James Heathers, Director (j@medicalevidenceproject.org)

Medical Evidence Project

Exercise as a Treatment for Depression?

Forensic analysis of seven problematic studies commonly included in meta-analyses on the subject

Report Ex-Dep

medicalevidenceproject.org

Medical
Evidence
Project

Exercise as a Treatment for Depression? Forensic analysis of seven problematic studies commonly included in meta-analyses on the subject

Citation: Heathers, James. 2026. “Exercise as a treatment for depression? Forensic analysis of seven problematic studies commonly included in meta-analyses on the subject.” Available at medicalevidenceproject.org/exercise-depression.

The Medical Evidence Project thanks Matthew Jané for analysis and investigation contributing to this report, Andrew Vigotsky for validation, formal analysis, and manuscript review, and Mark Ebell and Henry Barry for manuscript review. More information on this report is available at medicalevidenceproject.org/exercise-depression.

This report is a product of the Medical Evidence Project, which aims to reduce harm to patients and improve outcomes by employing forensic meta-analysis techniques to find errors in influential medical articles and guidelines. This work is made possible by a grant from Coefficient Giving. Learn more about our work at medicalevidenceproject.org/about.

The Medical Evidence Project is an endeavor of The Center for Scientific Integrity, whose mission is to promote transparency and integrity in science and scientific publishing and to disseminate best practices and increase efficiency in science. Information about the center’s leadership and finances is available at centerforscientificintegrity.org.

Published by The Center for Scientific Integrity, 121 W. 36th St., Suite 209, New York, NY 10018.

Sections

1 INTRODUCTION

The Medical Evidence Project (medicalevidenceproject.org), an initiative of The Center for Scientific Integrity, seeks to improve the reliability of medical research by applying forensic metascience to clinical evidence. Our analysts identify statistical and methodological problems in influential published studies and meta-analyses and share the results. The ultimate goal of the Medical Evidence Project is to ensure that clinical guidelines are based on robust, verifiable evidence with the aim of reducing morbidity and mortality.

This Medical Evidence Project report evaluates a series of primary studies contained in a recently published Cochrane review: "Exercise for depression" by Clegg et al. (2026), the seventh Cochrane review on the topic since 2000. The sixth, Cooney et al. (2013), has to date garnered 644 citations according to Scopus, including 21 citations within global medical guidelines. These metrics along with significant news reporting on the new Clegg et al. review indicate a consistently high level of professional and lay interest in the question of whether exercise is an effective treatment for depression.

That interest is not surprising; the World Health Organization estimates that, globally, over one in 20 adults suffer from depression (World Health Organization, 2023). Notably for purposes of this report, a review of prominent guidelines and national policies by the Medical Evidence Project suggests that exercise has sometimes been framed as a viable treatment option for depression, including in some cases as an effective alternative to medications and psychological therapies.

In the United Kingdom, for instance, the NICE (National Institute for Health and Care Excellence) guidelines were updated in 2022 to advise to "not routinely offer antidepressant medication as first-line treatment for less severe depression", listing cognitive behavioral therapy (CBT) and exercise as preferred alternatives (National Institute for Health and Care Excellence (NICE), 2022). The Australian Medicare Benefits Schedule allows exercise as a treatment for depression to be reimbursed (Australian Government Department of Health and Aged Care, 2025).

The use of exercise as a treatment for depression also appears in influential sources of advice to physicians. For example, UpToDate, a clinical decision support tool widely used by physicians in the U.S., cites the previous Cochrane review Cooney et al. (2013) in its recommendation of exercise as an effective "adjunctive therapy or monotherapy" for individuals with major depressive disorder (UpToDate, 2026).

Yet Medical Evidence Project analysts have been increasingly concerned about the quality of the evidence supporting the idea that exercise is an effective treatment for depression. In fact, when Clegg et al. (2026) became public in early January 2026, our analysts were already investigating several trials in this research domain due to earlier concerns raised regarding the quality and integrity of published research in the field.

Clegg et al. (2026) aggregates 73 randomized controlled trials (RCTs), 69 of which contributed data to the assembled meta-analytic results. This report concentrates on the 57 trials (2189 participants) comparing exercise with no-treatment or control intervention, and the nine trials (405 participants) with long-term follow-up, as these provide the headline results for the meta-analysis as a whole.

Of these 57 studies, we present forensic analysis of seven studies, some of which were also included in earlier Cochrane reviews and clinical guidelines. Our analysis of those studies reveals critical issues ranging from internal inconsistencies, unlikely baseline data, mathematically impossible formulae, physiological contradictions, and, in one case, a study that reports to have administered a product described with the name of a toothpaste instead of an anti-depressant.

Clegg et al. (2026) cautiously concludes that "Exercise may be moderately more effective than a

control intervention for reducing symptoms of depression. Exercise appears to be no more or less effective than psychological or pharmacological treatments, though this conclusion is based on a few small trials. Long-term follow-up was rare."

Our finding of significant reasons to distrust at least seven of the included studies in the review further weakens those already tempered conclusions. This has real-world implications because, as noted above, some clinical guidelines explicitly recommend exercise as a treatment for depression based on the assumption the evidence in favor of it is strong. In reality, the actual impact of exercise on depression remains largely unknown.

2 METHODOLOGICAL INTRODUCTION TO FORENSIC ANALYSIS

Meta-analysis is the statistical technique of combining the results of multiple studies on the same topic in an attempt to obtain better answers to a research question, such as whether exercise is an effective treatment for depression. But the quality of a meta-analysis is, of course, only as good as the studies it aggregates. Responsible meta-analysts have addressed this by enforcing appropriate inclusion criteria, stratifying conclusions in terms of measurable bias, integrating measurements of heterogeneity, sensitivity, fragility, etc.

Nevertheless, as we elaborate in the discussion section that concludes this report, the medical research community now recognizes that these traditional procedures may be insufficient to ensure that meta-analyzed literature is trustworthy. Forensic metascience is beginning to provide aid in this regard by introducing an extra layer of scrutiny. Forensic metascience practitioners analyze the numerical, statistical, and methodological statements within published medical studies in an effort to find data or conclusions that may be mathematically impossible, garbled, nonsensical, mistaken, misunderstood, or fraudulent (e.g., Carlisle, 2012; Bolland et al., 2021). This work can help prevent problematic studies and reports from influencing patient care.

There are many previously published statistical observations and applications which underlie forensic metascience (Heathers, 2025). Those used in the making of this report include:

- **Standard consistency checks**. Many statistical tests are reported in scientific papers as the difference between groups. These groups have defined means, standard deviations, and sample sizes. In addition, a test statistic (e.g., a t, χ^2, or U value) and a p-value are usually reported. Many of these tests can simply be recalculated to check for accuracy of the study's presentation in terms of its own math.

- **The GRIM and GRIMMER tests**. "GRIM" stands for Granularity-Related Inconsistency of Means, and "GRIMMER" for "Granularity-Related Inconsistency of Means Mapped to Error Repeats." When a mean is calculated from a small sample of whole integers, mathematically impossible values (called 'GRIM inconsistencies') exist wherein the presented mean cannot possibly be calculated from the presented underlying data (Brown and Heathers, 2017). For example, in the use of the Beck Depression Inventory (BDI; Beck, Ward, et al., 1961), participants rate 21 items with symptom scores from 0 to 3, which gives each participant a score from 0 to 63. If the mean BDI score is tabulated from n = 5 participants, then the mean BDI must end with 0.0, 0.2, 0.4, 0.6, or 0.8; those are the only mathematical possibilities. If the mean is presented as 10.36, then we know that something is wrong in the presentation. The GRIMMER test is slightly more complicated, but essentially calculates the same properties for the standard deviation (Anaya, 2016, Allard, 2018). GRIMMER contains three distinct failures for the standard deviations listed in a paper: Type 1, if the sum of squares is not a whole integer; Type 2, if the standard deviation cannot be back-calculated, and Type 3, if the sum of a group of integers and their squares do not share the same parity (see Allard, 2018 for details).

- **The GRIM-U test**. The Medical Evidence Project recently presented the GRIM-U test, which extends the principle of the GRIM test into ranked data. In this test, we see that (1) U values and ranked values calculated by the test have absolute maximums and minimums, and (2) as with the GRIM test, there are a restricted number of possible outcomes. This again allows us to see whether presented summary data is mathematically impossible, an outcome that raises concerns about the quality of a published study.

- **Table 1 tests**. Medical studies usually include a description of research participants in a single comprehensive table of data that describes, for example, their demographic details, baseline blood tests, functional capacity, history of illness, etc. These presentations of data often include group and total data which can be compared to check for whether they make

mathematical sense. For instance, the mean of a Medication group plus the mean of a Control group adjusted for their relevant denominators will directly determine the mean of the total sample in all cases. If these numbers do not check out mathematically, the study's presentation is obviously problematic.

- **Within-subjects data.** Studies that include two or more measurements from the same group across different time points provide another opportunity for forensic metascience, as the correlations between Time 1 and Time 2 data can be calculated, and so this gives us another opportunity to look for red flags. Firstly, correlations between Time 1 and Time 2 mathematically must fall between -1 and 1. Furthermore, the correlations should realistically indicate the same person is being measured twice (i.e., we have strong reason to expect, in this type of research, a moderate and positive correlation). When we find results that do not show realistic correlations between time points, it is reasonable to be concerned about the quality of the study.
- A variety of other **methodological inconsistencies** may also send up warning signals. These may include, for example, where two sections of a paper describe the same data point, procedure, or analysis differently, or where details necessary to interpret or understand the required results are missing or incorrect.

All of the forms of analysis described above can be combined, modified, or mathematically extended. When this is the case in the analysis used for this report, the necessary derivation is given in this document. For sections requiring longer code to explain our methods, for those who want to learn more, there is an accompanying .ipynb document in R which reproduces the results from this report.

3 FORENSIC ANALYSIS OF INDIVIDUAL STUDIES

Here we present brief forensic metascientific examinations of a subset of studies included in Clegg et al. (2026), the recent Cochrane review on "Exercise for depression." The seven studies examined below are included because they have been identified by Medical Evidence Project analysts as showing concerning anomalies in terms of data, methodology and, in some cases, research integrity. We identified several of these papers as problematic prior to the publication of Clegg et al. (2026). The others included below were identified after that publication as new to Cochrane review inclusion and as having a strong effect size along with problematic content. For each case, we provide an overview of the study's characteristics and claimed findings followed by an examination of the data each presents.

In Clegg et al. (2026), the calculated effect of exercise on depressive symptoms from 57 trials (the first primary forest plot presented on pg. 30; pooled SMD for depressive symptoms at the end of treatment, random effect model) was -0.67 (95% CI -0.82 to -0.52). When the seven trials discussed here are removed, the result falls from -0.67 to -0.59 (95% CI -0.73 to -0.45). This means a weaker effect than Clegg et al. (2026) identified in terms of exercise as a treatment for depression.

Clegg et al. (2026) also analyzed nine studies that included long-term followup. When a problematic study is removed, pooled data from the remaining eight falls similarly (from SMD -0.53, 95% CI -1.11 to 0.06, to SMD -0.21, 95% CI -0.52 to 0.1). Again, this suggests increased uncertainty about whether exercise has an impact on depression in the longer term.

The .ipynb document provided as an accompaniment to this report shows in Blocks 1 through 4 the two primary forest plots in Clegg et al. (2026) as they appear before and after the removal of the studies identified here as problematic.

3.1 Mutrie (1988)

Introduction

The conference proceedings paper Mutrie (1988) is based on a PhD thesis from 1986 (Mutrie, 1986) reporting on the investigation of the therapeutic effects of aerobic exercise on depression. While the 1988 conference paper, which was included in the meta-analysis in Clegg et al. (2026), clearly reports the same study as the 1986 thesis (nearly all the outcome data match), the description of the three study arms do not match from the thesis to the conference paper.

- In both the 1986 thesis and 1988 paper, Group A is described as including nine subjects assigned to "eight weeks of aerobic exercise, with strengthening and stretching exercise introduced after four weeks".
- In the 1986 thesis, Group B is described as eight subjects assigned to "eight weeks of strengthening and stretching exercise, with aerobic exercise introduced after four weeks" whereas the 1988 paper describes eight subjects who "received strengthening and stretching exercises for the second four weeks and a combination of aerobic, strengthening and stretching exercises for the next eight weeks." In other words, the 1986 thesis describes Group B participating for eight weeks while the 1988 paper describes Group B participating for 16 weeks.
- The 1986 thesis describes Group C as seven subjects who "received no treatment for four weeks and then received aerobic, strengthening, and stretching exercises for eight weeks." The 1986 thesis therefore indicates Group C participated for 12 weeks, which is in contrast to Group A (eight weeks) and Group B (eight or 16 weeks). The 1988 paper does not appear to describe Group C except to say those subjects "received no treatment" for the first four weeks.

- By contrast, other text in the 1986 thesis and 1988 conference paper suggest all the groups were in fact under study for the same period of time (namely eight weeks), as would normally be expected.

All the study groups were small (fewer than 10 subjects in each), with a total study population of 24.

Both manuscripts reported a very substantial reduction in depression symptoms for Group A, the group that undertook eight weeks of aerobic exercise. Indeed, despite intervening with a total of only 60-90 minutes of exercise per week, Mutrie reported one of the largest effect sizes recorded in the literature over the last 40 years. In any given meta-analysis on the topic, Mutrie (1988) typically reports the largest standardized effect seen (see, for instance, Figure 2, Lawlor and Hopker, 2001). It is included in Clegg et al. (2026) and notably shows approximately twice the standardized mean difference of the average provided in that most recent Cochrane review on this subject. Mutrie's study is consequently responsible for a substantial increase in the supposed overall effect size of exercise for depression, even though it contributes a small number of participants.

Like many studies of depression, this one administered the Beck Depression Inventory (BDI; Beck, Ward, et al., 1961) and Profile of Mood States scale (POMS; McNair et al., 1971) to participants before and after the interventions. The BDI is a 21-item self-report inventory for the characteristic symptoms of and attitudes in depression. The POMS is a standardized psychological scale for measuring transient moods (like tension, fatigue, or anger).

In Mutrie (1986) and Mutrie (1988), several of the means and standard deviations reported in these two scales are incompatible with the reported small sample sizes. In other words, the data does not make sense upon forensic analysis. We detail this further below and in the accompanying .ipynb file (Block 5).

GRIM failures in BDI outcomes

Three out of six mean values for the depression scale (BDI) in Mutrie's 1986 thesis fail the GRIM test (see Table 1, below; note the test results return the same for Mutrie, 1988).

