

Audit Clinical Evidence

AUDIT REPORT

ACE2026002

Audit of Singh et al. (1997)

A randomized controlled trial of progressive resistance training in depressed elders

PUBLISHED

2026-10-12

MODIFIED

2026-10-12

POPULATION

OLDER ADULTS

AUTHORS

TREATMENT

EXERCISE

Matthew B. Jané

University of Connecticut (Storrs, CT, USA)

matthew@auditclinicalevidence.org

OUTCOME

DEPRESSION

Kutenda F. Mvududu

University of Massachusetts Boston (Boston, MA, USA)

kutenda@auditclinicalevidence.org

DESIGN

RCT

VERSION DOI

<https://doi.org/10.5281/zenodo.23331179>

ALL VERSIONS

<https://doi.org/10.5281/zenodo.23331178>

COPYRIGHT

© 2026 Matthew B. Jané

LICENSE

Creative Commons Attribution 4.0 International (CC BY 4.0)

The views expressed herein are solely those of the authors and do not represent their affiliated organizations. Full disclaimers and editorial policy at auditclinicalevidence.org/disclaimers.

Audit of Singh et al. (1997)

A randomized controlled trial of progressive resistance training in depressed elders

AUTHORS

Matthew B. Jané¹ Kutenda F. Mvududu²

¹University of Connecticut (Storrs, CT, USA) ²University of Massachusetts Boston (Boston, MA, USA)

ABSTRACT

This report audits a randomized controlled trial by Singh, Clements, and Fiatarone (1997) that compared 10 weeks of progressive resistance training with an attention control condition in 32 community-dwelling adults aged 60 and over with major depression, minor depression, or dysthymia. Every table in the article labels the values reported in parentheses as standard deviations, yet the Methods state that they are standard errors. The article's test statistic values support the Methods (i.e., the parenthetical values are standard errors). The reported age ranges are impossible under the reported standard deviations. Additionally, the depression subgroup means in Table 4 are impossible under the whole-group standard deviations in Table 3. The largest post-test standardized mean difference that the article can support on any depression outcome is -0.72 . We located 24 meta-analyses that report a study-level effect size for this trial. Of these, 19 meta-analyses assign Singh (1997) a value larger than any possibly correct extraction of the effect size. The median assigned value is -2.74 and the largest is -5.04 . Four report exactly -1.75 , which is due to meta-analysts treating the parenthetical values of the BDI scores as standard deviations. One of these meta-analyses is a Cochrane review. Furthermore, five of the 10 categorical baseline p values can not be reproduced by any standard test of the reported contingency table cells. Later in the paper there is a reported subgroup mean that is impossible given its sample size. In total, this audit identifies 110 instances of Research Integrity Violations across two categories. Due to the pervasive contamination the misreporting has had on the evidence syntheses in this field, we recommend retraction.

SECTIONS

1	Executive summary	2
1.A	Description of audited study	2
1.B	Impact of audited study	2
1.C	Summary of findings	7
2	The values in parentheses are standard errors	7
2.A	The labels and the Methods disagree	7
2.B	Values that can not be standard deviations	8
2.C	Reported p values reproduce only as standard errors	10
3	Reported statistics that no reading can recover	13
3.A	Baseline p values that are not baseline comparisons	13
3.B	p values that can not be reproduced under any reading	16
3.C	Training intensity SDs that their ranges do not allow	18
3.D	GRIM failure in a subgroup mean	19
4	Conclusion	20
	References	22
	Version history	24

1 | Executive summary

1.A Description of audited study

This report audits Singh (1997), a highly cited RCT of progressive resistance training as a treatment for depression in older adults. The trial has two arms. The exercise arm received 10 weeks of high-intensity resistance training, three sessions per week ($n = 17$). The attention control arm attended an interactive health education program twice a week for 10 weeks ($n = 15$). Subjects were community-dwelling volunteers aged 60 to 84 who met DSM-IV criteria for major depression, minor depression, or dysthymia. Depression was measured at baseline and at 10 weeks (post-intervention) with the Beck Depression Inventory (BDI), the Hamilton Rating Scale for Depression (HRSD), the Geriatric Depression Scale (GDS), and a count of DSM-IV symptoms. The article also reports quality of life (SF-36), self-efficacy, morale, strength, and physical function. All 32 subjects completed the trial. The authors conclude that resistance training is an effective antidepressant in depressed elders.

The trial was one of the first RCTs of resistance training for depression. Meta-analyses of exercise and depression have included Singh (1997) for nearly three decades. The problems documented in this report concern how the summary statistics are labeled and reported. Those summary statistics are the values that meta-analyses extract.

1.B Impact of audited study

We call a work that cites the audited study *infected*, because its findings incorporate results that this report identifies as unreliable. According to OpenAlex (retrieved 2026-09-20), 440 articles cite the study (infected articles). Of these, 77 are meta-analyses or reviews (infected reviews). The infected reviews have themselves been cited 12450 times, which measures how far the audited results have spread through the evidence base. The audited study or an infected review is cited by 4 clinical guidelines or consensus statements (see Table 1).

Table 1. Citation counts for the audited study

Study	Infected articles	Infected reviews	Spread via reviews	Spread to guidelines
Singh (1997)	440	77	12450	4
Total	440	77	12450	4

Note. Counts are from OpenAlex, retrieved 2026-09-20, and depend on its coverage; other databases (Scopus, Web of Science, Google Scholar) give different totals. A work that cites an audited paper is described here as infected, meaning its findings incorporate that paper's results. *Infected articles* is the number of works indexed by OpenAlex that cite the paper. *Infected reviews* is the subset of those works that OpenAlex types as a review or whose title indicates a meta-analysis or review. *Spread via reviews* is the total number of citations received by the infected reviews in that row, as a measure of how far the paper's results have propagated; a review that cites more than one audited paper is counted in each row, and a work citing several of those reviews is counted once per review, so rows and the total can overlap. *Spread to guidelines* is the number of distinct works whose title indicates a clinical guideline, consensus, or position statement and that cite the paper directly or cite one of its infected reviews; only guidelines published in journals are indexed, so policy documents from agencies such as WHO or NICE are not counted, and a guideline reached from more than one audited paper appears in each row.

The values in parentheses in the article's tables are standard errors (SEs), although every table labels them as standard deviations (SDs), as Section 2 shows. An SD describes how widely the subjects' scores are spread around their group mean, whereas an SE describes how precisely that mean has been estimated and equals the SD divided by the square root of the group size, so in groups of 15 and 17 subjects the SE is about a quarter of the SD. A meta-analyst extracts the mean, the SD, and the sample size of each

group from a results table such as Table 3 of Singh (1997), whose footnote reads “Means and (SD)”, and computes a standardized mean difference (SMD), the difference between the two group means divided by their pooled SD, so that the effect is expressed in SD units. Taken at face value, the post-intervention BDI scores of 9.8 (2.4) and 13.8 (2.0) give an SMD of -1.75 (negative values favor exercise throughout). Read as SEs, the same scores give -0.44 . The two readings differ by a factor of about four on every depression measure, because dividing by an SE in place of an SD shrinks the denominator by $\sqrt{17}$ or $\sqrt{15}$, both close to four. Across the six depression outcomes and under Cohen’s d , Hedges’ g , or Glass’s delta, the three SMD formulas in common use, the largest post-test SMD that the article’s own numbers can support is -0.72 . A post-test effect size beyond -0.72 did not come from a correct reading of the article.

