.

(44) Dunn AL, Trivedi MH, Kampert JB, *et al.* The DOSE study: a clinical trial to examine efficacy and dose response of exercise as treatment for depression. *Control Clin Trials* 2002; 23(5):584-603.

(45) Dunn AL, Trivedi MH, Kaplan K, *et al.* Exercise treatment for depression. Efficacy and dose response. *Am J Prev Med* 2005; 28: 1-8.

(46) Herman S, Blumenthal JA, Babyak M, *et al.* Exercise therapy for depression in middle-aged and older adults: predictors of early dropout and treatment failure. *Health psychol* 2002; 21:553-563.

(47) Ho CWH, Chan SC, Wong JS, *et al.* Effect of aerobic exercise training on chinese population with mild to moderate depression in Hong Kong.

(48) Hoffman BM, Babyak MA, Craighead WE, Sherwood A, Doraiswamy PM, Coons MJ et al. Exercise and pharmacotherapy in patients with major depression: one-year follow-up of the SMILE study. *Psychosom Med* 2011; 73(2):127-133.

(49) Kerling A, Tegtbur U, Gutzlaff E, Kuck M, Borchert L, Ates Z et al. Effects of adjunctive exercise on physiological and psychological parameters in depression: A randomized pilot trial. *Rehabil Res Pract* 2014;2014:627376

(50) Khatri P, Blumenthal J, Babyak M, *et al.* Effects of exercise training on cognitive functioning among depressed older men and women. *J Aging Phys Act* 2001; 9: 43-57.

(51) Krogh J, Petersen L, Timmermann M, *et al.* Design paper: The DEMO trial: A randomized, parallel-group, observer-blinded clinical trail of aerobic versus non-aerobic versus relaxation training for patients with light to moderate depression. *Contemp Clin Trials* 2007; 28:79-89.

(52) Mather AS, Rodriguez C, Guthrie MF, *et al.* Effects of exercise on depressive symptoms in older adults with poorly responsive depressive disorder: randomised controlled trial. *Br J Psychiatry* 2002; 180:411-415.

(53) Mota-Pereira J, Silverio J, Carvalho S, *et al.* Moderate exercise improves depression parameters in treatment-resistant patients with major depressive disorder. *J Psychiatr Res* 2011; 45(8):1005-1011.

(54) Pfaff JJ, Alfonso H, Newton RU, *et al.* ACTIVEDEP: a randomised, controlled trial of a home-based exercise intervention to alleviate depression in middle-aged and older adults. *Br J Sports Med* 2014; 48(3):226-232.

(55) Pilu A, Sorba M, Hardoy M, *et al.* Efficacy of physical activity in the adjunctive treatment of major depressive disorders: preliminary results. *Clin Pract Epidemio Ment Health* 2007;3:8.

(56) Schuch FB, Vasconcelos-Moreno MP, Borowsky C, Fleck MP. Exercise and severe depression: preliminary results of an add-on study. *J Affect Disord* 2011; 133(3):615-618.

(57) Singh NA, Clements KM, Fiatarone MA. A randomized controlled trial of the effect of exercise on sleep. *Sleep* 1997; 20(2):95-101.

(58) Singh NA, Stavrinos TM, Scarbek Y, *et al.* A randomized controlled trial of high versus low intensity weight training versus general practitioner care for clinical depression in older adults. *J Gerontol A Biol Sci Med Sci* 2005; 60(6):768-776.

(59) Singh NA, Clements KM, Singh MA. The efficacy of exercise as a long-term antidepressant in elderly subjects: a randomized, controlled trial. *J Gerontol A Biol Sci Med Sci* 2001; 56(8):M497-M504.

(60) Singh N, Clements K, Fiatarone M. A randomized controlled trial of progressive resistance training in depressed elders. *Journal of Gerontology* 1997; 52:M27-M35.

(61) Toni G, Belvederi MM, Zanetidou S, *et al.* Physical exercise for late life Major Depression: The SEEDS study. *Eur J Prev Cardiol* 2014; 21:S107.

(62) Veale D, Le FK, Pantelis C, *et al.* Aerobic exercise in the adjunctive treatment of depression: a randomized controlled trial. *J R Soc Med* 1992; 85(9):541-544.

(63) Vieira JLL, Porcu M, da Rocha PGM. The regular physical exercise's practice as complementary therapy to the treatment of depressed women. *J Bras Psiquiatr* 2007;56(1):23-8.

(64) Epstein D. Reno: University of Nevada:Aerobic activity versus group cognitive therapy: An evaluative study of contrasting interventions for the alleviation of clinical depression *(dissertation),* 1986.

(65) Danielsson L, Papoulias I, Petersson E, *et al.* Exercise or basic body awareness therapy as add-on treatment for major depression: A controlled study. *J Affect Disord* 2014; 168:98-106.

(66) Carneiro L, Fonseca A, Vieira-Coelho M, *et al.* Effects of structured exercise and pharmacotherapy vs pharmactherapy for adults with depressive symptoms: A randomized clinical trial. *J Psychitr Res* 2015;(71):48-55.

(67) Legrand FD, Neff EM. Efficacy of exercise as an adjunct treatment for clinically depressed inpatients during the initial stages of antidepressant pharmacotherapy: An open randomized controlled trial. *J Affect Disord* 2016; 191:139-144.

(68) Doose M, Ziegenbein M, Hoos O, *et al*. Self-selected intensity exercise in the treatment of major depression: A pragmatic RCT. *Int J Psychiatry Clin Pract* 2015; 19(4):266-275.

(69) Salehi I, Hosseini SM, Haghighi M, *et al.* Electroconvulsive therapy (ECT) and aerobic exercise training (AET) increased plasma BDNF and ameliorated depressive symptoms in patients suffering from major depressive disorder. *J Psychiatr Res* 2016; 76:1-8.

(70) Huipeng F, Xiaohui Z. Curative effect and observation of exercise on patients with depression. *Chin Med Pharm* 2013; 3:187-188.

(71) Jinchun W, Bo Y, Rong X, *et al.* Effect of exercise therapy in improving social function and quality of life in patients with depression. *Nurs J Chin PLA* 2015; 32(21):24.

(72) Guifeng S. The role of setting - up exercise on rehabilitation in patients with depressive episode. *Chin J Health Psychol* 2015; 23:184-186.

(73) Fang D, Min R, Ai-qin W. The effect of aerobic exercise with different intensity and frequency on depressive symptoms. *Chin J Behav Med & Brain Sci* 2013; 22:830-832.

(74) Olson RL, Brush CJ, Ehmann PJ, *et al.* A randomized trial of aerobic exercise on cognitive control in major depression. *Clin Neurophysiol* 2017; 128(6):903-913.

(75) Euteneuer F, Dannehl K, Del RA, *et al.* Immunological effects of behavioral activation with exercise in major depression: an exploratory randomized controlled trial. *Transl Psychiatry* 2017; 7(5):e1132.

(76) Patten C, Bronars C, Douglas K, *et al.* Supervised, Vigorous Intensity Exercise Intervention for Depressed Female Smokers: A Pilot Study. *Nicotine Tob Res* 2017; 19:77-86.

(77) Pentecost C, Farrand P, Greaves CJ, *et al.* Combining behavioural activation with physical activity promotion for adults with depression: findings of a parallel-group pilot randomised controlled trial (BAcPAc). *Trials* 2015; 16:367.

(78) Farrand P, Pentecost C, Greaves C, *et al.* A written self-help intervention for depressed adults comparing behavioural activation combined with physical activity promotion with a self-help intervention based upon behavioural activation alone: study protocol for a parallel group pilot randomised controlled trial (BAcPAc). *Trials* 2014; 15:196.

(79) Martinsen EW, Medhus A, Sandvik L. Effects of aerobic exercise on depression: A controlled study. *BMJ* 1985; 291:109.

(80) Veale D, Fevre K, Pantalis C, *et al.* Aerobic exercise in the adjunctive treatment of depression: A randomised controlled trial. *J R Soc Med* 1992; 85:541-544.

(81) Ossip-Klein DJ, Doyne EJ, Bowman ED, *et al.* Effects of running or weight lifting on self-concept in clinically depressed women. *J Consult Clin Psychol* 1989; 57(1):158-161.

(82) Blumenthal JA, Babyak MA, Moore KA, *et al.* Effects of exercise training on older patients with major depression. *Arch Intern Med* 1999; 159(19):2349-2356.

(83) Belvederi MM, Amore M, Menchetti M, *et al.* Physical exercise for late-life major depression. *Br J Psychiatry* 2015; 207(3):235-242.

(84) Schuch FB, Vasconcelos-Moreno MP, Borowsky C, *et al.* Exercise and severe major depression: effect on symptom severity and quality of life at discharge in an inpatient cohort. *J Psychiatr Res* 2015; 61:25-32.

(85) Carta MG, Hardoy MC, Pilu A, *et al.* Improving physical quality of life with group physical activity in the adjunctive treatment of major depressive disorder. *Clin Pract Epidemiol Ment Health* 2008; 4:1.doi: 10.1186/1745-0179-4-1.

(86) Klein MH, Greist JH, Gurman AS, *et al.* A comparative outcome study of group psychotherapy vs. exercise treatments for depression. *Int J Ment Health* 1985; 13:148-177.

(87) Roberts I, Ker K, Edwards P, *et al.* The knowledge system underpinning healthcare is not fit for purpose and must change. *BMJ* 2015; 350:h2463.

