Mental Health and Physical Activity 8 (2015) 21–36

Contents lists available at ScienceDirect

Mental Health and Physical Activity

journal homepage: www.elsevier.com/locate/menpa

Honey, I shrunk the pooled SMD! Guide to critical appraisal of systematic reviews and meta-analyses using the Cochrane review on exercise for depression as example

Panteleimon Ekkekakis*

Department of Kinesiology, Iowa State University, USA

ARTICLE INFO

Article history:
Received 12 December 2014
Accepted 12 December 2014
Available online 8 January 2015

Keywords:
Critical appraisal
Inclusion criteria
Exclusion criteria
Methodological quality

ABSTRACT

Problem: In several countries, physical activity is now recommended in clinical practice guidelines as an option for the treatment of subthreshold, mild, and moderate adult depression. However, most physicians do not present this option to their patients, attributing their decision to the perception that the supporting research evidence is inadequate. To assist readers in developing a strategy for evaluating pertinent research evidence, the present analysis offers a critical appraisal of the Cochrane systematic review and meta-analysis examining the effects of exercise on depression. Remarkably, successive updates of this review have reported a gradual "shrinkage" of the pooled standardized mean difference associated with exercise by 44%, from −1.10 in 2001 to −0.62 in 2013.
Method: The analysis evaluated the inclusion and exclusion criteria, the uniformity of rules, the rationale behind protocol changes, the procedures followed in assessing methodological quality, and reporting errors.
Results: Inspection of the details of the review demystifies the "shrinkage" phenomenon, revealing that it is attributable to specific, questionable methodological choices and the fluidity of the review protocol. Reanalysis of the same database following rational modifications shows that the effect of exercise is large. Restricting the analysis to high-quality trials yields an effect size significantly different from zero.
Conclusions: Although the clinical value of the Cochrane review is questionable, its educational potential is undeniable. Clinicians, students, referees, editors, systematic reviewers, guideline developers, and policymakers can use the present analysis as a template for evaluating the influence of methodological choices on the conclusions of systematic reviews and meta-analyses.

> Who in this Brave New World is to peer review the reviewers? (Shapiro, 1995, p. 658)

In the burgeoning field of research investigating the effects of physical activity on mental health, studies focusing on depression are of exceptional importance, for two main reasons. First, the World Health Organization recognizes depressive disorders as the leading cause of disability and one of the costliest disorders worldwide. Therefore, physical activity, an intervention promising not only meaningful efficacy but also global accessibility, virtual absence of adverse side effects, and low cost, represents a very appealing option for health care systems and organizations. Second, in several countries that have adopted "stepped care" or "stepped collaborative care" models for treating depression, physical activity is recommended in clinical practice guidelines as one of the options that should be offered to patients with subthreshold depressive symptoms or mild to moderate levels of depression (i.e., the vast majority of patients with depressive symptoms in primary care). Depression is the first – and still the only – mental health disorder for which physical activity is recommended as an evidence-based treatment.

One evidence synthesis on physical activity and depression that is cited extensively, particularly in the medical literature, is an ongoing series of Cochrane systematic reviews, conceived as a periodic update of an earlier meta-analysis by Lawlor and Hopker (2001). The latest installment was published by Cooney et al.

* 237 Forker Building, Department of Kinesiology, Iowa State University, Ames, IA 50011, USA. Tel.: +1 515 294 8766; fax: +1 515 294 8740.
E-mail address: ekkekaki@iastate.edu.

http://dx.doi.org/10.1016/j.mhpa.2014.12.001
1755-2966/

(2013). Reflecting the tone of its conclusions, a news article in the *British Medical Journal* summarized the key findings under the title "Evidence that exercise helps in depression is still weak" (Kmietowicz, 2013). Cooney, Dwan, and Mead (2014) subsequently published a summary of their results as a "Clinical Evidence Synopsis" in the *Journal of the American Medical Association* (*JAMA*), one of the most prestigious and widely read medical journals in the world. In this high-profile summary, Cooney et al. (2014) concluded that the antidepressant effect of exercise "may be small" (p. 2432) since "analysis of high-quality studies alone suggests only small benefits" or even "no association of exercise with improved depression" (p. 2433). These conclusions, which seem to question the validity of clinical practice guidelines, gain added significance in light of the fact that many physicians report a reluctance to recommend physical activity to patients with depression, mainly due to the perception that the supporting research evidence is insufficient (Searle et al., 2012).

The Cochrane review (Cooney et al., 2013) is a large document of 160 pages and nearly 60,000 words, making it unlikely that most readers would be inclined to read it in its entirety. Judging from the articles that have cited this and previous editions of the review thus far, it appears that citing authors tend to echo the conclusions of the review without having subjected the document to a thorough, independent critical appraisal. However, especially given its impact on the literature, and its potential impact on clinical practice worldwide, it seems useful to attempt a critical dissection of the methods of the Cochrane review. This analysis may then serve as a template for the evaluation of other, past or future, similar reviews.

Indeed, the Cochrane review should be of interest not only to readers interested in the effects of exercise on depression but also to the broader Evidence-Based Medicine community. This is because updates of this review published over a period of only 12 years have resulted in the remarkable stepwise reduction of the pooled standardized mean difference (SMD) by 44%, from −1.10 in 2001, to −0.82 in 2009, to −0.67 in 2012, to −0.62 in 2013.

Thus, the present analysis has a dual purpose. First, this is the first in-depth critique of the Cochrane series of systematic reviews and meta-analyses examining the effects of exercise on depression. As an unintended consequence of the rising global interest in this topic, the strength and quality of the evidence are presently clouded by controversy, confusion, and polarized opinions. Therefore, patients, clinicians, and policymakers may benefit from a critical evaluation of the Cochrane review, arguably the most comprehensive and highest-profile synthesis of the research evidence conducted to date. In particular, emphasis is placed on elucidating the causes of the intriguing gradual "shrinkage" of the pooled SMD reported in successive updates of the review. Second, the present analysis was also conceived as an example-based, step-by-step guide for critically appraising other systematic reviews and meta-analyses in physical activity and mental health. Although credible introductory guides on reading evidence syntheses abound (e.g., Murad et al., 2014), they are written with a generic scope. To help readers identify elements that may be especially susceptible to bias, and thus warrant closer scrutiny, the present analysis uses the following aspects of the Cochrane review as examples, to illustrate how methodological decisions can influence the results and conclusions of a systematic review and/or meta-analysis: (a) the choice of inclusion and exclusion criteria, (b) the uniform-versus-selective application of rules, (c) the rationale behind protocol changes, (d) the lesser known implications of the random-effects meta-analytic model, (e) the complexities involved in appraising the methodological quality of randomized controlled trials (RCTs), and (f) reporting errors.

1. Beware of wayward inclusion criteria resulting in an "apples and oranges" problem

Ioannidis (2010) cautioned that "inclusion/exclusion criteria are a magnificent tool for selecting the data that we like, and for reaching the conclusions that we have already reached before running an analysis" (p. 170). Thus, inclusion and exclusion criteria should be a top target for readers engaged in the critical appraisal of systematic reviews and/or meta-analyses.

An index of the extent of heterogeneity of the effects is an essential and integral element of meta-analyses performed within the context of Cochrane reviews (Deeks, Higgins, & Altman, 2008). Yet, remarkably, in the *JAMA* Clinical Evidence Synopsis, Cooney et al. (2014) did not report an index of heterogeneity. However, their main analysis of 35 trials, which yielded a "medium" pooled SMD of −0.62 (95% CI from −0.81 to −0.42), also revealed significant heterogeneity, $\tau^2 = 0.19$; $\chi^2\,(34) = 91.35$, $p < 0.00001$, $I^2 = 63\%$. Similarly, the analysis restricted to the six "high-quality" trials, which yielded a "small" pooled SMD of −0.18 (95% CI from −0.47 to 0.11), was also characterized by significant heterogeneity, $\tau^2 = 0.07$; $\chi^2\,(5) = 11.76$, $p = 0.04$, $I^2 = 57\%$. According to the Cochrane handbook, this level of heterogeneity is "substantial" (Deeks et al., 2008, p. 278). In such cases, researchers are urged to consider not performing a meta-analysis since this level of heterogeneity indicates an "apples and oranges" problem. At a minimum, researchers must investigate the sources of the heterogeneity and consider either excluding the studies that cause the heterogeneity or analyzing them separately.

Although this was not highlighted in either the Cochrane review or the *JAMA* Clinical Evidence Synopsis, a major source of heterogeneity was the type of comparator used. According to the Cochrane handbook, if "there is a mix of comparisons of different treatments with different comparators," it is "nonsensical to combine all included studies in a single meta-analysis" (Deeks et al., 2008, pp. 246–247). The only statements made by Cooney et al. (2013) regarding this issue, which is undoubtedly critical for the interpretation of the results, were either vague or pointed in the wrong direction: (a) "the type of control intervention may influence effect sizes" (p. 33), (b) "there was substantial heterogeneity; this might be explained by a number of factors including variation in the control intervention" (p. 34), and (c) "we explored the influence of the type of control intervention; this suggests that exercise may be no more effective than stretching/meditation or relaxation on mood" (p. 35). Closer inspection of the role of different comparators, however, proves exceptionally illuminating.

1.1. Studies without control groups

The categorization by the type of comparator revealed two discrepant categories, associated with pooled SMDs whose 95% confidence intervals included zero (−0.15 and −0.24; see Fig. 1). The category yielding the lowest average effect ($n = 6$, pooled SMD −0.15, 95% CI from −0.42 to 0.13) consisted of studies that compared the effectiveness of exercise combined with either medication or psychotherapy to that other treatment alone. In other words, the studies in this category, comprising nearly a fifth of the total number of studies in the meta-analysis, did not involve a comparison between an "exercise" and a "control" group. Rather, in these comparative effectiveness trials, the groups labeled by Cooney et al. as "control" received a recognized form of therapy for depression (i.e., pharmacotherapy or psychotherapy) and the groups labeled "exercise" did not participate in exercise as monotherapy but rather received a combination of exercise and another form of therapy, thus creating the potential for unpredictable cross-treatment interactions and confounds. As one example of such

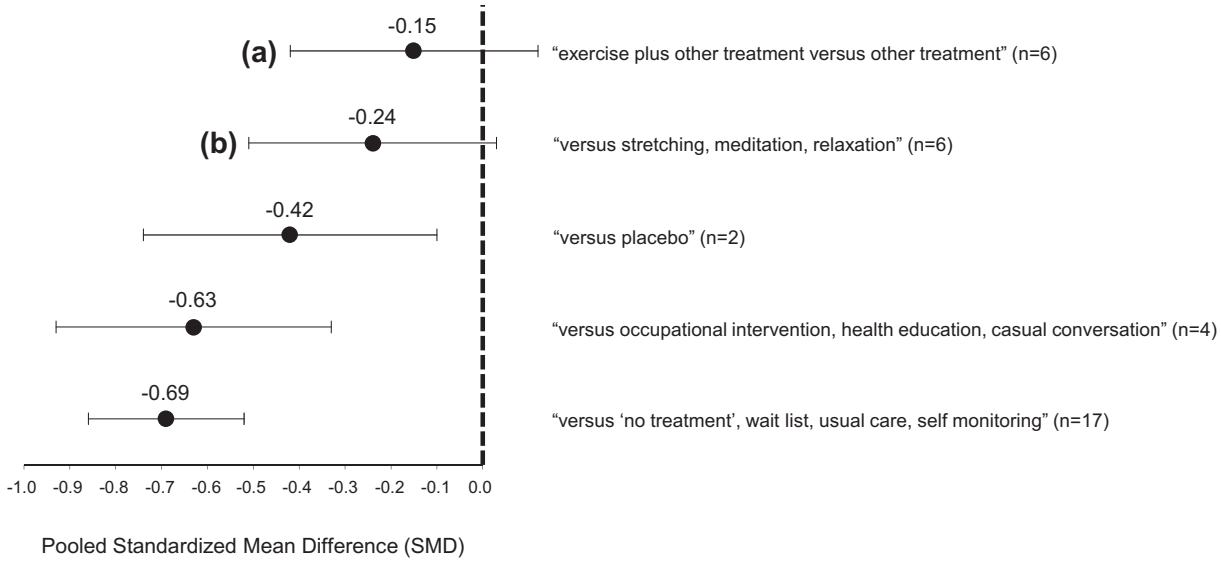


Fig. 1. The moderating effect of the type of comparator on effect sizes. Two categories of studies yield pooled SMDs whose 95% confidence intervals include zero: (a) studies without control groups (i.e., exercise plus other treatment versus other treatment alone); and (b) studies with active-treatments (e.g., meditation, stress management) or low-dose exercise groups as comparators (i.e., stretching and toning, yoga). Whiskers represent 95% confidence intervals.

interactions, animal studies have shown that selective serotonin reuptake inhibitors reduce locomotor activity and spontaneous running behavior (Marlatt, Lucassen, & van Praag, 2010; Weber et al., 2009).