We located 24 meta-analyses that include Singh (1997) and report a study-level effect size for the trial (see Table 2 and Figure 1). Most extracted post-test effects, which compare the two groups’ scores at 10 weeks. Three, Gordon (2018), Dai (2024), and Li (2026), extracted change-score SMDs, which compare how much each group’s scores changed between baseline and 10 weeks. A change-score SMD can be larger than the post-test SMD from the same trial, here because the exercise group started with higher depression scores and fell further. Gordon’s value of -1.03 is the difference between the groups in GDS change divided by the pooled baseline SD. Dai’s value of -1.28 is the difference in BDI change divided by the SD of the change scores, which depends on the correlation between each subject’s baseline and 10-week scores, at a correlation of 0.8. Both reproduce exactly when the values in parentheses are read as SEs. Li does not report the correlation it assumed, but its value of -1.62 is consistent with a change-score SMD under the correct reading at a high correlation. We count all three as defensible, together with the two post-test values within the bound (Lawlor 2001; Liu 2011). The other 19 meta-analyses extracted post-test effects and assign values beyond -0.72 , which no correct post-test extraction can produce. A change-score extraction would not rescue most of them either. A change score standardized by the baseline SD can reach at most -1.72 on any outcome, whereas one standardized by the SD of the change can reach at most -1.89 at a correlation of 0.8. Of the 19, 15 are beyond both bounds. The other four report exactly -1.75 , the value that the post-test misreading produces, as the next paragraph shows. The median effect assigned to this trial is -2.74 and the largest is -5.04 (Yu 2022).

The values that recur across meta-analyses show how the error was made. Four meta-analyses (Cooney 2013; Heinzl 2015; Rhyner 2016; Kvam 2016), one of them the Cochrane review, report exactly -1.75 , the Hedges’ g that results when a meta-analyst treats the values in parentheses for the post-intervention BDI scores as SDs. The values of -2.70 and -2.77 (Krogh 2017; Silveira 2013) are Hedges’ g and Cohen’s d for the post-intervention HRSD scores under the same misreading. In these cases the arithmetic is correct and only the input is wrong, because the meta-analysts did what the table told them to do. The other values beyond the bound can not be reproduced from any single outcome in Table 3 under either reading, so their inputs can not be identified from the values alone. Four of them (Schuch 2016a, 2016b; Hooper 2018; Rossi 2024) report -3.11 or -3.10 . The network meta-analysis of 218 trials by Noetel (2024) publishes its extracted data, in which the baseline and 10-week values of this trial appear with the values in parentheses entered as SDs (1.8, 2.4, 1.7, and 2.0 for the BDI). Its model is arm-based, with each arm’s change from baseline divided by that arm’s baseline SD and corrected for small samples, and gives the exercise arm a standardized change of -6.06 on the BDI against -2.55 for the control arm, a difference of -3.52 (-5.03 on the HRSD). All of these values are larger than any reading of the article’s tables can produce.

Table 2. Effect sizes assigned to Singh, Clements, and Fiatarone (1997) by meta-analyses that include it, against the largest value the article's own numbers support.

Meta-analysis	SMD assigned	Defensible	DOI
Yu (2022)	-5.04	No	10.1186/s40798-022-00529-5
Cunha (2024)	-3.98	No	10.1016/j.psychres.2024.115746
Nebiker (2018)	-3.83	No	10.3389/fpsy.2018.00305
Noetel (2024)	-3.52	No	10.1136/bmj-2023-075847
Heissel (2023)	-3.46	No	10.1136/bjsports-2022-106282
Dauwan (2021)	-3.38	No	10.1007/s00415-019-09493-9
Schuch (2016a)	-3.11	No	10.1016/j.jpsychires.2016.02.023
Schuch (2016b)	-3.11	No	10.1590/1516-4446-2016-1915
Hooper (2018)	-3.11	No	ProQuest (thesis)
Rossi (2024)	-3.10	No	10.1016/j.psychres.2024.115805
Rosenbaum (2014)	-3.05	No	10.4088/JCP.13r08765
Silveira (2013)	-2.77	No	10.1159/000345160
Krogh (2017)	-2.70	No	10.1136/bmjopen-2016-014820
Stubbs (2016)	-2.42	No	10.1007/s40279-015-0441-5
Rethorst (2009)	-2.20	No	10.2165/00007256-200939060-00004
Kvam (2016)	-1.75	No	10.1016/j.jad.2016.03.063
Cooney (2013)	-1.75	No	10.1002/14651858.CD004366.pub6
Heinzel (2015)	-1.75	No	10.1024/1662-9647/a000133
Rhyner (2016)	-1.75	No	10.1123/japa.2015-0146
Li (2026)	-1.62	Yes	10.3389/fpsyg.2026.1815160
Dai (2024)	-1.28	Yes	10.3390/healthcare12151498
Gordon (2018)	-1.03	Yes	10.1001/jamapsychiatry.2018.0572
Liu (2011)	-0.52	Yes	10.3109/09638288.2010.487145
Lawlor (2001)	-0.45	Yes	10.1136/bmj.322.7289.763

Note. SMD assigned is the effect size each meta-analysis reports for this trial in its forest plot or data table, signed so that negative values favor exercise. Defensible is Yes when the assigned value is not beyond -0.72, the largest post-test SMD that any of the article's six depression outcomes can produce under the correct reading, or when the meta-analysis extracted a change-score SMD that the correct reading can produce (Gordon 2018, Dai 2024, Li 2026). Noetel 2024 is an arm-based network meta-analysis. Its value is the exercise arm's standardized change from baseline on the BDI minus the control arm's, from the published data file.