(88) Gluud C, Jakobsen J, Imberger G, *et al.* Re: The knowledge system underpinning healthcare is not fit for purpose and must change - Responses to the opposing viewpoints of Roberts and colleagues and Tovey and colleagues. *BMJ* 2015; 350(h2463).

(89) Thorlund K, Engstrøm J, Wetterslev J, *et al.* User manual for Trial Sequential Analysis (TSA). *Copenhagen Trial Unit ,Centre for Clinical Intervention Research, Copenhagen, Denmark P1-115, Available from https://www ctu dk/tsa* 2011.

(90) Krogh J, Hjorthoj CR, Jakobsen JC, *et al.* DEPERROR: Risks of systematic errors in drug and non-drug randomized clinical trials assessing intervention effects in patients with unipolar depression. *J Affect Disord* 2015; 179:121-127.

(91) Savovic J, Jones HE, Altman DG, *et al.* Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. *Ann Intern Med* 2012; 157(6):429-438.

(92) Poole C, Greenland S. Random-effects meta-analyses are not always conservative. *Am J Epidemiol* 1999; 150(5):469-475.

(93) Ioannidis JP, Greenland S, Hlatky MA, *et al*. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet* 2014; 383(9912):166-175.

(94) Chan AW, Song F, Vickers A, *et al.* Increasing value and reducing waste: addressing inaccessible research. *Lancet* 2014; 383(9913):257-266.

(95) Moher D, Glasziou P, Chalmers I, *et al.* Increasing value and reducing waste in biomedical research: who's listening? *Lancet* 2016; 387(10027):1573-1586.

(96) Glasziou P, Altman DG, Bossuyt P, *et al.* Reducing waste from incomplete or unusable reports of biomedical research. *Lancet* 2014; 383(9913):267-276.

(97) Chalmers I, Bracken MB, Djulbegovic B, *et al.* How to increase value and reduce waste when research priorities are set. *Lancet* 2014; 383(9912):156-165.

(98) Al-Shahi SR, Beller E, Kagan J, *et al.* Increasing value and reducing waste in biomedical research regulation and management. *Lancet* 2014; 383(9912):176-185.

(99) Garattini S, Jakobsen JC, Wetterslev J, *et al.* Evidence-based clinical practice: Overview of threats to the validity of evidence and how to minimise them. *Eur J Intern Med* 2016; 32:13-21.

(100) Sterne JA, Sutton AJ, Ioannidis JP, *et al.* Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ* 2011; 343:d4002.

(101) Slade SC, Dionne CE, Underwood M, Buchbinder R. Consensus on Exercise Reporting Template (CERT): Explanation and Elaboration Statement. *Br J Sports Med* 2016.

(102) Stubbs B, Rosenbaum S, Vancampfort D, *et al.* Exercise improves cardiorespiratory fitness in people with depression: A meta-analysis of randomized control trials. *J Affect Disord* 2016; 190:249-253.

(103) Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. *Ann Intern Med* 2010; 152(11):726-732.

(104) Chan AW, Tetzlaff JM, Gotzsche PC, *et al.* SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. *BMJ* 2013; 346:e7586.

(105) Skoog M, Saarimäki J, luud C, heinin M, rlandsson K, Aamdal S. Transparency and registration in clinical research in the Nordic countries. Nordic Trial Alliance, *NordForsk* 2015; 1-108

Figure legends

Figure 1. Effect of exercise on depression severity in patients diagnosed with depression

Table 1. Characteristics of trials assessing exercise for patients diagnosed with depression

Author, first Country of origin	Participants	Severity of depression at baseline	N at baseline (included in trial efficacy analysis)	Type of intervention	Frequency	Duration
Klein 1985 USA	Outpatients Mean age: 30 (SD 7) 72% female	SCL-D: 2.4 (SD 1)	50 (22)	*Aerobic exercise:* Supervised individual running. *Control group*: Supervised meditation in groups	2 sessions per week Control group: 1 session per week	12 weeks
Martinsen 1985 Norway	Inpatients Mean age: 40 (range 17-60 Distribution of sex not reported	BDI: 28.0 (SD 9)	49(43)	*Aerobic exercise*: Supervised group exercise. *Control group*: Occupational therapy.	3 sessions per week Control group: 3 sessions per week	9 weeks
Epstein 1986 USA	Outpatients Mean age: 39 (range 24 to 60) (NR) % female	BDI: 23.4 (SD 7)	21 (17)	*Aerobic exercise:* Supervised group exercise. *Control group*: Waitlist control.	3 sessions per week	8 weeks
Doyne 1987 USA	Outpatients Mean age: 29 (SD 4) 100 % female	HAM-D_{17}: 13.0 (SD 7)	52 (25)	*Aerobic exercise OR weightlifting*: Supervised individual exercise. *Control group*: Waiting list.	4 sessions per week	8 weeks
Veale 1992 UK	Outpatients Mean age: 35 (range 19-58) 64% female	BDI: 24.5 (SD 6)	83 (65)	*Aerobic exercise*: Supervised group exercise. *Control group*: Standard treatment from psychiatric services.	3 sessions per week	12 weeks
Singh 1997 USA	Outpatients Recruited from a register of volunteers Mean age: 71 (SD 1)	BDI: 19.9 (SD 2.3)	32 (32)	*Progressive resistance training*: Supervised group exercise. *Control group*: Attended seminars on health.	3 sessions per week Control group: 2 sessions per week	10 weeks
Blumenthal 1999 USA	Outpatients Mean age: 57 (SD 7) 71.8% female	HAM-D_{17}: Not reported	103 (103)	*Aerobic exercise*: Supervised exercise plus antidepressant medication (sertraline). *Control group*: Antidepressant medication (sertraline).	3 sessions per week	16 weeks
Mather 2002 UK	Outpatients Treatment resistant Mean age: 65 (range 53-91) 69% female	HAM-D_{17}: 17.1 (SD 6)	86 (85)	*Mixed aerobic and non-aerobic exercise*: Supervised group exercise. *Control group*: Attended health seminars.	2 sessions per week Control group: 2 seminars per week	10 weeks
Dunn 2005 USA	Outpatients Mean age: 36 (SD 6)	HAM-D_{17}: 19.4 (SD 2)	80 (80)	*Aerobic exercise*: Individually supervised	Group (1) and (2): 3 sessions	12 weeks

	75% female			exercise with (1) low energy expenditure (EE) OR (2) high EE OR (3) low EE OR (4) high EE. *Control group*: Flexibility exercise.	per week Group (3) and (4): 5 sessions per week Control group: 3 sessions per week	
Singh 2005 Australia	Outpatients Mean age: 69 (SD 6) 55% female	HAM-D_{17}: 18.9 (SD 4.2)	60 (54)	*Progressive resistance training (PRT)*: (1)Low intensity PRT OR (2) high intensity PRT. *Control group*: Standard GP care.	Group (1) and (2): 3 sessions per week	8 weeks
Pilu 2007 Italy	Outpatients Treatment resistant Age between 40 and 60 100% female	HAM-D_{17}: 19.7 (SD 6)	30 (30)	*Resistance exercise*: Supervised group sessions. *Control group*: Standard treatment.	2 sessions per week	32 weeks
Viera 2007 Brazil	Outpatients Mean age 43.66 (SD NR) 100% female	HAM-D_{21}: 31.9 (SD 3)	18 (18)	*Aerobic exercise:* Supervised water aerobics. *Control group*: Standard GP care.	2 sessions per week	12 weeks
Blumenthal 2007 USA	Outpatients Mean age: 52 (SD 8) 75.8% female	HAM-D_{17}: 16.7 (SD 4)	153 (153)	*Aerobic exercise*: (1) Supervised group exercise OR (2) home-based exercise. *Control group*: Placebo medication.	(1) and (2): 3 sessions per week	16 weeks
Krogh 2009 Denmark	Outpatients Mean age: 39 (SD 9) 74% female	HAM-D_{17}: 17.8 (SD 4)	165 (165)	*Exercise*: (1) Aerobic supervised group exercise OR (2) supervised group resistance training *Control group*: relaxation and stretching exercise.	(1)and (2): 2 sessions per week Control group: 2 sessions per week	16 weeks
Mota-Pereira 2011 Portugal	Outpatients Treatment resistant Mean age: 47.5 (SD 3) 65.5% female	HAM-D_{17}: 17.1 (SD 3)	33 (29)	*Aerobic exercise:* Homebased exercise + supervised. *Control group*: Attention control.	4 home-based sessions/week. 1 supervised session/week Control group: 1 supervised session/week	12 weeks
Krogh 2012 Denmark	Outpatients Mean age: 42 (SD 11) 67% female	HAM-D_{17}: 18.9 (SD 4)	115 (115)	*Aerobic exercise*: Supervised group exercise. *Control group*: Supervised stretching exercise in groups.	3 sessions per week Control group: 3 sessions per week	12 weeks
Chalder 2012 UK	Outpatients Mean age: 40 (SD 13) 66% female	BDI: 32.1 (SD 9)	361 (361)	*Exercise*: Participants received individually tailored support and encouragement to engage in physical activity. *Control group*: Standard GP care.	Individual	16 weeks