Table 1: GRIM failures in Mutrie (1986)

Time	Group	n	Mean	SD	Consistent	Reason
Intake	A	9	22.44	6.82	TRUE	Passed all
Intake	B	8	21.86	4.21	**FALSE**	GRIM inconsistent
Intake	C	7	23.00	5.80	TRUE	Passed all
Week 4	A	9	9.46	4.28	**FALSE**	GRIM inconsistent
Week 4	B	8	14.63	7.63	TRUE	Passed all
Week 4	C	7	21.42	5.26	**FALSE**	GRIM inconsistent

This indicates the data presented lacks internal consistency, raising questions about the presented results. There are four primary possibilities for these failures:

1. The descriptive statistics are not accurately calculated given the sample size.
2. The samples have unreported attrition (which would be concerning in a study where every study arm has fewer than 10 subjects).
3. Data from the BDI was somehow recorded not in whole numbers but with sub-units. We note, however, this would be an extremely unusual feature of data collection, as the BDI is typically measured by participants indicating which individual checklist item best reflects their internal experience, and these are scored in whole numbers.
4. Data is misreported in the thesis.

The following table shows the results of GRIM tests run on the data presented in Mutrie (1986). A

"true" result in the "Consistent" column indicates the data is not throwing up red flags. A "false" result indicates the data is showing inconsistent and/or mathematically impossible results.

We note that Mutrie (1988) reports two of these numbers differently, but the changed numbers do not appear to constitute corrections. The standard deviation for Group C's mean BDI score at intake is given as 5.80 in Mutrie (1986) and as 5.26 in Mutrie (1988), and the standard deviation for Group C's mean BDI score at Week Four is given as 5.26 in Mutrie (1986) but as 25.26 in Mutrie (1988), the paper included by Clegg et al. (2026). For these aberrant values, for Intake and Week 4 values in Group C ($n = 7$), both SDs pass the GRIMMER test.

Given that the data changes from 21.4**2**(5.26) in 1986 to 21.4(**2**5.26) in 1988, this is likely a transposition error – this would explain both the nature of the change, and the fact that a standard deviation of 25.26 is vanishingly unlikely in BDI scores corresponding to patients with 'mild to moderate' depression. While mathematically possible, an example of group scores that would return this SD would contain both very low and very high BDI scores (e.g., [0, 0, 8, 11, 17, 51, 63]).

There is a further curious feature in Mutrie (1988) located on pg.101, concerning the ANOVA used when the data are collapsed to measure the main effect over time for all groups. This is stated as $F(1.26) = 32.59, p < 0.001$. This is both an extremely substantial F-value and also misreported, as only one value is presented. It is unlikely to represent $F(1, 26)$ as this would be measuring 2 groups, and the degrees of freedom df_2 would be higher than the number of participants, which is impossible.

A far more likely explanation is: as Mutrie (1988) used SPSS, which will apply Mauchly's Test of Sphericity and adjust the degrees of freedom automatically if there is a sphericity violation, $F(1.26)$ is a misreporting of the Greenhouse-Geisser adjusted degrees of freedom.

As we have 3 timepoints and 24 participants, the *uncorrected* F should be ($df_1 = Time - 1 = 2$; $df_2 = (N - 1)(Time - 1) = 46$) i.e., $F(2, 46)$. This implies the Greenhouse-Geisser ϵ was 0.63, and therefore $df_1 = 2 \times 0.63 =$ **1.26** and $df_2 = 46 \times 0.63 =$ **28.98**

In other words, the reported $F(1.26) = 32.59$ is truncated from $F(1.26, \mathbf{28.98}) = 32.59$.

This means the p-value is correctly reported (as $p < 0.001$) but also represents an extremely large and unlikely overall effect as $p = 8.68 \times 10^{-7}$.

GRIM/MER failures in POMS subscales

Another unusual set of numerical features appears later in Mutrie (1986) and Mutrie (1988), specifically within the means and standard deviations for the POMS subscales: tension, depression, anger, vigor, fatigue, confusion, and total mood disturbance (TMD). The results of the tests on this data are displayed in our Table 2, with GRIM and GRIMMER failures identified.

Again, these inconsistencies in isolation may not be indicative of reporting, numerical, or statistical errors, as it is technically possible that half-units were included without notation in the scale measurements and against the instructions of the scale.

Conclusion

For 40 years, the small study reported first in Mutrie 1986 and later as Mutrie 1988 has been included in meta-analyses on exercise and depression, and it has been having a measurable effect on the reviews' conclusions because the thesis purports to show exercise has a powerful impact on depression. In fact, the effect size claimed by Mutrie represents the largest standardized mean difference recorded by Clegg et al. (2026) over the 45 years of assembled research in the primary analysis (see the Block 1 forest plot in the .ipynb file). The study is and has always been an outlier, and the anomalies present cannot be easily explained.

Table 2: GRIM/GRIMMER results for POMS subscales at week four in Mutrie (1986)

Outcome	Group	n	Mean	SD	Consistent	Reason
Tension	A	9	13.78	7.45	TRUE	Passed all
Depression	A	9	10.33	10.86	TRUE	Passed all
Anger	A	9	6.22	3.73	TRUE	Passed all
Vigor	A	9	13.11	9.24	TRUE	Passed all
Fatigue	A	9	9.33	4.74	TRUE	Passed all
Confusion	A	9	6.89	3.92	TRUE	Passed all
TMD	A	9	33.4	31.75	TRUE	Passed all
Tension	B	8	14.63	6.69	**FALSE**	GRIMMER inconsistent (test 3)
Depression	B	8	18.75	12.3	TRUE	Passed all
Anger	B	8	14.75	9.59	TRUE	Passed all
Vigor	B	8	11.50	6.28	TRUE	Passed all
Fatigue	B	8	11.26	3.66	**FALSE**	GRIM inconsistent
Confusion	B	8	9.13	5.44	TRUE	Passed all
TMD	B	8	57.0	35.26	TRUE	Passed all
Tension	C	7	19.57	3.91	TRUE	Passed all
Depression	C	7	18.43	13.65	TRUE	Passed all
Anger	C	7	7.00	5.83	TRUE	Passed all
Vigor	C	7	9.42	4.16	**FALSE**	GRIM inconsistent
Fatigue	C	7	16.57	4.66	**FALSE**	GRIMMER inconsistent (test 1)
Confusion	C	7	13.28	4.46	**FALSE**	GRIM inconsistent
TMD	C	7	66.86	29.12	TRUE	Passed all

3.2 Mota-Pereira et al. (2011)

Introduction

Mota-Pereira et al. (2011) report an RCT among patients with treatment- resistant major depressive disorder (MDD). Following an initial screening of 150 patients, 33 patients being treated with usual pharmacotherapy were randomized to a moderate-intensity exercise (n=22) or a control group (n=11). The intervention "consisted of home-based 30-45 min/day walks, 5 days/week, for 12 weeks, being 1 walk per week supervised." Depression was measured via the Hamilton Depression Rating Scale (HAMD17; Hamilton, 1960) and Beck Depression Inventory (BDI; Beck, Ward, et al., 1961).

This published study includes multiple inconsistencies, errors, and anomalies in statistical reporting that raise concerns about the integrity of the results and their suitability for inclusion in meta-analysis. Some of the issues in this paper have already been documented on PubPeer by Meyerowitz-Katz (2023). The journal in which it was published appears to have taken no action.

These problems appear in both baseline reporting and outcome analyses, and they persist across text, figures, and tables. Issues are described below, with recalculations provided to illustrate the magnitude of the problems. The study information investigated are reproduced below.

Baseline p-values do not match reported means/SDs

As noted above, reports included in meta-analyses should be expected to show internal consistency with regard to studies' data presentations. In other words, recalculations using the presented material should produce the same numbers as those presented in a report. If, when we check the math, we cannot arrive at the same calculations, something appears to be wrong in the published data.

Table 1 of Mota-Pereira et al. (2011) presents baseline demographic and clinical characteristics for the control (medication only) and experimental intervention (medication + exercise) groups.

According to the table notes, the p-values come from independent-samples t-tests for continuous variables and from Fisher's exact test for categorical variables. Yet when we re-compute the t-tests directly from the reported means, standard deviations, and sample sizes, none of the p-values align with those published.

One possible explanation is that the authors mislabeled standard errors (SEs) as standard deviations (SDs) in the table. Indeed, if the values are reinterpreted as SEs, the recalculated p-values do move closer to the reported ones. However, even under this generous assumption, the results still do not reconcile, and the discrepancies go in both directions: in some cases the reported p-values are smaller than expected, while in others, they are larger. Crucially, the mismatches persist regardless of whether the numbers are treated as SDs (as labeled) or as SEs.

These errors indicate that the baseline inferential statistics in Mota-Pereira et al. (2011)'s Table 1 cannot be reproduced from the data as presented (see our Table 3 for details, and Block 6 of the .ipynb document for this and the below derivations). Both the original and the recalculated p-values are inconsistent with randomization.

Table 3: T-test p-values based on summary statistics vs. reported in Mota-Pereira et al. (2011)

		Student		Welch	
Variable	**Reported p**	**Possible bounds**	**Consistent**	**Possible bounds**	**Consistent**
Assuming reported SD is correct					
Age	0.385	[0.0055, 0.0059]	**FALSE**	[0.0091, 0.0096]	**FALSE**
BMI	0.247	[0.0023, 0.0025]	**FALSE**	[0.0066, 0.0070]	**FALSE**
HAMD17	0.014	[0.0000, 0.0000]	**FALSE**	[0.0000, 0.0000]	**FALSE**
BDI	0.031	[0.0000, 0.0000]	**FALSE**	[0.0000, 0.0000]	**FALSE**
GAF	0.003	[0.0000, 0.0000]	**FALSE**	[0.0000, 0.0000]	**FALSE**
CGI-S	0.010	[0.0000, 0.0000]	**FALSE**	[0.0000, 0.0000]	**FALSE**
Assuming reported SD is actually SE					
Age	0.385	[0.3919, 0.3963]	**FALSE**	[0.3952, 0.3996]	**FALSE**
BMI	0.247	[0.3139, 0.3197]	**FALSE**	[0.3247, 0.3303]	**FALSE**
HAMD17	0.014	[0.0078, 0.0083]	**FALSE**	[0.0079, 0.0084]	**FALSE**
BDI	0.031	[0.0220, 0.0230]	**FALSE**	[0.0226, 0.0236]	**FALSE**
GAF	0.003	[0.0022, 0.0023]	**FALSE**	[0.0026, 0.0027]	**FALSE**
CGI-S	0.010	[0.0074, 0.0103]	TRUE	[0.0080, 0.0110]	TRUE

Fisher's exact test p-values are incorrect

Mota-Pereira et al. (2011) reports using Fisher's exact test to calculate group differences in the male vs. female composition of the two study groups at baseline in Table 1, and in the patient status (comparing no response to treatment vs. response to treatment vs. remission of depression) in Table 4. A response to treatment was defined as a decrease from baseline HAMD17 score by 50%, and remission defined as a HAMD17 total score ≤ 7 by study conclusion.

In the study's Table 1, the published p-value for male (2 control, 8 exercise) vs. female (8 control, 11 exercise) Fisher's test is 0.110, but recalculating the test from the implied contingency table yields a value of p = 0.414. Applying a continuity correction does not account for this discrepancy. In Table 4, Fisher's exact tests comparing remission/response vs. no response are reported. Recomputing the p-values from the cell counts produces mismatched results. For remission vs. no response , we can reproduce the reported p = 0.061, but for response vs. no response (0/10 control, 4/10 exercise), the reported p = 0.094 cannot be reproduced; the calculated value is p = 0.114. Applying a continuity correction also does not account for this discrepancy.

These results are displayed below in our Table 4. Again, the reported data lacks internal consistency.

Table 4: Reproducing Fisher's exact test p-values from Mota-Pereira et al. (2011).

Source	Context	a	b	c	d	Reported	Reproduced	Consistent
Table 1	Baseline female vs male	8	2	11	8	0.110	0.414	**FALSE**
Table 4	Remission vs no response	5	10	0	10	0.061	0.061	TRUE
Table 4	Response vs no response	4	10	0	10	0.094	0.114	**FALSE**

Baseline values in the study's Table 4 do not match those in the study's Table 1

In Mota-Pereira et al. (2011), Table 4 reports baseline and follow-up HAMD17 means and SDs for patients treatment status split across three categories (no response to treatment, response to treatment, or remission of depression). In the control group, every subject was labeled a non-responder; therefore the baseline means and SDs of the non-responder group should be identical to the baseline means of the control group in Table 1. But they do not align. Specifically, the control group baseline HAMD mean is 13.00 (1.42, n = 10) in Table 1 but becomes 14.00 (4.76, n = 10) in the study's Table 4.

The study's Table 4 also mislabeled the exercise group as having 10 subjects, when it is otherwise said to have 19.

In-text claim contains inconsistent values

On p.1008, the authors of Mota-Pereira et al. (2011) state:

> *although not statistically significant, participants that showed remission also had lower HAMD17 total scores at baseline than participants that showed only response or no response (13.00 $\pm$ 4.6, 22.75 $\pm$ 5.3 and 17.64 $\pm$ 7.1, respectively, p > 0.05) (Fig. 3).*

But the study's Figure 3 does not contain HAMD17 values; it displays values from other rating scales (BDI, GAF, and CGI-S). Table 4 contains the values for baseline for remission, response, and no response, however, only the means for 'remission' and 'response' are presented correctly (13.00 and 22.75, respectively). The 'no response' value is listed as 17.64 — yet in the table it is given as 21.10. It is possible that this is meant to indicate the pooled mean of both n=10 groups – Control *and* Exercise non-responders – but that would also be incorrect as $(14.00 + 21.10)/2 = 17.55$.

The text of Mota-Pereira et al. (2011) therefore references the wrong figure and, more importantly, misreports numerical values.

Multiple outcomes fail the GRIM test

A GRIM test confirms multiple mathematical concerns in Mota-Pereira et al. (2011)'s Table 1. For example, the severity subscale of the Clinical Global Impression scale (CGI-S; Guy, 1976) is a 1–7 integer scale and, for the control group (n = 10), the reported mean is 2.67. No combination of 10 integers from 1 to 7 can average to 2.67; it is arithmetically impossible. This again tells us something is amiss in the reported data.

The baseline mean values of the depression scales presented in Mota-Pereira et al. (2011)'s Table 1 and the change score means reported in Table 2 were also GRIM-tested. (It is possible to GRIM-test the contents of Table 2 because the difference between two integers is still an integer.) The results of the GRIM test are displayed in our Table 5.