Nevertheless, Cooney et al. (2013, 2014) maintained that the studies in this category can be used to construct an "exercise" versus "no-treatment control" comparison. This claim is presumably based on the assumption that one can add and subtract therapeutic efficacies in algebraic fashion, as in (Exercise + Medication) − Medication = Exercise. It should be emphasized that, due to the absence of a control group, such studies have not been included in other relevant reviews and meta-analyses (e.g., Bridle, Spanjers, Patel, Atherton, & Lamb, 2012; Danielsson, Noras, Waern, & Carlsson, 2013; Josefsson, Lindwall, & Archer, 2014; Silveira et al., 2013). It should also be noted that, in the meta-analysis that served as the forerunner to the Cochrane review, Lawlor and Hopker (2001) similarly reported that the scope of their review included studies in which "exercise was an adjunct, with both treatment and control groups receiving an identical established treatment" (p. 2). However, even in that case, such studies (e.g., Blumenthal et al., 1999; Fremont & Craighead, 1987) were not included in the main meta-analysis comparing the effects of exercise against "no-treatment" control groups. They were only used, appropriately, in secondary analyses comparing exercise to "established treatments" (i.e., psychotherapy, pharmacotherapy). Thus far, the only other meta-analysis in which studies with established treatments as comparators have been included in the calculation of the overall pooled SMD is the review by Krogh, Nordentoft, Sterne, and Lawlor (2011).

The assumption that therapeutic efficacies can be added and subtracted algebraically is demonstrably unsound. Although it is conceivable that two therapeutic modalities working via different mechanisms (e.g., pharmacotherapy and psychotherapy) may exhibit synergistic or additive effects (e.g., Cuijpers, Sijbrandij, et al., 2014; von Wolff, Hölzel, Westphal, Härter, & Kriston, 2012), combination or augmentation strategies working via the same or overlapping mechanisms are unlikely to achieve levels of efficacy fully equivalent to the summated effects of the respective monotherapies. This is especially so if patients are not resistant to one of the two treatments as monotherapy (i.e., if they respond well to each treatment administered separately). For example, studies have shown that (a) the combination of exercise and a mechanistically distinct therapeutic modality, such as electroconvulsive therapy, is better than either modality alone (Salehi et al., 2014) and (b) exercise can be a useful augmentation therapy for patients who had responded poorly or partially to a previous regimen of pharmacotherapy (Trivedi et al., 2011; Trivedi, Greer, Grannemann, Chambliss, & Jordan, 2006). However, even when (a) exercise and the other parallel treatment do not target the same mechanism and (b) the patients are not good responders to the other treatment, it is still unlikely that the additive effect of exercise would be comparable to the effect exercise would have had if administered as monotherapy.

Exercise is theorized to treat depression in part through the same mechanism as antidepressant drugs (i.e., by enhancing serotonergic neurotransmission and stimulating neurogenesis) and in part through the same mechanism as psychotherapy (i.e., by enhancing perceived coping ability and self-appraisals). Indeed, evidence suggests that the concurrent administration of a selective serotonin reuptake inhibitor does not augment the exercise-induced adaptations in serotonergic neurotransmission (MacGillivray, Reynolds, Rosebush, & Mazurek, 2012) or its neurogenic effect (Bjørnebekk, Mathé, & Brené, 2010). Moreover, the concurrent administration of pharmacotherapy or psychotherapy may undermine, rather than augment, the exercise-induced sense of self-efficacy because therapeutic effects may be attributed to the external agent (drug or therapist) instead of internal factors such as personal effort and control (Babyak et al., 2000). For example, in one of the studies, in which a group received exercise in combination with a full regimen of sertraline, "during treatment, several [patients] in the combined group mentioned spontaneously that the medication seemed to interfere with the beneficial effects of the exercise program" (Babyak et al., 2000, p. 636).

To illustrate, consider the two studies from this group in which the other treatment was also offered as monotherapy in one of the arms. In the case of Fremont and Craighead (1987), the comparison between the combination of exercise and counseling to counseling-alone yields an SMD of 0.23 (95% CI from −0.45 to 0.90) in favor of counseling. In contrast, the comparison between exercise-alone and counseling-alone yields an SMD of −0.27 (95% CI from −0.98 to 0.44) in favor of exercise. In the case of Blumenthal et al. (1999),

the comparison between the combination of exercise and sertraline versus sertraline-alone yields a small effect in favor of sertraline (SMD 0.14, 95% CI from −0.25 to 0.52), whereas the comparison between exercise-alone and sertraline-alone reduces the difference to zero (SMD 0.06, 95% CI from −0.33 to 0.45).

1.2. Exercise versus exercise studies

The category yielding the second lowest average effect ($n = 6$, pooled SMD −0.24, 95% CI from −0.51 to 0.04) included studies that used comparators engaged in either a form of therapy (combination of stress management, meditation, and yoga) in one case or other modalities of exercise in the remaining five cases. This is in spite of the fact that Cooney et al. (2013) reportedly excluded all studies that had "no non-exercising comparison group" (p. 10). Again, neither type of treatment can be deemed an inert "no-treatment control," so these studies should also be labeled, more accurately, as comparative effectiveness trials rather than treatment-control trials.

In the study by Klein et al. (1985), the treatment that was entered by Cooney et al. (2014) as "control" was characterized by the researchers as "meditation-relaxation therapy," was delivered by therapists, was designed to "incorporate some of the body awareness and mastery aspects of running," and participants "were taught to concentrate and relax as a means of reducing stress" (p. 155) through a range of breathing techniques and yoga exercises. At the 12-week endpoint, this group fared slightly better than both the exercise group and the group engaged in interpersonal and cognitive group psychotherapy. By treating this therapy group as "control," Cooney et al. (2014) entered the study with an SMD of 0.24 in favor of the "control." Had they considered the other therapy group (i.e., interpersonal and cognitive) as the "control," the study would have been entered with an SMD of −0.22 in favor of exercise.

Of the studies that used groups engaged in different modalities of exercise as comparators, the DEMO trial by Krogh, Saltin, Gluud, and Nordentoft (2009) is the most influential because of its large sample size. It compared (a) an "aerobic" exercise group ($n = 55$), (b) a "strength-training" group ($n = 55$), and (c) a group reportedly engaged in "relaxation" ($n = 55$). Of these, Cooney et al. (2014) selected the "aerobic exercise" group as "exercise" and the "relaxation" group as "control." Labels notwithstanding, however, the groups differed minimally as exercise stimuli. The "strength-training" group engaged in circuit training, a potent form of combined resistance and cardiovascular (aerobic) exercise, resulting in an improvement of aerobic capacity by 8%, just short of the 11% found in the "aerobic" group (which engaged in short intervals of running, cycling, rowing, etc.). Paradoxically, the activities for the "relaxation" group were reportedly designed to "avoid muscular contractions" but the participants were told to do so by "alternating muscle contraction and relaxation in different muscle groups" (p. 792). Likewise, the activities were reportedly designed to avoid "stimulation of the cardiovascular system" but the participants were told to do so by raising their level of perceived exertion up to 12 on a 6–20 scale (between the anchors "fairly light" and "somewhat hard"), a rating within the range recommended by the American College of Sports Medicine (ACSM, 2013) for the improvement of cardiorespiratory fitness. As a result, the "relaxation" group exhibited gains in muscular strength by 10–17% and aerobic capacity by 6% within 4 months. Not surprisingly, since all three groups engaged in exercise in sufficient doses to improve fitness, all three reduced their depression scores to a similar extent. The "aerobic" group had a slightly higher postintervention average depression score, resulting in an effect size in favor of "control" and against "exercise" (SMD 0.25).

Interestingly, Cooney et al. (2013) reported that the follow-up DEMO-II trial (Krogh, Videbech, Thomsen, Gluud, & Nordentoft, 2012) is "unlikely to fulfil inclusion criteria, as the control arm also received exercise" (p. 101). In this newer trial, the comparator was labeled "stretching exercise" rather than "relaxation." It included 10 min of low-intensity warm-up on a stationary bike, 20 min of stretching, and 15 min of "low intensity exercises such as throwing and catching balls" (Krogh et al., 2012, p. 3). While Cooney et al.'s (2013) assessment that this intervention constitutes "exercise" is correct, this "exercise" reduced aerobic capacity by 4% whereas the "relaxation" intervention in Krogh et al. (2009) increased it by 6%.

While the "stretching exercise" comparator in Krogh et al. (2012) was correctly deemed ineligible ("as the control arm also received exercise"), other trials that used stretching groups as comparators were deemed eligible. This was the case despite the fact that flexibility exercise (i.e., stretching) is explicitly identified by the ACSM (2013) as a form of "exercise" (p. 186). These exercise groups were used as "non-exercising comparison groups" even when the researchers specifically used the term "exercise" to describe the activities performed by these groups (e.g., Dunn, Trivedi, Kampert, Clark, & Chambliss, 2005; Knubben et al., 2007).

In fact, most researchers who used groups engaged in different modalities of exercise, such as stretching or yoga, as comparators have remarked that these appear to be active interventions. For example, according to Foley et al. (2008), the inclusion of the stretching group "was intended to differentiate between the effects of aerobic and non-aerobic physical activity" (p. 72). They noted that "it may be that, rather than acting as an exercise 'placebo', the stretching program contained active components which contributed to the decrease in depressive symptoms and improvements in coping efficacy" (p. 72). Similarly, Chu, Buckworth, Kirby, and Emery (2009) initially described their stretching group as "a stretching exercise contact control group" (p. 38) engaging in "supervised stretching and flexibility exercise" (p. 39). However, in their discussion, they pointed out that the group in fact engaged in "yoga-based stretching exercise" (p. 42). The authors concluded that this type of exercise, "may actually be an effective activity treatment for depressive symptoms" (p. 42).

1.3. What was the impact of these inclusions?

As shown in Fig. 2 (panel a), of the studies included in the main analysis by Cooney et al. (2013), six of the eight effect sizes that mostly weaken the pooled SMD, and all five effect sizes favoring the so-called "control" groups, belong in the aforementioned categories (i.e., studies without control groups, studies with exercising comparison groups). Removing the six studies without control groups raises the pooled SMD to −0.66 (95% CI from −0.85 to −0.47) while reducing heterogeneity, $\tau^2 = 0.13$; χ^2 (28) = 59.89, $p = 0.0004$, $I^2 = 53\%$. Removing the additional six studies with exercising comparison groups raises the pooled SMD further, to −0.72 (95% CI from −0.91 to −0.53), while also reducing heterogeneity, $\tau^2 = 0.09$; χ^2 (22) = 38.55, $p = 0.02$, $I^2 = 43\%$.

Cooney et al. (2014) also stated that "analyzing only the 6 trials with adequate allocation concealment, intention-to-treat analysis, and blinded outcome assessment ($n = 464$) showed no association of exercise with improved depression" (SMD −0.18, 95% CI from −0.47 to 0.11) (p. 2432). However, this conclusion must also be questioned, not only because heterogeneity was again "substantial," $\tau^2 = 0.07$; χ^2 (5) = 11.76, $p = 0.04$, $I^2 = 57\%$, but also because three of the six studies belong in the aforementioned categories (i.e., one without a control group and two with exercising comparison groups). Removing these leaves two trials with placebo controls and an additional trial involving a health education control but conducted with older patients (53–91 years) who had not responded to therapeutic doses of antidepressant therapy for at

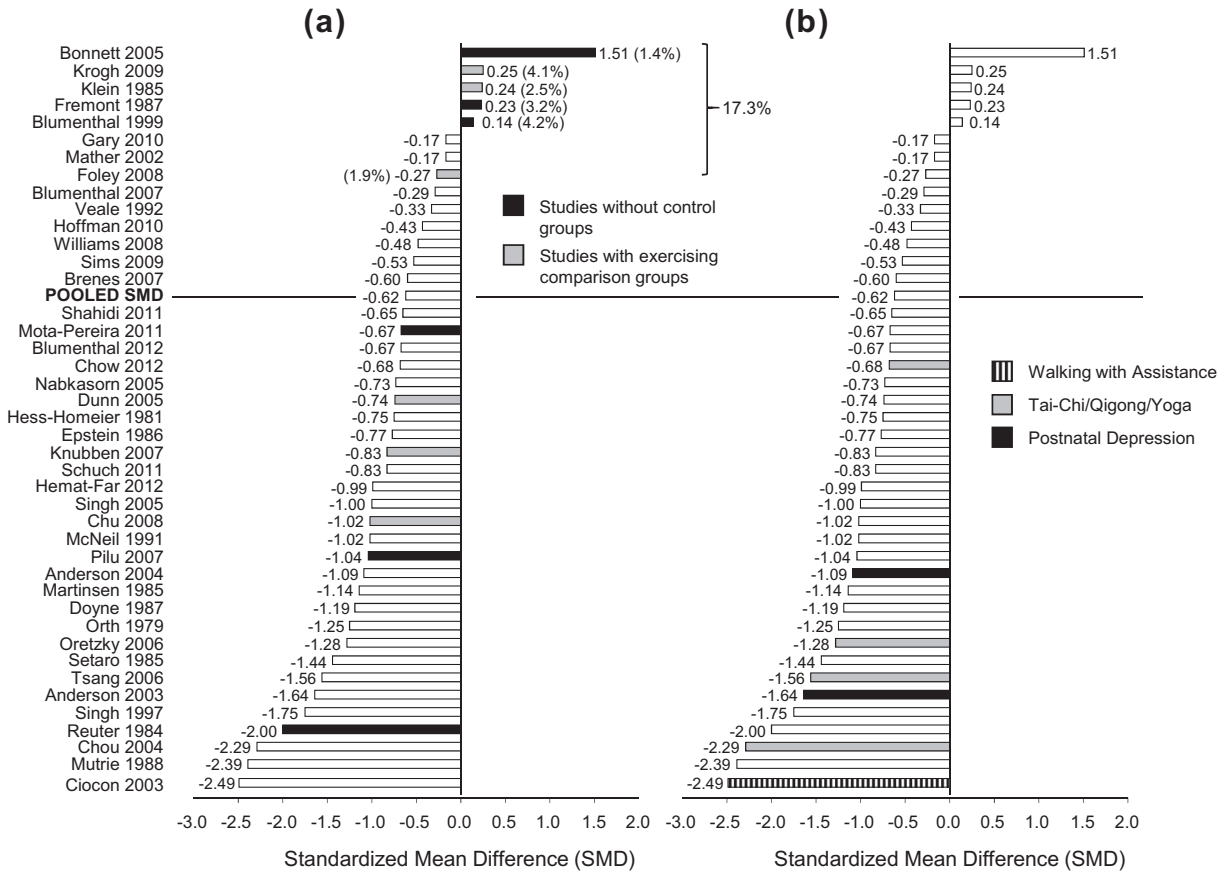


Fig. 2. Influence of questionable inclusion (panel a) and exclusion (panel b) criteria used by Cooney et al. (2013). Panel (a) shows that studies that should have been excluded (i.e., studies without control groups or with exercising comparison groups) contributed six of the eight effect sizes that mostly weaken the pooled SMD (adding a total weight of 17.3% to one tail of the distribution), and all five effect sizes favoring the so-called "control" groups. Removing the six studies without control groups would raise the pooled SMD to −0.66 (95% CI from −0.85 to −0.47). Removing the additional six studies with exercising or active-treatment comparison groups would raise the pooled SMD to −0.72 (95% CI from −0.91 to −0.53). Panel (b) shows that all studies excluded on the basis of questionable arguments (i.e., walking with assistance, yoga, tai-chi, qigong, postnatal depression) would have strengthened the pooled SMD in favor of exercise. Inclusion of these studies would have resulted in a "large" pooled SMD of −0.77 (95% CI from −0.97 to −0.58). Combining the removal of studies that should not have been included and the restoration of studies that should not have been excluded would result in a pooled SMD of −0.90 (95% CI from −1.11 to −0.69).

least six weeks. With the analysis restricted to these three high-quality trials, the pooled SMD was significantly different from zero (SMD −0.33, 95% CI from −0.59 to −0.07).