Fang 2013 China	Inpatients Mean age: 44 (SD 14) 66.9% female	HAM-D_{24}: 29.2 (SD 5)	90 (90)	*Aerobic exercise:* Group 1 and 2 had supervised group exercise, high intensity. *Control group:* 15 min stretching	Group 1 and 2 had 3 and 5 sessions per week, respectively Control group: 3 sessions per week	6 weeks
Huipeng 2013 China	Inpatients Mean age: 30 (SD 5) 100% female	HAM-D_{17}: 28 (SD 5)	68 (68)	*Aerobic exercise:* Jogging *Control group:* Standard treatment	5 sessions per week	6 weeks
Cassandra 2014 Honkong	Inpatients Mean age: 46 (SD 12) 67.3% female	MADRS: 19 (10)	52 (52)	*Aerobic exercise:* Supervised exercise. *Control group*: 10 min stretching.	5 sessions per week	3 weeks
Danielsson 2014 Sweden	Outpatients Mean age: 45 (SD 13) 76% female	MADRS: 24.0 (SD 5)	42 (42)	*Mixed aerobic and non-aerobic exercise*: First two weeks individual supervised exercise then supervised group exercise. *Control group*: One session with advice on physical activity.	2 sessions per week	10 weeks
Pfaff 2014 Australia	Outpatients Mean age: 61 (SD 8) 63% female	MADRS: 21.3 (SD NR)	200 (200)	*Resistance exercise*: Supervised home-based exercise *Control group*: Standard GP care	3 sessions per week	12 weeks
Guifeng 2015 China	Inpatients Mean age: 33 (SD 14) 70% female	HAM-D_{24}: 25.9 (SD 4)	70 (70)	*Aerobic exercise:* Supervised group exercise *Control group:* Standard treatment	5 sessions per week	8 weeks
Junchin 2015 China	Inpatients Mean age: 28 (SD 7) 61% female	HAM-D_{24}: 25.8 (SD 3)	70 (70)	*Aerobic exercise:* Supervised aerobic exercise of the patients own choice Control group: Standard treatment	5 sessions per week	8 weeks
Schuch 2015 Brazil	Inpatients Mean age: 40 (SD 11) 74% female	HAM-D_{17}: 26.7 (SD 2)	50 (50)	*Aerobic exercise:* Supervised individual exercise. *Control group*: Standard treatment.	3 sessions per week	2 weeks
Kerling 2015 Germany	Inpatients Mean age: 43 (SD 10)	MADRS: 24.0 (SD 9)	42 (42)	*Aerobic exercise:* Supervised exercise. *Control group*: Standard treatment.	3 sessions per week	6 weeks
Belvederi 2015 Italy	Outpatients Mean age: 75 (SD 6) 71% female	HAM-D_{17}: 20.1 (SD 3)	121 (121)	*Aerobic exercise:* (1) Sertraline + supervised non-progressive exercise OR (2) sertraline + supervised progressive aerobic	3 sessions per week	24 weeks

				exercise. *Control group*: Sertraline.		
Carneiro 2015 Portugal	Outpatients Mean age: 50.16 (SD 12) 100% female	BDI: 48.8 (SD 10)	26 (19)	*Aerobic exercise:* Supervised exercise *Control group:* Standard treatment	3 sessions per week	16 weeks
Doose 2015 Germany	Outpatients Mean age: 47.9 (SD 11) 63% female	HAM-D_{17}: 14.2 (SD 3)	46 (46)	*Aerobic exercise:* Supervised aerobic exercise *Control group*: Standard treatment	3 sessions per week	8 weeks
Pentecost 2015 UK	Outpatients Mean age: 44.4 (SD 14) 48% female	PHQ-9: 16.5 (SD 4)	60 (44)	*Exercise:* Behavioral activation plus physical activity promotion *Control group:* Behavioral activation	Individual	12 weeks
Salehi 2016 Iran	Inpatients Mean age: 30.0 (SD 6) 35% female	HAM-D_{21}: 43.4 (SD 8)	40 (40)	*Aerobic exercise + ECT:* Supervised aerobic exercise *Control group:* ECT	3 sessions per weeks Control group 3 ECTs per week	4 weeks
Legrand 2016 France	Inpatients Mean age: 46.9 (SD 13) 67% female	BDI: 36.0 (SD 6)	24 (24)	*Aerobic exercise:* Supervised aerobic exercise *Control group:* Standard treatment	10 sessions in 10 consecutive days	10 days
Euteneuer 2017 Germany	Outpatients Mean age: 37.1 (SD 12) 52% female	BDI: 27.2 (SD 9)	71 (68)	*Exercise:* CBT + PA promotion *Control group:* CBT + low energy activities	Individual	16 weeks
Olson 2017 Ireland	Outpatients Mean age: 21.1 (SD 2) 80% female	BDI: 24.2 (SD 12)	50 (30)	*Aerobic exercise:* Supervised aerobic exercise *Control group:* Stretching exercise	3 sessions per week 3 sessions per week	8 weeks
Patten 2017 USA	Outpatients Mean age: 37.5 (SD 11) 100% female	PHQ-9: 11.7 (SD 5)	30 (26)	*Aerobic exercise:* Supervised aerobic exercise *Control group:* Health education	3 sessions per week	12 weeks

SCL-D: Symptom Check List, depression subscale; HAM-D_{17}: Hamilton Depression Scale, 17 items; BDI: Beck's Depression Inventory; SD: Standard deviation; ECT: Electroconvulsive therapy; PHQ-9: Patient Health Questionnaire; CBT: Cognitive Behavioral Therapy

Table 2. Risk of bias in trials assessing exercise for patients diagnosed with depression

Author, Year of publication	Sequence generation	Allocation concealment	Blinding of participants and trial personnel assessors	Blinding of outcome assessors	Incomplete outcome data	Selective outcome reporting	For profit bias	Other bias	Comment on 'Other bias'
Klein 1985	Unclear	Unclear	High	High	High	Low	Low	Low	
Martinsen 1985	Unclear	Unclear	High	High	High	Low	High	Low	
Epstein 1986	Unclear	Unclear	High	High	High	Low	Unclear	High	Baseline difference
Doyne 1987	Unclear	Unclear	High	Low	High	Low	Unclear	High	Baseline difference
Veale 1992	Unclear	Unclear	High	High	High	Low	Low	High	Baseline difference
Singh 1997	Low	Unclear	High	Low	Low	Low	Low	High	Baseline difference
Blumenthal 1999	Unclear	Unclear	High	Low	High	Low	High	Low	
Mather 2002	Low	Low	High	Low	High	Low	Low	Low	
Dunn 2005	Low	Low	High	Low	High	High	High	Low	
Singh 2005	Low	Low	High	Low	High	Low	Unclear	Low	
Pilu 2007	Unclear	Unclear	High	Unclear	Low	Low	Unclear	Low	
Viera 2007	Unclear	Unclear	High	Unclear	Low	Low	Unclear	Low	
Blumenthal 2007	Low	Low	High	Low	High	High	Low	Low	
Krogh 2009	Low	Low	High	Low	Low[1]	High	High	High	Baseline difference
Mota-Pereira 2011	Unclear	Unclear	High	Low	High	Low	High	High	Baseline difference
Krogh 2012	Low	Low	High	Low	Low	Low	Low	Low	
Chalder 2012	Low	Low	High	High	Low	Low	Low	Low	
Fang 2013	Unclear	Unclear	High	Unclear	Unclear	High	Unclear	Low	
Huipeng 2013	Unclear	Unclear	High	Unclear	Low	Low	Unclear	Low	
Cassandra 2014	Low	Unclear	High	Low	High	Low	Low	Low	
Danielsson 2014	Unclear	Low	High	Low	High	Low	Low	Low	
Pfaff 2014	Low	Low	High	Low	Low[1]	Low	Low	High	Baseline difference
Guifeng 2015	Unclear	Unclear	High	Unclear	Low	Low	Unclear	Low	
Jinchun 2015	Unclear	Unclear	High	Unclear	Low	Low	Unclear	Low	
Schuch 2015	Unclear	Low	High	Low	Low	Low	Low	Low	
Kerling 2015	Unclear	Unclear	High	Unclear	Low	Low	Low	Low	
Belvederi 2015	Low	Low	High	Low	High	Low	Low	High	Post-hoc sample size
Carneiro 2015	Unclear	Low	High	High	Unclear	Low	Low	Low	
Doose 2015	Unclear	Unclear	High	High	High	Low	Low	High	No sample size calc.
Pentecost 2015	Low	Low	High	High	High	Low	Low	Low	
Salehi 2016	High	High	High	Low	Unclear	Low	Low	High	Baseline difference
Legrand 2016	Low	High	High	High	High	Low	Unclear	Low	
Euteneuer 2017	Low	Unclear	High	High	High	Low	Low	Low	
Olson 2017	Low	Unclear	High	High	High	Low	Low	Low	
Patten 2017	Unclear	Unclear	High	High	High	Low	Low	Low	

Table 3. Heterogeneity of effect estimates for trials assessing the effect of exercise for patients diagnosed with depression explored by comparing sub-groups.