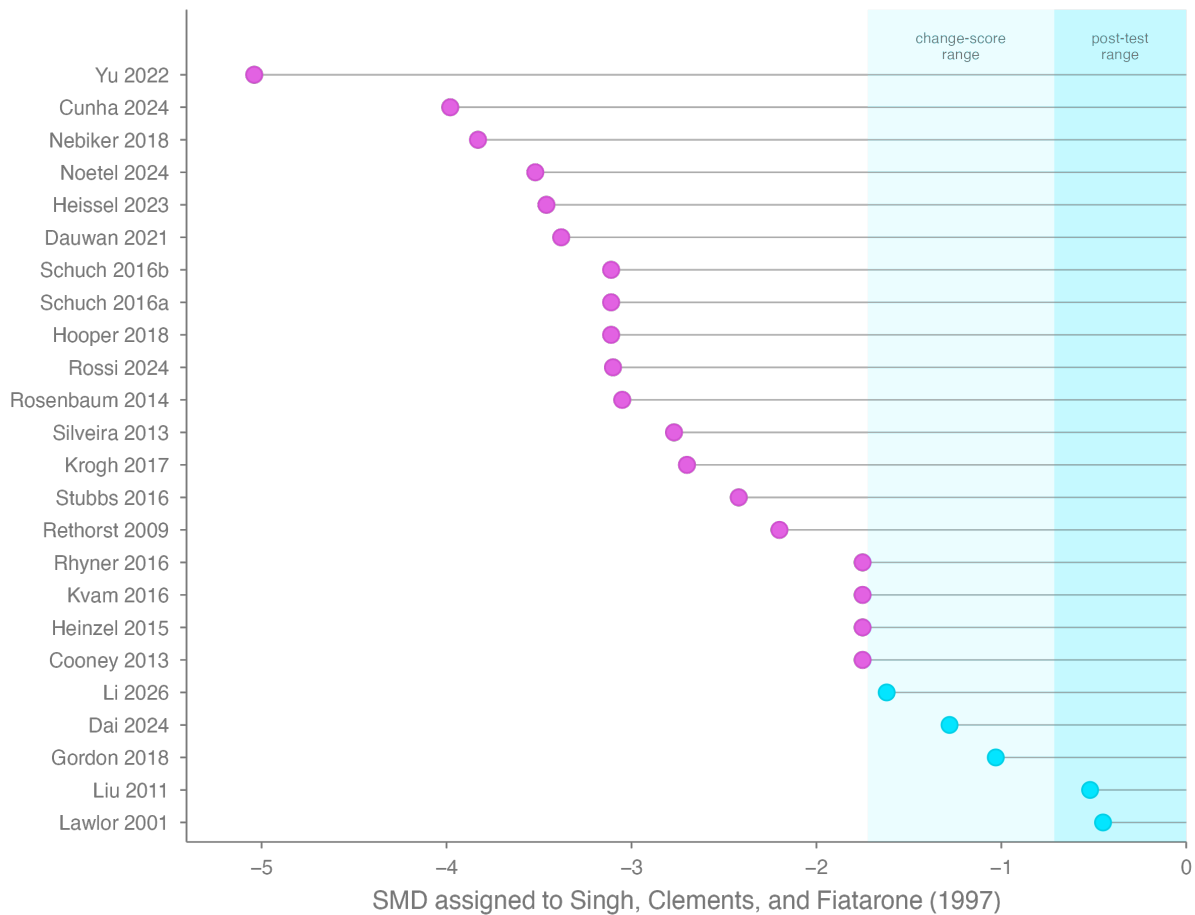


Figure 1. The effect size assigned to Singh, Clements, and Fiatarone (1997) by each of the meta-analyses in Table 2. Negative values favor exercise. The blue ribbon covers the range of post-test SMDs (0 to -0.72) that the article's own numbers support on any depression outcome. The lighter ribbon extends that range to the largest change-score SMD standardized by the baseline SD (-1.72). Blue points are defensible extractions. Magenta points fall beyond any defensible extraction.

Two forest plots, the figures in which a meta-analysis lists each study's effect size beside the pooled estimate, show where this trial's effect sits inside a published synthesis. In the meta-analysis of exercise for depression in older adults by Cunha (2024), this trial enters three times, as "Singh et al., 1997a", "1997b.1", and "1997b.2", with effect sizes of -3.02, -3.55, and -3.98 (Figure 2). These are the three largest effects in the analysis, where the next largest is -1.34 and the pooled estimate is -0.94. In the meta-analysis of neuromuscular exercise interventions by Nebiker (2018), this trial has an effect of -3.83 and a control group of 19 subjects, four more than the trial had (Figure 3). This entry is the largest effect in the analysis, more than twice the next largest.