The test failures again tell us something is amiss with the presented data in Mota-Pereira et al. (2011).

Implausibly high correlations in pre- and post-test measurements

As we noted in the introduction to our methods, reports that include two or more measurements from the same group across different time points provide meta-scientists an opportunity to look for red flags. For example, we can look at the pre- and post-test correlations and to see if the results

Table 5: GRIM test of baseline characteristics in Mota-Pereira et al. (2011)

			Baseline		Change scores	
Variable	**Arm**	**n**	**Mean**	**Consistent**	**Mean**	**Consistent**
HAMD17	Control	10	13.00	TRUE	0.60	TRUE
HAMD17	Exercise	19	19.32	TRUE	-6.84	TRUE
BDI	Control	10	17.83	**FALSE**	4.30	TRUE
BDI	Exercise	19	24.68	TRUE	-6.47	TRUE
GAF	Control	10	66.50	TRUE	-5.44	**FALSE**
GAF	Exercise	19	53.84	TRUE	8.05	TRUE
CGI-S	Control	10	2.67	**FALSE**	0.33	**FALSE**
CGI-S	Exercise	19	3.89	TRUE	-0.89	TRUE

are implausibly high or low.

In Mota-Pereira et al. (2011), the reported correlations between pre- and post-test measures exceed 0.98 in both the control and intervention groups. That either group — no less both groups — would show such high correlations is implausible in terms of real-world results. Human behavior and measurement of it inevitably involve variation and unpredictability, and the results in this report show very, very little.

How little? We can deduce the pre-post correlations for all of the participants in Mota-Pereira et al. (2011) with the following two-step procedure for the experimental group:

(1) Table 4 lists all the subgroups in the experimental group (19 participants), which can be combined; *non-responders* (baseline mean = 21.10, SD = 7.64, n = 10; end-point mean = 18.20, SD = 5.63, n = 10), *responders* (baseline mean = 22.75, SD = 5.31, n = 4; end-point mean = 8.75, SD = 0.96, n = 4), and *remitters* (baseline mean = 13.00, SD = 4.64, n = 5; end-point mean = 4.00, SD = 1.23, n = 5). This gives us the pooled sample means and SDs for all 19 participants with a high degree of precision for the baseline (mean = 19.316, SD = 7.358) and end-point (mean = 12.474, SD = 7.588, n = 19). Then,

(2) as Table 2 gives us the mean *difference* between these groups, and the SD of that difference (6.84 and 1.47 respectively), we can use equations from Jané et al. (2024) to calculate the pre-post correlation from the baseline and end-point means and SDs.

Carrying out this analysis, we find the pre-post correlations for the exercise group is approximately .981 (the control group, similarly calculated, is .984). This is an almost-perfect correlation of pre- and post-tests for these two groups, meaning that the baseline scores almost perfectly predict the progress a patient will make over the twelve weeks of treatment. If we compare this result to previous meta-analyses of test-retest reliability of the HAMD17, Mota-Pereira et al. (2011) appears to present the highest test-retest correlation of all time in this area of research (see Trajković et al., 2011). It would be extremely unusual in any scenario for a treatment effect to contain a higher correlation than a test's reliability. This observation has been explored in previous empirical criticism, most notably in Vul et al. (2009).

Conclusion

Mota-Pereira et al. (2011) was noted three years ago to contain a range of inconsistencies, and our analysis confirms reasons to be concerned about inclusion of this paper in a meta-analysis. We find inconsistent data, incorrectly calculated statistical tests, and baseline data that almost perfectly predict the eventual outcome data.

3.3 Barclay et al. (2014)

Introduction

Barclay et al. (2014) report a small pilot study which investigated the effects of a six-week mixed exercise intervention on depression symptoms in 11 participants. The intervention group included five subjects who exercised 1-2 hours per week, and the control group included six subjects who were not assigned to exercise. Unusually, the study also attempted to investigate the physiological correlates for improvements in depression using electroencephalography (EEG) and blood neurotransmitter levels as physiological markers.

The authors reported that the exercise group experienced significant reductions in depression according to BDI-II scores (the updated Beck Depression Inventory; Beck, Steer, & Brown, 1991), favorable changes in their brain electrical activity (specifically, in their frontal alpha asymmetry), and improved serotonin levels compared to the control group.

Several features of this study raise concerns: the physiological measures appear to be reported incompletely or are nonsensical; the methodology is not provided; and the variation in baseline physiological measures is implausible.

Retrospective clinical trial registration

The proper registration of a clinical trial involves recording the key details (like the primary analyses of interest, number of participants to be recruited, etc.) in a public repository before the study begins. Registration of this type is designed to reduce publication bias and analytical flexibility and lower "researcher degrees of freedom," making studies more rigorous from the outset.

In 2004, the International Committee of Medical Journal Editors (ICMJE) announced that clinical trial registration would become mandatory starting in 2005. Contrary to those standards, Barclay et al. (2014) was registered *retrospectively*. The study text and the clinicaltrials.gov entry specify that it was begun in March 2013, conducted until May 2013, and registered after it concluded, in December 2013.

While this issue alone does not render a study report invalid, it does constitute a negative mark. And, in the case of Barclay et al. (2014), it is not the only reason for concern about inclusion in a meta-analysis.

Uninterpretable physiological methods

Physiological and biological methods can add additional insight into questions in the behavioral sciences. For example, if a stressful experimental intervention is measured subjectively via self-report, the inclusion of heart rate or heart rate variability can simultaneously offer mechanistic insight into the nature or magnitude of that stress. However, physiological measurement is challenging and requires carefully controlled experimental methods and precise reporting. A close reading of Barclay et al. (2014) shows a concerning absence of clear methods and indeed turns up passages that are at best classified as uninterpretable.

For example, on p.33, the authors provide the following results for their EEG measurements:

> *the exercise group showed a significant decrease in frontal alpha activity measured at F7 in comparison to the control group, which is a 15.7% decrease from pretest (5.1Hz) to posttest (4.3Hz) [Figure 2].*

This passage appears to encapsulate a number of misunderstandings or misrepresentations. In this experiment, *theta* waves (which are present in the ranges indicated, from 4Hz to 8Hz) were measured and reported as alpha waves. Additionally, the "decrease in activity" referred to is not specified as to its nature. The paper is presumably not describing a decrease in power (i.e., the height of the waves present), but a decrease in mean or median frequency (i.e., the distance

between the peaks). But, as no further details are present in the methods about how the EEG was conducted, it is not possible to analyze further. It seems fair to characterize these "results" as uninterpretable.

Barclay et al. (2014))'s descriptions of other primary physiological measurement are also incomplete, as the authors provide numerical values for the catecholamines (e.g., 406.4, 10.6) but never the units (e.g., pg/mL, ng/L). Nor do they explain how the physiological analysis was conducted (e.g., HPLC, ELISA, etc.). There is no specification of how these fairly fragile measurements taken were managed (e.g., the use of chilled EDTA tubes, preservatives like metabisulfite, centrifugation, freezing, etc.). No specific kit manufacturer is cited, so none of the above can be inferred.

Additionally, serotonin (one of the neurotransmitters measured) is typically not measured in serum, as it was in Barclay et al. (2014), but in plasma, as it is released from degrading blood platelets at high and variable levels. Finally, postural, dietary, time-of-day, etc., controls are not mentioned, even though these factors are known to strongly affect some catecholamine measurements.

Incomplete (non-specified) power calculation

Experimental power calculations attempt to determine the minimum sample size necessary for an experiment to reliably reject the null hypothesis if the true effect is as specified. Researchers make this calculation to know how many participants they must enroll in a study to obtain meaningful results.

The putative power calculation provided by Barclay et al. (2014) does not actually calculate statistical power. As a consequence, the study report does not explain why the researchers proceeded with only 11 participants. On p.32, the text states:

> *Although a sample size of 20 was expected for achieving a level of power that is greater than 0.70, due to time constraints, only a sample of 11 was able to be obtained but power was still found to be sufficient for a pilot study.*

This is not really a power calculation as it does not specify an expected difference between exercise and control participants. If calculated with 11 participants split into two groups, the degrees of freedom are $n = 9$, and thus the critical t-value is 2.262. If this were calculated correctly, 70% power (as the authors indicate they seek) would be achieved when $d \approx 1.69$ (where $\delta \approx 2.79$). This means the researchers were seeking an effect size (from a sample of 11 people) more than twice the effect size ultimately calculated and published in Clegg et al. (2026), and twice the traditional value for a 'large' effect size ($d = 0.8$).

We also note the more traditional threshold for false negative control is 80% power, which is achieved when $d \approx 1.91$ (where $\delta \approx 3.15$). In short, there is critical information missing from this report, and the information that is provided is scientifically abnormal.

That being said, a power analysis in the context of a pilot study is somewhat redundant; typically, a pilot is performed to assess the feasibility of a method, not to establish the expected magnitude or direction of an effect.

Statistical anomalies

Very little categorical data is presented in Table 1 of Barclay et al. (2014), as most of the measurements are (presumably) continuous values drawn from analysis of serum and EEG. GRIM and GRIMMER tests cannot be conducted on these figures, but the data on circulating catecholamines show significant and confusing heterogeneity that can be quantified with other tests.

In particular, there is a substantial difference in baseline variance between the two samples; an

F-test for the equality of variances (below, F-test EoV) reveals extreme heterogeneity of variance ($\approx$ 27x for Norepinephrine, $\approx$ 57x for Serotonin) (see below our Table 6). As the F-test is highly sensitive to outliers, Levene's test is preferred but we cannot calculate it in the absence of the underlying data. An approximate visualization of this heterogeneity is given in the accompanying .ipynb file as an appendix.

Table 6: Baseline means, standard deviations, and statistical comparisons for neurotransmitters from Barclay et al. 2014.

Variable	Exercise Mean (SD)	Control Mean (SD)	Mean Diff	Welch's p	F-test p (EoV)
Norepinephrine	406.4 (322.4)	286.2 (61.1)	120.20	0.4551	**0.0026**
Serotonin	10.6 (7.0)	44.8 (52.9)	-34.20	0.1755	**0.0016**
Epinephrine	45.6 (24.1)	43.2 (39.5)	2.40	0.9044	0.3597
Dopamine	14.2 (7.8)	11.3 (3.8)	2.90	0.4784	0.1466

It is helpful that the mean and SD *of the individual changes* are reported for every measurement in both groups over time. These can be easily translated into pre-post correlations (i.e. the correlation between any measurement before and then after the conclusion of the experiment).

We begin with the derivation provided in Jané et al. (2024) :

$$r = \frac{SD_{Pre}^2 + SD_{Post}^2 - SD_{Gain}^2}{2 \cdot SD_{Pre} \cdot SD_{Post}}$$

Then we can calculate all the implied r-values between the two time points (see our Table 7). Reproduction of our Table 6 and Table 7 are given in the accompanying .ipynb file (Block 7).

Table 7: Implied pre-post correlations (r) for all variables in Barclay et al. 2014.

Variable	Exercise Group (r)	Control Group (r)
Depression	**0.022**	0.725
Norepinephrine	0.810	0.557
Epinephrine	**-0.544**	0.850
Serotonin	0.117	0.946
Dopamine	**0.002**	0.167
F1	0.397	0.731
F3	0.312	0.888
F7	0.705	0.852
F2	**0.042**	0.779
F4	0.333	0.647
F8	0.469	0.739

Both depression and sympathetic tone should be expected to be trait-like, and therefore stable. Individuals with high baseline catecholamine levels would typically continue to have high levels relative to the cohort even if the absolute mean decreases. A near-zero or a substantial negative correlation, as shown above in bold, strongly implies an irregular measurement process. For the epinephrine measurements in particular, the participants with the highest baseline epinephrine levels tended to become those with the lowest post-intervention. While this is from a small sample, this is implausible.

Conclusion

Barclay et al. (2014) is a retrospectively registered clinical trial without a functional power analysis. It shows multiple anomalies in its physiological methods. Its baseline data shows the groups included are curiously different, and the data over the experiment has reduced consistency between timepoints.

3.4 Prakhinkit et al. (2014)

Introduction

Prakhinkit et al. (2014) reports a study aimed at determining the effects of exercise and walking meditation on depression, functional fitness, and vascular reactivity. The study involved randomization of 45 elderly participants into three groups of 15 participants each: a sedentary control group (CON), a traditional walking exercise group (TWE) and a Buddhist walking meditation group (BWM).

Most of the measurements in the study were typical physical performance markers from exercise tests commonly used to evaluate elderly populations (e.g., the sit-to-stand test, sit-and-reach test, 6-minute walk test, etc.). Depression was measured by a Thai translation of the Geriatric Depression Scale, a self-report mood scale for older adults.

As shown below, this report contradicts itself in its descriptions of methods; it indicates the groups were balanced for age and physical ability at baseline, but, on recalculation, the groups are shown to be substantially different at baseline, sufficient to question whether adequate randomization took place. Additionally, the central analysis is vague and upon recalculation shows improbable outcomes.

Substantial differences in baseline data and questions about randomization

According to p.412 in Prakhinkit et al. (2014):

> *The eligible subjects were stratified based on age and depression levels (mild-to-moderate depression) and then were randomly allocated into three groups using the random number table generated by the computer...*

Contrary to this claim, our analysis indicates the participants were unlikely to be stratified then randomly distributed into the three groups. The mean age of the sedentary CON (control) group appears to be approximately 1 SD larger than that of the TWE and BWM groups, as reported in Prakhinkit et al. (2014)'s Table 1. The data for age at baseline is reported as mean and standard errors (SE): CON 81.0 $\pm$ 1.7; TWE: 74.8 $\pm$ 1.7; BWM 74.0 $\pm$ 1.9).

The SE values are reported to 1 decimal place, and this requires us to report a range and not single values for SE, as SE is converted to standard deviation (SD) by multiplying by the square root of n. So, in this case, the reported SE of the control group (1.7) may lie anywhere in the range from 1.65 to 1.75. If we convert this to SD, the value for the SD is from 6.39 to 6.78. To recalculate the values given in Prakhinkit et al. (2014)'s Table 1, we can separately test the largest possible difference (where the means differ the most with the smallest SDs) and the smallest possible difference (where the means differ the least with the largest SDs).

The derivation of the p-values for a one-way analysis of variance (ANOVA) from summary statistics using R is given in the accompanying .ipynb file (Block 8). From it, we can calculate the lowest and highest possible p-values, both of which are statistically significant (see our Table 8).