Subgroups	Number of Trials (participants)	Random effects meta-analysis SMD (95% CI., p, I^2)	Subgroup explains heterogeneity P value
Risk of bias			
Less than high risk of bias[1]	4 (530)	-0.11 (-0.41 to 0.18; p = 0.45; I^2 = 62%)	<0.001
High risk of bias	31 (1968)	-0.75 (-0.98 to -0.52; p < 0.001; I^2 = 81%)	
Age			
Old (>59 years)	5 (492)	-0.77 (-1.34 to -0.19; p = 0.009; I^2 = 87%)	0.78
Young (<59 years)	30 (2006)	-0.68 (-0.90 to -0.45; p < 0.001; I^2 = 83%)	
Exercise context			
Group exercise	26 (1785)	-0.75 (-1.01 to -0.50; p < 0.001; I^2 = 83%)	0.30
Individual exercise	9 (713)	-0.52 (-0.88 to -0.16; p = 0.005; I^2 = 73%)	
Duration			
Less than 10 weeks	15 (721)	-0.92 (-1.09 to -0.74; p < 0.001; I^2 = 14%)	0.007
10 weeks or more	20 (1777)	-0.49 (-0.75 to -0.23; p < 0.001; I^2 = 83%)	
Attention control			
Attention control	10 (733)	-0.56 (-0.98 to -0.15; p = 0.008; I^2 = 85%)	0.91
Waitlist	2 (47)	-0.67 (-2.48 to 1.13; p = 0.47; I^2 = 88%)	
Pharmacotherapy			
Add-on	11 (734)	-0.92 (-1.38 to -0.46; p < 0.001; I^2 = 86%)	0.82
No medication	6 (318)	-0.82 (-1.58 to -0.06; p = 0.03; I^2 = 88%)	
Somatic comorbidity			
Somatic co-morbidity	0	N/A	
No co-morbidity	35 (2331)	N/A	
Minor depression			
Incl. minor depression	6 (350)	-0.90 (-1.65 to -0.15; p = 0.02; I^2 = 86%)	0.53
No minor depression	25 (2148)	-0.65 (-0.87 to -0.43; p < 0.001; I^2 = 81%)	
Patient setting			
Inpatients	10 (549)	-0.88 (-1.07 to -0.70; p < 0.001; I^2 = 6%)	0.07
Outpatients	21 (1782)	-0.60 (-0.85 to -0.35; p < 0.001; I^2 = 83%)	
Trial size			
Trials n ≤ 50	18 (578)	-1.11 (-1.52 to -0.72; p < 0.001; I^2 = 78%)	0.001
Trials n > 50	17 (1920)	-0.37 (-0.57 to -0.18; p < 0.001; I^2 = 75%)	
Increase in exercise capacity			
VO2max > 2.8 ml/kg/min	5 (340)	-0.48 (-1.08 to 0.13; p = 0.12; I^2 = 86%)	0.65
VO2max ≤ 2.8 ml/kg/min	6 (661)	-0.32 (-0.61 to 0.02; p = 0.03; I^2 = 68%)	

Table 4. Summary of findings

Exercise compared to control or treatment as usual for depression

Patient or population: depression

Setting: In- or out-patients

Intervention: exercise

Comparison: control or treatment as usual

Outcomes	Anticipated absolute effects* (95% CI)		Relative effect (95% CI)	№ of participants (studies)	Quality of the evidence (GRADE)	Comments
	Risk with control or treatment as usual	Risk with exercise				
Severity of depression	-	**0.66 SMD lower** (0.46 lower to 0.86 lower)	-	2498 (35 RCTs)	⊕◯◯◯ VERY LOW [1]	Lower depression scores indicate improvement. SMD of 0.3 is considered clinically relevant.
Lack of remission	**Study population**		**RR 0.78** (0.68 to 0.90)	1639 (19 RCTs)	⊕◯◯◯ VERY LOW [2]	Remission is, with minor variations, defined as not full-filling the criteria for depression.
	646 per 1000	**504 per 1000** (426 to 594)				
Serious adverse events	**Study population**		**RR 2.21** (0.24 to 20.21)	335 (3 RCTs)	⊕⊕◯◯ LOW [3]	
	0 per 1000	**0 per 1000** (0 to 0)				
Quality of life	-	**0.40 SMD higher** (0.03 lower to 0.83 higher)	-	827 (9 RCTs)	⊕◯◯◯ VERY LOW [4]	Quality of life was assessed using a number of different methods. Higher score indicates improved quality of life. Seven of 24 trials reported on this outcome
Depression severity after the intervention	-	**0.06 SMD lower** (0.25 lower to 0.14 higher)	-	713 (7 RCTs)	⊕⊕◯◯ LOW [5]	Lower depression scores indicate improvement. SMD of 0.3 is considered clinically relevant.
Lack of remission after the intervention	**Study population**		**RR 0.95** (0.82 to 1.11)	777 (5 RCTs)	⊕⊕◯◯ LOW [6]	
	469 per 1000	**446 per 1000** (385 to 521)				
Depression severity. Restricted to trials with less than high risk of bias.	-	**0.11 SMD lower** (0.41 lower to 0.18 higher)	-	530 (4 RCTs)	⊕⊕◯◯ LOW [7]	Lower depression scores indicate improvement. SMD of 0.3 is considered clinically relevant.

***The risk in the intervention group** (and its 95% confidence interval) is based on the assumed risk in the comparison group and the **relative effect** of the intervention (and its 95% CI). **CI:** Confidence interval; **SMD:** Standardised mean difference; **RR:** Risk ratio

GRADE Working Group grades on evidence
High quality: We are very confident that the true effect lies close to that of the estimate of the effect
Moderate quality: We are moderately confident in the effect estimate. The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different
Low quality: 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 the effect

1. Downgraded by 3: risk of bias, inconsistency and publication bias
2. Downgraded by 3: risk of bias, inconsistency and publication bias
3. Downgraded by 2: imprecision and publication bias

4. Downgraded by 3: risk of bias, inconsistency and imprecision
5. Downgraded by 2: risk of bias and imprecision
6. Downgraded by 2: risk of bias and imprecision
7. Downgraded by 2: inconsistency and imprecision

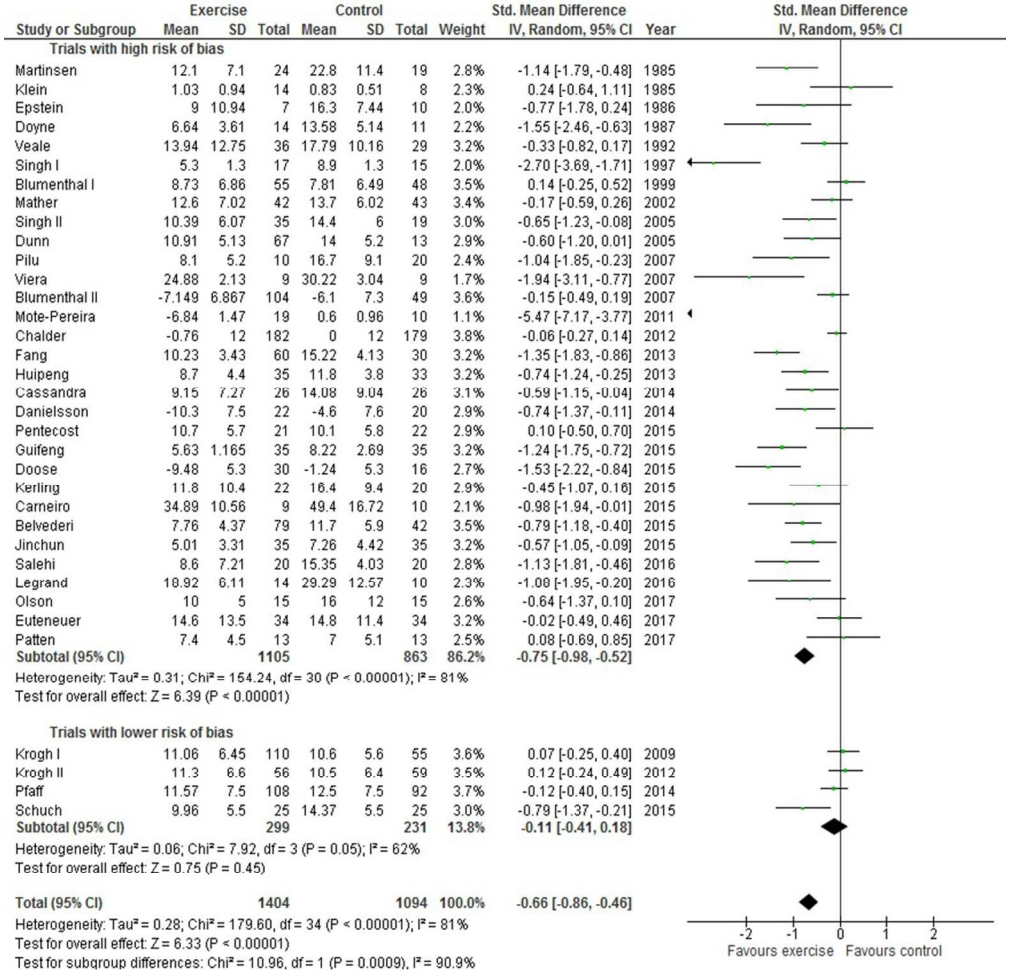


Figure 1

94x91mm (300 x 300 DPI)

Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary Table S1

Table S1. Imputation of missing data for trials assessing exercise for patients diagnosed with depression