1.4. Beware of future inclusions

Caution should also be used in evaluating future updates of the Cochrane review. Cooney et al. (2013) announced that "for future updates, we will include data from trials that reported subgroups with depression" (p. 33). The authors acknowledged that this change of protocol was specifically prompted by the publication of one "large, high-quality, cluster-randomized trial" (p. 33), namely the OPERA trial (Underwood et al., 2013). It should be emphasized that this trial, which yielded null results, was characterized by a failure to implement the previously planned intervention protocol (Ellard, Thorogood, Underwood, Seale, & Taylor, 2014). This failure impacted nearly every aspect of the intervention, from the population that was targeted (participants were older than anticipated and too frail to exercise), to the level of exercise that could be performed (low-intensity, seated exercises instead of the planned standing, moderate-intensity aerobic activities), to influencing the culture of the randomized facilities (contrary to expectations, there was no evidence of raising the level of physical activity beyond the minimal number of scheduled exercise sessions).

From a meta-analytic standpoint, the OPERA trial will further exacerbate the "apples and oranges" problem, as it is clearly methodologically distinct from all other trials. First, it is the only trial that used a cluster-randomized design. Second, OPERA is the only trial in this literature in which medical contraindications to exercise were (remarkably) not an exclusion criterion; participants were included as long as they could communicate their responses to the depression questionnaire. Third, OPERA is also the only trial in which participants diagnosed with dementia were not excluded despite the fact that the outcome measure of depression, which relied entirely on self-reports, was not validated for respondents with dementia.

Exacerbating the potential for bias, if OPERA is added to the data set, it will have the largest weight (almost 5%). This is because although randomization was done at the level of nursing homes, the analyses were carried out at the level of individuals. Thus, the sample size is larger than that of other trials in this literature, all of which involved participant recruitment and randomization at the level of individuals. The subgroup characterized as "depressed" at baseline (on the basis of a previously untested cutoff score, without verification by a diagnostic clinical interview) included 374 participants. Thus, the below-average effect associated with this trial (SMD −0.07, 95% CI from −0.31 to 0.18) will have a disproportionate attenuating effect on the pooled SMD. However, it should be clear

that, given the aforementioned failure to implement the planned intervention, this effect cannot be reasonably deemed as representing "exercise."

2. Beware of wayward exclusion criteria resulting in selective "trimming" of the evidence

Besides questionable inclusion criteria, the analysis by Cooney et al. (2014) also involved changes in protocol resulting in the selective exclusion of several studies yielding large effects favoring exercise. Readers should contemplate whether these exclusions were principled by questioning whether the arguments used to justify the exclusions are based on sound reasoning.

2.1. Definitions are essential

In the section entitled "Comparison of findings with current practice guidelines" of the *JAMA* Clinical Evidence Synopsis, Cooney et al. (2014) argued that "the UK National Institute for Health and Clinical Excellence recommends structured exercise, 3 times a week for 10–14 weeks, for the treatment of mild to moderate depression" (p. 2433; also see same statement on p. 8 in Cooney et al., 2013). This statement contains crucial inaccuracies. In actuality, the National Institute for Health and Care Excellence (NICE) recommends "a structured group physical activity program" for "people with persistent subthreshold depressive symptoms or mild to moderate depression" (National Collaborating Centre for Mental Health and National Institute for Health and Clinical Excellence, 2010, p. 213). Therefore, the first inaccuracy is that, contrary to the claim by Cooney et al. (2013, 2014) that NICE recommends "exercise," the guideline specifies "physical activity." Many readers, especially those without academic backgrounds in exercise science or kinesiology, might not have noticed the difference. However, this difference had a profound impact on study selection criteria (as detailed in the following sections) and, ultimately, on the results of the Cochrane review. The second inaccuracy is that the NICE guideline encompasses individuals with subthreshold depressive symptoms, thus extending the potential application of physical activity over a broader segment of the population. Therefore, the section entitled "Comparison of findings with current practice guidelines" in Cooney et al. (2014) should be viewed cautiously. It should be clear that there is no agreement between the review and the guidelines in either the nature of the treatment ("exercise" versus "physical activity") or the condition being treated ("depression" defined as meeting diagnostic criteria versus "depression" that spans the range from "persistent subthreshold depressive symptoms" to moderate depression).

Although NICE had used the term "exercise" in its now-defunct National Clinical Practice Guideline 23, first issued in 2004, it changed the term to "physical activity" in its current guideline, National Clinical Practice Guideline 90, issued in 2010 (National Collaborating Centre for Mental Health and National Institute for Health and Clinical Excellence, 2010). The change was made with the explicit purpose of encompassing a broader range of activities under the umbrella of "physical activity" than those subsumed under the rubric of "exercise."

Cooney et al. (2013) introduced the requirement that the studies included in the review must focus "on exercise defined according to ACSM criteria" (p. 8). ACSM (2013) defines "physical activity," very broadly, as "any bodily movement produced by the contraction of skeletal muscles that results in a substantial increase in caloric requirements over resting energy expenditure" (p. 2). On the other hand, ACSM defines "exercise," more narrowly, as "a type of physical activity consisting of planned, structured, and repetitive bodily movement done to improve and/or maintain one or more components of physical fitness" (p. 2). The NICE definition of "physical activity" shares the breadth of scope of the ACSM definition, encompassing aerobic (e.g., training of cardiorespiratory capacity) and anaerobic types of activities (e.g., training of muscular strength and endurance), as well as, very importantly, flexibility, coordination, and even "relaxation" activities. The only notable difference is that the NICE definition adds the condition that the activities must be "structured … with a recommended frequency, intensity and duration" (National Collaborating Centre for Mental Health and National Institute for Health and Clinical Excellence, 2010, p. 191). Despite this addition, however, the definition is still different from what ACSM considers "exercise" since the crucial element of purpose (i.e., "to improve and/or maintain one or more components of physical fitness") is absent. Therefore, the suggestion that the Cochrane review, whose scope is allegedly delimited to ACSM-defined "exercise," can be used as an empirical test of the NICE guideline, which explicitly refers to "physical activity," should be regarded as dubious. Some of the implications of this crucial point are highlighted in the following sections.

2.2. "Walking with assistance" is not "exercise"

Cooney et al. (2013) used the argument that certain interventions failed to fulfill the ACSM definition of "exercise" as the basis for excluding several studies, even some that had been included in earlier versions of the Cochrane review. In perhaps the most controversial case, Cooney et al. (2013) excluded the study by Ciocon and Galindo-Ciocon (2003) because, reportedly, "further scrutiny led us to conclude that the intervention did not fulfil the definition of 'exercise'" (p. 14) and the "intervention appeared not to be exercise according to ACSM definition" (p. 94). The intervention, which was applied to long term residents in a nursing home, consisted of "getting residents up, walk with assistance, every shift for at least 20 min" (Ciocon & Galindo-Ciocon, 2003). It is unclear which of the elements of the ACSM definition this intervention failed to match, since the program was apparently planned, the activity was structured, and walking involves repetitive bodily movement done with the purpose of restoring several components of fitness (e.g., muscular strength, endurance). What is clear, however, is that this study would have yielded a large SMD (−2.49, 95% CI from −3.24 to −1.74) in favor of exercise. With a weight of 2.8%, inclusion of this study in the analysis would have changed the pooled SMD from −0.62 (95% CI from −0.81 to −0.42) to −0.68 (95% CI from −0.90 to −0.47).

2.3. Tai-chi, qigong, and yoga are not "exercise"

Cooney et al. (2013) excluded all RCTs involving tai-chi, qigong, or yoga, which had been included in earlier versions of the review. This decision was also based on the assertion that these activities do not constitute "exercise." As noted, "exercise" is defined by the ACSM (2013) as "a type of physical activity consisting of planned, structured, and repetitive bodily movement done to improve and/or maintain one or more components of physical fitness" (p. 2). What is perhaps not widely appreciated is that the components of physical fitness include, besides the well-known factors of cardiovascular endurance and muscular strength, such factors as flexibility, agility, coordination, and balance. Indeed, the ACSM (2013) has specifically identified tai-chi, qigong, and yoga as examples of "neuromotor exercise, resistance exercise, and flexibility exercise" (p. 189), defining "neuromotor exercise" as the type of training that involves balance, coordination, gait, agility, and proprioceptive training ("sometimes called functional fitness training"). Therefore, it is again unclear how a planned, structured, and purposeful

program of tai-chi, qigong, or yoga would fail to fulfill the ACSM criteria of "exercise."

It should also be pointed out that this questionable exclusion impacted the important analysis restricted to "high-quality" trials. Specifically, Cooney et al. (2013) excluded an RCT that investigated the effects of a tai-chi intervention on older (>60 years) partial responders to escitalopram (Lavretsky et al., 2011). Mysteriously, this study was reportedly excluded not because the intervention consisted of tai-chi but rather because "control is health education, an active intervention" (Cooney et al., 2013, p. 96). This exclusion criterion, however, is perplexing since the review already includes other trials with health education control groups (e.g., Mather et al., 2002; Singh, Clements, & Fiatarone, 1997). Lavretsky et al. (2011) described their tai-chi treatment as purposeful (a "health management intervention" which, besides lowering depression, was done for the purpose of "helping older adults cope with fatigue"; p. 841). The intervention was further described as consisting of repetitive bodily movement ("physical activity" consisting of "repetitious, nonstrenuous, slow-paced movement"; p. 842), with planning and structure [each session lasted "120 min and also included 10 min of warm-up (e.g., stretching and breathing) and 5 min of cool-down exercises"; p. 842]. Therefore, these characteristics match the aforementioned defining attributes of "exercise."

Moreover, according to Lavretsky et al. (2011), the study satisfied all three of the criteria for high methodological quality used by Cooney et al. (2013), namely adequate allocation concealment, blinded outcome assessment, and intention-to-treat analysis: (a) "randomization was performed by using a computer-generated schedule, independent of treatment personnel" and "allocation concealment was implemented by using sealed, sequentially numbered boxes that were identical in appearance for the two treatment groups"; (b) "all assessments were performed by the raters blinded to the treatment group assignment" and "subjects were asked not to disclose their group assignment to the raters" (p. 841); and (c) "all outcome results used intent-to-treat analyses" (p. 843). It is, therefore, noteworthy that the effect size of this trial (SMD −0.40) was more than twice the average from high-quality trials reported by Cooney et al. (2013, 2014).

2.4. Postnatal depression is not "depression"

Cooney et al. (2014) excluded two RCTs of postnatal depression (Armstrong & Edwards, 2003, 2004), which, although small, both had large SMDs in favor of exercise (−1.64, 95% CI from −2.68 to −0.59; and −1.09, 95% CI from −2.07 to −0.11). These exclusions were made without providing any justification (e.g., differences in pathophysiology or treatment options). Indeed, nothing in the description of depressive disorders "with peripartum onset" (pp. 186–187) in the Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association, 2013) suggests that this type of major depression represents a distinct disorder in terms of pathophysiology or treatment options.

2.5. What was the impact of these exclusions?

As shown in Fig. 2 (panel b), through several selective exclusions (also see the next section), Cooney et al. (2014) eliminated 5 of the 9 (and 6 of the 13) strongest effect sizes in favor of exercise. Inclusion of the excluded studies described in this section (i.e., assisted walking, tai-chi, qigong, yoga, postnatal depression) in the main analysis raises the pooled SMD to −0.77 (95% CI from −0.98 to −0.57). Including these studies while also excluding studies with questionable comparators (i.e., studies without control groups, studies with exercising comparison groups) results in a large pooled SMD of −0.90 (95% CI from −1.11 to −0.69).

3. Beware of cherry-picking and selective application of rules

Systematic reviews and meta-analyses involve numerous decisions that can potentially (especially in the aggregate) influence the outcome. Therefore, it is essential to ensure not only that any rules are principled and fully documented but also that they are applied uniformly.