Table 8: Analysis of age randomization (with rounding) in Prakhinkit et al. (2014)

Scenario	Group	Mean Age	Imputed SE	Calculated SD	ANOVA p-Value
Max Difference (Best case)	CON	81.05	1.65	6.39	**0.0103**
	TWE	74.75	1.65	6.39	
	BWM	73.95	1.85	7.17	
Min Difference (Worst case)	CON	80.95	1.75	6.78	**0.0199**
	TWE	74.85	1.75	6.78	
	BWM	74.05	1.95	7.55	

The baseline age data presented by Prakhinkit et al. (2014) are incompatible with random allocation, let alone age-stratified randomization. We cannot conclude that randomization did not occur, but a randomization process would be extremely unlikely to produce imbalances this extreme.

Given this difference in the mean ages of the three groups, it is reasonable to ask if there is a similar difference in the relative functional capacity of those groups. Consequently, the data presented on p.414 for functional fitness measures were similarly tested via one-way ANOVA to establish baseline consistency (see our Table 9).

Table 9: Recalculation of baselines taken from 'Changes in Functional Fitness Measures During the Interventions' in Prakhinkit et al. (2014)

Variable	**Control** *Mean (SD)*	**Trad. Walking** *Mean (SD)*	**Walk. Med.** *Mean (SD)*	**Recalculated** p
Arm Curl	8.2 (2.5)	11.3 (2.9)	12.2 (2.6)	0.001
Chair Stand	5.6 (2.5)	7.0 (2.5)	9.0 (3.0)	0.009
Back Scratch	-21.2 (14.1)	-19.1 (23.1)	-10.3 (9.4)	0.196
Sit & Reach	-4.2 (7.2)	-8.7 (12.6)	-2.0 (7.5)	0.185
Timed Up-Go	14.8 (5.4)	12.2 (2.9)	9.9 (2.2)	0.006
6-Min Walk Test	149 (86.5)	221 (46.9)	164 (48.6)	0.015

This data shows the three groups had an appreciable difference in physical function at baseline.

The overall difference (reproduced in our Table 9) can be quantified by combining the ANOVA p-values to make an omnibus p-value. Traditionally, authors are warned against testing baseline data in this manner (see, e.g., Altman, 1985). That is because if randomization has taken place, the p-values should be uniformly distributed and thus if enough baseline results are reported, the probability of a small p-value is substantial but essentially meaningless. Here, rather than asking about group comparability (on each variable) following randomization, we are asking how compatible the group differences are with randomization.

As readers might be unfamiliar with the calculation of an omnibus p-value, we present it here. In the above table, we can convert each p-value to a Z-score using the inverse normal cumulative distribution function ($\Phi^{-1}(1-p)$):

$Z_1 = \Phi^{-1}(0.999) \approx 3.09$
$Z_2 = \Phi^{-1}(0.991) \approx 2.37$
$Z_3 = \Phi^{-1}(0.804) \approx 0.86$
$Z_4 = \Phi^{-1}(0.815) \approx 0.90$
$Z_5 = \Phi^{-1}(0.994) \approx 2.51$
$Z_6 = \Phi^{-1}(0.985) \approx 2.17$

In the tradition of Carlisle (2012), we can calculate Stouffer's Z-score, where:

$Z_{combined} = \frac{\sum Z_i}{\sqrt{k}}$, where $k = 6$ (number of tests), making the sum of Z-scores:

$$3.09 + 2.37 + 0.86 + 0.90 + 2.51 + 2.17 = 11.90$$

and therefore the combined Z:

$$Z_{combined} = \frac{11.90}{\sqrt{6}} = \frac{11.90}{2.449} \approx \mathbf{4.86}$$

Converting this back into probability using the standard conversion for Z gives $Z = 4.86$ as:

$$p(Z > 4.86) \approx \mathbf{5.9 \times 10^{-7}}$$

Assuming the subjects were properly randomized and the tests are independent (see below), the probability of observing baseline functional fitness imbalances at least this extreme across these six measures is approximately 1 in 1.7 million (see .ipynb Block 9). This magnitude of baseline systemic imbalance strongly indicates an irregularity in the randomization process.

This figure should be regarded as a rough estimate, because the values tested are *not* independent. (In fact, it would be strange if group tests of basic gross motor movement and motion were independent.) In this case, the variance can be adjusted according to Strube (1985), conservatively assuming a strong inter-dependence between the various tests of $r = 0.5$:

The variance of the sum of k Z-scores with an average correlation $\bar{r}$ is:

$$Var(\sum Z_i) = k + k(k-1)\bar{r}$$

As we have $k = 6$ tests and have set our inter-correlation of $\bar{r} = 0.5$:

$$Var(\sum Z_i) = 6 + 6(5)(0.5) = 6 + 15 = \mathbf{21}$$

(Note: compared to the independent variance, which is only $k = 6$).

We then divide our sum of Z-scores (11.8906, truncated above to 11.9) by the square root of this new, larger variance:

$$Z_{adjusted} = \frac{11.8906}{\sqrt{21}} = \frac{11.8906}{4.5826} = \mathbf{2.595}$$

We then look up the p-value for a Z-score of 2.595:

$$p \approx \mathbf{0.0047}$$

This adjusted omnibus probability sits at roughly 1 in 211, meaning the baseline imbalance remains statistically significant even after conservative correction for dependence.

To summarize the above, it seems as if the Control (CON) group in Prakhinkit et al. (2014) was both older and substantially less physically capable than either of the intervention groups, even though the baseline data was supposed to be stratified to prevent such a difference.

Unlikely central ANOVA results

Given the unusual features of the baseline data in Prakhinkit et al. (2014), it would be helpful to recreate the ANOVA which provides the central result extracted by Clegg et al. (2026). This result is presented on pp.413-414:

> *The depression score decreased* $(p < 0.05)$ *in the BWM group* $(16.8 \pm 0.9$ *vs* $8.6 \pm 0.6U)$ *and the magnitude of reductions in depression was significantly greater in the BWM than in the TWE group. No significant changes in depression scores were obtained in the TWE* $(17.3 \pm 1.0$ *vs* $15.5 \pm 0.9U)$ *and sedentary control group* $(17.9 \pm 0.7$ *vs* $18.6 \pm 0.6U)$.

(Note: U appears to be a symbol for units.)

The method section states:

> *Two-way (group · time) analyses of variance (ANOVA), followed by Bonferroni multiple comparisons were used to determine the significant differences among group means.*

But if this analysis was conducted, it was not reported in Prakhinkit et al. (2014). There are also no reported F values, no mention of an interaction, no values reported with any precision beyond one decimal place, no mention of how follow-up testing was conducted, and no differentiation between 'differences' that might be follow-up or main effects.

Nevertheless, we can estimate the magnitude of some of the data present. Given we can be reasonably sure that the Time 1 data is correlated with the Time 2 data, we can use the same identity from the above analysis of Barclay et al. (2014) given in Jané et al. (2024) to reverse-engineer the presented values:

$$r = \frac{SD_{Pre}^2 + SD_{Post}^2 - SD_{Gain}^2}{2 \cdot SD_{Pre} \cdot SD_{Post}}$$

The accompanying .ipynb file (Block 10) contains a simulation of this ANOVA. This is possible as we know the Time 1 and Time 2 means and standard deviations, and we can use the identity above to test a plausible range of r values, i.e., the strength of the correlation between Time 1 and Time 2 data. If we test realistic values (between r=0.3 and r=0.8), a Monte-Carlo simulation gives a modal value of a group effect at p = 10^{-5}, and an interaction effect of p = 10^{-7}. This represents an extremely strong effect driven by the large increase in depressive symptoms observed in the BWM (walking meditation) group.

Concentrating on that group in isolation, the Time 1 values (16.8 ± 0.9) decrease dramatically to Time 2 (8.6 ± 0.6) (n=14). To find a reasonable upper bound for this p-value, we can set the value of the correlation between these two timepoints (r) at 0 — i.e. where baseline depression is completely uncorrelated to the post-treatment depression.

First, we retrieve the SDs:

$$SD_{Pre} = 0.9 \times \sqrt{14} \approx 3.37$$
$$SD_{Post} = 0.6 \times \sqrt{14} \approx 2.24$$

Then we can simply solve when r=0:

$$SD_{Gain} = \sqrt{SD_{Pre}^2 + SD_{Post}^2}$$
$$SD_{Gain} = \sqrt{3.37^2 + 2.24^2}$$
$$SD_{Gain} = \sqrt{11.36 + 5.02}$$
$$SD_{Gain} \approx 4.05$$

This can be transformed back into the Standard Error of the Gain(SE_{gain}):

$$SE_{gain} = \frac{SD_{gain}}{\sqrt{n}}$$
$$SE_{gain} = \frac{4.05}{\sqrt{14}} \approx 1.08$$

And, finally, we retrieve t by dividing the mean difference by the standard error of the gain SE_{gain}.

$$t = \frac{\text{Mean Difference}}{SE_{gain}}$$
$$t = \frac{8.2}{1.08} \approx 7.58$$

As $df = 13$ and $t = 7.58$, the *realistic upper bound* of the p-value is ≈ 0.000004. Like the ANOVA simulation presented above, this is an extremely strong difference for a behavioral intervention, especially given that p would be a great deal smaller with a reasonable assumption for r (e.g. 0.5).

Finally, if we test the mathematically possible but physiologically unrealistic upper bound of $r = -.99$, where $df = 13$ and $t = 5.48$, the p-value is still extremely small ($p \approx 0.0001$).

If a researcher found any of the possible results outlined in this section, it would be a strange decision to report the accompanying p-value as simply $p < 0.05$.

Conclusion

Prakhinkit et al. (2014) describes unexplained and substantial differences in the samples at the study's baseline, with the control group appearing to have been significantly older and less functionally capable than the intervention groups. The report also suffers from an absence of analytical details, such that the central results cannot be statistically reproduced, and it purports to show an extremely substantial (i.e., unbelievable) change in depressive symptomatology.

3.5 Abdollahi et al. (2017)

Introduction

The RCT reported by Abdollahi et al. (2017) assessed the incremental benefit of exercise as an adjunct to cognitive behavioral therapy (CBT). In the study, subjects were randomized between 12 weeks of CBT (control) or CBT + exercise 3x per week (CBT + EX), and 54 completed the protocol. The report indicated the addition of exercise contributed to a significant reduction in depression.

As we detail below, Abdollahi et al. (2017) includes data points that do not match from one part of the report to another, the data appear to lack internal mathematical consistency, and the lead author has had work retracted for authorial ethics violations.

Inconsistent claims about the subject population and related GRIM failures

On p.58, the report's abstract states, "In a randomized clinical trial, 54 mildly to moderately depressed patients… were assigned" to the two arms of the study. In contrast to this claim of randomization of 54 patients, p.60 refers to "70 individuals who were randomized" of whom 54 "completed the intervention and provided complete data." This seems to confuse the number of subjects entering the study (i.e., being assigned to a study arm) and the number completing the study.

This confusion might appear trivial were it not for the fact that it appears to carry into Table 2 of Abdollahi et al. (2017), where forensic analysis turns up a curious pattern of GRIM failures in the presented results.

In the table in question, six baseline figures are presented (the two arms each assessed with three scales: depression, suicidal ideation, and activities of daily living), and all of these numbers pass the GRIM test for n=70. However, *every* corresponding post-study figure fails the GRIM test if the baseline remains at n=35 for each arm. When we administer a GRIM test using the number of completed participants (n=29 for CBT only, n=25 for CBT + EX), the results pass.

We can further confirm the confusion regarding the number of subjects at the outset by splitting the presented totals (the rightmost column in the report's Table 2) into subgroups in proportion (i.e., assuming n=35/35 or n=29/25; see our Table 10).

It therefore seems very likely that the confusion between the abstract and the methods section in Abdollahi et al. (2017) is continued in the report's Table 2, which mixes together the *complete* baseline dataset of 70 subjects with the *completed* data set of 54 subjects for analysis. This ambiguity makes it challenging to recalculate some of the presented statistics.

We went on to analyze the presented data for the "completed" vs. "dropout" groups. By combining the properties of the changing sample sizes and the reported totals, a somewhat complex derivation of the mean data can be reconstructed (see this report's Appendix for details).

This derivation reveals two inconsistencies. First, the baseline differences in suicidal ideation given on p.60 is reported as 0.84, but recalculated comes to 0.85. Second, the age means for the Completed and Dropout groups are recalculated at 49.46 (N=54) and 50.37 (n=16) respectively. This directly contradicts the abstract, which states the mean age of assigned subjects as 48.25. These discrepancies tell us something is amiss in the reported data in Abdollahi et al. (2017).

Table 10: GRIM test of mixed baseline and post-test sample sizes in Abdollahi et al. (2017)

Measure	Reported Mean	Test N=35 (ITT)	Test N=29/25 (Completed)	Conclusion
BDI-II (Depression)				
CBT Pre	20.51	TRUE	**FALSE**	Pre-test uses N=35
CBT Post	14.48	**FALSE**	TRUE (N=29)	Post-test uses N=29
CBT + EX Pre	20.89	TRUE	**FALSE**	Pre-test uses N=35
CBT + EX Post	9.68	**FALSE**	TRUE (N=25)	Post-test uses N=25
BSSI (Suicidal Ideation)				
CBT Pre	14.77	TRUE	**FALSE**	Pre-test uses N=35
CBT Post	12.72	**FALSE**	TRUE (N=29)	Post-test uses N=29
CBT + EX Pre	14.29	TRUE	**FALSE**	Pre-test uses N=35
CBT + EX Post	7.08	**FALSE**	TRUE (N=25)	Post-test uses N=25
ADL (Activities)				
CBT Pre	15.03	TRUE	TRUE (N=29)	Ambiguous (Matches Both)
CBT Post	15.21	**FALSE**	TRUE (N=29)	Post-test uses N=29
CBT + EX Pre	14.60	TRUE	TRUE (N=25)	Ambiguous (Matches Both)
CBT + EX Post	17.12	**FALSE**	TRUE (N=25)	Post-test uses N=25

These are not large anomalies, and it is hard to know what they represent. Two likely options are (1) typographical errors, or (2) the fact that the presented statistics were calculated from continuous data, which would be a possible explanation if it was simulated *but not restricted to integers*.

However, the inconsistencies of how data was handled in this case make the data describing the main study effects in Table 2 and multi-level model presented in Table 3 of Abdollahi et al. (2017) uninterpretable. Several of the relevant re-calculable effects are implausibly small, e.g., the suicidality scores in Table 2 are 14.29 (2.66) at baseline, falling to 7.08 (1.26) post-study. To approximate the magnitude of this difference, we can assume a reasonable pre-post correlation of $r = 0.5$ and recalculate the t-test – in this case, $t = 15.64$, $p = 4.34 \times 10^{-14}$, which is an implausibly large main effect.