Outcome	Result from review	Best/worse (1SD)	Best/worse (2SD)	Worse/best (1SD)	Worse/best (2SD)
Depression SMD (95% CI)	-0.66 (-0.86 to -0.45) p < 0.001; I^2 = 81%	-0.77 (-1.00 to -0.54) p < 0.001; I^2 = 86%	-0.78 (-1.02 to -0.55) p < 0.001; I^2 = 86%	-0.60 (-0.81 to -0.39) p < 0.001; I^2 = 84%	-0.57 (-0.78 to -0.36) p < 0.001; I^2 = 84%)
		Good Outcome	**Poor outcome**	**Good/poor outcome**	**Poor/good outcome**
Lack of remission (95% CL)	RR 0.78 (0.68 to 0.90) p < 0.001; I^2 = 69%	RR 0.75 (0.64 to 0.89) p = 0.0008; I^2 = 73%	RR 0.88 (0.83 to 0.94) p = 0.0002; I^2 = 69%	RR 0.71 (0.61 to 0.81) p < 0.001; I^2 = 68%	RR 0.86 (0.71 to 1.04) p = 0.12; I^2 = 83%
Serious adverse events (95% CL)	RR 2.21 (0.24 to 20.21) p = 0.48; I^2 = 0%	RR 2.19 (0.23 to 20.76) p = 0.50, I^2 = 50%	RR 0.92 (0.37 to 2.30) p = 0.86, I^2 = 60%	RR 0.08 (0.02 to 0.34) p = 0.001, I^2 = 5.4%	RR 19.17 (2.64 to 139.2) p = 0.004, I^2 = 0%

Imputation of missing data for continuous outcome: 'best-worst' - assumed that all participants lost to follow-up in the intervention group had a beneficial outcome (the group mean minus 1 or 2 SD), and all participants lost to follow-up in the placebo group have had a harmful outcome (the group mean plus 1 SD and 2 SD). The reverse 'worst-best-case' scenario is the reverse of the 'best-worst' scenario.

Missing data for the 'remission' outcome was imputed according to the following scenarios: 1) poor outcome analysis: none of the drop-outs/participants lost from both arms experienced the outcome; 2) good outcome analysis: all of the drop-outs/participants lost from both arms experienced the outcome; 3) extreme case analysis favouring the experimental intervention, all of the drop-outs/participants lost from the experimental arm, but none of the drop-outs/participants lost from the control arm experienced the outcome; and 4) extreme case analysis favouring the control: all drop-outs/participants lost from the experimental arm, but none from the control arm experienced the outcome. Missing data for 'serious adverse events' was calculated with the reverse assumptions.

Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary Table S2

Table S2. Trials characteristics for exploration of heterogeneity in trials assessing the effect of exercise in patients diagnosed with depression

Trial	Lower risk of bias	Age > 60	Group vs. individual	Duration	Attention control waitlist	Exercise as add on to drugs vs. exercise alone	Within-study dose exercise	Increase in VO2max [1]	Somatic disease vs. only MD	Trial Includes minor depression
Klein 1985	No	Young	Individual	12 weeks	Other	Exercise alone	No	No	No	Yes
Martinsen 1985	No	Young	Group	9 weeks	Attention control	Unclear	No	11[a]	No	No
Epstein 1986	No	Young	Group	8 weeks	Waitlist	Unclear	No	No	No	Yes
Doyne 1987	No	Young	Individual	8 weeks	Waitlist	Exercise alone	No	No	No	Yes
Veale 1992	No	Young	Group	12 weeks	Other	Unclear	No	No	No	No
Singh 1997	No	Old	Group	10 weeks	Attention control	Exercise alone	No	N/A	No	Yes
Blumenthal 1999	No	Young	Group	16 weeks	Other	Add on	No	2.3	No	No
Mather 2002	No	Old	Group	10 weeks	Attention control	Add on	No	No	No	No
Dunn 2005	No	Young	Individual	12 weeks	Attention control	Exercise alone	Yes	No	No	No
Singh 2005	No	Old	Group	8 weeks	Other	Exercise alone	Yes	N/A	No	Yes
Pilu 2007	No	Young	Group	24 weeks	Other	Add on	No	No	No	No
Viera 2007	No	Young	Group	12 weeks	Other	Add on	No	No	No	No
Blumenthal 2007	No	Young	Group	16 weeks	Other	Add on	No	2.0[a]	No	No
Krogh 2009	Yes	Young	Group	16 weeks	Attention control	No	No	2.9	No	No
Mota-Pereira 2011	No	Young	Group	12 weeks	Other	Add on	No	No	No	No
Krogh 2012	Yes	Young	Group	12 weeks	Attention control	Exercise alone	No	3.4	No	No
Chalder 2012	No	Young	Individual	32 weeks	Other	No	No	No	No	No
Fang 2013	No	Young	Group	6 weeks	Attention control	No	Yes	No	No	No
Huipeng 2013	No	Young	Group	6 weeks	Other	No	No	No	No	No
Cassandra 2014	No	Young	Group	3 weeks	Other	Add on	No	No	No	No
Danielsson 2014	No	Young	Group	10 weeks	Other	Add on	No	2.4	No	No
Pfaff 2014	Yes	Old	Group	12 weeks	Other	No	No	1.5	No	Yes
Guifeng 2015	No	Young	Group	8 weeks	Other	No	No	No	No	No
Jinchun 2015	No	Young	Group	8 weeks	Other	No	No	No	No	No
Schuch 2015	Yes	Young	Individual	2 weeks	Other	No	No	No	No	No
Kerling 2015	No	Young	Group	6 weeks	Other	No	No	2.8	No	No

Belvederi 2015	No	Old	Group	24 weeks	Other	Add on	Yes	0.3[a]	No	No
Carneiro 2015	No	Young	Group	16 weeks	Other	Add on	No	No	No	No
Doose 2015	No	Young	Group	8 weeks	Other	No	No	3.2	No	No
Pentecost 2015	No	Young	Individual	12 weeks	Other	No	No	No	No	No
Legrand 2016	No	Young	Individual	10 days	Other	No	No	No	No	No
Salehi 2016	No	Young	Individual	4 weeks	Other	Add on	No	No	No	No
Euteneuer 2017	No	Young	Individual	16 weeks	Attention control	No	No	No	No	No
Olsen 2017	No	Young	Group	8 weeks	Attention control	No	No	No	No	No
Patten 2017	No	Young	Group	12 weeks	Other	No	No	5.0	No	No

[1]Increase in VO2max is based on increase in intervention group, if [a] then value is based on an estimate from text or figures.

Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary Table

Table S3. Heterogeneity of effect estimates for trials assessing the effect of exercise for patients diagnosed with depression on lack of remission.