3.1. "Exercising" versus "non-exercising" comparators

Cooney et al. (2013) declared that they excluded studies that had "no non-exercising comparison group" (p. 10). On the basis of this rule, for example, the authors excluded the study by Bosscher (1993) because it reportedly involved a comparison between "different types of exercise with no non-exercising control group" (Cooney et al., 2013, p. 93). In this study, depressed inpatients were "randomly assigned to short-term running therapy or to a treatment-as-usual with mixed physical and relaxation exercises" (Bosscher, 1993, p. 170). Participants allocated to the treatment-as-usual control group engaged in (a) two 50-min sessions per week of "relaxed, low-intensity physical activity," including ball games, jumping on trampolines, and gymnastics, and (b) one 50-min session per week of "relaxation and breathing exercises" (Bosscher, 1993, p. 176).

In contrast, Cooney et al. (2013) decided to include the aforementioned DEMO trial by Krogh et al. (2009). In that study, participants allocated to the "relaxation training" group did 20–30 min of "exercises on mattresses," including exercises with inflatable balls, "followed by light balance exercises for 10–20 min and by relaxation exercises with alternating muscle contraction and relaxation in different muscle groups while lying down for 20–30 min" (Krogh et al., 2009, p. 792). As noted earlier, these exercises increased maximal aerobic capacity by 6% and the strength of different muscle groups by 10–17%. By comparison, participants in the "aerobic training" group exhibited increases in aerobic capacity that were slightly larger (11%) but gains in muscular strength that were smaller (3–10%).

The *excluded* study by Bosscher (1993) would have yielded an SMD of −1.22 (95% CI from −2.25 to −0.19) in favor of "exercise." In contrast, the *included* study by Krogh et al. (2009) yielded an SMD of 0.25 (95% CI from −0.17 to 0.66) in favor of "control."

3.2. "Active treatment" versus "control"

According to Cooney et al. (2013), the Cochrane review had two distinct objectives: (a) to determine the effectiveness of exercise compared with no treatment (no intervention or control) and (b) "to determine the effectiveness of exercise compared with other interventions (psychological therapies, alternative interventions such as light therapy, pharmacological treatment)" (p. 10). The operational definition of an "other intervention" (also termed "active treatment" on p. 10) was that the "aim of the treatment was to improve mood." Reportedly, this category includes "pharmacological treatments, psychological therapies, or other alternative treatments" (p. 10).

Presumably based on this operational definition, Cooney et al. (2013) considered "light therapy" as an "active treatment" or "alternative intervention" (p. 9). Thus, a trial comparing exercise to light therapy (Pinchasov, Shurgaja, Grischin, & Putilov, 2000) was not included in the main analysis since "light therapy" presumably did not qualify as "control." Instead, the study was considered in a separate comparison (Analysis 3.1). It should be noted, however, that, even though "light therapy" may be used with the purpose of improving mood, NICE has determined that there is no clear evidence that light therapy is efficacious for the treatment of

depression (National Collaborating Centre for Mental Health and the National Institute for Health and Clinical Excellence, 2010, p. 450).

On the other hand, as noted earlier, the "meditation-relaxation therapy" group in Klein et al. (1985), in which psychotherapists taught participants "to concentrate and relax as a means of reducing stress" (p. 155) was not considered an "active treatment" or "alternative intervention" but rather as a "no treatment" control. This was despite the fact that the treatment was labeled as "therapy," was delivered by therapists, and was presumably done with the purpose of improving mood, since participants were encouraged to use the techniques "to reduce tension in their daily lives" (p. 155).

The study by Pinchasov et al. (2000), which was *excluded* from the main analysis, would have yielded an SMD of −1.48 (95% CI from −2.56 to −0.41) in favor of exercise. In contrast, the *included* study by Klein et al. (1985) yielded an SMD of 0.24 (95% CI from −0.64 to 1.11) in favor of the "control."

3.3. "Depressed" versus "non-depressed" participants

Cooney et al. (2013) declared that they included trials "if the participants were defined by the author of the trial as having depression (by any method of diagnosis and with any severity of depression)" (p. 9). In contrast, they "excluded trials that randomized people both with and without depression, even if results from the subgroups of participants with depression were reported separately" (p. 9).

On the basis of this rule, Cooney et al. excluded the study by Kerse et al. (2010) because the participants "did not all have diagnosis of depression to enter trial" (p. 96). Indeed, Kerse et al. (2010) recruited participants on the basis of a previously validated three-question screen for depression rather than a formal diagnosis. After entering the trial, it was found that 27% of those participants met criteria for a diagnosis of major depression and 53% had either elevated scores on a depression questionnaire or met at least one of the criteria for depression from a standardized diagnostic interview.[1]

On the other hand, Cooney et al. (2013) included a trial by Blumenthal et al. (2012), in which participants (coronary heart disease patients) also did not have to have depression to enter the study. The inclusion criteria only specified "elevated score (≥7)" on the Beck Depression Inventory II (p. 1055). This score, however, is well within the range indicating "none or minimal" depressive symptoms and is only half of the recommended cutoff score for "mild" depressive symptoms (i.e., 14). For a sample of patients with heart disease, the inclusion of several somatic symptoms in the Beck Depression Inventory II (e.g., perceived level of energy, tiredness, sleep, appetite, libido) makes a score of 7 likely to reflect primarily the physical consequences of heart disease rather than depression (Thombs et al., 2010). Indeed, upon entry, fewer than half (47%) of randomized participants met diagnostic criteria for depression.

Because the effect of exercise on depression is likely to be smaller in individuals with a lower baseline level of symptoms, basing the analysis on the entire sample biases the estimate of the exercise effect downward. Changing the effect size associated with Blumenthal et al. (2012) from the entire sample to the subsample with a depression diagnosis would strengthen the SMD from −0.67 (95% CI from −1.23 to −0.12) to −0.94 (95% CI from −1.82 to −0.07).

[1] Computing an effect size using the values reported by Kerse et al. (2010) yields a large SMD in favor of exercise (−2.74, 95% CI from −3.14 to −2.34). Contact with the authors, however, revealed that the values inadvertently reported as standard deviations were, in fact, standard errors, thus reducing the SMD to −0.28 (95% CI from −0.57 to 0.00). The authors were unaware of this error. Until they were contacted for this article, they had not been asked to confirm the reported figures by other investigators.

4. Beware of the reasoning behind protocol changes

According to the Cochrane handbook, "changes in the protocol should not be made on the basis of how they affect the outcome of the research study. *Post hoc* decisions made when the impact on the results of the research is known, such as excluding selected studies from a systematic review, are highly susceptible to bias and should be avoided" (Green & Higgins, 2008, p. 12). In previous updates of the Cochrane review on exercise and depression, trials that included multiple exercise arms were represented in the analysis by the exercise arm that showed the largest effect. This decision was based on the premise that there was no a priori basis for selecting one type or amount of exercise versus another as being "optimal." Cooney et al. (2014) affirmed that this remains the case today, stating that "the optimal type, intensity, frequency, and duration of exercise for depression remain unclear" (p. 2432). Yet, despite this acknowledgment, they introduced a new change in their review protocol, now selecting "the exercise arm which provides the biggest 'dose' of exercise." They declared that this was deemed necessary because doing otherwise "may overestimate the effect of exercise" (Cooney et al., 2013, p. 12).

The decision to select the trial arms that received the "biggest dose" was unaccompanied by a conceptual rationale or empirical evidence and, as such, appears arbitrary. As with any dose–response relationship, the "biggest dose" of a treatment may not be optimal and may, in fact, be toxic. The highest intensity of exercise, for example, may be intimidating or exhausting and the highest frequency or duration may be inconvenient or unrealistic, leading to perceptions of failure and possible exacerbation of depressive symptoms. Therefore, if the dose–response curve is hormetic (i.e., inverted-J or -U), as many dose–response curves are, "the biggest dose" may well be detrimental.

Nevertheless, Cooney et al. (2013) applied this protocol change, supplementing it with a sensitivity analysis (p. 26), reportedly aimed to compare the pooled SMD when using the "biggest dose" (−0.62, 95% CI from −0.81 to −0.42) to the pooled SMD when using the "smallest dose" (−0.44, 95% CI from −0.55 to −0.33). They concluded that using the "smallest dose" also results in "a moderate clinical effect in favor of exercise" (p. 26).

Closer inspection of this analysis, however, leads to some bewildering observations. In actuality, the sensitivity analysis did not include effect sizes derived from trial arms with the "smallest doses" but rather a peculiar mix of treatments, including some treatments that did not involve exercise or physical activity. In fact, out of 10 effect sizes that were changed to so-called "smallest doses" for the sensitivity analysis, only three involved lower doses of exercise. Specifically, one involved lower-intensity resistance exercise (20%, with SMD −0.32, versus 80% of one-repetition maximum, with SMD −1.75; Singh et al., 2005), one involved lower-intensity aerobic exercise (40–55%, with SMD −0.38, versus 65–75% of oxygen uptake reserve, with SMD −1.02; Chu et al., 2009), and one involved lower total energy expenditure (7.0 kcal/kg/week over 3 days per week, with SMD −0.41, versus 17.5 kcal/kg/week over 5 days per week, with SMD −0.74; Dunn et al., 2005).

For the remaining seven studies, the effect sizes used for the sensitivity analysis included switching (a) from supervised to home exercise but with identical exercise prescriptions (Blumenthal et al., 2007), (b) from an aerobic exercise arm to a resistance exercise arm but without any evidence that the latter represented a smaller "dose" of exercise (Doyne et al., 1987; Krogh et al., 2009; Mutrie, 1986), (c) from walking to a combination of walking, strength, and flexibility but for the same duration per session (Williams & Tappen, 2008), (d) from jogging to any "self-chosen activity" which was a "non-physical activity approximately two-thirds of the time" (65%; Orth, 1979, p. 233), and (e) in arguably the most controversial case,

from an instructor-led aerobic dance class to "an arts and crafts class" engaged in "the fabrication and painting of ceramic arts" (Setaro, 1985, p. 113). Therefore, given these choices, this "sensitivity analysis" cannot be deemed informative with respect to the effects of different doses of exercise, nor can it reveal the true impact of selecting the arms with the "largest dose" of exercise.

4.1. Which studies did the change(s) in protocol affect?

In actuality, the decision to select the trial arms with the "biggest dose" of exercise altered the entry of only one study, namely the DOSE trial by Dunn et al. (2005). However, this was an important study since it was one of only six RCTs comprising the subgroup of "high-quality" trials, the analysis of which reportedly showed "no association of exercise with improved depression" (Cooney et al., 2014, p. 2433).

The DOSE trial used a 2 × 2 factorial design, plus a stretching-and-flexibility exercise comparison group. The two factors that were manipulated were (a) the total weekly energy expenditure (a "low dose" of 7 kcal/kg/week, roughly equivalent to 80 min of moderate-intensity activity per week, and a "public health dose" of 17.5 kcal/kg/week, roughly equivalent to 180 min of moderate-intensity activity per week) and (b) frequency (3 days per week and 5 days per week). The authors explained that the factors were crossed, such that "each energy expenditure group was divided into 3- or 5-day/week groups" (p. 2). In other words, the participants allocated to the "public health dose over 5 days per week" group did the same amount (dose) of exercise as those allocated to the "public health dose over 3 days per week" group; the difference was that the amount of exercise was distributed over five days as opposed to three.

This point was misconstrued in all previous versions of the Cochrane review, in which the trial was erroneously described as having compared "four different 'doses' of aerobic exercise" and the results were incorrectly summarized as having shown that "high-intensity exercise was more effective than low-intensity exercise" (Mead et al., 2009, p. 9). These errors prompted the lead researcher of the DOSE trial, Dr Andrea Dunn, to contact the reviewers and request a correction, reiterating that the study did not involve a manipulation of intensity but rather "the two factors that were manipulated were frequency of exercise and total energy expenditure" (Cooney et al., 2013, p. 154). While the reviewers noted that "feedback received from a trialist … was addressed" (Cooney et al., 2013, p. 155), the text of the review continued to incorrectly describe the DOSE trial as having "compared four different 'doses' of aerobic exercise" until the 2012 update.

Cooney et al. (2013) finally corrected this mistake, describing the DOSE trial as having compared "4 different aerobic exercise programs, that varied in total energy expenditure (7.0 kcal/kg/week or 17.5 kcal/kg/week) and frequency (3 days per week or 5 days per week)" (p. 59). Even though this correction seems to suggest that the design of the trial was properly understood, Cooney et al. (2013) incorrectly selected the "public health dose over 5 days per week" as the arm that provided "the biggest 'dose' of exercise." It should be clear, however, that the "public health dose over 5 days per week" group and the "public health dose over 3 days per week" group received the same "dose" of exercise.

In making this erroneous selection, the SMD associated with the Dunn et al. (2005) study was reduced from −1.16 ("public health dose over 3 days per week") to −0.74 ("public health dose over 5 days per week"). As noted, the importance of this error is exacerbated by the fact that the Dunn et al. (2005) trial was also one of the six trials deemed to be of high quality (i.e., with adequate allocation concealment, intention-to-treat analysis, and blinded outcome assessment). Dunn et al. (2005) noted that, although to a non-significant degree ($p = 0.46$), adherence was lower among the groups that had to exercise 5 days per week (65%) than those who had to exercise on 3 days per week (78%). Since all exercise took place "in the laboratory" (p. 2), the added transportation burden for participants who had to divide their exercise dose over 5 days each week could account for this difference in adherence. This exemplifies the problems that could result from the decision to select "the biggest dose" of exercise as a way of reducing bias.