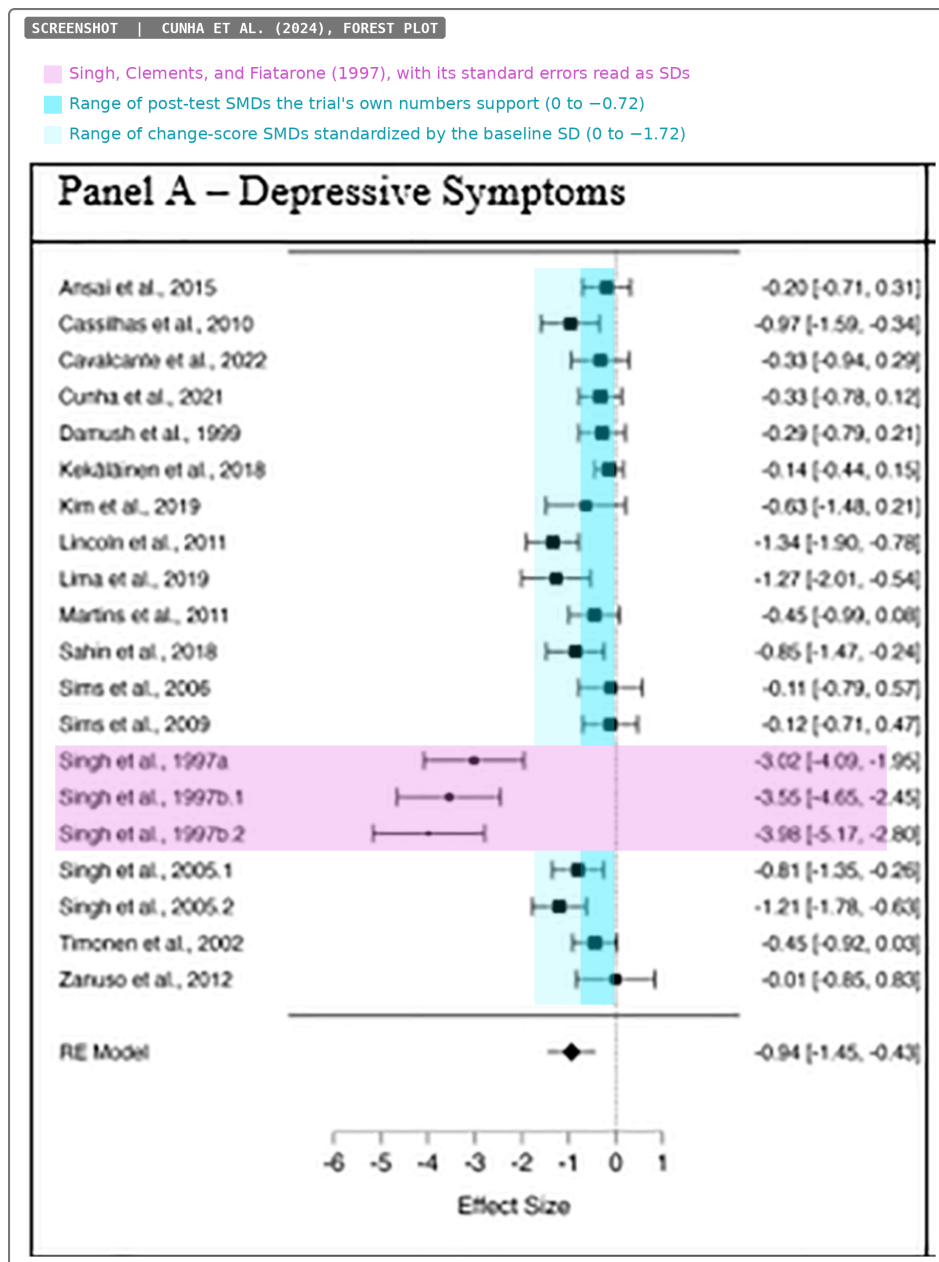


Figure 2. The forest plot from Cunha (2024) for depressive symptoms, annotated. The three entries for this trial are highlighted. The darker band marks the range of post-test SMDs that a correct extraction of the article can reach, and the lighter band the range that a change-score SMD standardized by the baseline SD can reach (0 to -1.72).

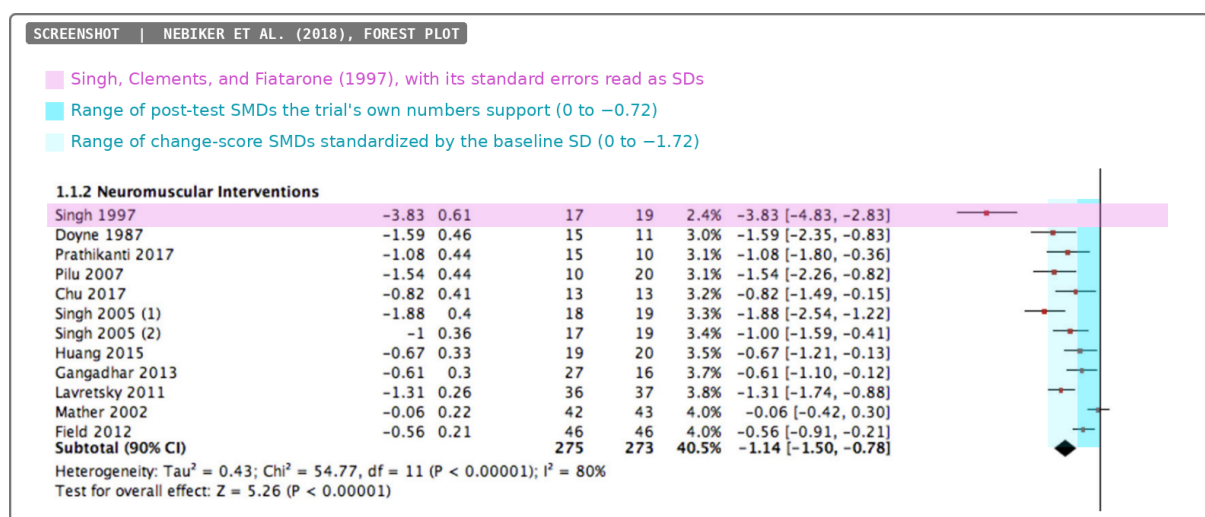


Figure 3. The neuromuscular interventions subgroup of the forest plot from Nebiker (2018), annotated in the same way. This trial is the first row.

For nearly three decades, the evidence base on exercise as a treatment for depression has carried this small trial as one of its largest effects. Read correctly, the trial shows a moderate effect, but its tables label the standard errors as standard deviations and meta-analysts extracted them as labeled.

1.C Summary of findings

The values in parentheses throughout the article are standard errors, although every table labels them as standard deviations. The article's own age ranges, subgroup means, and test statistics show this (Section 2). Read as labeled, the trial's effect on depression is about four times larger than the data support. Of the 24 meta-analyses we located, 19 carry the inflated effect. Separately, the column labeled "Baseline p -value" in the outcome tables does not report baseline comparisons, 8 other reported p values can not be reproduced under any reading, two reported SDs are impossible given their ranges, and one subgroup mean is impossible given its sample size (Section 3). In total, this audit identifies 110 instances of Research Integrity Violations. We recommend retraction (Section 4).

2 | The values in parentheses are standard errors

Findings in this report are labeled using version 1.0 of the ACE Research Integrity Violations taxonomy (ACE-RIV). A reported value that is wrong as labeled but consistent with the rest of the article once the statistic is read as what it actually is (here, a standard error labeled as a standard deviation) is recorded under RIV 2.4, Data Miscalculation or Erroneous Reporting. A reported value that no reading makes consistent with the article's other numbers, such as a p value that no test of the reported data can return or an SD larger than its range allows, is recorded under RIV 2.2, Statistical Inconsistencies and Impossible Data Sets. Every finding in this section is recorded under RIV 2.4.

2.A The labels and the Methods disagree

Tables 1, 2, 3, 4, 6, and 7 of Singh (1997) report the summary statistics for the baseline characteristics and outcomes. The footnote of each of these tables states that the values are "Means and (SD)". Yet the statistical analysis section of the Methods states that they are standard errors (see Figure 4).

Continuous data are described as the mean (standard error) or median and range as appropriate, except for quality of life measures which are presented as mean (standard deviation) to allow comparison with reported normal values.

Both statements can not be true, but the article reports enough to determine which one is correct.

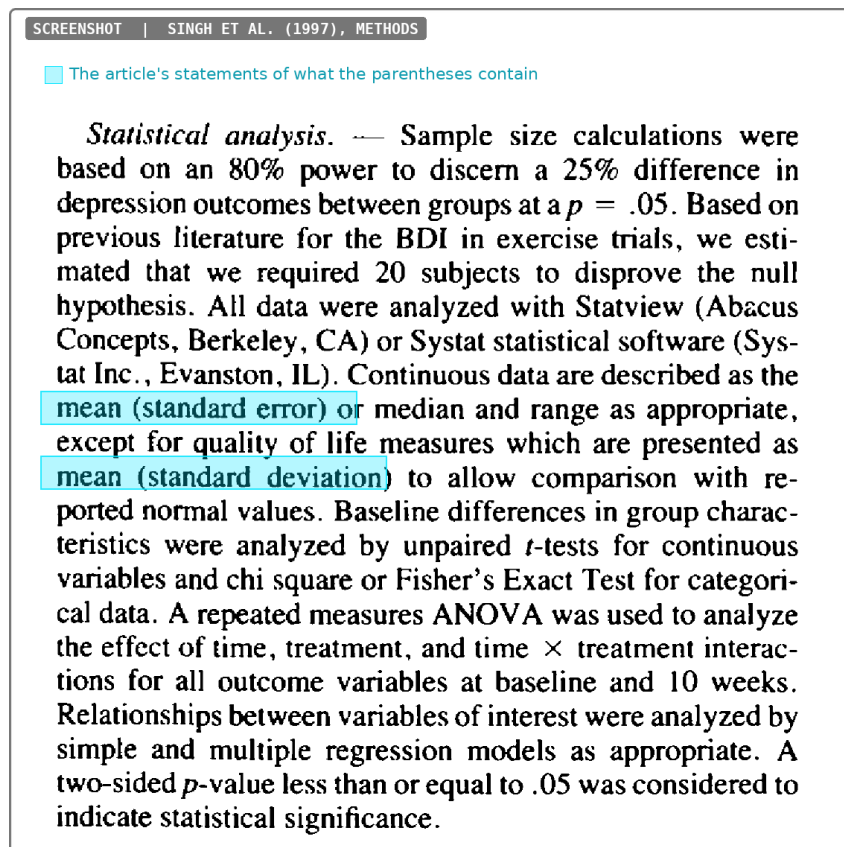


Figure 4. The statistical analysis paragraph of the Methods of Singh (1997). Blue marks the two statements of what the parentheses contain, standard errors for continuous data and standard deviations for the quality of life measures.