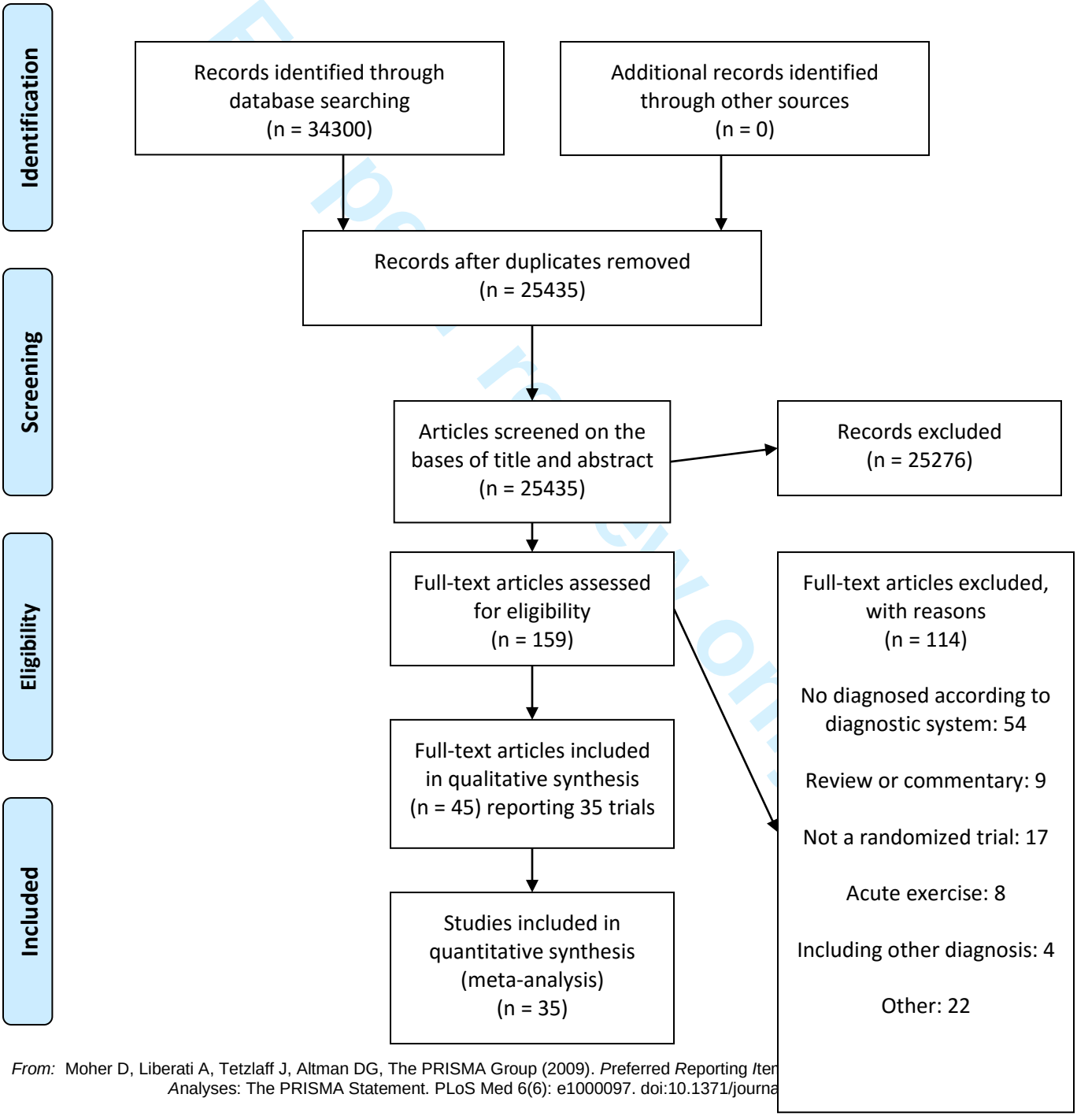
Subgroups	Number of Trials (participants)	Random effects meta-analysis RR (95% CI., p, I^2)	Subgroup explains heterogeneity P value
Risk of bias			
Less than high risk of bias[1,2]	2 (165)	0.95 (0.74 to 1.23; p = 0.70; I^2 = 20%)	0.18
High risk of bias	17 (1474)	0.77 (0.64 to 0.92; p = 0.003; I^2 = 75%)	
Age			
Old (>59 years)	3 (299)	0.61 (0.21 to 1.02; p = 0.37; I^2 = 91%)	0.62
Young (<59 years)	16 (1340)	0.81 (0.70 to 0.93; p = 0.003; I^2 = 64%)	
Exercise context			
Group exercise	14 (1156)	0.80 (0.66 to 0.96; p = 0.02; I^2 = 72%)	0.69
Individual exercise	5 (483)	0.74 (0.52 to 1.04; p = 0.08; I^2 = 77%)	
Duration			
Less than 10 weeks	8 (393)	0.63 (0.51 to 0.77; p < 0.001; I^2 = 40%)	0.004
10 weeks or more	11 (1246)	0.93 (0.78 to 1.10; p = 0.39; I^2 = 69%)	
Attention control			
Attention control	4 (364)	0.91 (0.73 to 1.12; p = 0.38; I^2 = 42%)	0.07
Waitlist	1 (25)	0.44 (0.21 to 0.93; p = 0.03; I^2 = 0%)	
Pharmacotherapy			
Add-on	7 (540)	0.72 (0.54 to 0.96; p = 0.03; I^2 = 69%)	0.62
No medication	4 (252)	0.75 (0.52 to 1.09; p = 0.13; I^2 = 66%)	
Somatic comorbidity			
Somatic co-morbidity	0	N/A	
No co-morbidity	19 (1639)	N/A	
Minor depression			
Incl. minor depression	3 (203)	0.63 (0.21 to 1.89; p = 0.41; I^2 = 87%)	0.69
No minor depression	16 (1436)	0.79 (0.68 to 0.92; p = 0.002; I^2 = 69%)	
Patient setting			
Inpatients	6 (322)	0.71 (0.60 to 0.84; p < 0.001; I^2 = 0%)	0.21
Outpatients	13 (1317)	0.84 (0.69 to 1.01; p = 0.07; I^2 = 77%)	
Trial size			
Trials n ≤ 52	9 (358)	0.62 (0.50 to 0.76; p < 0.001; I^2 = 45%)	0.002
Trials n > 52	10 (1281)	0.95 (0.80 to 1.12; p = 0.52; I^2 = 68%)	

[1]Trials potentially having less bias than trials with high risk of bias.

Article:
Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplemental

Figure S1. Flow diagram for identification of trials assessing the effects of exercise for patients with depression.



From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). *P*referred *R*eporting *I*ter
*A*nalyses: The PRISMA Statement. PLoS Med 6(6): e1000097. doi:10.1371/journa

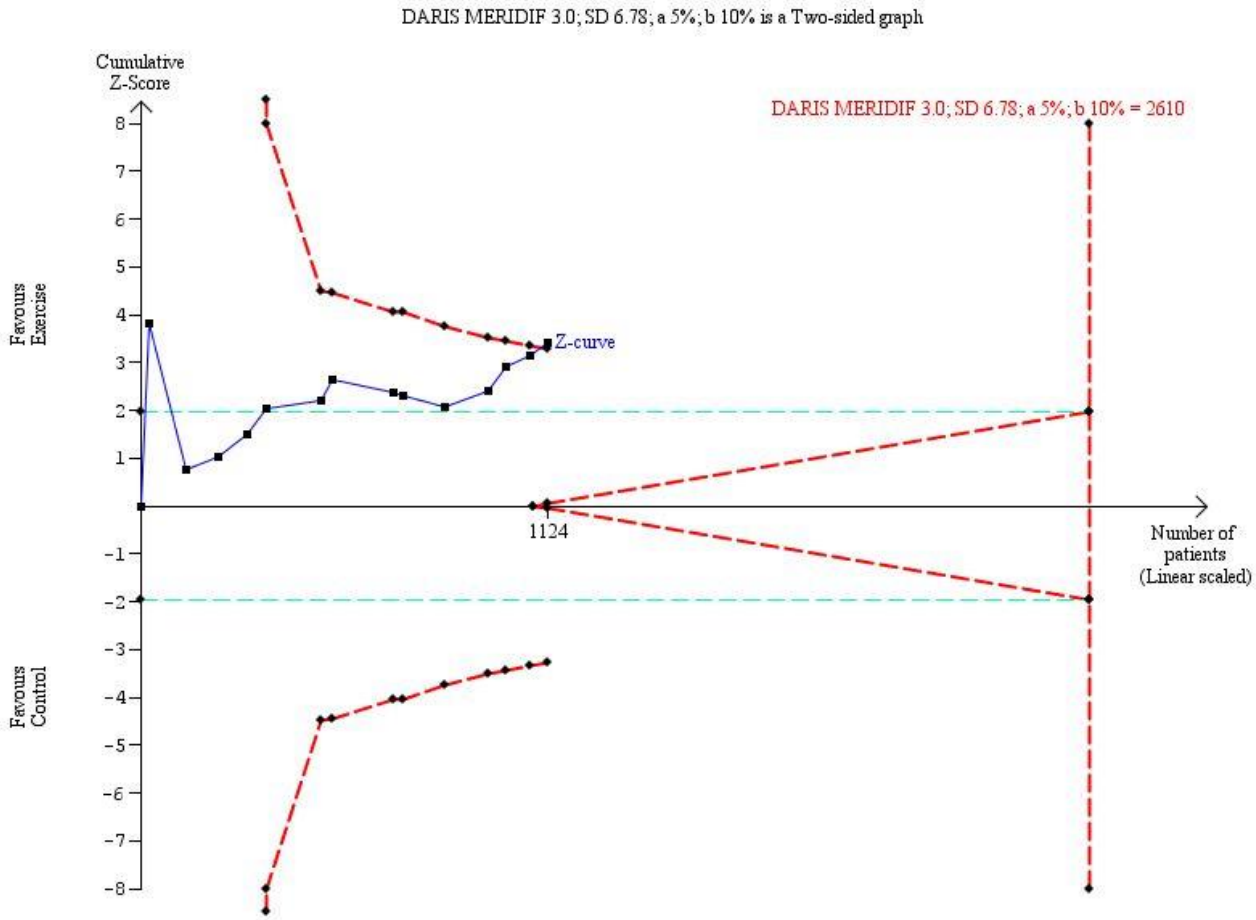
For more information, visit www.prisma-statement.org.

Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary Figure

Figure S2. Trial Sequential Analysis and required information size for the effect of exercise for depressive symptoms including four-teen trials reporting on HAM-D_{17}.