4.2. Aerobic exercise as the de facto exercise archetype

Closely associated with the decision to select the "biggest dose" of exercise as opposed to the empirically established optimal dose, Cooney et al. (2013) also chose to designate as "exercise" the aerobic exercise arm for those trials that included both an aerobic-exercise and a resistance-exercise arm. What was different in this case is that this decision was never explicitly stated in the text of the review and, thus, was unsupported by either conceptual arguments or empirical evidence.

Nevertheless, this decision influenced both the overall analysis and the analysis restricted to "high-quality" trials. Specifically, while the strength and aerobic exercise arms from the Doyne et al. (1987) study (weight 2.5%) yielded similar effects compared to the waitlist control (−1.19), the larger Krogh et al. (2009) study (weight 4.1%) yielded 0.25 from the aerobic-exercise arm in favor of the so-called "relaxation" (i.e., alternative-modality exercise) comparator, whereas the resistance-exercise arm would have yielded −0.10 in favor of "exercise".

5. Beware of the tricky consequences of the random-effects model

The random-effects model of meta-analysis is commonly used because, unlike the restrictive assumption of the fixed-effects model that all effect sizes in a data set estimate the same underlying treatment effect, the random-effects model assumes that the effects from different studies follow a distribution (some are smaller, some are larger). This is a more realistic assumption that allows for the possibility of heterogeneity in the sample of studies (though using the random-effects model does not absolve researchers of the obligation to investigate the sources of heterogeneity). In the random-effects model, a measure of the extent of heterogeneity among the effect sizes from different studies (an estimate of between-study variance) is added to the standard errors associated with these effects. This measure of heterogeneity is termed τ^2. When the data set is homogeneous, then $\tau^2 = 0$ and the results of the fixed-effects and random-effects models are identical. On the other hand, with higher heterogeneity, the magnitude of τ^2 grows and so do the differences in the results from random- and fixed-effects models.

A relatively underappreciated consequence of the random-effects model is that "by adding a constant number (τ^2) to the weight of each study, the relative contributions of each trial will become more equal" (Moayyedi, 2004, p. 2297). Consequently, "small studies will therefore become more prominent and larger trials will become less to the overall effect estimate compared to fixed effects models" (p. 2297).

5.1. Can the inclusion of questionable small studies be consequential?

The most extreme SMD favoring the "control" group reported by Cooney et al. (2013) was 1.51. It was associated with the doctoral dissertation by Bonnet (2005). The decision to include this study in the meta-analysis is puzzling for several reasons. First, the study

was not conceived as an RCT but rather as a case series. As such, it reported no aggregate-level statistics (i.e., "utilized a single subject design in which all subjects served as their own control" and "charted progression of self-report measures was used to analyze the data, which is the typical method of reviewing single-subject research"; Bonnet, 2005, p. 51). Second, the study had no control group. It involved a comparison between cognitive therapy ($n = 6$) and cognitive therapy plus walking ($n = 5$). This was converted to an "exercise versus no-treatment" comparison by Cooney et al. (2013) based on the aforementioned dubious assumption that treatment effects can be added and subtracted in algebraic fashion. Third, the "exercise" intervention consisted of "walking on a treadmill for 20 min at a moderate intensity of four miles per hour, twice weekly, for six weeks" (Bonnet, 2005, pp. ii–iii). This amount of physical activity is only 27% of the minimum recommended for health promotion (240 min of moderate activity over six weeks instead of the recommended 900 min). Fourth, due to the small sample size, randomization was heavily biased, resulting in 37% difference in depression scores between the groups at baseline (with the participants in the cognitive behavioral therapy group scoring lower). This degree of bias makes any between-group comparison of post-intervention scores essentially uninterpretable.

Nevertheless, Cooney et al. (2013) included this study in their meta-analysis by calculating the means and standard deviations of the two groups while also imputing the missing values (by "carrying forward baseline data for those who dropped out"; p. 55). As noted in the Cochrane handbook, "inflating the sample size of the available data up to the total numbers of randomized participants is not recommended as it will artificially inflate the precision of the effect estimate" (Higgins, Deeks, & Altman, 2008, p. 492). Despite this admonition, it is not entirely uncommon for meta-analysts with access to individual-level data to intervene in this manner by performing an "intention-to-treat" analysis that the original authors had failed to perform. However, it should be clear that intervening in this manner can introduce considerable bias when (a) dropout was 36% (4 of 11 participants) and (b) as noted earlier, there was severe baseline imbalance as a result of unsuccessful randomization. This intervention by Cooney et al. (2013) resulted in a 529% increase in the SMD associated with the study by Bonnet (2005), in favor of the "control," from 0.24 (without imputation of the missing values) to 1.51 after the last-observation-carried-forward imputation that capitalized on the severe baseline imbalance.

In a fixed-effects model, this problem would have had very limited impact on the analysis given the very small sample size. However, because of the use of a random-effects model, which inflates the relative contribution of smaller studies while attenuating the relative contribution of larger ones, this study, despite accounting for only 0.8% of the total sample size (11 of 1353), was entered in the analysis with a weight of 1.4% (75% inflation). In contrast, for example, the high-quality, placebo-controlled RCT by Blumenthal et al. (2007), with a sample size of 100 (i.e., 7.4% of 1353), was entered in the analysis with a weight of 4.2% (i.e., 43% attenuation).

6. Beware of blanket negative statements about methodological weaknesses

Cooney et al. (2014) noted that "many" of the trials in the meta-analysis "had methodological weaknesses" (p. 2433). This negative assessment reflects similar statements in the Cochrane review (e.g., "uncertainties remain regarding how effective exercise is for improving mood in people with depression, primarily due to methodological shortcomings" in Cooney et al., 2013, p. 33). The assertion about the low methodological quality of this body of evidence is crucial. In the oft-cited meta-analysis that inspired the subsequent series of Cochrane reviews, Lawlor and Hopker (2001) first used the argument that "most studies were of poor quality" (p. 3) as justification for deflecting attention away from the large effect size (pooled SMD −1.10, 95% CI from −1.50 to −0.60) and interpreting the results as indicating uncertainty: "it is not possible to determine from the available evidence the effectiveness of exercise in the management of depression" (p. 6). The claim about the poor quality of the evidence has since been prominently featured in all editions of the Cochrane review. Closer inspection, however, shows that the picture is more complicated than these blanket statements suggest.

It is important to note that nearly half (16 of 37, 43%) of the studies reviewed by Cooney et al. (2013) were published in 2000 or earlier, before the original Consolidated Standards of Reporting Trials (CONSORT) had been broadly implemented in the literature (Begg et al., 1996). For these earlier studies, evaluating methodological quality based on whether the reports reflect the standard phraseology that was introduced with the CONSORT guidelines and checklist results in biased risk assessments. It should also be emphasized that the problem of incomplete reporting of methodological details during the pre-CONSORT era is certainly not unique to the line of research examining the effects of exercise on depression. This problem characterized most of the medical literature of the pre-CONSORT era (Moher, Jones, & Lepage, 2001) and indeed served as the impetus for the development of CONSORT.

6.1. Concealment of group allocation

In deciding whether allocation was adequately concealed, Cooney et al. (2013) used the typical approach of the post-CONSORT era of searching for specific catchphrases (see Begg et al., 1996, p. 638) that are taken to signify adequate concealment (i.e. randomization at a site remote from the study; computerized allocation with records kept in a locked file; drawing of sequentially numbered, sealed, and opaque envelopes). There are at least two problems with this approach.

First, as alluded earlier, the odds of finding a pre-CONSORT study that uses these exact catchphrases are small. Thus, this decision summarily, yet possibly unfairly, precludes studies prior to 2000 from being considered "low risk" and, therefore, of high quality. Critical reviews have called attention to this problem, showing that the absence of these catchphrases in earlier reports should not be taken as evidence that group allocation was necessarily inadequately concealed (Devereaux et al., 2004). As one example, in the study by Mutrie (1986), "the random assignment procedure was carried out by a predetermined assignment of subjects' intake number by a computer program" (p. 59). The randomization took place at a university site remote from the National Health Service surgeries where the participants were recruited. Although a reasonable reading of these methods suggests that the possibility of violating the concealment of group allocation was low, concealment was designated as "inadequate" and thus the study was placed in the "high risk of bias" category by Cooney et al. (2013, p. 79).

Second, Cooney et al. (2013) characterized the risk of selection bias (i.e., from violation of allocation concealment) as "unclear" in cases in which the reports provided no pertinent information and authors either did not respond or could not provide this information. However, for unspecified reasons, this rule was not applied to those studies that were included in the earlier meta-analysis by Lawlor and Hopker (2001). In these 10 cases, Cooney et al. (2013) carried over the risk assessments made by Lawlor and Hopker (2001). Lawlor and Hopker (2001) reported that they categorized studies as having (a) adequate concealment (i.e., use of the CONSORT catchphrases), (b) inadequate concealment (i.e., open list or

Table 1
Example of the plasticity of meta-analytic evidence summarized by Cooney et al. (2014), focusing on "high-quality" trials.

Study	SMD as entered	Comment	SMD, scenario 1	SMD, scenario 2
Blumenthal et al., 1999	Weight: 19.6% SMD: 0.14 [−0.25, 0.52]	The study did not include a control group. Cooney et al. (2013) constructed an "exercise versus no treatment" comparison by entering the exercise-plus-medication group as "exercise" and the "medication-alone" group as "control." The study is a comparative effectiveness trial, not a treatment-control trial, and, as such, it should be excluded.	Weight: 0.0% SMD: 0.14 [−0.25, 0.52]	Weight: 0.0% SMD: 0.14 [−0.25, 0.52]
Blumenthal et al., 2007	Weight: 19.3% SMD: −0.29 [−0.68, 0.11]	This trial compared a supervised exercise group to a placebo control group. No changes proposed.	Weight: 39.4% SMD: −0.29 [−0.68, 0.11]	Weight: 20.0% SMD: −0.29 [−0.68, 0.11]
Blumenthal et al., 2012	Weight: 14.3% SMD: −0.67 [−1.23, −0.1]	Participants entered the study if they had "elevated score (≥7)" on the BDI-II, which is in the "none or minimal" range and only half of the cutoff for "mild" depressive symptoms (i.e., 14). Only 47% of randomized participants met diagnostic criteria for depression. The trial should have been excluded because the Cochrane review reportedly "excluded trials that randomized people both with and without depression" (Cooney et al., 2013, p. 9). Restricting the analysis to patients with depression at baseline lowers the sample size but raises the SMD. The trial should either be excluded or the SMD for the subsample of patients with depression should be used.	Weight: 0.0% SMD: −0.94 [−1.82, −0.07]	Weight: 7.0% SMD: −0.94 [−1.82, −0.07]
Dunn et al., 2005.	Weight: 9.9% SMD: −0.74 [−1.50, 0.02]	The trial included an "exercise placebo control group," the members of which engaged in 3 days/week of stretching and flexibility exercise for 15–20 min per session. Thus, the trial should have been excluded since the Cochrane review reportedly "excluded studies comparing two different types of exercise with no non-exercising comparison group" (Cooney et al., 2013, p. 10). Moreover, Cooney et al. (2013) erroneously designated one group (public health dose over 5 days/week) as having received the "biggest dose" of exercise. In actuality, the "dose" was the same as in another group (public health dose over 3 days/week), which yielded a higher SMD (−1.16 versus −0.74). The trial should either be excluded for having an exercise comparator or the "biggest dose" with the optimal frequency (3 days/week) should be used.	Weight: 0.0% SMD: −1.16 [−1.94, −0.37]	Weight: 8.3% SMD: −1.16 [−1.94, −0.37]
Krogh et al., 2009	Weight: 18.6% SMD: 0.25 [−0.17, 0.66]	The trial included a so-called "relaxation" group which engaged in "exercises on mattresses or Bobath Balls," "light balance exercises," and "relaxation exercises with alternating muscle contraction and relaxation in different muscle groups" up to a Rating of Perceived Exertion of 12, which is within the range recommended by the ACSM for the improvement of cardiorespiratory fitness. Indeed aerobic fitness in this group increased by 6% and muscular strength by 10–17%. Moreover, contrary to the decision to use the arms providing "the biggest 'dose' of exercise" (Cooney et al., 2013, p. 12), Cooney et al. designated the "aerobic" exercise group, as opposed to the "circuit training" exercise group, as "exercise." However, the latter yielded a much stronger overall training effect (+8% in aerobic fitness, +25–29% in strength) than the former (+11% in aerobic fitness, +3–10% in strength). This changed the SMD from −0.10 in favor of "exercise" to 0.25 in favor of "control." The trial must either be excluded for not using a "non-exercising comparison group," as per the Cooney et al. (2013) criteria, or the group receiving the "biggest dose" of exercise should be used.	Weight: 0.0% SMD: −0.10 [−0.51, 0.32]	Weight: 19.0% SMD: −0.10 [−0.51, 0.32]
Mather et al., 2002	Weight: 18.3% SMD: −0.17 [−0.59, 0.26]	The trial involved older adults (≥53 years) who had been on a therapeutic dose of antidepressants for at least 6 weeks with no evidence of response. There was an exercise group (45 min of endurance, strengthening, stretching, twice per week for 10 weeks) and a no-exercise control (health education). No changes proposed.	Weight: 34.1% SMD: −0.17 [−0.59, 0.26]	Weight: 18.7% SMD: −0.17 [−0.59, 0.26]
Knubben et al., 2007	Weight: 0.0% SMD: −0.83 [−1.49, −0.16]	This trial was not considered of high quality because of a minor discrepancy in the reporting (according to the text, 39 patients fulfilled the inclusion criteria; 38 are cited in the table and abstract). However, the authors state that "the study was carried out following the 'intention to treat' principle" with missing values imputed "using the worst rank assumption" (p. 31).	Weight: 0.0% SMD: −0.83 [−1.49, −0.16]	Weight: 10.7% SMD: −0.83 [−1.49, −0.16]
Lavretsky et al., 2011	Weight: 0.0% SMD: −0.40 [−0.88, 0.08]	This trial was excluded by Cooney et al. (2013) reportedly because "control is health education, an active intervention" (Cooney et al., 2013, p. 96). This exclusion criterion is not appropriate since the review contains other trials with health education control groups (Mather et al., 2002; Singh et al., 1997). Cooney et al. (2013) also excluded all trials of tai chi, qigong, and yoga for not satisfying the criteria for "exercise" (i.e., planning, structure, repetitive bodily movement, purpose). The Tai Chi treatment of Lavretsky et al. (2011) cannot be excluded on this basis. The treatment was labeled a "health management intervention" designed to help older adults "cope with fatigue [and] perceived physical limitations" (p. 841). It was described as "physical activity" consisting of "repetitious" movement and had structure (120 min, including 10 min of warm-up, 5 min of cool-down exercises, p. 842). The trial examined older adults (>60 years) with a current episode of depression who had not achieved remission after 6 weeks of treatment with a therapeutic dose of escitalopram.	Weight: 26.5% SMD: −0.40 [−0.88, 0.08]	Weight: 16.3% SMD: −0.40 [−0.88, 0.08]
Pooled SMD [95% CI]	Weight: 100.0% SMD: −0.18 [−0.47, 0.11]		Weight: 100.0% SMD: −0.28 [−0.52, −0.03]	Weight: 100.0% SMD: −0.42 [−0.68, −0.17]
Heterogeneity	$\tau^2 = 0.07$ χ^2 (5) = 11.76 $p = 0.04$ $I^2 = 57\%$		$\tau^2 = 0.00$ χ^2 (2) = 0.50 $p = 0.78$ $I^2 = 0\%$	$\tau^2 = 0.05$ χ^2 (6) = 9.96 $p = 0.13$ $I^2 = 40\%$