2.B Values that can not be standard deviations

Beneath the mean age of each group, Table 1 reports the age range, 61 to 82 years in the exercise group and 60 to 84 years in the control group. A standard deviation limits how far apart the youngest and oldest subjects can be. The widest range a group can have for a given SD arises when one subject sits at each end of the range and every other subject sits exactly at the mean, because that arrangement spreads the ages as far as possible while adding the least to the SD. For a group of n subjects with standard deviation s , that widest possible range is $s\sqrt{2(n-1)}$. A reported range wider than this can not come from a group with that SD. If the value of 1.5 for the exercise group is an SD, even one as large as 1.55 before rounding, the ages of its 17 subjects could span at most 8.8 years, against a reported span of 21 years. If the value of 2.0 for the control group is an SD, the ages of its 15 subjects could span at most 10.8 years, against a reported 24 years. If the value of 1.2 in the abstract is an SD, the ages of all 32 subjects could span at most 9.8 years, against a reported 24 years. The values in parentheses can not be SDs. If they are standard errors (SEs), the implied SDs are 6.2, 7.7, and 6.8 years, which fit the reported ranges.

SCREENSHOT | SINGH ET AL. (1997), TABLES 3 AND 4

Reported SDs that are smaller than the floor the Table 4 subgroup means impose

The article's statements of what the parentheses contain

A

Table 3. Depression Outcomes

Variable	Exercise (n = 17)	Control (n = 15)	Baseline p-value	Time p-value	Group × Time p-value
BDI (total)					
Pre	21.3 (1.8)*	18.4 (1.7)	.84	<.0001	.002
Post	9.8 (2.4)	13.8 (2.0)			
% reduction	59 (7)	29 (6)			
BDI (psychological subscale)					
Pre	13 (1.4)	11.9 (1)	.80	<.0001	.07
Post	6.3 (2.0)	8.5 (1.6)			
BDI (somatic subscale)					
Pre	8.2 (0.6)	6.5 (0.4)	.98	<.0001	<.0001
Post	3.5 (0.7)	5.5 (0.8)			
HRSD					
Pre	12.3 (0.9)	11.4 (1.0)	.30	<.001	.008
Post	5.3 (1.3)	8.9 (1.3)			
% reduction	59 (8)	20 (9)			
GDS					
Pre	16.9 (1.6)	13.9 (1.4)	.90	<.0001	.0004
Post	8.6 (1.8)	12.0 (1.8)			
% reduction	53 (8)	15 (8)			
DSM-IV Symptoms					
Pre	5.2 (0.4)	4.6 (0.4)	.63	<.001	.0003
Post	2.0 (0.4)	3.3 (0.6)			
% reduction	66 (8)	34 (8)			

Notes: BDI = Beck Depression Inventory; HRSD = Hamilton Rating Scale of Depression; GDS = Geriatric Depression Scale; DSM-IV = *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.).

*Means and (SD).

B

Table 4. Minor and Major Depression Outcomes by Diagnostic Group

Variable	Exercise	Control	Baseline p-value	Time p-value	Group × Time p-value
Minor Depression					
BDI					
Pre	16.7 (1.7)*	15.6 (1.0)	.13	<.0001	.001
Post	3.7 (1.3)	11.3 (1.9)			
HRSD					
Pre	9.3 (0.9)	10.6 (1.0)	.03	.0004	.19
Post	2.7 (0.8)	7.0 (1.3)			
Major Depression					
BDI					
Pre	24.6 (2.6)	21.5 (3.8)	.87	.002	.34
Post	14.9 (3.8)	16.3 (4.7)			
HRSD					
Pre	14.8 (0.8)	13.6 (2.9)	.34	.003	.04
Post	7.2 (2.2)	13.5 (2.5)			

Notes: Minor depression: 7 exercise and 10 control subjects; Major depression: 9 exercise and 4 control subjects. BDI = Beck Depression Inventory; HRSD = Hamilton Rating Scale of Depression.

*Means and (SD).

Figure 5. Panel A shows Table 3 of Singh (1997), the depression outcomes for the whole groups. Panel B shows Table 4, the same outcomes by diagnostic subgroup. Pink marks the BDI and HRSD parentheticals in Table 3, every one of which is below the smallest standard deviation that the subgroup means in Table 4 allow. Blue marks the footnotes stating what the parentheses contain.

The outcome tables can also be checked against each other, because Table 4 breaks the BDI and HRSD outcomes of Table 3 down by diagnostic subgroup, with 7 exercise and 10 control subjects with minor depression and 9 exercise and 4 control subjects with major depression (see panel B of Figure 5). The one dysthymic subject in each group is in neither subgroup. The variance of a whole group (the square of its SD) equals the variance within the subgroups plus the variance between the subgroup means. Neither part can be negative, so the spread of the subgroup means alone puts a lower bound on the SD of the whole group, whatever the spread within each subgroup. For the exercise group at baseline, the subgroup BDI means of

16.7 and 24.6 around a whole-group mean of 21.3 require a whole-group SD of at least 3.87. Table 3 reports 1.8. This bound holds with every reported mean free to fall anywhere within its rounding interval (the values that round to the reported one, 21.25 to 21.35 for a reported 21.3), in whichever direction lowers the bound, with the reported value taken at the top of its own interval, and with the dysthymic subject ignored. The same calculation for all 8 BDI and HRSD values in Table 3 gives lower bounds from 1.31 to 5.51, and every reported value, from 0.9 to 2.4, is below its bound. None of these values can be SDs. Read as SEs, the implied SDs, from 3.7 to 9.9, are all above their bounds.