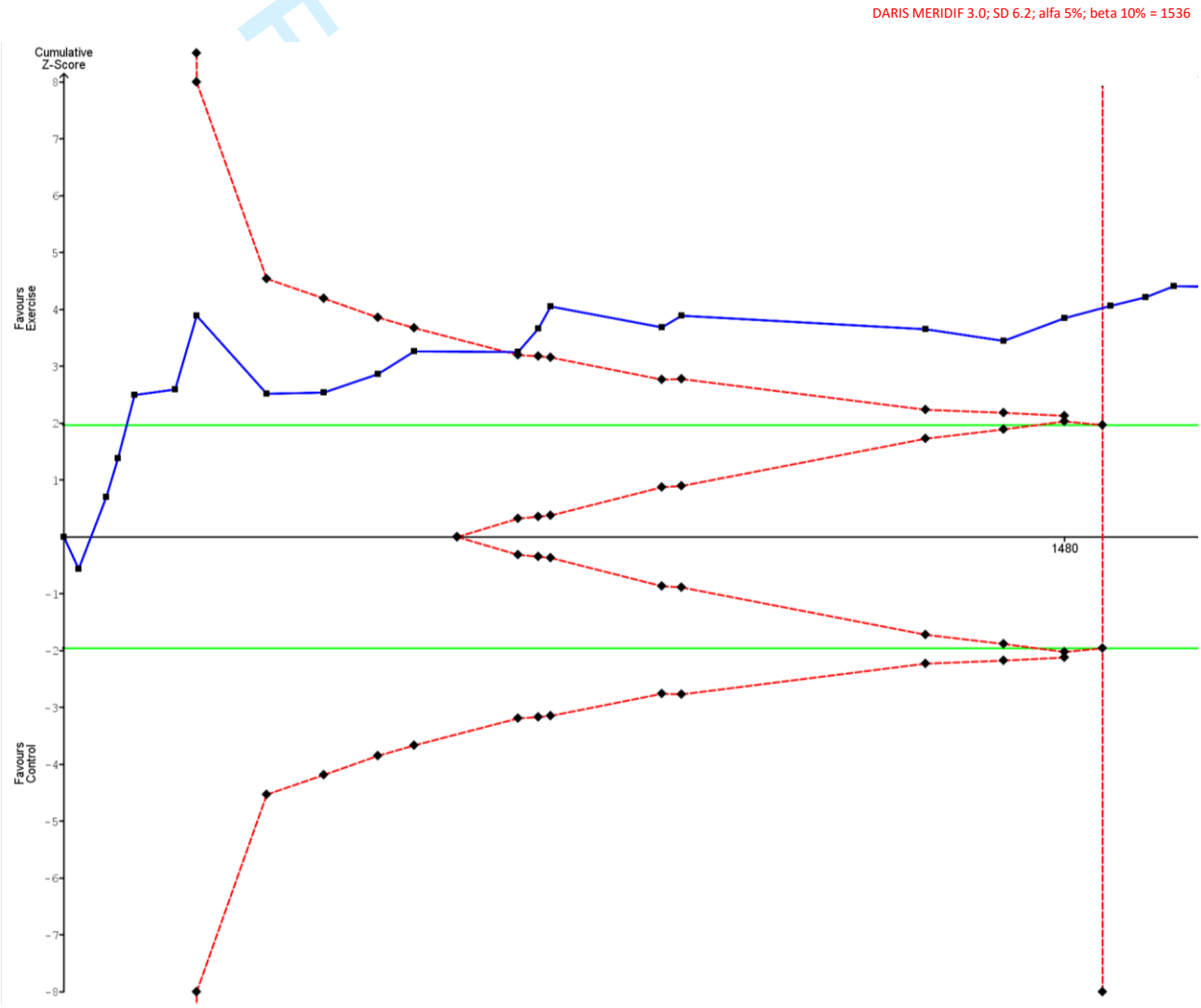


Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary Figure

Figure S3. Trial Sequential Analysis and required information size for the effect of exercise for depressive symptoms including 35 trials 'converted' to a HAM-D_{17} scale.



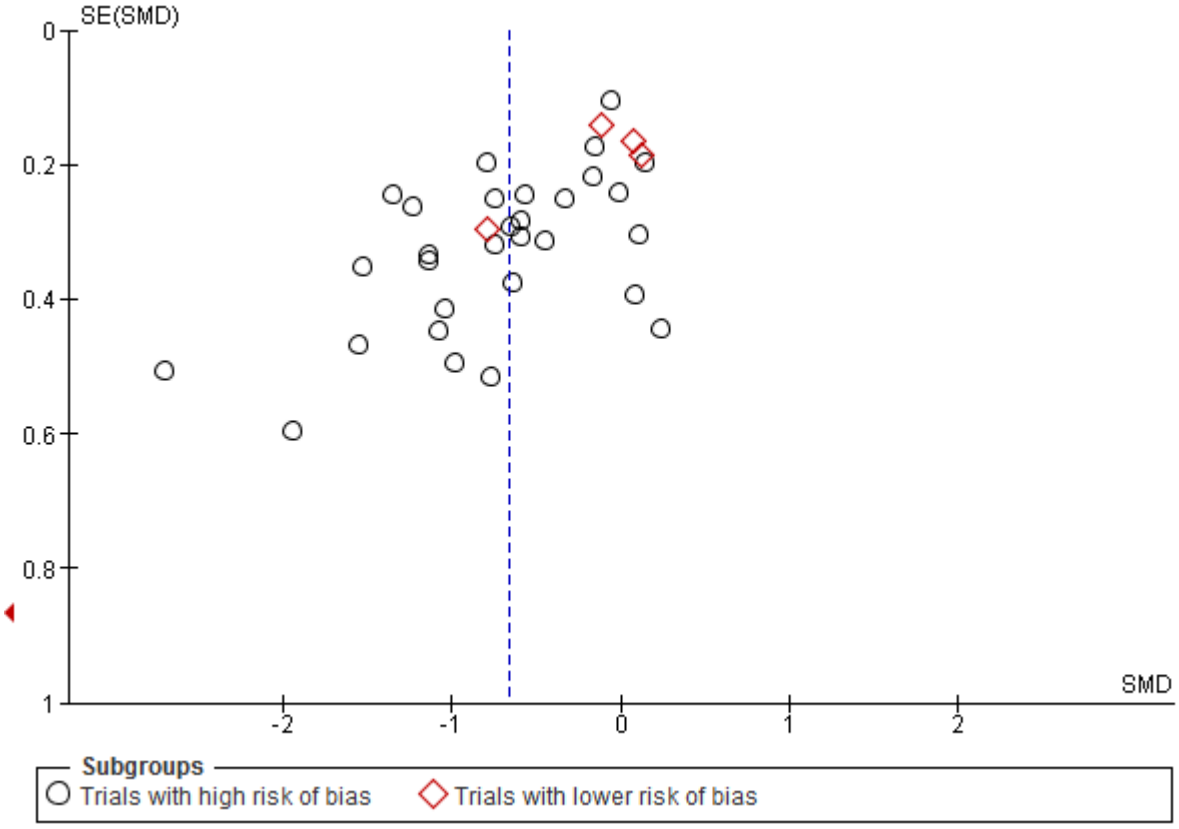
Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary Figure

Figure S4.

Funnel plot of 35 trials assessing the antidepressant effect of exercise as a continuous outcome

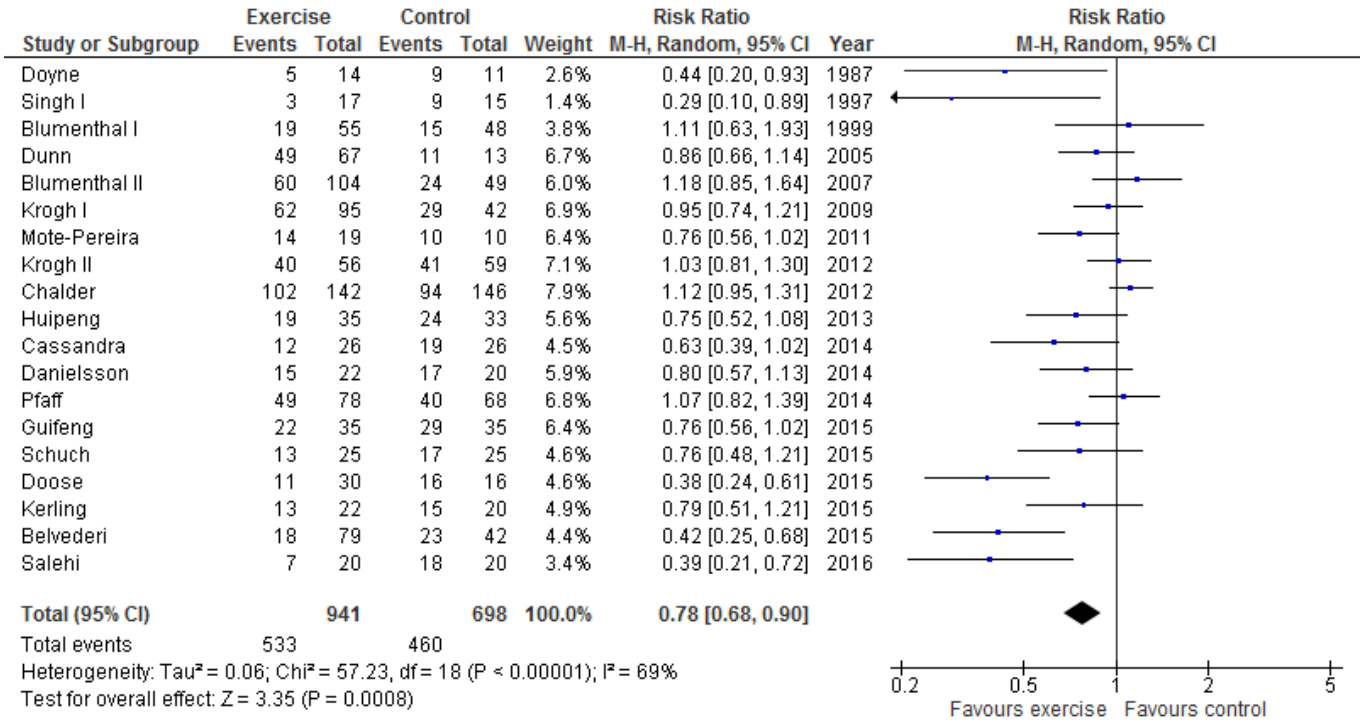


Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary Figure

Figure S5. Effect of exercise on lack of remission for patients diagnosed with depression