tables of random numbers; open computer systems; drawing of non-opaque envelopes), or (c) unclear concealment (no information in the report and the authors either did not respond or could not provide information). However, as can be seen in their Table 1 (p. 4), Lawlor and Hopker (2001) did not adhere to this rule and, instead, summarily labeled all studies that did not contain information on the methods of concealment as having "no" concealment. Since these assessments were carried over to the Cochrane review, this resulted in the interesting phenomenon of all 10 of the studies listed by Cooney et al. (2013) as having "high risk" of selection bias having come from Lawlor and Hopker (2001). However, as explained, these would have been labeled as having "unclear" risk if Cooney et al. (2013) had applied their own criteria (i.e., lack of specific information in the report). In other words, the distinction between "unclear risk" and "high risk" of selection bias in Cooney et al. (2013) does not reflect an actual assessment of risk but rather whether the study was assessed by Cooney et al. (2013) or by Lawlor and Hopker (2001).

In at least one pre-2000 case, contact with a researcher (Blumenthal et al., 1999) revealed that the categorization of allocation concealment as "inadequate" by Lawlor and Hopker (2001) was erroneous. As acknowledged by Cooney et al. (2013), "further information from the author has enabled us to change this to low risk" (p. 52). What remains unknown is for how many additional cases the initial "high risk" designation by Lawlor and Hopker (2001) was erroneous.

6.2. Intention-to-treat analysis

Cooney et al. (2013) reported that "when [they] could not obtain information either from the publication or from the authors, [they] classified the trial as 'not intention-to-treat'" (p. 13). However, it should be evident that automatically assuming the worst-case scenario without evidence can be a source of bias. For example, even though Cooney et al. (2013) cited the published journal article by Chu et al. (2009), which stated that "intent-to-treat analysis of all randomized participants was conducted" and "missing data were imputed by carrying forward the last recorded observation" (p. 39), they based their quality assessment on Chu's earlier unpublished dissertation, in which the analysis was indeed not by "intention to treat." Thus, the study was counted as "not intention-to-treat" (p. 57) and, therefore, "high risk." Interestingly, the intention-to-treat analysis slightly strengthened the effect size associated with that study in favor of exercise (−1.13 from −1.02) while also slightly increasing its weight (3.0% from 2.7%).

There were additional errors. For example, of two cases in which Cooney et al. (2013) used individual-level data to calculate means and standard deviations after carrying the last observations forward, they counted one study (Orth, 1979) as "intention to treat" but not the other (Bonnet, 2005). In another case (Martinsen, Medhus, & Sandvik, 1985), even though the authors stated that "for patients who stopped treatment between weeks 6 and 9 their score at week 9 was taken as being the same as their score at week 6" (p. 109), Cooney et al. (2013), determined that the "analysis [was] not intention-to-treat" (p. 74).

Finally, Cooney et al. (2013) seem to have penalized researchers for a probable typographical error and, as a result, did not consider a study with a large effect in favor of exercise (SMD −0.83, 95% CI from −1.49 to −0.16) as being of high quality. Knubben et al. (2007) wrote that (a) randomization was "on the basis of a computer-generated number list" and the "study collaborators contacted the randomization center by telephone" to allocate each participant (p. 30); (b) "all patients were rated by the same investigator, who was unaware of the participants' group assignment" (pp. 30–31); and (c) "the study was carried out following the 'intention to treat' principle" with missing values imputed "using the worst rank assumption" (p. 31). However, Knubben et al. (2007) mentioned that "39 of 45 patients who fulfilled the inclusion criteria agreed to participate in the study and were recruited (Table 1)" (p. 30), whereas their Table 1 and abstract state 38. Because of this one-person inconsistency, the study was not considered by Cooney et al. (2013) as "intention-to-treat" despite the explicit statements by the authors in the article that the analysis had followed the "intention-to-treat" principle.[2]

6.3. Blinding

Regarding the rules used to determine whether outcome assessments were "blind," Cooney et al. (2013) made contradictory statements. Initially, they stated that "in exercise trials, participants cannot be blind to the treatment allocation" but they "were uncertain what effect this would have on bias" (p. 21). Later, however, they changed their position that the inability to blind participants to treatment allocation has *uncertain* effects on bias, stating instead that this problem can only *overestimate* the treatment effect. Thus, in their discussion, Cooney et al. (2013) argued that, because "it is generally not possible to blind participants or those delivering the intervention to the treatment allocation," this entails that "if the primary outcome is measured by self-report, this is an important potential source of bias" that "may lead to an *overestimate* of treatment effect sizes" (p. 34, italics added). Based on this assertion, they summarily designated all studies in which the primary outcome measure was a depression questionnaire (i.e., the majority of studies) as entailing "high risk of bias."

Although this mechanism of bias is possible, indiscriminately condemning all studies in which depression was assessed by questionnaire to the "high-risk" category appears unjustified. While clinician-administered standardized interviews base part of their scoring on expert behavioral observations, the bulk of the score still depends on self-reports of symptoms provided by the respondents themselves. Since there is no way around the problem of research participants being aware of whether or not they were exercising (just like there is no way to blind participants to whether or not they received psychotherapy), the blinding of the person administering the outcome measure (interview or questionnaire) is arguably a more relevant possible determinant of bias. This, however, was not considered.

As one example, for the study by Mutrie (1986), Cooney et al. (2013) decided that the "outcome assessment [was] not blind" (p. 79) because the outcome measure was a questionnaire. However, Mutrie took steps to ensure that the self-report of participants would not be affected in any way by external circumstances: (a) "all questionnaires were completed in private and put in sealed envelopes which were returned to the investigator"; (b) "these questionnaires were not scored until the eight-week exercise program had been completed"; and (c) "the data were then confidentially stored in accordance with The Pennsylvania State University guidelines regarding the protection of human subjects" such that "there was no possibility of the consultants being aware of subjects' scores" (Mutrie, 1986, p. 71). It is unclear how these safeguards result in "high" bias whereas a clinical interview by a blinded assessor automatically entails "low" bias.

It should also be noted that some studies used both clinician-administered interviews and questionnaires to assess depression. Using multiple methods of assessing the outcome is common practice in RCTs, as it represents an acknowledgment that each method (e.g., interview, questionnaire) has relative advantages and

[2] The corresponding author did not respond to a request for clarification.

disadvantages and can, therefore, provide non-redundant information. This is an important point for a patient-reported outcome such as depression for which there is no gold-standard blood assay or imaging test. However, in the pre-CONSORT era, outcome measures were often listed in arbitrary order (even in alphabetical order), as it was uncommon to designate one outcome measure as "primary" and others as "secondary." Other meta-analysts get around this problem by averaging the effect sizes derived from different measures of the same outcome and, thus, presumably calculating a more robust estimate of the overall effect (e.g., Cuijpers, Turner, et al., 2014, see p. 688). Instead, Cooney et al. (2013) decided that, if the designation of a measure as "primary" was missing, they would assign this designation to either (a) the "outcome reported in the abstract" or (b) the "first outcome reported in the Results section" (p. 12).

As a result of choosing this strategy, in the study by Singh et al. (1997), in which both the Beck Depression Inventory and the Hamilton Rating Scale for Depression were designated "primary outcomes" (p. M30), Cooney et al. (2013) selected the Beck Depression Inventory (because it happened to be listed first) and thus automatically designated the study as "high risk" (p. 89). Had they selected the Hamilton Rating Scale, the study would have been designated "low risk" since, according to Singh et al. (1997), "all outcome measures … were performed by a blinded assessor" (p. M28). To illustrate the frivolity of this type of quality assessments, in a follow-up study, Singh et al. (2005) similarly labeled both the Hamilton Rating Scale for Depression and the Geriatric Depression Scale (a questionnaire) as "primary outcomes" but, for whatever reason, happened to list the Hamilton first in the Methods (they later reversed the order in the Results). In conjunction with the fact that "a blinded psychiatrist performed all outcome measures" (p. 769), the study was designated as "low risk" by Cooney et al. (2013, p. 90). Likewise, the study by Doyne et al. (1987), in which the Hamilton Rating Scale for Depression was administered by raters who "were not informed of subjects' condition assignments" (p. 749), was placed in the "outcome assessment not blind" and, therefore, "high risk" category by Cooney et al. (2013, p. 59) because they designated the Beck Depression Inventory as the "primary outcome" (since it happened to be listed first).

7. Beware of errors, especially where it counts the most

Readers commonly assume that, perhaps as a function of the good reputation of the overseeing organization (such as the Cochrane Collaboration) or the scientific prestige of the journal (such as *JAMA*), the data reported in systematic reviews and meta-analyses have undergone several layers of rigorous peer review and can thus be trusted without the need for independent verification. In actuality, this romanticized view of the stringency of the peer review system frequently proves fallacious. As unnecessarily laborious as it seems, it behooves the readers to verify the integrity of the reported data.

Arguably the most crucial piece of information in understanding and evaluating the results of the Cooney et al. (2013) meta-analysis was Analysis 5.5 (pp. 137–138), which examined the effects of different types of comparators. Out of 23 analyses reported in the review, however, the table presenting the results of Analysis 5.5 was wrong (it displayed mean differences instead of standardized mean differences, despite the use of different outcome measures). Researchers interested in evaluating the statement that the substantial heterogeneity in the main analysis "might be explained by a number of factors including variation in the control intervention" (p. 34) are thus precluded from doing so. It is unfortunate that the amount of work required to perform the analysis independently would likely discourage most readers.

Likewise, the most eye-catching and clinically interesting piece of information in the *JAMA* Clinical Evidence Synopsis (Cooney et al., 2014) was the figure showing the effect sizes associated with the different studies using the Beck Depression Inventory. However, the figure was also wrong (the order in which the studies were listed was the upside-down mirror-image of the order in which the effect sizes were listed, such that none matched). Thus, most readers would likely become frustrated trying to make sense of the data.

Making matters worse, although the figure is supposed to display the studies in which the Beck Depression Inventory was the primary outcome measure, Cooney et al. (2014) included the study by Blumenthal et al. (1999), in which the primary outcome measure was the Hamilton Rating Scale for Depression (the Beck Depression Inventory was also used, as a secondary outcome, but Cooney et al. used the data from the Hamilton Rating Scale). Because the study by Blumenthal et al. (a) carried a large weight in determining the pooled SMD (8.9%) and (b) as explained earlier, the study did not have a control group (it involved a comparison between exercise-plus-medication versus medication-alone), its near-zero mean difference (0.92 units) had a considerable attenuating influence on the overall effect. Removal of that one study that was incorrectly included in the tally would have changed the mean difference from below 5 points (−4.76, 95% CI from −6.99 to −2.53) to over 5 points (−5.34, 95% CI from −7.50 to −3.19) while reducing heterogeneity from $I^2 = 74\%$ [$\tau^2 = 12.71$; $\chi^2(15) = 58.08$, $p < 0.00001$] to $I^2 = 66\%$ [$\tau^2 = 9.96$; $\chi^2(14) = 41.60$, $p = 0.0001$]. This is an important difference since 5 points on the Beck Depression Inventory is a commonly considered criterion of clinical efficacy.