2.C Reported *p* values reproduce only as standard errors

Every reported *p* value in these tables can be recalculated from the reported means, the values in parentheses, and the group sizes. We recalculated each one twice, once reading the values in parentheses as standard deviations (SDs) and once as standard errors (SEs). Every reported value is rounded, so each mean and parenthetical value may sit anywhere in its rounding interval. Table 1 and Table 2 report unpaired comparisons, which we tested with Student's and Welch's *t* tests. Each outcome table reports a Group × Time *p* value for every variable, which tests whether the two groups changed by different amounts over the 10 weeks. In a repeated measures ANOVA with two groups and two time points, that *p* value is the *t* test comparing the change scores (each subject's 10-week score minus their baseline score) of the two groups. The SD of a change score depends on the correlation between the baseline and 10-week scores, which the article does not report. We therefore let that correlation take any value from -1 to 1, separately in each group, which gives the widest range of *p* values each reading could produce. Allowing every correlation can only make the test more lenient. Under either reading the range runs from near zero, where the correlation is close to 1 and the change scores are tightly clustered, up to a ceiling set by a correlation of -1, where they are most spread out. A reading is ruled out only when the reported *p* value lies above that ceiling, so that no correlation could have produced it. Each range is rounded outward to the precision of the reported value. A reading is possible when the reported *p* value falls inside its range (see Table 3 and Table 4). Table 5 serves as a check on the method, because the Methods identify its parenthetical values as SDs. Read as SDs, all 8 of its Group × Time *p* values reproduce, whereas three of them do not reproduce with the values read as SEs.

For 25 variables in Tables 1, 2, 3, 4, 6, and 7, no reading of the parenthetical values as SDs can return the reported *p* value, whereas every one of them reproduces with the values read as SEs. For medications per day in Table 1, the SD reading gives .00–.01 against a reported .36, and the SE reading gives .25–.46. For HRSD in Table 3, the SD reading gives .000–.001 against a reported .008. For lifting self-efficacy in Table 6, it gives .00–.55 against a reported .59. For the other variables the test does not separate the two readings, either because the reported *p* value is very small or very large or because the range under the SD reading is wide enough to contain it. ADL in Table 7 reproduces only as SDs, because the control group's baseline parenthetical value of 0.1 is far smaller than the other three values in the row (0.8, 0.8, and 0.9). It is described but not counted.

The values reported in parentheses are standard errors, as the Methods state, yet every table labels them as standard deviations. Each parenthetical value whose label a reported *p* value shows to be wrong counts once. These values are correct once read as SEs, so they are 86 instances of **Data Miscalculation or Erroneous Reporting (RIV 2.4)**. Values whose test does not separate the two readings are described here but not counted. So are the “% reduction” and “% change” rows, which have no test, and the values the abstract and the Results give without saying whether they are SDs or SEs. The three reported *p* values that reproduce under neither reading are counted in Section 3.B.

Table 3. Reported p-values in Tables 1 through 4 against the p-values each reading of the parenthetical values can produce.

Variable	Test	Reported <i>p</i>	Reproduced <i>p</i>		Possible as SD	Possible as SE
			As SD	As SE		
Table 1, baseline characteristics						
Age (years)	Unpaired	.34	.00–.14	.21–.71	No	Yes
Education (years)	Unpaired	.65	.02–.18	.55–.73	No	Yes
Ethanol (g per day)	Unpaired	.34	.02–.07	.56–.63	No	No
Medications per day	Unpaired	.36	.00–.01	.25–.46	No	Yes
Chronic diseases	Unpaired	.10	.00–.01	.01–.17	No	Yes
Mini-Mental State Exam	Unpaired	.70	.00–.50	.48–.87	No	Yes
Duration of symptoms (months)	Unpaired	.45	.01–.06	.50–.62	No	No
Table 2, adverse events						
Visits to a health professional	Unpaired	.90	.16–1.00	.72–1.00	Yes	Yes
Minor illness	Unpaired	.50	.00–.07	.37–.64	No	Yes
Hospital days	Unpaired	.40	.00–.05	.41–.60	No	No
Musculoskeletal pain	Unpaired	.80	.20–.71	.74–.93	No	Yes
Table 3, depression outcomes						
BDI total	Group × Time	.002	.000–.001	.000–.258	No	Yes
BDI psychological	Group × Time	.07	.00–.04	.00–.59	No	Yes
BDI somatic	Group × Time	<.0001	.0000–.0001	.0000–.0794	Yes	Yes
HRSD	Group × Time	.008	.000–.001	.000–.207	No	Yes
GDS	Group × Time	.0004	.0000–.0001	.0000–.2108	No	Yes
DSM-IV symptoms	Group × Time	.0003	.0000–.0001	.0000–.2370	No	Yes
Table 4, diagnostic subgroups						
Minor depression, BDI	Group × Time	.001	.000–.001	.000–.075	Yes	Yes
Minor depression, HRSD	Group × Time	.19	.00–.02	.00–.41	No	Yes
Major depression, BDI	Group × Time	.34	.00–.34	.02–.72	Yes	Yes
Major depression, HRSD	Group × Time	.04	.00–.01	.00–.25	No	Yes

Note. Unpaired is Student's and Welch's two-sample *t* tests. Group × Time is the *t* test on change scores, which is the interaction of a two-group, two-time repeated measures ANOVA. Its range allows any pre-post correlation from –1 to 1 in each group. Every mean and parenthetical value may sit anywhere in its rounding interval, and each range is rounded outward to the precision of the reported value. As SD uses the parenthetical values as reported. As SE multiplies each by the square root of its group size (exercise 17, control 15, or 14 in Table 7, and the subgroup sizes in Table 4). Possible is No when the reported *p*-value falls outside the range that reading can produce.

Table 4. Reported p-values in Tables 5 through 7 against the p-values each reading of the parenthetical values can produce. Table 5 is the control, the one table whose values the Methods identify as standard deviations.

Variable	Test	Reported <i>p</i>	Reproduced <i>p</i>		Possible as SD	Possible as SE
			As SD	As SE		
Table 5, quality of life (SDs according to the Methods)						
Physical functioning	Group × Time	.22	.00–.54	.07–.88	Yes	Yes
Role physical	Group × Time	.16	.00–.46	.02–.86	Yes	Yes
Bodily pain	Group × Time	.001	.000–.109	.000–.680	Yes	Yes
General health	Group × Time	.06	.00–.51	.23–.87	Yes	No
Vitality	Group × Time	.002	.000–.234	.016–.763	Yes	No
Social functioning	Group × Time	.008	.000–.100	.000–.674	Yes	Yes
Role emotional	Group × Time	.02	.00–.27	.00–.78	Yes	Yes
Mental health	Group × Time	.24	.00–.66	.32–.92	Yes	No
Table 6, self-efficacy and morale						
Lifting	Group × Time	.59	.00–.55	.00–.88	No	Yes
Climbing	Group × Time	.63	.00–.43	.00–.85	No	Yes
Pushups	Group × Time	.48	.00–.28	.00–.79	No	Yes
Jogging	Group × Time	.25	.00–.05	.00–.61	No	Yes
Walking	Group × Time	.14	.00–.02	.00–.53	No	Yes
Philadelphia Geriatric Morale Scale	Group × Time	.03	.00–.01	.00–.39	No	Yes
Agitation	Group × Time	.17	.00–.04	.00–.60	No	Yes
Attitude toward aging	Group × Time	.04	.00–.01	.00–.33	No	Yes
Loneliness	Group × Time	.30	.00–.45	.00–.85	Yes	Yes
Table 7, function and performance						
ADL	Group × Time	.36	.08–1.00	.63–1.00	Yes	No
IADL	Group × Time	.80	.00–1.00	.00–1.00	Yes	Yes
SIP total	Group × Time	.84	.00–.90	.43–.98	Yes	Yes
Gait speed	Group × Time	.32	.00–1.00	.00–1.00	Yes	Yes
6-minute walk	Group × Time	.36	.00–.31	.04–.80	No	Yes
Chair rise, single	Group × Time	.06	.00–.01	.00–.39	No	Yes
Chair rise, 5 times	Group × Time	.10	.00–.01	.03–.47	No	Yes
Strength	Group × Time	<.0001	.0000–.0002	.0000–.2633	Yes	Yes