Authorial concerns

Recent scholarship on the assessment of the trustworthiness of scientific reports recommends that, in reaching an overall judgment on a publication's reliability, assessors consider authors' track records with regard to retractions, expressions of concern, association with authorship-for-sale schemes, and so on.

We note that the lead author of Abdollahi et al. (2017), Abbas Abdollahi (Scopus ID 56030239100), has had his work subject to three retractions. The retraction notices list the reason as "a serious breach of [the publisher's] authorship policies and of publication ethics." Frontiers journal group retracted two of Abdollahi's papers (DOIs 10.3389/fpsyt.2023.1285013 and 10.3389/fpsyt.2023.1285087) during its 2023 purge of authorship-for-sale articles, as acknowledged in the publisher's official press release "Frontiers Implements New Policy to Counter 'Authorship-for-Sale'" (2023).

Abdollahi has also co-authored papers with individuals identified as having been involved with paper milling, including Maria Jade Catalan Opulencia (Scopus ID 57204649842; Retraction Watch coverage, 28 retractions); Yasser Fakri Mustafa (Scopus ID 57203725947; Retraction Watch coverage, 21 retractions); and Walid Kamal Abdelbasset (Scopus ID 57208873763; 21 retractions).

Conclusion

Abdollahi et al. (2017) confuses data on the number of subjects randomized vs. those who completed the study, reports extremely small p-values, includes statistics on age and suicidality that show a lack of internal consistency, and the publication record of the lead author includes three retractions and co-authorship with individuals who have substantive histories of retractions for authorial ethics.

3.6 Wang and Li (2022)

Introduction

Wang and Li (2022) describes an RCT of hospitalized patients with depression, comparing a control group receiving standard drug therapy against an experimental group receiving drug therapy plus a supplementary sports intervention. The researchers assessed depression severity using the Hamilton Depression Rating Scale (HAM-D) every two weeks for eight weeks. The reported results indicate an unusually large benefit for the exercise group: the HAM-D score in the exercise intervention arm drops from an average score of approximately 34 to 8. This result is included in Clegg et al. (2026) in Figure 6, where it is shown to be a substantial outlier.

Forensic analysis reveals several implausible or unlikely features in Wang and Li (2022): the main effect is not directly calculable but implausibly large; the two equations presented are not recognizable as mathematically accurate; there is a very unlikely pattern in the depression data; and the treatment efficacy calculation is incorrect.

Sample sizes are unclear

The sample sizes in this study are difficult to establish. In the abstract (p.174), Wang and Li (2022) report a total participation of 60 patients, but the methods section on p.175 mentions that 68 participants were recruited and that 4 withdrew from the exercise intervention arm i.e., 64 patients contributed at baseline.

The group sizes after randomization are not reported, making it difficult to assess claims about the two arms of the study. Wang and Li (2022)'s Table 1 (p.175) reports the group sizes as 30 each, but the 'Control' group is then reported as 7 + 6 + 5 + 13, which comes to 31. In the absence of a flow diagram, it is uncertain where the variously listed 68, 64, 61, or 60 patients participated.

Unusually high baseline depression scores

Baseline scores (which reflect subjects' states prior to an experiment's commencement) warrant attention because they tell us something about the people in a study. In some circumstances, they can also provide insight into the trustworthiness of a reported study.

In Wang and Li (2022), the baseline HAM-D scores for the control and intervention groups are given as means of 35 and 34, respectively. Admission to a clinical trial for treatment of a depressive disorder typically begin at a HAM-D score of 17 to 20. (Wang and Li (2022) used 17 as a minimum criteria to enter their reported study.) The onset of severe symptoms is typically understood to begin at a score between 23 and 25 (Zimmerman et al., 2013).

A *mean* HAM-D score in the mid-30s with a standard deviation of approximately 3 (as reported here) indicates a significant portion of the studied participants entered the study in a severe depressive crisis of the type that would make successful participation in an exercise study unlikely. Only four subjects were reported as having dropped out of the exercise intervention arm.

At no point does Wang and Li (2022) mention if the data presented (e.g., 34 ± 3.16 for the baseline HAM-D in the exercise group) is reporting SD or SE values. However, it is impossible for the presented measures of dispersion to be SE values, as this would make the SD $\approx$ 17, and almost all the participants would score either the minimum HAM-D score to enter the study (17) or the maximum (52).

Simple GRIM test failures

The control and exercise intervention in Wang and Li (2022) each include 30 participants according to the first column the report's Table 1, as noted above, and the abstract refers to a total of 60 participants (30 x 2). Each participant has a HAM-D score that is a whole number. Thus, the mean of

HAM-D scores for each arm should be expressed in units of 1/30th. However, the 10 means reported for baseline and 2-, 4-, 6-, and 8-week progress are all reported as whole numbers (all the included SDs or SEs are expressed to two decimal places). The probability of this happening from numbers expressed in 1/30th units chosen at random is $(1/30)^{10}$, meaning p=0.0000000000000016935. We can therefore consider this to amount to a failure of believability within the presented data.

The drug used for treatment is misnamed in the report

According to this published study, both the control and exercise intervention groups were administered "the depression drug Biyoujie" at a dose of 20mg/day. We cannot find any reference to this drug in any known pharmacopoeia. We presume this should read Baiyoujie, which is a generic brand of fluoxetine (Prozac) produced for the Chinese market. Biyoujie appears to be a brand of toothpaste.

A previous meta-analysis on exercise and depression (Mavranezouli et al., 2024) excluded Wang and Li (2022) from inclusion in part because it could not ascertain the identity of the treatment drug (at pp.19-20).

Presentation of inconsistent equations with no apparent relevance

One section of Wang and Li (2022) unaccountably delves into sports biomechanics and purports to outline two equations that respectively "get the muscle movement state of different training stages", and subsequently use "the [previous] formula fused with neural network theory [to] build a muscle influence model under sports training" (see our Figure 1 below).

Figure 1: Equations 1 and 2 from Wang and Li (2022)

$$M^{*}_{(W)} = \frac{f_2 \otimes e_{(f)}}{[\chi^{(t)} \square g^{(t)}]} \vartheta(\omega) \times Q(k) \qquad \psi(o,j) = \frac{[\partial(j,p) * k(p,o)]}{\varphi(\Im) * \gamma(j,p)} \otimes \xi$$

Neither equation is recognizable from any related calculation in biomechanics. Both mix together seemingly random symbols — a tensor product, omega (the traditional symbol for angular velocity), a replacement character or a 'tofu,' a Boltzmann constant, and other mathematical symbols apparently chosen from time-domain analysis, physics, and signal processing.

Having presented these equations, Wang and Li (2022) go on to cite other papers which are clinical studies of exercise and depression but not in any way related to biomechanical theory. Then, having stated the equations, the authors never refer to, use, or develop them anywhere else in the paper.

Incorrect categorical analysis

As above (see 'Sample sizes are unclear') the group sizes overall are uncertain, but the analysis on p.175 splits all the completed participants of both the exercise and control groups into categorical classifications, in this case (e.g. the 'Control group' is reported as 'Get well' (n=7), 'Markedly effective' (n=6), 'Get better' (n=5), and 'Invalid' (n=13)).

The analysis method used in Wang and Li (2022)'s Table 1 is unstated, but the comparison between these treatment categories within the outcomes of the two groups is reported as 'X2=0.012, P<0.05', which presumably means a chi-squared test (χ^2) was used. If this result is recalculated from the data presented on p.175, our calculations return a different result (χ^2=6.4425, p = .09196).

Conclusion

Wang and Li (2022) provide unclear information about the numbers of participants in their study, detail implausibilities around the participants' severity of condition at baseline, report depression

scores over time that are so unlikely as to be unbelievable, misname the drug used in the study, list equations which are not internally consistent or mathematically comprehensible (or relevant), and miscalculate the categorical treatment category data.

3.7 Bademli et al. (2023)

Introduction

Bademli, N. Lök, and S. Lök (2023) reports a study which randomized a group of 62 caregivers of patients with schizophrenia to either an exercise group (31 people who underwent a structured exercise program) or a control group (31 people who were asked not to alter their normal routine) for 12 weeks. Depression was measured by the BDI, and the burden on caregivers was measured by the Zarit Caregiver Burden scale (a 22-item scale with answers from 0 to 4 for each item; see Zarit, Reever, and Bach-Peterson, 1980).

Forensic analysis finds the sample sizes reported are internally inconsistent, and, when we recalculate key values based on the data provided, we cannot recreate the reported results.

Indeterminate sample sizes and power analysis

Bademli, N. Lök, and S. Lök (2023) provides the following inconsistent numbers with regard to study participation:

- Figure 1 describes 124 caregivers being assessed for study eligibility, with 48 deemed ineligible, leaving 76 participants to be "put into the randomization process."
- But the abstract, methods section, Figure 1, and Tables 2 and 3 of the paper all describe randomization of two groups of 31 caregivers, for a total of 62 study participants.
- Meanwhile, Table 1 (different from Figure 1) describes data from two groups of 30 participants, for a total of 60 participants.
- In addition, the abstract indicates one participant did not complete the post-intervention questionnaire, but this is not reflected in figures given elsewhere in the paper.

Bademli, N. Lök, and S. Lök (2023) present a power analysis to determine the number of participants needed in the study's two groups to reach an acceptably low false negative rate (i.e., an 80% chance of calculating a difference between the included groups if the true effect is as defined). If we assume control and intervention groups of 31 people each, the power analysis described in Bademli, N. Lök, and S. Lök (2023) is correctly calculated.

However, on p.1110, Bademli et al. list the predicted magnitude of the difference between the two groups as d=0.73, citing Rebar et al. (2015). This figure does not appear in Rebar et al., 2015, which lists a standardized mean difference (SMD) of 0.5.

If we use 0.5 instead, the power calculation in Bademli, N. Lök, and S. Lök (2023) would be:

$$\text{Parameters:} \quad \alpha = 0.05, \quad \text{Power:} (1 - \beta) = 0.80, \quad d = 0.50$$

$$\text{Formula:} \quad n = \frac{2(Z_{\alpha/2} + Z_{\beta})^2}{d^2}$$

$$\text{Calculation:} \quad n = \frac{2(1.96 + 0.84)^2}{0.50^2} = \frac{2(7.84)}{0.25} = 62.72$$

In other words, using the power analysis from the cited Rebar et al. (2015) paper, this comparison would require 63 participants per group, or 126 participants in total. As such, Bademli, N. Lök, and S.

Lök (2023) is substantially underpowered to detect the effect of exercise on depression. It is unclear where a SMD of 0.73 was derived, and no explanation is given beyond the citation.

Unusual U test values

The Medical Evidence Project recently published an explication of the GRIM-U test, which allows a user to establish minimum and maximum parameters for rank-sum values returned from the Mann-Whitney U test or, by extension, the Wilcoxon Signed Rank test. The data presented in Bademli, N. Lök, and S. Lök (2023)'s Table 2 was reanalyzed with this test, which essentially allows us to look for presented data that include mathematically impossible conclusions.

Table 2 indicates 31 participants in each of the two arms. Devolving the Mann-Whitney U test in the matter of the GRIM-U test, we can establish the minimum and maximum rank sums that are possible from two groups of 31. If two samples of 31 are compared with a rank-sum test, the maximum separation is when ranks are totally distinct — when the highest value in Group 1 is lower than the lowest value in Group 2. This gives Group 1 the ranks of 1:31 and Group 2 the ranks of 32:62. These correspond to the lowest and highest possible rank sums of 496 $(1+2+3+4+....+31)$ and 1457 $(32+33+34+35+...+62)$ respectively, and thus the Mann-Whitney U test returns values that range from 0 to 961. Both p-values for these extremes represent the smallest possible p-values from the test (in this case, 1.3×10^{-11}). From Bademli, N. Lök, and S. Lök (2023)'s Table 2 reporting of the baseline group, $U = 125, p = 0.76$ is incongruent. If $U = 125, p = 5.6\times10^{-7}$. However, if $p = 0.76, U = 502$ or 502.5.

If we apply the same principle to the other Mann-Whitney U values presented by Bademli, N. Lök, and S. Lök (2023), we find a series of other discrepancies (see our Table 11). All of this can be calculated from trivial modifications of the Medical Evidence Project's Excel spreadsheet for GRIM-U calculations. (Note that there may be several possible U values for the same p-values, especially when p-values are small, and this table gives single examples only.)

Table 11: The relationship between U and p-values in Bademli et al. (2023).

Measure	Reported U	Reported p	p if U	U if p (low)	U if p (high)
Baseline, Burden	125.0	.76	.00000057	459	502
Post, Burden	275.0	.005	.0038	281	680
Baseline, BDI	175.0	.92	.000017	473	488
Post, BDI	650.0	.001	.017	247	714

The derivations above are not modified meaningfully if we assume the groups are n=30 or n=31.