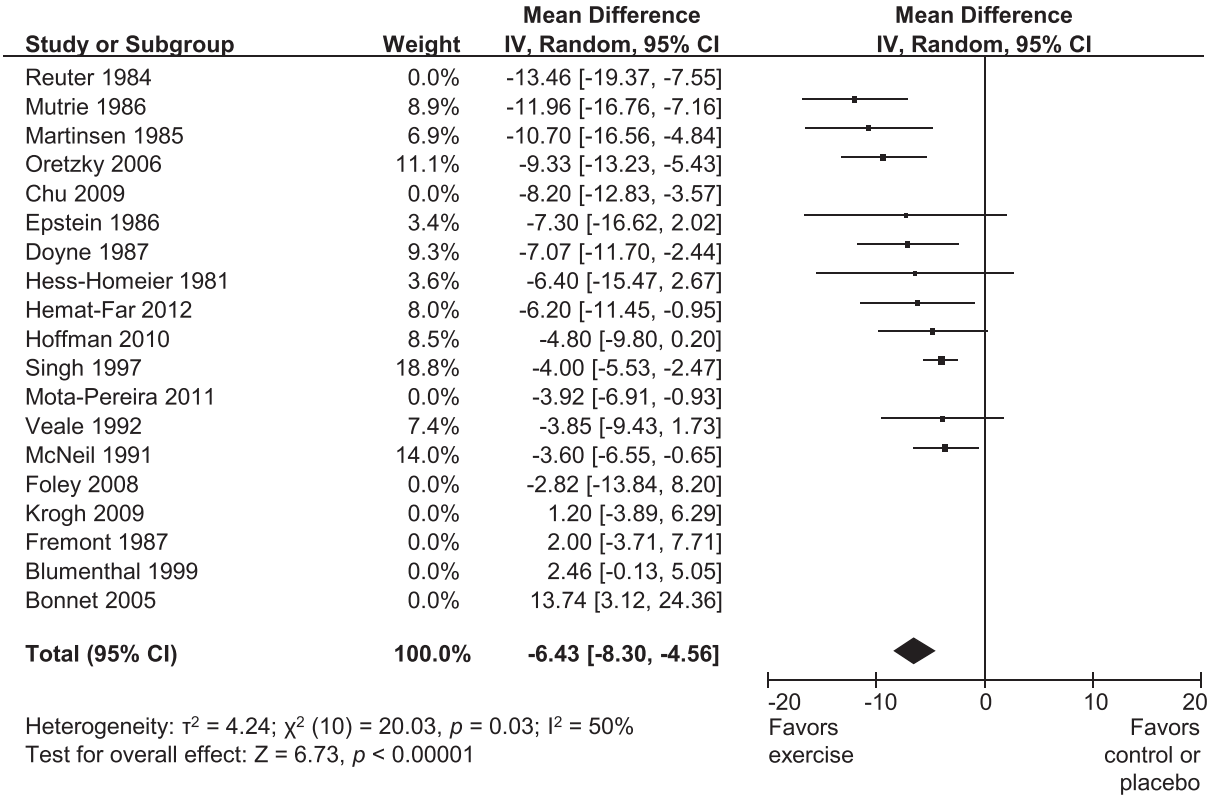
8. Shaking the magic picture: what is the effect of exercise on depression?

The present analysis illustrated that, although systematic reviews and meta-analyses are commonly assumed to be structured, rigorous, and based on uniform application of rules, in actuality they involve numerous decisions that allow ample flexibility. In turn, this flexibility creates substantial potential for bias. Using the Cochrane review on exercise for depression (Cooney et al., 2013) as an example, it was shown that changing some of these decisions based on well supported arguments can crucially alter the essential conclusions of the review. Specifically, the following changes are proposed: (a) studies without a control group should be excluded; (b) studies with active treatments as comparators (e.g., relaxation, meditation, stress management) should be excluded or considered separately; (c) studies with exercising groups as comparators (e.g., stretching and toning, yoga) should be excluded; (d) studies of postnatal depression should be included since there is no scientific basis for their exclusion; and (e) studies of tai-chi, qigong, and yoga should be included as long as they satisfy the ACSM definition of "exercise" (i.e., they consist of planned, structured, and repetitive bodily movement done to improve and/or maintain one or more components of physical fitness, including flexibility, agility, coordination, and balance).

By applying these rules to the database of Cooney et al. (2013), the pooled SMD is raised from "medium" (−0.62, 95% CI from −0.81 to −0.42) to "large" (−0.90, 95% CI from −1.11 to −0.69). Even after removing the two studies with the strongest effects in favor of exercise (i.e., Ciocon & Galindo-Ciocon, 2003, with SMD −2.49, 95% CI from −3.24 to −1.74; and Mutrie, 1986, with SMD −2.39, 95% CI from −3.76 to −1.02), the pooled SMD remains "large" (−0.80, 95% CI from −0.98 to −0.62) and heterogeneity is $\tau^2 = 0.12$; $\chi^2(28) = 58.05$, $p = 0.0007$; $I^2 = 52\%$. Examination of the mean differences from those studies that used the Hamilton Rating Scale for

Depression and the Beck Depression inventory (see Fig. 3), as either primary or secondary outcome, shows that the effects exceed commonly used criteria of clinical efficacy (i.e., 3 and 5 points, respectively).

The conclusion by Cooney et al. (2013, 2014) that the analysis of high-quality trials shows only a "small" and statistically non-significant effect of exercise must also be questioned since it relies on questionable inclusion and exclusion criteria. Table 1



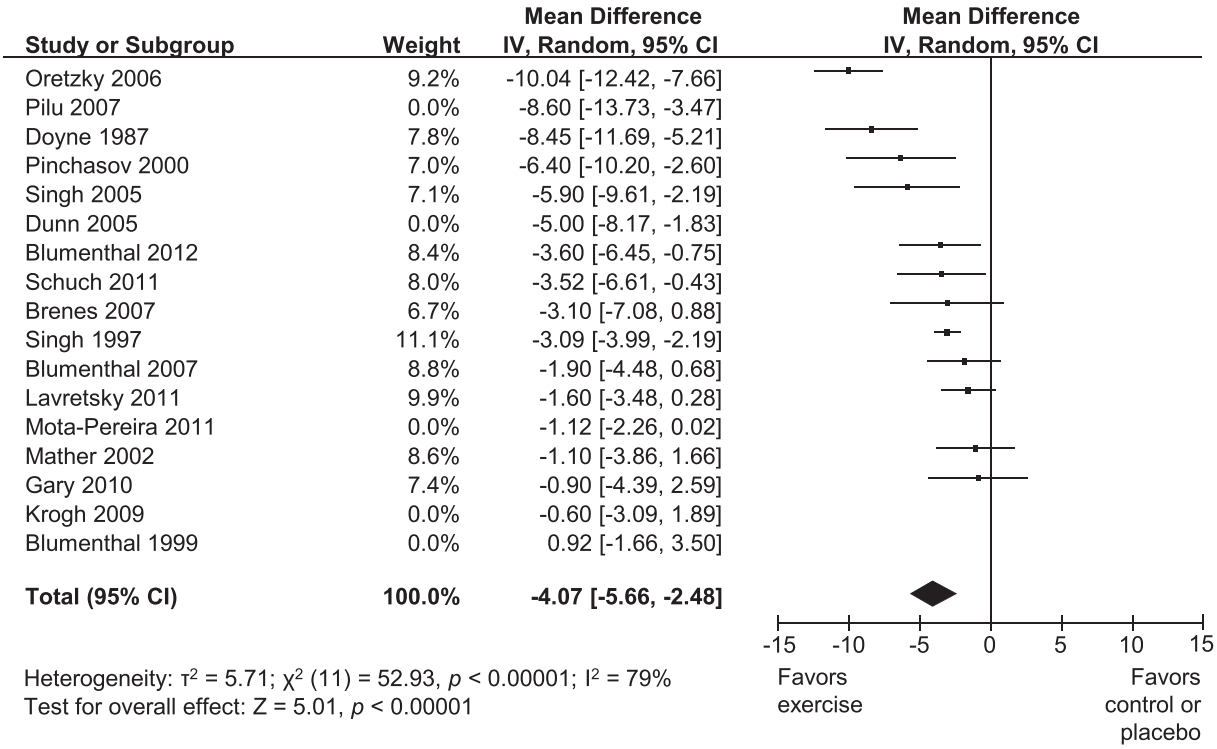


Fig. 3. Mean differences associated with exercise interventions in the Beck Depression Inventory (top panel) and Hamilton Rating Scale for Depression (bottom panel), used as either primary or secondary outcome measures. The pooled mean differences exceed commonly considered thresholds for demonstrating clinical efficacy, namely 5 and 3 points, respectively. Studies whose effects do not appear in the forest plots were excluded for having inappropriate comparators, as explained in the text.

examines the robustness of this conclusion under two scenarios. Under the more stringent scenario 1: (a) the Blumenthal et al. (1999) study must be excluded since it did not include a control group; (b) the Blumenthal et al. (2012) study must be excluded because the sample consisted of individuals with and without depression; (c) the Dunn et al. (2005) and Krogh et al. (2009) studies must both be excluded because they did not have a non-exercising comparison group; (d) the Lavretsky et al. (2011) study must be included as the intervention satisfied the criteria for "exercise." Under this scenario, the effect size is between "small" and "medium" and significantly different from zero (pooled SMD −0.28, 95% CI from −0.52 to −0.03).

Scenario 2 retains more studies but corrects certain methodological choices by Cooney et al. (2013) that were erroneous or unjustified: (a) it excludes Blumenthal et al. (1999) for not having a control group; (b) it includes Lavretsky et al. (2011) since there was no reason to exclude it; (c) it changes the effect size associated with the study by Dunn et al. (2005), since the effect selected by Cooney et al. (2013) was based on the false premise that the "public health dose over 5 days per week" represented a "bigger dose" of exercise compared to "public health dose over 3 days per week"; (d) it considers the circuit training group from Krogh et al. (2009), instead of the "aerobic exercise" group, as the group that received the "biggest dose" of exercise based on overall fitness gains; (e) it considers the effect of exercise among those participants in Blumenthal et al. (2012) who had depression at baseline; and (f) it includes Knubben et al. (2007) as a high-quality trial, accepting the statement of the authors that the analysis was done by intention-to-treat. This analysis also yields an effect size between "small" and "medium" that is significantly different from zero (pooled SMD −0.42, 95% CI from −0.68 to −0.17).

Limiting the analysis to only the two high-quality trials with pill placebo controls (using the fixed-effects model) yields a pooled SMD of −0.42 (95% CI from −0.74 to −0.09) when including the entire sample from Blumenthal et al. (2012) and −0.40 (95% CI from −0.76 to −0.04) when considering only the patients with depression diagnosis at baseline. Although the number of studies with pill placebo control groups is still very small, it should be pointed out that these preliminary figures are considerably higher than those from the 10 placebo-controlled trials of psychotherapy (pooled SMD −0.25; Cuijpers, Turner, et al., 2014). They also surpass the average effect size from published (pooled SMD −0.37) and unpublished (pooled SMD −0.15) placebo-controlled trials of antidepressant drugs (Turner, Matthews, Linardatos, Tell, & Rosenthal, 2008). Even authors with multiple disclosed ties to the pharmaceutical industry concede that, at best, the average effect of antidepressants compared to placebo does not exceed |0.32| to |0.34| (Fountoulakis, Veroniki, Siamouli, & Möller, 2013).

In closing, it could be argued that, while the clinical value of the systematic review and meta-analysis by Cooney et al. (2013, 2014) is questionable, its educational value is undeniable. The review represents an excellent example of how methodological decisions by researchers can crucially alter the outcome. The main lesson for clinicians, students, referees, editors, systematic reviewers, guideline developers, and policymakers is that the mechanisms that can alter the outcome are not necessarily complex, esoteric, or outside the capabilities of individuals equipped with at least a modicum of knowledge and analytic skill. Therefore, perhaps the most essential prerequisite for effective critical appraisal is the realization that published conclusions, no matter how definitively stated and regardless of the prestige of the journal in which they appear, must always be carefully scrutinized rather than passively accepted.

References

American College of Sports Medicine. (2013). *ACSM's guidelines for exercise testing and prescription* (9th ed.). Philadelphia, PA: Lippincott Williams & Wilkins.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

Armstrong, K., & Edwards, H. (2003). The effects of exercise and social support on mothers reporting depressive symptoms: a pilot randomized controlled trial. *International Journal of Mental Health Nursing, 12*(2), 130–138.

Armstrong, K., & Edwards, H. (2004). The effectiveness of a pram-walking exercise programme in reducing depressive symptomatology for postnatal women. *International Journal of Nursing Practice, 10*(4), 177–194.

Babyak, M., Blumenthal, J. A., Herman, S., Khatri, P., Doraiswamy, M., Moore, K., et al. (2000). Exercise treatment for major depression: maintenance of therapeutic benefit at 10 months. *Psychosomatic Medicine, 62*(5), 633–638.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., et al. (1996). Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *Journal of the American Medical Association, 276*(8), 637–639.

Bjørnebekk, A., Mathé, A. A., & Brené, S. (2010). The antidepressant effects of running and escitalopram are associated with levels of hippocampal NPY and Y1 receptor but not cell proliferation in a rat model of depression. *Hippocampus, 20*(7), 820–828.

Blumenthal, J. A., Babyak, M. A., Doraiswamy, P. M., Watkins, L., Hoffman, B. M., Barbour, K. A., et al. (2007). Exercise and pharmacotherapy in the treatment of major depressive disorder. *Psychosomatic Medicine, 69*(7), 587–596.

Blumenthal, J. A., Babyak, M. A., Moore, K. A., Craighead, W. E., Herman, S., Khatri, P., et al. (1999). Effects of exercise training on older patients with major depression. *Archives of Internal Medicine, 159*(19), 2349–2356.