Note. Unpaired is Student's and Welch's two-sample *t* tests. Group × Time is the *t* test on change scores, which is the interaction of a two-group, two-time repeated measures ANOVA. Its range allows any pre-post correlation from –1 to 1 in each group. Every mean and parenthetical value may sit anywhere in its rounding interval, and each range is rounded outward to the precision of the reported value. As SD uses the parenthetical values as reported. As SE multiplies each by the square root of its group size (exercise 17, control 15, or 14 in Table 7, and the subgroup sizes in Table 4). Possible is No when the reported *p*-value falls outside the range that reading can produce.

3 | Reported statistics that no reading can recover

The values in this section are wrong under either reading of the parentheses, so each is recorded under RIV 2.2.

3.A Baseline p values that are not baseline comparisons

Tables 3 through 7 of Singh (1997) each contain three p value columns labeled “Baseline”, “Time”, and “Group $\times$ Time”. The Methods state that a repeated measures ANOVA was used “to analyze the effect of time, treatment, and time $\times$ treatment interactions”, which is three effects. A reader would take the first column to be a comparison of the groups at baseline. The Methods name three effects for three columns, so the first column most likely reports the treatment main effect of that ANOVA, the average difference between the groups across both time points, which is not a comparison of the groups at baseline. Table 5 is the place to test whether the column is a baseline comparison, because it is the only table where the Methods identify the values in parentheses as SDs (see Figure 6).

SCREENSHOT | SINGH ET AL. (1997), TABLE 5

☐ Baseline p values that no test of the baseline scores can return (seven of the eight entries)
☐ The article's statement of what the parentheses contain

Table 5. Quality of Life Outcomes

Variable	Normal Values (65–74 yr)	Exercise (n = 17)	Control (n = 15)	Baseline p -value	Time p -value	Group $\times$ Time p -value
Physical Functioning						
Pre	69.4 (26.3)*	65.9 (23.5)	64.3 (24.9)	.34	.01	.22
Post		82.6 (18.4)	70.3 (27.8)			
Role Physical						
Pre	65.4 (41.3)	52.9 (34.0)	63.3 (44.2)	.95	.20	.16
Post		73.5 (42.8)	61.7 (44.2)			
Bodily Pain						
Pre	68.5 (26.4)	56.8 (26.9)	72.3 (31.0)	.98	.51	.001
Post		74.2 (19.0)	58.5 (28.6)			
General Health						
Pre	62.6 (22.4)	54.0 (17.5)	64.9 (24.8)	.49	.12	.06
Post		63.9 (26.5)	63.5 (23.9)			
Vitality						
Pre	59.9 (22.1)	37.6 (17.9)	45.3 (24)	.73	.0001	.002
Post		61.7 (25.3)	48.7 (27.5)			
Social Functioning						
Pre	80.6 (25.6)	51.5 (25.0)	69.2 (26.2)	.66	.002	.008
Post		82.4 (19.3)	70.8 (26.2)			
Role Emotional						
Pre	81.4 (34.6)	43.1 (35)	60.0 (42.2)	.91	.007	.02
Post		76.5 (32.8)	62.2 (43.4)			
Mental Health						
Pre	76.9 (18)	55.5 (20.3)	57.0 (18.7)	.80	.0002	.24
Post		71.8 (26.5)	66.1 (22.6)			

*Means and (SD).

Figure 6. Table 5 of Singh (1997), quality of life outcomes on the SF-36. Pink marks the 7 “Baseline p -value” entries that no test of the baseline scores can return. Blue marks the footnote. This is the one table whose parenthetical values the Methods identify as standard deviations.

We recalculated an unpaired t test on the baseline scores of each of the 8 SF-36 subscales from the reported means, SDs, and sample sizes (see Table 5). With every reported mean and SD free to sit anywhere within its rounding interval, 7 of the 8 reported p values fall outside the range that either t test can return, each by more than 0.2. The exception is mental health, where the groups start and end close together. Reading the SDs as SEs does not reproduce any of the 8. For bodily pain, the baseline means of 56.8 and 72.3 with SDs

of about 30 give .14, against a reported p value of .98. For social functioning, the baseline means of 51.5 and 69.2 give .06, against a reported .66. For role emotional, the baseline means of 43.1 and 60.0 give .23, against a reported .91.

The mislabel matters most where a baseline comparison matters. For bodily pain, social functioning, and role emotional, the groups start 15 to 18 points apart and end close together. A reader of Table 5 would take the reported values of .98, .66, and .91 to mean that the groups started out alike.

SCREENSHOT | SINGH ET AL. (1997), TABLE 3

Baseline p values that do not reproduce as baseline comparisons under the standard error reading

Table 3. Depression Outcomes

Variable	Exercise ($n = 17$)	Control ($n = 15$)	Baseline p -value	Time p -value	Group $\times$ Time p -value
BDI (total)					
Pre	21.3 (1.8)*	18.4 (1.7)	.84	<.0001	.002
Post	9.8 (2.4)	13.8 (2.0)			
% reduction	59 (7)	29 (6)			
BDI (psychological subscale)					
Pre	13 (1.4)	11.9 (1)	.80	<.0001	.07
Post	6.3 (2.0)	8.5 (1.6)			
BDI (somatic subscale)					
Pre	8.2 (0.6)	6.5 (0.4)	.98	<.0001	<.0001
Post	3.5 (0.7)	5.5 (0.8)			
HRSD					
Pre	12.3 (0.9)	11.4 (1.0)	.30	<.001	.008
Post	5.3 (1.3)	8.9 (1.3)			
% reduction	59 (8)	20 (9)			
GDS					
Pre	16.9 (1.6)	13.9 (1.4)	.90	<.0001	.0004
Post	8.6 (1.8)	12.0 (1.8)			
% reduction	53 (8)	15 (8)			
DSM-IV Symptoms					
Pre	5.2 (0.4)	4.6 (0.4)	.63	<.001	.0003
Post	2.0 (0.4)	3.3 (0.6)			
% reduction	66 (8)	34 (8)			

Notes: BDI = Beck Depression Inventory; HRSD = Hamilton Rating Scale of Depression; GDS = Geriatric Depression Scale; DSM-IV = *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.).
*Means and (SD).