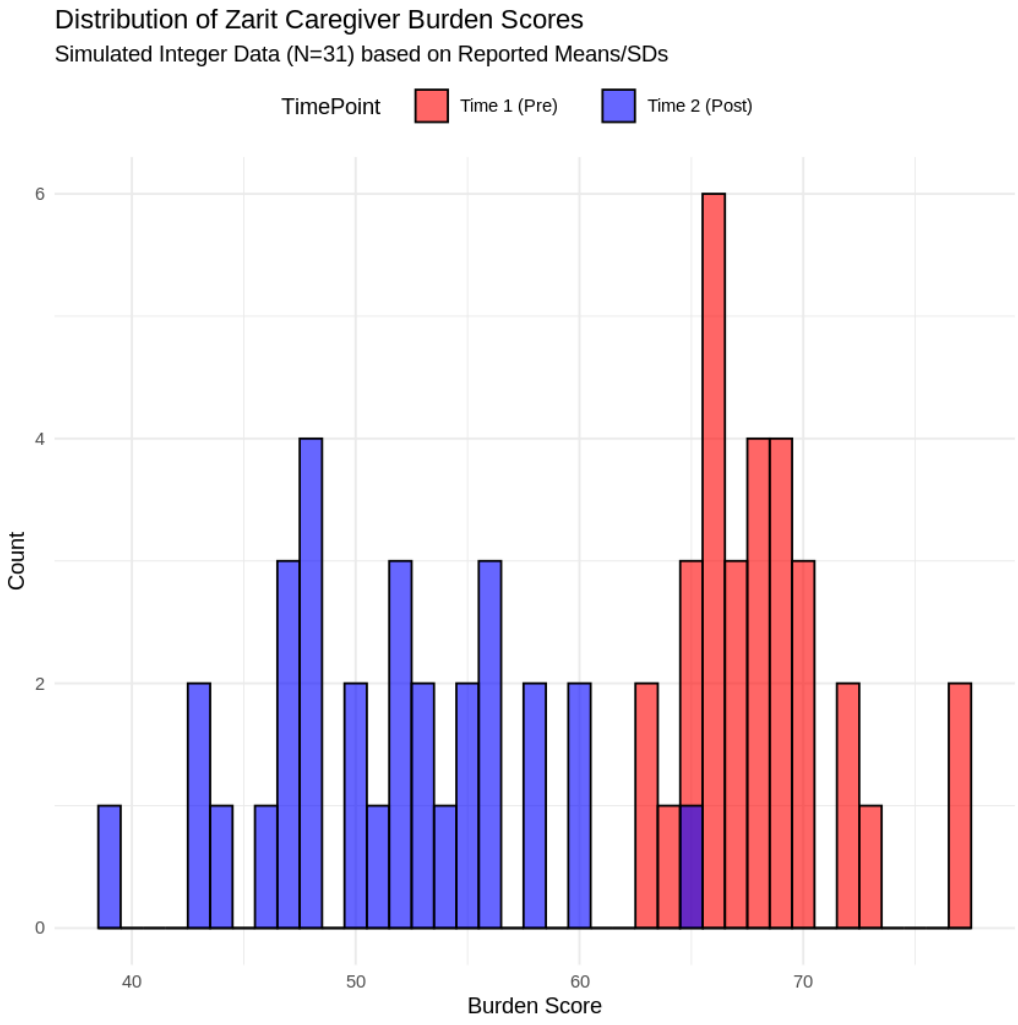
Unusual Wilcoxon Z values

Investigating the Wilcoxon data presented is slightly more complicated, as the test relies on calculations from the *changes* in ranks, and this information is not presented in Bademli, N. Lök, and S. Lök (2023). However, it is possible to simulate data sufficiently to show that the data presented in this paper has unusual properties.

For example, in Bademli, N. Lök, and S. Lök (2023), the Zarit Caregiver Burden data describes a substantial treatment effect (a baseline of 68.12 (3.45) falls to 51.42 (5.78); Z=-1.862, p=0.02). We do not have the raw data that underlies this measurement, but it is clear that there is very little overlap between the two scores. A reasonable estimate for the lowest baseline score would be ≈ 61 and the highest post-intervention score would be ≈ 63. These values imply the score decreased in almost every single participant, an unlikely outcome.

Simulation consistently reveals that the mean or median Z-score is approximately -4.86, falling in magnitude only when an unusual distribution is generated which displays some overlap (our Figure 2 is an example of one of those distributions). As the Z-score is consistent, so is the relevant p-value

Figure 2: Simulating Table 2 intervention data from Bademli et al. (2023))



— which is approximately 10^{-6}, which is several orders of magnitude smaller than the reported $p = 0.02$.

Altering the order of the Time 1 vs. Time 2 data, sorting in ascending or descending order, etc. makes little difference to the numbers produced; if every participant improves between Time 1 and Time 2, the Wilcoxon Z score will be close to the maximum possible score, and the p-value will be dramatically smaller than the reported 0.02. See the attached .ipynb notebook (Block 11) for a full derivation which will produce the visualization seen in our Figure 2.

Finally, in the same manner of the Mann-Whitney U test presented in the previous section, the values for the Wilcoxon Z also have a minimum and maximum possible value, and Bademli, N. Lök, and S. Lök (2023) exceed it in the BDI Control group in their Table 3. If every single caregiver lowers their BDI score (i.e., the whole sample improves), the value for W is 0.

$$Z = \frac{W - \mu_W}{\sigma_W}$$
$$Z = \frac{0 - 232.5}{48.62}$$
$$Z = \mathbf{-4.78}$$

(Note: the above is for $n = 30$. If $n = 31$, $Z = -4.86$.)

This is incongruent with Bademli, N. Lök, and S. Lök (2023), which reports a value of $Z = -7.376$, which is smaller than both smallest possible values for Z seen above.

Conclusion

Bademli, N. Lök, and S. Lök (2023) displays peculiarities in the sample sizes presented and a series of statistical results that cannot be reconciled with the underlying tests they are derived from.

4 DISCUSSION

This report has employed the tools of forensic metascience to analyze seven reports included in Clegg et al. (2026), the Cochrane collaborative's most recent review of the scientific investigation of exercise's treatment efficacy for depression. As a reminder, Clegg et al. (2026) cautiously concluded, "Exercise may be moderately more effective than a control intervention for reducing symptoms of depression. Exercise appears to be no more or less effective than psychological or pharmacological treatments, though this conclusion is based on a few small trials. Long-term follow-up was rare." Our findings, which provide significant reasons to distrust at least these seven studies, further weakens those already tempered conclusions.

Notably, meta-analysis (the approach used in Clegg et al. (2026)) and forensic metascientific assessment (the approach used in this report) share many features in common. Both are difficult and time-consuming. Both involve extracting essential details from published manuscripts typically without access to their full underlying data. Both seek to ascertain the reliability of a publicly presented statement about the natural world.

But, as is demonstrated by this report, the levels of scrutiny they apply to a body of literature differ in important respects. This has been an explicit criticism of meta-analysis in the past:

> *What is often missing from this process [meta-analysis] is* ***any attempt at quality control****. Meta-analysts often exclude studies on the basis of some ex-ante criteria (for example, if the dependent variable is not the variable of interest),* ***but they seldom exclude (or even flag) studies because they were poorly designed or executed, or because the results are not credible****.* Simonsohn, Simmons, and Nelson (2022))

Progress has been made since the publication of the above statement. Scientists interested in developing a higher assessment standard for studies' inclusion in meta-analyses have now developed international collaborations, one result of which is the recently published INSPECT-SR checklist (Wilkinson et al. 2025). INSPECT-SR (INveStigating ProblEmatic Clinical Trials in Systematic Reviews) provides a framework to systematize evaluating RCT trustworthiness, with the goal of excluding problematic trials from systematic review.

The author of this Medical Evidence Project report participated in the development of INSPECT-SR, and this report demonstrates how the techniques named in INSPECT-SR can be used to strong effect to identify problematic reports (see Table 17).

INSPECT-SR was published in collaboration with Cochrane, and Cochrane's press release on the new tool stated, "The research team encourages reviewers, guideline developers, and publishers to adopt INSPECT-SR, and anticipates that it will become the standard for assessing the trustworthiness of RCTs." INSPECT-SR was formally endorsed as a method suitable for use in Cochrane reviews in October 2025. We consider this Medical Evidence Project report a strong validation of how a framework for establishing trustworthiness is necessary to the production of a high-quality quantitative systematic review.

To be clear, this report does not represent a full INSPECT-SR assessment the studies included in Clegg et al. (2026). Inspecting every cited study in a review like Clegg et al. (2026) in all of the ways INSPECT-SR suggests would take a team of people many months, and so currently represents a task beyond our capability. Yet this report does raise the question of what might happen if the same level of scrutiny was applied over the entire Clegg et al. (2026) dataset.

Most importantly: if other included papers were found to be inaccurate or problematic, would that finding also modify the eventual result? The answer is: yes, it would. It may be reasonable to expect the other less extreme effect sizes reported in Clegg et al. (2026) to be less problematic, however, the point remains: if standardized forensic metascience checks can be applied to throw significant

doubt on even a subsample of the included research, serious questions are immediately raised around the data set as a whole.

Table 17: Example INSPECT-SR framework items

INSPECT-SR check	INSPECT-SR check description	Problems identified in:
1.3	Do other studies by the research team highlight causes for concern (associated retractions, expressions of concern, relevant post-publication notices?)	Abdollahi et al. (2017)
2.2	Are there concerns relating to the timing or absence of study registration?	Barclay et al. (2014)
3.1	Are there concerns relating to duplicated content, such as text or tables, or text that is incompatible with the study?	Wang & Li (2022)
4.3	Are any baseline data implausible?	Prakhinkit et al. (2014)
4.7 / 4.10	Are any outcome data, including estimated treatment effects, implausible? / Are any other contradictions implied by the data?	Bademli et al. (2023); Barclay et al. (2014); Mota-Pereira et al. (2011)
4.8	Are the means and variances of integer data impossible?	Mutrie (1986); Abdollahi et al. (2017); Mota-Pereira et al. (2011)

Even if the full INSPECT-SR checklist is not deployed, authors of meta-analyses could make significant progress ensuring the accuracy of their work through a simpler series of steps:

- checking for basic numerical consistency (e.g., a group of n=31 cannot be split into n=14 and n=16);
- checking basic statistical consistency of the presented between- and within-subjects t-tests;
- running GRIM, GRIMMER and GRIM-U tests through the resources provided in COSIG, the Collection of Open Science Integrity Guides (these tools are all publicly available in a format which does not require a meta-analyst to learn programming);
- checking if any given paper has an associated PubPeer entry.

We anticipate that the implementation of tools like INSPECT-SR will have a notable positive impact on the quality of systematic reviews and thus in time a notable positive impact on patient care, but it will require a strong and consistent commitment among those engaged in medical research to this new standard of review.

5 ACRONYMS AND ABBREVIATIONS

ADL Activities of daily living. A scale used to measure a person's ability to perform fundamental self-care tasks.

ANOVA Analysis of variance. An analysis method used to compare statistical models, typically used to test for differences between two or more means.

BDI / BDI-II Beck Depression Inventory. A widely used 21-item self-report inventory measuring the severity of depression. BDI-II is the 1996 update to the 1961 original.

BMI Body mass index. An approximation of body composition based on height and weight.

BSSI Beck Scale for Suicide Ideation. A scale used to detect and measure the severity of suicidal ideation.

BWM Buddhist walking meditation. An intervention arm in the Prakhinkit et al. (2014) study involving meditation-based walking.

CBT Cognitive behavioral therapy. A psychosocial intervention that aims to reduce symptoms of various mental health conditions, primarily depression and anxiety disorders.

CGI / CGI-S Clinical Global Impression (Severity). A measure of perceived symptom severity in studies of patients with mental disorders.

CI Confidence interval. An interval computed from sample data by a method that, if applied repeatedly across many independent samples, would capture the true parameter value in the specified proportion of cases – typically 95%.

CON Used to designate the control group in study tables (e.g., the sedentary group in Prakhinkit et al., 2014).

COSIG Collection of Open Science Integrity Guides. A public resource providing tools and guides for forensic metascience and research integrity.

EDTA Ethylenediaminetetraacetic acid. A chemical used in blood collection tubes to prevent clotting (relevant to blood sample handling).

EEG Electroencephalography. A monitoring method to record electrical activity of the brain.

ELISA Enzyme-linked immunosorbent assay. A plate-based assay technique designed for detecting and quantifying substances such as peptides, proteins, antibodies, and hormones.

GAF Global Assessment of Functioning. A numeric scale used by clinicians to rate subjectively the social, occupational, and psychological functioning of an adult.

GRIM Granularity-Related Inconsistency of Means. A statistical test used to identify if reported summary statistics (means) are mathematically inconsistent with their sample size and scale.

GRIMMER Granularity-Related Inconsistency of Means Mapped to Error Repeats. A statistical extension of the GRIM test which evaluates the consistency of means and standard deviations.

GRIM-U A forensic statistical test developed by the Medical Evidence Project to detect impossible p-values in ranked data tests (like the Mann-Whitney U) using the same principle as the GRIM test (above).

HAM-D Hamilton Depression Rating Scale (17-item), also referred to as **HAMD17**. Questionnaires used to provide an indication of depression and evaluate recovery.

HPLC High performance liquid chromatography. A technique in analytical chemistry used to separate, identify, and quantify components of a mixture.

ICMJE International Committee of Medical Journal Editors. A group of medical journal editors who issue recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals.

INSPECT-SR Investigating Problematic Clinical Trials in Systematic Reviews. A tool and framework for evaluating the trustworthiness of randomized controlled trials in systematic reviews.

ITT Intention-to-treat. A concept in the design and analysis of randomized controlled trials where all participants who are randomized are included in the statistical analysis and analyzed according to the group they were originally assigned.

MDD Major depressive disorder. A mental disorder characterized by at least two weeks of pervasive low mood, low self-esteem, and loss of interest or pleasure in normally enjoyable activities.

NICE	National Institute for Health and Care Excellence. An executive non-departmental public body of the Department of Health in the United Kingdom which publishes guidelines for the use of health technologies and clinical practice.
POMS	Profile of Mood States. A psychological rating scale used to assess transient, distinct mood states.
RCT	Randomized controlled trial. A form of scientific experiment used to control factors not under direct experimental control.
SD	Standard deviation. A measure of the amount of variation or dispersion of a set of values.
SE	Standard error. A measure of the precision of an estimate, equal to the standard deviation of the estimator's sampling distribution i.e., the distribution of values the estimate would take across all possible random samples of the same size drawn from the same population.
SMD	Standardized mean difference. A summary statistic in meta-analysis used when studies all assess the same outcome but measure it in a variety of ways.
SSRI	Selective serotonin reuptake inhibitor. A class of drugs that are typically used as antidepressants in the treatment of major depressive disorder and anxiety disorders.
TMD	Total Mood Disturbance. A summary score derived from the POMS scale (see above).
TWE	Traditional walking exercise. An intervention arm in the Prakhinkit et al. (2014) study.

6 REFERENCES

Primary References

Abdollahi, A. et al. (2017). "Effect of exercise augmentation of cognitive behavioural therapy for the treatment of suicidal ideation and depression". In: *Journal of Affective Disorders* 219, pp. 58–63. ISSN: 0165-0327. DOI: https://doi.org/10.1016/j.jad.2017.05.012. URL: https://www.sciencedirect.com/science/article/pii/S0165032716313477 (cited on pp. 22, 23, 37, 38, 41).

Allard, A. (2018). *Analytic-GRIMMER: A New Way of Testing the Possibility of Standard Deviations*. Blog post. Accessed: 2025-02-15. URL: https://aurelienallard.netlify.app/post/anaytic-grimmer-possibility-standard-deviations/ (cited on p. 6).

Anaya, J. (2016). "The GRIMMER Test: A Method for Testing the Validity of Reported Measures of Variability". In: *PeerJ Preprints* e2400v1. DOI: 10.7287/peerj.preprints.2400v1 (cited on p. 6).