Blumenthal, J. A., Sherwood, A., Babyak, M. A., Watkins, L. L., Smith, P. J., Hoffman, B. M., et al. (2012). Exercise and pharmacological treatment of depressive symptoms in patients with coronary heart disease: results from the UPBEAT (understanding the prognostic benefits of exercise and antidepressant therapy) study. *Journal of the American College of Cardiology, 60*(12), 1053–1063.

Bonnet, L. H. (2005). *Effects of aerobic exercise in combination with cognitive therapy on self-reported depression* (Unpublished doctoral dissertation). Hempstead, NY: Hofstra University.

Bosscher, R. J. (1993). Running and mixed physical exercises with depressed psychiatric patients. *International Journal of Sport Psychology, 24*, 170–184.

Bridle, C., Spanjers, K., Patel, S., Atherton, N. M., & Lamb, S. E. (2012). Effect of exercise on depression severity in older people: systematic review and meta-analysis of randomised controlled trials. *British Journal of Psychiatry, 201*(3), 180–185.

Chu, I.-H., Buckworth, J., Kirby, T. E., & Emery, C. F. (2009). Effect of exercise intensity on depressive symptoms in women. *Mental Health and Physical Activity, 2*(1), 37–43.

Ciocon, J. O., & Galindo-Ciocon, D. (2003, August 19). *Loneliness and depression in nursing home setting: The effect of a restorative program*. Paper presented at the 11th International Congress of the International Psychogeriatric Association. Chicago, IL.

Cooney, G. M., Dwan, K., Greig, C. A., Lawlor, D. A., Rimer, J., Waugh, F. R., et al. (2013). Exercise for depression. *Cochrane Database of Systematic Reviews, 9*, CD004366.

Cooney, G., Dwan, K., & Mead, G. (2014). Exercise for depression. *Journal of the American Medical Association, 311*(23), 2432–2433.

Cuijpers, P., Sijbrandij, M., Koole, S. L., Andersson, G., Beekman, A. T., & Reynolds, C. F., 3rd (2014). Adding psychotherapy to antidepressant medication in depression and anxiety disorders: a meta-analysis. *World Psychiatry, 13*(1), 56–67.

Cuijpers, P., Turner, E. H., Mohr, D. C., Hofmann, S. G., Andersson, G., Berking, M., et al. (2014). Comparison of psychotherapies for adult depression to pill placebo control groups: a meta-analysis. *Psychological Medicine, 44*(4), 685–695.

Danielsson, L., Noras, A. M., Waern, M., & Carlsson, J. (2013). Exercise in the treatment of major depression: a systematic review grading the quality of evidence. *Physiotherapy Theory and Practice, 29*(8), 573–585.

Deeks, J. J., Higgins, J. P. T., & Altman, D. G. (2008). Analysing data and undertaking meta-analyses. In J. P. T. Higgins, & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions* (pp. 243–296). Hoboken, NJ: John Wiley & Sons.

Devereaux, P. J., Choi, P. T., El-Dika, S., Bhandari, M., Montori, V. M., Schünemann, H. J., et al. (2004). An observational study found that authors of randomized controlled trials frequently use concealment of randomization and blinding, despite the failure to report these methods. *Journal of Clinical Epidemiology, 57*(12), 1232–1236.

Doyne, E. J., Ossip-Klein, D. J., Bowman, E. D., Osborn, K. M., McDougall-Wilson, I. B., & Neimeyer, R. A. (1987). Running versus weight lifting in the treatment of depression. *Journal of Consulting and Clinical Psychology, 55*(5), 748–754.

Dunn, A. L., Trivedi, M. H., Kampert, J. B., Clark, C. G., & Chambliss, H. O. (2005). Exercise treatment for depression: efficacy and dose response. *American Journal of Preventive Medicine, 28*(1), 1–8.

Ellard, D. R., Thorogood, M., Underwood, M., Seale, C., & Taylor, S. J. (2014). Whole home exercise intervention for depression in older care home residents (the OPERA study): a process evaluation. *BMC Medicine, 12*, 1.

Foley, L. S., Prapavessis, H., Osuch, E. A., De Pace, J. A., Murphy, B. A., & Podolinsky, N. J. (2008). An examination of potential mechanisms for exercise as

a treatment for depression: a pilot study. *Mental Health and Physical Activity, 1*(1), 69–73.
Fountoulakis, K., Veroniki, A. A., Siamouli, M., & Möller, H. J. (2013). No role for initial severity on the efficacy of antidepressants: results of a multi-meta-analysis. *Annals of General Psychiatry, 12*(1), 26.
Fremont, J., & Craighead, L. W. (1987). Aerobic exercise and cognitive therapy in the treatment of dysphoric moods. *Cognitive Therapy and Research, 11*(2), 241–251.
Green, S., & Higgins, J. P. T. (2008). Preparing a Cochrane review. In J. P. T. Higgins, & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions* (pp. 11–30). Hoboken, NJ: John Wiley & Sons.
Higgins, J. P. T., Deeks, J. J., & Altman, D. G. (2008). Special topics in statistics. In J. P. T. Higgins, & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions* (pp. 481–529). Hoboken, NJ: John Wiley & Sons.
Ioannidis, J. P. A. (2010). Meta-research: the art of getting it wrong. *Research Synthesis Methods, 1*(3–4), 169–184.
Josefsson, T., Lindwall, M., & Archer, T. (2014). Physical exercise intervention in depressive disorders: meta-analysis and systematic review. *Scandinavian Journal of Medicine and Science in Sports, 24*(2), 259–272.
Kerse, N., Hayman, K. J., Moyes, S. A., Peri, K., Robinson, E., Dowell, A., et al. (2010). Home-based activity program for older people with depressive symptoms, DeLLITE: a randomized controlled trial. *Annals of Family Medicine, 8*(3), 214–223.
Klein, M. H., Greist, J. H., Gurman, A. S., Neimeyer, R., Lesser, D. P., Bushnell, N. J., et al. (1985). A comparative outcome study of group psychotherapy vs. exercise treatments for depression. *International Journal of Mental Health, 13*(3–4), 148–176.
Kmietowicz, Z. (2013). Evidence that exercise helps in depression is still weak, finds review. *British Medical Journal, 347*, f5585.
Knubben, K., Reischies, F. M., Adli, M., Schlattmann, P., Bauer, M., & Dimeo, F. (2007). A randomised, controlled study on the effects of a short-term endurance training programme in patients with major depression. *British Journal of Sports Medicine, 41*(1), 29–33.
Krogh, J., Nordentoft, M., Sterne, J. A., & Lawlor, D. A. (2011). The effect of exercise in clinically depressed adults: systematic review and meta-analysis of randomized controlled trials. *Journal of Clinical Psychiatry, 72*(4), 529–538.
Krogh, J., Saltin, B., Gluud, C., & Nordentoft, M. (2009). 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. *Journal of Clinical Psychiatry, 70*(6), 790–800.
Krogh, J., Videbech, P., Thomsen, C., Gluud, C., & Nordentoft, M. (2012). DEMO-II trial. Aerobic exercise versus stretching exercise in patients with major depression: a randomised clinical trial. *PLoS One, 7*(10), e48316.
Lavretsky, H., Alstein, L. L., Olmstead, R. E., Ercoli, L. M., Riparetti-Brown, M., Cyr, N. S., et al. (2011). Complementary use of tai chi chih augments escitalopram treatment of geriatric depression: a randomized controlled trial. *American Journal of Geriatric Psychiatry, 19*(10), 839–850.
Lawlor, D. A., & Hopker, S. W. (2001). The effectiveness of exercise as an intervention in the management of depression: systematic review and meta-regression analysis of randomised controlled trials. *British Medical Journal, 322*(7289), 763–767.
MacGillivray, L., Reynolds, K. B., Rosebush, P. I., & Mazurek, M. F. (2012). The comparative effects of environmental enrichment with exercise and serotonin transporter blockade on serotonergic neurons in the dorsal raphe nucleus. *Synapse, 66*(5), 465–470.
Marlatt, M. W., Lucassen, P. J., & van Praag, H. (2010). Comparison of neurogenic effects of fluoxetine, duloxetine and running in mice. *Brain Research, 1341*, 93–99.
Martinsen, E. W., Medhus, A., & Sandvik, L. (1985). Effects of aerobic exercise on depression: a controlled study. *British Medical Journal, 291*(6488), 109.
Mather, A. S., Rodriguez, C., Guthrie, M. F., McHarg, A. M., Reid, I. C., & McMurdo, M. E. (2002). Effects of exercise on depressive symptoms in older adults with poorly responsive depressive disorder: randomised controlled trial. *British Journal of Psychiatry, 180*, 411–415.
Mead, G. E., Morley, W., Campbell, P., Greig, C. A., McMurdo, M., & Lawlor, D. A. (2009). Exercise for depression. *Cochrane Database of Systematic Reviews, 2008*(4), CD004366.
Moayyedi, P. (2004). Meta-analysis: can we mix apples and oranges? *American Journal of Gastroenterology, 99*(12), 2297–2301.
Moher, D., Jones, A., & Lepage, L. (2001). Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. *Journal of the American Medical Association, 285*(15), 1992–1995.
Murad, M. H., Montori, V. M., Ioannidis, J. P., Jaeschke, R., Devereaux, P. J., Prasad, K., et al. (2014). How to read a systematic review and meta-analysis and apply the results to patient care. *Journal of the American Medical Association, 312*(2), 171–179.
Mutrie, N. (1986). *Exercise as a treatment for depression within a national health service* (Unpublished doctoral dissertation). University Park, PA: Pennsylvania State University.
National Collaborating Centre for Mental Health and National Institute for Health and Clinical Excellence. (2010). *The treatment and management of depression in adults* (Updated ed.). Leicester and London: The British Psychological Society and The Royal College of Psychiatrists.
Orth, D. K. (1979). *Clinical treatments of depression* (Unpublished doctoral dissertation). Morgantown, WV: West Virginia University.
Pinchasov, B. B., Shurgaja, A. M., Grischin, O. V., & Putilov, A. A. (2000). Mood and energy regulation in seasonal and non-seasonal depression before and after midday treatment with physical exercise or bright light. *Psychiatry Research, 94*(1), 29–42.
Salehi, I., Hosseini, S. M., Haghighi, M., Jahangard, L., Bajoghli, H., Gerber, M., et al. (2014). Electroconvulsive therapy and aerobic exercise training increased BDNF and ameliorated depressive symptoms in patients suffering from treatment-resistant major depressive disorder. *Journal of Psychiatric Research, 57*, 117–124.
Searle, A., Calnan, M., Turner, K. M., Lawlor, D. A., Campbell, J., Chalder, M., et al. (2012). General practitioners' beliefs about physical activity for managing depression in primary care. *Mental Health and Physical Activity, 5*(1), 13–19.
Setaro, J. L. (1985). *Aerobic exercise and group counseling in the treatment of anxiety and depression* (Unpublished doctoral dissertation). College Park, MD: University of Maryland.
Shapiro, S. (1995). Systematic reviews. *Journal of the American Medical Association, 274*(8), 657–658.
Silveira, H., Moraes, H., Oliveira, N., Coutinho, E. S., Laks, J., & Deslandes, A. (2013). Physical exercise and clinically depressed patients: a systematic review and meta-analysis. *Neuropsychobiology, 67*(2), 61–68.
Singh, N. A., Clements, K. M., & Fiatarone, M. A. (1997). A randomized controlled trial of progressive resistance training in depressed elders. *Journal of Gerontology, 52*(1), M27–M35.
Singh, N. A., Stavrinos, T. M., Scarbek, Y., Galambos, G., Liber, C., & Fiatarone Singh, M. A. (2005). A randomized controlled trial of high versus low intensity weight training versus general practitioner care for clinical depression in older adults. *Journal of Gerontology, 60A*(6), 768–776.
Thombs, B. D., Ziegelstein, R. C., Pilote, L., Dozois, D. J., Beck, A. T., Dobson, K. S., et al. (2010). Somatic symptom overlap in Beck depression inventory II scores following myocardial infarction. *British Journal of Psychiatry, 197*(1), 61–66.
Trivedi, M. H., Greer, T. L., Church, T. S., Carmody, T. J., Grannemann, B. D., Galper, D. I., et al. (2011). Exercise as an augmentation treatment for non-remitted major depressive disorder: a randomized, parallel dose comparison. *Journal of Clinical Psychiatry, 72*(5), 677–684.
Trivedi, M. H., Greer, T. L., Grannemann, B. D., Chambliss, H. O., & Jordan, A. N. (2006). Exercise as an augmentation strategy for treatment of major depression. *Journal of Psychiatric Practice, 12*(4), 205–213.
Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A., & Rosenthal, R. (2008). Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine, 358*(3), 252–260.
Underwood, M., Lamb, S. E., Eldridge, S., Sheehan, B., Slowther, A. M., Spencer, A., et al. (2013). Exercise for depression in elderly residents of care homes: a cluster-randomised controlled trial. *Lancet, 382*(9886), 41–49.
Weber, M., Talmon, S., Schulze, I., Boeddinghaus, C., Gross, G., Schoemaker, H., et al. (2009). Running wheel activity is sensitive to acute treatment with selective inhibitors for either serotonin or norepinephrine reuptake. *Psychopharmacology, 203*(4), 753–762.
Williams, C. L., & Tappen, R. M. (2008). Exercise training for depressed older adults with Alzheimer's disease. *Aging and Mental Health, 12*(1), 72–80.
von Wolff, A., Hölzel, L. P., Westphal, A., Härter, M., & Kriston, L. (2012). Combination of pharmacotherapy and psychotherapy in the treatment of chronic depression: a systematic review and meta-analysis. *BMC Psychiatry, 12*, 61.