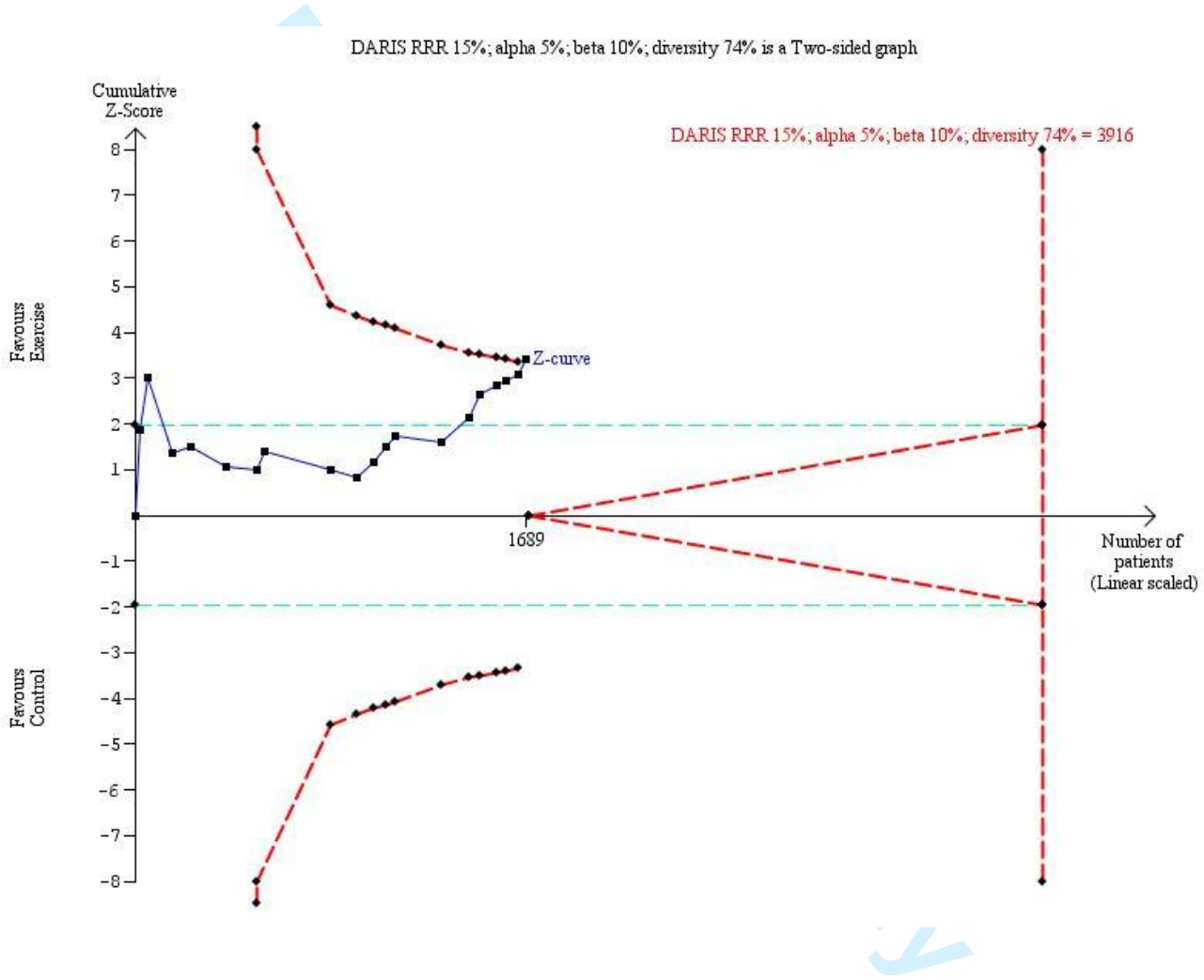


Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary Figure

Figure S6. Trial Sequential Analysis and required information size for the effect of exercise on lack of remission.



Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary Figure S7

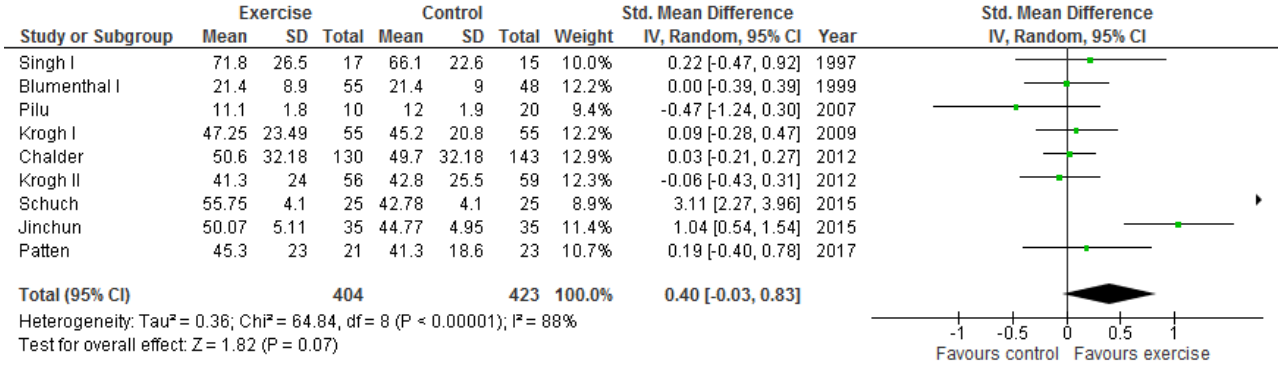
Figure S7. Effect of exercise on risk of serious adverse events for patients diagnosed with depression

Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary Figure S8-S10

Figure S8. The effect of exercise on quality of life in patients diagnosed with depression



Quality of life was assessed using different scales: Singh I, Chalder and Patten used the SF-36, Blumenthal used Life Satisfaction Index, Pilu and Schuch used the WHOQOL, Krogh I and Krogh II used the WHO-Five Well-being Scale, and Jinchun used the GQOLI-74.

Figure S9. The effect of exercise on depression severity after the intervention in patients diagnosed with depression

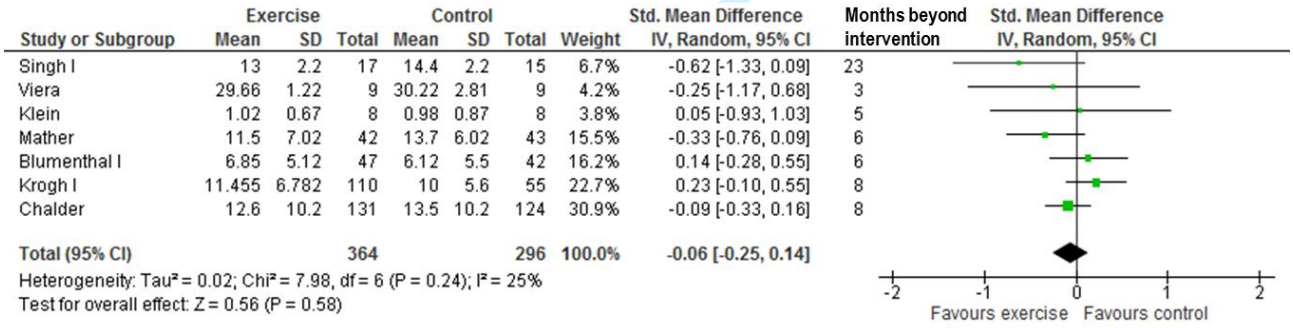
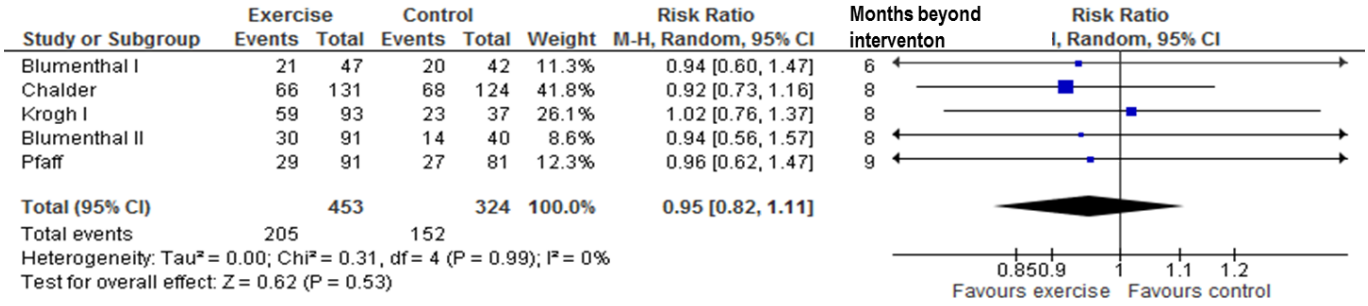


Figure S10. The effect of exercise on risk of lack of remission after the intervention in patients diagnosed with depression



Article:

Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential-analysis

Supplementary material (S1)

An example of bibliographical search for PubMEd

#1 Depression [MeSH]

#2 Depresive disorder [MeSH]

#3 Exercise [Text Word]

#4 Aerobic [Text Word]

#5 Non-aerobic [Text Word]

#6 Physical activity [Text Word]

#7 Physical fitness [Text Word]

#8 Walking [MeSH]

#9 Jogging [MeSH]

#10 Running [MeSH]

#11 Bicycling [MeSH]

#12 Swimming [MeSH]

#13 Strength [Text Word]

#14 Resistance [Text Word]

#15 #1 OR #2

#16 #3 OR #4 OR #5 OR #6 OR #7 OR #8 OR #9 OR #10 OR #11 OR #12 OR #13 OR #14

#17 #15 AND #16

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	5
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	6
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	1
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	7
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	6
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	6
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	6
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	8
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	7
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	8
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	9
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	9

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	8
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	11
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	Fig. 1
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Table 1
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	Table 2
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	Fig 2
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	Fig 3-fig8
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	Table 2
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	Table 3
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	21
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	21
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	22
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	26

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: www.prisma-statement.org.