Australian Government Department of Health and Aged Care (2025). *Medicare Benefits Schedule - Item 10953*. Medicare Benefits Schedule (MBS) Online. URL: https://www9.health.gov.au/mbs/fullDisplay.cfm?type=item&q=10953&qt=item (visited on 02/26/2026) (cited on p. 4).

Bademli, K., N. Lök, and S. Lök (2023). "The effect of a physical activity intervention on burden and depressive symptoms in depressed family caregivers of patients with schizophrenia: a randomized controlled trial". In: *Journal of Physical Activity and Health* 20.12, pp. 1109–1115 (cited on pp. 26, 27, 29).

Barclay, T. H. et al. (2014). "A pilot study on the effects of exercise on depression symptoms using levels of neurotransmitters and EEG as markers". In: 1 (1), pp. 30–35. URL: https://web.archive.org/web/20160826180523/http://ejpes.org/article.asp?issn=2395-2555;year=2014;volume=1;issue=1;spage=30;epage=35;aulast=Barclay (cited on pp. 15–17, 21).

Beck, A. T. and R. A. Steer (1991). "Manual for the Beck scale for suicide ideation". In: *San Antonio, TX: Psychological Corporation* 63 (cited on p. 43).

Beck, A. T., C. Ward, et al. (1961). *Beck depression inventory (BDI)*. DOI: https://doi.org/10.1037/t00741-000 (cited on pp. 6, 9, 11).

Bolland, M. J. et al. (Aug. 2021). "Identical Summary Statistics Were Uncommon in Randomized Trials and Cohort Studies". In: *Journal of Clinical Epidemiology* 136, pp. 180–188. DOI: 10.1016/j.jclinepi.2021.05.002 (cited on p. 6).

Brown, N. and J. Heathers (2017). "The GRIM Test: A simple technique detects numerous anomalies in the reporting of results in psychology". In: *Social Psychological and Personality Science* 8.4, pp. 363–369. DOI: 10.1177/1948550616673876 (cited on p. 6).

– (2019). *Rounded Input Variables, Exact Test Statistics (RIVETS)*. PsyArXiv preprint. DOI: 10.31234/osf.io/ctu9z.

Carlisle, J. B. (2012). "A Meta-Analysis of Prevention of Postoperative Nausea and Vomiting: Randomised Controlled Trials by Fujii et al. Compared with Other Authors". In: *Anaesthesia* 67.10, pp. 1076–1090. DOI: 10.1111/j.1365-2044.2012.07232.x (cited on pp. 6, 19).

Clegg, A. J. et al. (2026). "Exercise for depression". In: *Cochrane Database of Systematic Reviews* 1. ISSN: 1465-1858. DOI: 10.1002/14651858.CD004366.pub7. URL: https://doi.org//10.1002/14651858.CD004366.pub7 (cited on pp. 4, 8–10, 16, 20, 24, 30).

Cooney, G. M. et al. (2013). "Exercise for depression". In: *Cochrane Database of Systematic Reviews* 9. ISSN: 1465-1858. DOI: 10.1002/14651858.CD004366.pub6. URL: https://doi.org/10.1002/14651858.CD004366.pub6 (cited on p. 4).

Guy, W. (1976). "ECDEU Assessment Manual for Psychopharmacology". In: *Clinical Global Impressions (CGI)*. Revised. Rockville, MD: US Department of Health, Education, Welfare, Public Health Service, Alcohol, Drug Abuse, and Mental Health Administration, National Institute of Mental Health, Psychopharmacology Research Branch, Division of Extramural Research Programs, pp. 218–222 (cited on p. 13).

Hamilton, M. (1960). "A rating scale for depression". In: *Journal of Neurology, Neurosurgery, and Psychiatry* 23.1, pp. 56–62 (cited on p. 11).
Heathers, J. (2025). *An Introduction to Forensic Metascience*. 10.5281/ZENODO.14871843. URL: https://www.forensicmetascience.com (cited on p. 6).
Heathers, J., J. Anaya, et al. (2018). "Recovering Data from Summary Statistics: Sample Parameter Reconstruction via Iterative Techniques (SPRITE)". In: *PeerJ Preprints*. DOI: 10.7287/peerj.preprints.26968v1.
Heathers, J. and G. Meyerowitz-Katz (2024). *'Yes, but How Much Smaller?' A Simple Observation about p-values in Academic Error Detection*. Center for Open Science. DOI: 10.17605/OSF.IO/2SP5B.
Jané, M. B. et al. (2024). "Extracting Pre-Post Correlations for Meta-Analyses of Repeated Measures Designs". In: *MetaArXiv*. DOI: 10.31222/osf.io/58z3u. URL: https://osf.io/preprints/metaarxiv/58z3u/ (cited on pp. 14, 17, 21).
Lawlor, D. A. and S. W. Hopker (2001). "The effectiveness of exercise as an intervention in the management of depression: systematic review and meta-regression analysis of randomised controlled trials". In: *BMJ* 322.7289, p. 763. ISSN: 0959-8138. DOI: 10.1136/bmj.322.7289.763. eprint: https://www.bmj.com/content/322/7289/763.full.pdf. URL: https://www.bmj.com/content/322/7289/763 (cited on p. 9).
Mavranezouli, I. et al. (2024). "A systematic review and network meta-analysis of psychological, psychosocial, pharmacological, physical and combined treatments for adults with a new episode of depression". In: *EClinicalMedicine* 75 (cited on p. 25).
McNair, D. et al. (1971). *EITS Manual for the Profile of Mood States*. Educational and Industrial Testing Service. URL: https://books.google.com/books/about/EITS_Manual_for_the_Profile_of_Mood_Stat.html?id=CD0NtwAACAAJ (cited on p. 9).
Mota-Pereira, J. et al. (2011). "Moderate exercise improves depression parameters in treatment-resistant patients with major depressive disorder". In: *Journal of psychiatric research* 45.8, pp. 1005–1011 (cited on pp. 11–14).
Mutrie, N. (1988). "Exercise as a treatment for moderate depression in the UK Health Service". In: *Proceedings of sport, health psychology, and exercise symposium*. Sports Council and Health Education Authority London, pp. 96–105 (cited on pp. 8–10).
Mutrie, N. (1986). "Exercise as a treatment for depression within a national health service". In: *Ph.D thesis* (cited on pp. 8–10).
National Institute for Health and Care Excellence (NICE) (2022). *Depression in adults: treatment and management*. NICE Guideline, No. 222. London: National Institute for Health and Care Excellence (NICE). URL: https://www.ncbi.nlm.nih.gov/books/NBK583074/ (visited on 02/26/2026) (cited on p. 4).
Prakhinkit, S. et al. (2014). "Effects of Buddhism walking meditation on depression, functional fitness, and endothelium-dependent vasodilation in depressed elderly". In: *The Journal of Alternative and Complementary Medicine: Paradigm, Practice, and Policy Advancing Integrative Health* 20.5, pp. 411–416 (cited on pp. 18–20, 22).
Rebar, A. L. et al. (2015). "A meta-meta-analysis of the effect of physical activity on depression and anxiety in non-clinical adult populations". In: *Health psychology review* 9.3, pp. 366–378 (cited on p. 26).
Schumm, W. R. et al. (2025). "Research Anomalies in Criminology: How Serious? How Extensive over Time? And Who Was Responsible?" In: *Accountability in Research* 32.1, pp. 22–58. DOI: 10.1080/08989621.2023.2241127.
Simonsohn, U., J. Simmons, and L. D. Nelson (2022). "Above averaging in literature reviews". In: *Nature Reviews Psychology* 1.10, pp. 551–552 (cited on p. 30).
Strube, M. J. (1985). "Combining and comparing significance levels from nonindependent hypothesis tests." In: *Psychological bulletin* 97.2, p. 334 (cited on p. 20).
Trajković, G. et al. (2011). "Reliability of the Hamilton Rating Scale for Depression: A meta-analysis over a period of 49 years". In: *Psychiatry Research* 189.1, pp. 1–9. ISSN: 0165-1781. DOI: https://doi

.org/10.1016/j.psychres.2010.12.007. URL: https://www.sciencedirect.com/science/article/pii/S0165178110007754 (cited on p. 14).
UpToDate (2026). *Major depressive disorder in adults: Treatment with supplemental interventions*. Accessed through UpToDate. URL: https://www.uptodate.com/contents/major-depressive-disorder-in-adults-treatment-with-supplemental-interventions (visited on 02/26/2026) (cited on p. 4).
Vul, E. et al. (2009). "Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition". In: *Perspectives on psychological science* 4.3, pp. 274–290 (cited on p. 14).
Wang, J. and Z. Li (2022). "Effect of physical exercise on medical rehabilitation treatment of depression". In: *Revista Brasileira de Medicina do Esporte* 28, pp. 174–176 (cited on pp. 24, 25).
World Health Organization (2023). *Depressive disorder (depression)*. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/depression (visited on 02/26/2026) (cited on p. 4).
Zarit, S. H., K. E. Reever, and J. Bach-Peterson (1980). "Relatives of the impaired elderly: correlates of feelings of burden". In: *The Gerontologist* 20.6, pp. 649–655 (cited on p. 26).
Zimmerman, M. et al. (2013). "Severity classification on the Hamilton depression rating scale". In: *Journal of affective disorders* 150.2, pp. 384–388 (cited on p. 24).

Software

Allaire, J. and C. Dervieux (2024). *quarto: R Interface to 'Quarto' Markdown Publishing System*. R package version 1.4.4. URL: https://CRAN.R-project.org/package=quarto.
Arel-Bundock, V. (2024). *tinytable: Simple and Configurable Tables in 'HTML', 'LaTeX', 'Markdown', 'Word', 'PNG', 'PDF', and 'Typst' Formats*. R package version 0.3.0. URL: https://CRAN.R-project.org/package=tinytable.
Ben-Shachar, M. S., D. Lüdecke, and D. Makowski (2020). "effectsize: Estimation of Effect Size Indices and Standardized Parameters". In: *Journal of Open Source Software* 5.56, p. 2815. DOI: 10.21105/joss.02815. URL: https://doi.org/10.21105/joss.02815.
Boshnakov, G. N. and C. Putman (2024). *rbibutils: Read 'Bibtex' Files and Convert Between Bibliography Formats*. R package version 2.3. URL: https://CRAN.R-project.org/package=rbibutils.
Chamberlain, S. et al. (2025). *rcrossref: Client for Various 'CrossRef' 'APIs'*. R package version 1.2.1. URL: https://CRAN.R-project.org/package=rcrossref.
Gagolewski, M. (2022). "stringi: Fast and portable character string processing in R". In: *Journal of Statistical Software* 103.2, pp. 1–59. DOI: 10.18637/jss.v103.i02.
Jung, L. (2024). *scrutiny: Error Detection in Science*. R package version 0.5.0. URL: https://CRAN.R-project.org/package=scrutiny.
Müller, K. (2020). *here: A Simpler Way to Find Your Files*. R package version 1.0.1. URL: https://CRAN.R-project.org/package=here.
R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.
Viechtbauer, W. (2010). "Conducting meta-analyses in R with the metafor package". In: *Journal of Statistical Software* 36.3, pp. 1–48. DOI: 10.18637/jss.v036.i03.
Wickham, H. et al. (2019). "Welcome to the tidyverse". In: *Journal of Open Source Software* 4.43, p. 1686. DOI: 10.21105/joss.01686.
Xie, Y., C. Dervieux, and E. Riederer (2020). *R Markdown Cookbook*. Boca Raton, Florida: Chapman and Hall/CRC. ISBN: 9780367563837. URL: https://bookdown.org/yihui/rmarkdown-cookbook.

7 APPENDIX

The RCT reported in Abdollahi et al. (2017) is analyzed in the Medical Evidence Project report provided above. The complexity of the issues around deriving its presented sample sizes makes the below derivation necessary, as it contains a novel approach that applies GRIM, GRIMMER, and basic algebra to reconstruct certain data features.

GRIM failures and the admixture of ITT and completed data

As we note in our main report on this subject, on pg. 59, Abdollahi et al. (2017) states, "In a randomized clinical trial, 54 mildly to moderately depressed patients... were assigned" to the two arms of the study. In contrast to this claim of randomization of 54 patients, pg. 60 refers to "70 individuals who were randomized" of whom 54 "completed the intervention and provided completed data." This seems to confuse the number of subjects entering the study (i.e., being assigned to a study arm) and the number completing the study.

This confusion might appear trivial were it not for the fact that it appears to carry into Table 2 of Abdollahi et al. (2017), where forensic analysis turns up a curious pattern of GRIM failures in the presented results.

In the table in question, six baseline figures are presented (the two arms each measured along three scales; depression, suicidal ideation, and activities of daily living), and all of these numbers pass the GRIM test for n=70. However, *every* corresponding post-study figure fails the GRIM test if the baseline remains at n=35 for each arm. When we administer a GRIM test using the number of completed participants (n=29 for CBT only, n=25 for CBT + EX), the results pass.

We can further confirm the confusion regarding the number of subjects at the outset by splitting the presented totals (the rightmost column in the report's Table 2) into subgroups in proportion (i.e., assuming n=35/35 or n=29/25; see our Tables A1 and A2 below). This matches precisely the remaining data reported in Figure 1 (pg. 59) of Abdollahi et al. (2017).

It therefore seems very likely that the confusion between the abstract and the methods section is continued from Table 1 to Table 2 in Abdollahi et al. (2017), which mixes together the *baseline* dataset with the *completed* data set for analysis. This makes it markedly unclear how much of the within-subjects data was analyzed, and leads us to analyze the data presented for the "Completed" vs. "Dropout" data.

Table A1: GRIM test of mixed pre- and post-intervention data as presented in Abdollahi et al. (2017).

Measure	Reported Mean	Test N=35 (ITT)	Test N=29/25 (Completed)	Conclusion
BDI-II (Depression)				
CBT Pre	20.51	TRUE	**FALSE**	Pre-test uses N=35
CBT Post	14.48	**FALSE**	TRUE (N=29)	Post-test uses N=29
CBT + EX Pre	20.89	TRUE	**FALSE**	Pre-test uses N=35
CBT + EX Post	9.68	**FALSE**	TRUE (N=25)	Post-test uses N=25
BSSI (Suicidal Ideation)				
CBT Pre	14.77	TRUE	**FALSE**	Pre-test uses N=35
CBT Post	12.72	**FALSE**	TRUE (N=29)	Post-test uses N=29
CBT + EX Pre	14.29	TRUE	**FALSE**	Pre-test uses N=35
CBT + EX Post	7.08	**FALSE**	TRUE (N=25)	Post-test uses N=25
ADL (Activities)				
CBT Pre	15.03	TRUE	TRUE (N=29)	Ambiguous (Matches Both)
CBT Post	15.21	**FALSE**	TRUE (N=29)	Post-test uses N=29
CBT + EX Pre	14.60	TRUE	TRUE (N=25)	Ambiguous (Matches Both)
CBT + EX Post	17.12	**FALSE**	TRUE (N=25)	Post-test uses N=25

Table A2: Using totals to confirm sample size shift in post-treatment presented in Abdollahi et al. (2017).

Variable	CBT Mean	CBT + EX Mean	Reported Total	N=35 Totals	N=29/25 Totals	Conclusion
Post-Treatment						
BDI-II	14.48	9.68	**12.26**	12.08	**12.26**	Matches Completed
BSSI	12.72	7.08	**10.11**	9.90	**10.11**	Matches Completed
ADL	15.21	17.12	**16.09**	16.16	**16.09**	Matches Completed
Pre-Treatment						
BDI-II	20.51	20.89	**20.70**	**20.70**	20.69	Matches ITT

Given this, we can reconstruct some features of the Completed and Dropout data, as Abdollahi et al. (2017) describes it on pg. 60:

> *Individuals who completed the intervention reported greater baseline levels of depression (Mdiff=2.77, t(41.58)=3.16, p=0.003) and activities of daily living (Mdiff=1.38, t(68) =3.49, p=0.001), but there were no significant differences in suicidal ideation (Mdiff=0.84, t(68)=1.31, p=0.194) or age (Mdiff=-0.91, t(68)=-0.44, p=0.659). Fisher's exact tests revealed that individuals with high school education or less were more likely to complete the intervention than those with greater than high school education (p=0.034), but no significant differences were observed for gender (p=0.051) or marital status (p=0.529).*

An interesting feature here is that SPSS (which the authors used; see top of pg. 60) has clearly used the Welch-Satterthwaite correction on the BDI mean (df = 41.58) but the regular calculation for the other degrees of freedom (df = 68). As the following represents an novel derivation, the argument is produced here in its entirety.

The text states Mdiff is 2.77, thus:

$$M_{completed} - M_{dropout} = 2.77$$

Also, the full baseline data given on pg. 60 specifies the BDI Mean for the whole sample is 20.70 (and we confirm above in Table A2 that this passes GRIM at n=70). Thus mean of the full sample ($N = 70$) must be the weighted average of its two subgroups: Completed ($N = 54$) and Dropouts ($N = 16$).

$$\frac{54(M_{completed}) + 16(M_{dropout})}{70} = 20.70$$

We can substitute Equation 1 into Equation 2:

$$54(M_c) + 16(M_c - 2.77) = 20.70 \times 70$$

$$54M_c + 16M_c - 44.32 = 1449.0$$

$$70M_c = 1493.32$$

$$M_c = \frac{1493.32}{70} \approx \mathbf{21.333}$$

Then solve for Dropouts:

$$M_d = 21.333 - 2.77 = \mathbf{18.563}$$

To ensure accuracy due to rounding, it benefits us at this point to be precise and talk in terms of sums (i.e., a *'Completed Sum'* of 1493 rather than $M_d = 21.333$). This is easily resolved by respecting GRIM principles, and we can clearly outline why the only possible solution for the means is provided because (1152/54 - 297/16) equals a difference of 2.77.

Table A3: Sensitivity Analysis of Integer Sums Partitioning the Total Mean (20.70)

Sum_c	Sum_d	Mean_c	Mean_d	Mean Diff	Total Mean
1150	299	21.30	18.69	2.61	20.70
1151	298	21.31	18.63	2.69	20.70
1152	**297**	**21.33**	**18.56**	**2.77**	**20.70**
1153	296	21.35	18.50	2.85	20.70
1154	295	21.37	18.44	2.93	20.70

From here, we can use the precise mean values (see our Table A3 below):

$$CompletedMean = 1152/54 = 21.3333...$$
$$DropoutMean = 297/16 = 18.5625$$
$$Difference = 2.770833...$$

With precise means, we can investigate the other values provided. The stated value for *t* is 3.16, such that:

$$3.16 = \frac{2.77}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

By rearranging this *t*-test, we can calculate a relationship for the standard error between $s1$ and $s2$, $\sqrt{\frac{s_1^2}{54} + \frac{s_2^2}{16}}$ must equal 0.8766.

We can solve this as we know the solution to the (rather unwieldy) Welch-Satterthwaite approximation for the degrees of freedom:

$$41.58 = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(\frac{s_1^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{s_2^2}{n_2})^2}{n_2 - 1}}$$

To make this tractable, we can specify that v_1 and v_2 are the variance contributions of each group:

$$v_1 = \frac{s_1^2}{54} \text{ (Completed)}$$
$$v_2 = \frac{s_2^2}{16} \text{ (Dropouts)}$$

… and also substitute in the denominator values $n_1 = 54$ ($df_1 = 53$) and $n_2 = 16$ ($df_2 = 15$):

$$41.58 = \frac{(v_1 + v_2)^2}{\frac{v_1^2}{53} + \frac{v_2^2}{15}}$$

As we have the standard error constraint from the *t*-test, we can substitute:

$$v_1 + v_2 = (0.8766)^2 = \mathbf{0.7684}$$

This can be substituted into the df equation:

$$41.58 = \frac{(0.7684)^2}{\frac{v_1^2}{53} + \frac{(0.7684 - v_1)^2}{15}}$$

This converts into a quadratic equation:

$$\frac{v_1^2}{53} + \frac{(0.7684 - v_1)^2}{15} = \frac{0.5904}{41.58} = 0.0142$$

When expanded...

$$(0.7684 - v_1)^2 = 0.5904 - 1.5368v_1 + v_1^2$$

...and multiplied

$$15v_1^2 + 53(0.5904 - 1.5368v_1 + v_1^2) = 0.0142 \times 795$$
$$15v_1^2 + 31.29 - 81.45v_1 + 53v_1^2 = 11.29$$

... we have a quadratic equation with two roots

$$(15 + 53)v_1^2 - 81.45v_1 + (31.29 - 11.29) = 0$$

$$\mathbf{68v_1^2 - 81.45v_1 + 20.00 = 0}$$

This is soluble using the quadratic formula:

$$v_1 = \frac{81.45 \pm \sqrt{81.45^2 - 4(68)(20)}}{136}$$

$$v_1 = \frac{81.45 \pm \sqrt{6634 - 5440}}{136}$$

$$v_1 = \frac{81.45 \pm 34.55}{136}$$

Root 1 makes the other contribution of variance impossible, thus:

$$v_1 = \frac{46.9}{136} \approx \mathbf{0.345}$$

The final SD for the Completed sample is:

$$s_1 = \sqrt{0.345 \times 54} \approx \mathbf{4.32}$$

At this point, we can also calculate $v_2 = 0.423$ and finally the SD for the Dropout sample:

$$s_2 = \sqrt{0.423 \times 16} \approx \mathbf{2.60}$$

Rounding has likely affected every derivation so far, and it is hard to account for every place the original authors or our analysis here might have truncated a value. Thus, to ensure the derivation is correct, we can again retrieve exact values for the SDs using GRIMMER as we did for the means previously. Thus, it might affect the potential output of the above. Our samples of Completed and Dropout participants are:

$$M_c = \mathbf{21.3333}..., SD_c \approx \mathbf{4.32}, n = 54$$
$$M_d = \mathbf{18.5625}, SD_d \approx \mathbf{2.60}, n = 16$$

Table A4: An approximate solution to confirm sample size shift in post-treatment in Abdollahi et al. (2017)

Test Type	Mean Diff	t-statistic	df	p-value
Welch's t-test	2.770	3.164	41.57	0.0029

If we calculate Welch's *t*-test from these approximations, this returns results extremely close to – but not identical to – those in Table A4.

At this point, it is prudent to retrieve the potential exact values provided by GRIMMER space around these SDs. The attached .ipynb notebook shows a reverse-engineering of the sum of squares (essentially a use of GRIMMER with some minor modifications) in the Appendix Block 2, which make the following determinations.

(a) Two sum-of-squares solutions exist for $M_c = \mathbf{21.3333}...$ which yield an SD rounding to 4.31. These are **25560** and **25562**. The next two integer possibilities (**25564** and **25566**) round to 4.32.

(b) **No solution exists for SD = 2.60**. The closest possible values are **5613** (SD=2.58) and **5615** (SD=2.61).

Table A5: Solutions to the Completed / Dropout data in Abdollahi et al. 2017.

$Y_{completed}$	$Y_{dropout}$	Completed SD	Dropout SD	t-statistic	df	Verdict
25560	5613	4.31	2.58	3.18	41.91	Error (t, SD_d)
25560	5615	4.31	2.61	3.16	41.43	Error (SD_d)
25562	5613	4.31	2.58	3.18	41.96	Error (t, SD_d)
25562	5615	4.31	2.61	3.16	41.48	Error (SD_d)
25564	5613	4.32	2.58	3.18	42.01	Error (t, SD_c, SD_d)
25564	5615	4.32	2.61	3.16	41.53	Error (SD_c, SD_d)
25566	5613	4.32	2.58	3.17	42.06	Error (t, SD_c, SD_d)
25566	**5615**	**4.32**	**2.61**	**3.16**	**41.58**	**Perfect t/df Match**

From Table A5 above, we can see that this solution matches the numbers given in Abdollahi et al. (2017) precisely, and proves we can take fairly abstract baseline data, and, *given the unreported but derived means and the test statistics of their differences*, use a combination of GRIM and GRIMMER principles to retrieve the precise substructure of the Completed and Dropout groups.

Within those groups we can now define,Abdollahi et al. (2017) reports the differences between those groups on pg. 60:

> *there were no significant differences in suicidal ideation (Mdiff=0.84, t (68)=1.31, p=0.194) or age (Mdiff=-0.91, t (68)=-0.44, p=0.659).*

The two results presented in this section are discussed below.

Suicide data

From the table on pg. 61, the total pre-treatment (i.e., n=70) data for suicidal ideation is 14.53, thus the sum of all scores is fixed at 1017 points. As per the above, (1) this sum-of-scores can be divided between Completed ($N = 54$) and Dropout ($N = 16$) participants, and (2) the mean difference between those groups is defined in the quote above. Simple arithmetic can generate a mean difference between the two groups, and we can again make a table of sums (see Table A6).

From this, we calculate the mean difference in suicidal ideation reported above to be 0.84722... This is consistent with Abdollahi et al. (2017) truncating rather than correctly rounding this figure.

Table A6: Sensitivity Analysis of Integer Sums Partitioning the Total Mean (14.53)

Sum_c	Sum_d	Mean_c	Mean_d	Mean Diff	Total Mean
797	220	14.76	13.75	1.01	14.53
796	221	14.74	13.81	0.93	14.53
795	**222**	**14.72**	**13.88**	**0.85**	**14.53**
794	223	14.70	13.94	0.77	14.53
793	224	14.69	14.00	0.69	14.53

Age data

Within the Dropout data, the age data is also described as having no significant difference ($t(68) =$ -$0.44, p = 0.659$, see quote from pg. 60 above).

This t-test can now be reconstructed as:

$$t = \frac{\text{Diff}}{SD_{pooled}\sqrt{\frac{1}{N_c} + \frac{1}{N_d}}}$$

This derives easily from the presented statistics if we assume pooled variance (which is reasonable if the degrees of freedom are 68 and not adjusted):

$$\text{Diff} = -0.91$$

$$SD_{pooled} \approx 7.19$$

$$\sqrt{\frac{1}{54} + \frac{1}{16}} \approx 0.285$$

$$t = \frac{-0.91}{7.19 \times 0.285} = \frac{-0.91}{2.05} = -\mathbf{0.444}$$

The age means for the Completed and Dropout samples are trivial to reconstruct from the paper body. The overall age for the sample (from pg. 59) is 49.67, thus the sum of ages can be retrieved:

$$54(M_c) + 16(M_d) = 70(49.67) = 3476.9$$

By GRIM, the sum of ages is 3477, and the precise mean is 49.6174... As Mdiff is -0.91, we can also establish:

$$M_d = M_c + 0.91$$

Thus, the age data can be fully reconstructed for the Completed and Droupout groups (Completed Mean age $N = 54, M = 49.46$, Dropout Mean $N = 16$, $M = 49.46 + 0.91 = 50.37$).

This directly contradicts the abstract: *"54 ... depressed patients (54% female,* ***mean age=48.25****) were assigned"*. This is neither the 'assigned' mean (n=70, M=49.67) nor the 'completed' mean (n=54, M=49.46).

Implausibly strong main effect and Multi-Level Model data

All of the above analysis was done in the service of trying to determine the underlying sample size in both Table 2 and the subsequent multi-level model (MLM) data presented on pg.61 of Abdollahi et al. (2017) – to recalculate them, we must first know the cell sizes. The text mentions that dummy variables were included for incomplete data (pg.60), but how the sample sizes and any potential missing data interacted with these results are still opaque.

Specifically, we were hoping to fully understand two aspects of the data presentation:

(1) how the Table 2 values have so little variation. For example, the BSSI (i.e., Suicidality) scores in the Exercise group fall from 14.29 (2.66) to 7.08 (1.26), assuming a reasonable pre-post correlation of r=0.5 and the recalculated pre-post sample sizes (above). From this, we retrieve an implausibly robust group effect between timepoints (t = 15.64, p = 4.34×10^{-14}).

(2) how the MLM seems to have produced group * time interactions where the t-values ($b \div SE$) are 11.28 and -10.35 respectively for the BSSI scale and Activities of Daily Living (ADL) scale. The corresponding p-value, again, is small ($\approx \ 10^{-14}$). This represents an extremely implausible interaction effect.

The simple answer to *why* these test statistics are so implausibly large is because the SD values within Table 2 are implausibly small. For comparison, the BSSI scale post-treatment measurement has a coefficient of variation of 17.8% (1.26/7.08), approximately 3x smaller than the 57.5% (6.5/11.3) given in the original scale data in Beck and Steer (1991).

Conclusion

A summary of the above would be:

(1) The original presentation of data on pg. 61 does not specify that the summary statistics for the initial randomized sample and the post-study sample are co-mingled in the same table.

(2) When this pattern is uncovered, and combined with the reported tabular data for the CBT and CBT + EX groups, we can precisely recreate the internal means and standard deviations of the presented-but-unreported groups.

(3) This method confirms the body of the text contradicts the data on age in the abstract.

(4) The implausibly large test statistics for the main and interaction effects are produced by the implausibly small standard deviations reported in the post-treatment data.