Figure 7. Table 3 of Singh (1997). Pink marks the six “Baseline p -value” entries, none of which reproduce as a comparison of the baseline scores.

The same column appears in the depression outcomes in Table 3 (see Figure 7). We recalculated the baseline t test for each of its six measures, treating the values in parentheses as SEs (see Section 2 and Table 5). None of the six reported p values can be reproduced under either reading of the parentheticals. One of these discrepancies changes the article’s account of the trial. On the somatic subscale of the BDI, the exercise group started at 8.2 and the control group at 6.5, with SEs of 0.6 and 0.4. The baseline t test returns a p value of .029 (.026 with Welch’s test), against .98 in Table 3. The Results state that “there were no significant differences at baseline in any variables measured.” The somatic subscale is also the outcome with the strongest reported interaction ($p < .0001$), on which the group with the higher baseline score improved the most. A group that starts with more symptoms has more room to improve, so a baseline difference on this subscale bears on the strongest result in the article.

The column is mislabeled in all five tables. The values in it can not be recovered as baseline comparisons under any reading of the parentheticals. Each one that no test of the baseline scores can return is an instance, 7 in Table 5 and 6 in Table 3, for 13 instances of **Statistical Inconsistencies and Impossible Data Sets (RIV 2.2)**. Tables 4, 6, and 7 contain the same column and are described here but not counted. The article reports no baseline comparison for any outcome, although it states that none differed at baseline. For at least one outcome, the somatic subscale of the BDI, the article’s own numbers show a baseline difference.

Table 5. The Baseline p-value column of Tables 5 and 3 against a test of the baseline scores. Table 5 is read as the Methods describe it (standard deviations), Table 3 as standard errors.

Variable	Exercise	Control	Reported <i>p</i>	Reproduced <i>p</i>		Reproduces
				Student	Welch	
Table 5, baseline values read as standard deviations						
Physical functioning	65.9 (23.5)	64.3 (24.9)	.34	.84–.87	.84–.87	No
Role physical	52.9 (34)	63.3 (44.2)	.95	.45–.47	.45–.48	No
Bodily pain	56.8 (26.9)	72.3 (31)	.98	.13–.15	.13–.16	No
General health	54 (17.5)	64.9 (24.8)	.49	.13–.18	.14–.20	No
Vitality	37.6 (17.9)	45.3 (24)	.73	.29–.33	.30–.34	No
Social functioning	51.5 (25)	69.2 (26.2)	.66	.05–.07	.05–.07	No
Role emotional	43.1 (35)	60 (42.2)	.91	.20–.25	.21–.25	No
Mental health	55.5 (20.3)	57 (18.7)	.80	.76–.90	.76–.90	Yes
Table 3, baseline values read as standard errors						
BDI total	21.3 (1.8)	18.4 (1.7)	.84	.22–.29	.22–.29	No
BDI psychological	13 (1.4)	11.9 (1)	.80	.33–.77	.32–.77	No
BDI somatic	8.2 (0.6)	6.5 (0.4)	.98	.01–.06	.01–.06	No
HRSD	12.3 (0.9)	11.4 (1)	.30	.43–.58	.43–.58	No
GDS	16.9 (1.6)	13.9 (1.4)	.90	.14–.21	.14–.20	No
DSM-IV symptoms	5.2 (0.4)	4.6 (0.4)	.63	.16–.45	.16–.44	No

Note. Baseline means and parenthetical values as reported, exercise (*n* = 17) and control (*n* = 15). Student is Student's two-sample *t* test (pooled variance) and Welch is Welch's two-sample *t* test (unequal variances), each applied to the baseline scores. Each gives the lowest and highest *p*-value the test can return once every reported value is allowed to sit anywhere in its rounding interval, rounded outward to the reported precision. Reproduces is Yes when the reported value falls inside either range.

3.B p values that can not be reproduced under any reading

Three of the unpaired comparisons in Table 3 reproduce under neither reading of the parentheticals. For ethanol intake in Table 1, the reported p value is .34, against .02–.07 as SDs and .56–.63 as SEs. For the duration of depressive symptoms, it is .45, against .01–.06 and .50–.62. For hospital days in Table 2, it is .40, against .00–.05 and .41–.60. All three variables are skewed, with SEs close to the size of the means, which points to a few subjects with large values. Such variables may have been summarized and tested differently (the Methods allow for medians and ranges “as appropriate”). As reported, they are three instances of **Statistical Inconsistencies and Impossible Data Sets (RIV 2.2)**.

Table 1 of Singh (1997) contains a p value for each of 10 categorical baseline characteristics (see Figure 8). The article states that p values for categorical data come from either a chi-square test or Fisher’s exact test, without saying which variable uses which. The Results identify the test through three comparisons of proportions reported with their p values: response to treatment in 10/17 exercisers vs 4/15 controls ($p = .067$), remission of minor depression in 6/7 vs 4/10 ($p = .06$), and remission of major depression in 8/9 vs 2/4 ($p = .12$). Pearson’s chi-square test without a continuity correction reproduces all three (.067, .059, and .125, respectively). Fisher’s exact test reproduces none of them (.087, .134, and .203, respectively). The authors therefore used an uncorrected chi-square test, even for tables with expected cell counts well below five, where the chi-square approximation is unreliable and Fisher’s exact test is the usual choice.

SCREENSHOT | SINGH ET AL. (1997), TABLE 1

Reported p values that neither the chi-square test nor Fisher's exact test reproduces

Table 1. Baseline Characteristics

Variable	Exercise ($n = 17$)	Control ($n = 15$)	p -value
Age (years)	70 (1.5)*	72 (2.0)	.34
Range	61–82	60–84	
Gender			
Male	5	7	.71
Female	12	8	
Marital Status			
Married	9	7	.76
Widowed	3	5	
Divorced	2	2	
Never married	2	1	
Income Annually			
<\$10,000	4	3	.95
\$10,000–30,000	7	6	
>\$30,000	6	6	
Residence			
House	13	12	.63
Apartment	4	3	
Employment			
Full time	3	0	.63
Part time	4	4	
Unemployed	10	11	
Education (years)	14.6 (0.8)*	14.1 (0.7)	.65
Smoking			
Lifelong nonsmoking	13	11	.78
Past smoker, noncurrent	4	4	
Current smoker	0	0	
Ethanol (avg daily intake in gms in last month)	4.8 (1.4)*	7.4 (4.7)	.34
Medications per day (no.)	4.5 (0.6)*	3.7 (0.6)	.36
Chronic diseases (no.)	2.9 (0.3)*	2.1 (0.3)	.10
Mini-Mental State Exam score (0–30)	28.4 (0.3)*	28.6 (0.4)	.70
DSM-IV diagnosis			
Minor depression	7	10	.40
Major depression mild	7	2	
Major depression moderate	2	2	
Dysthymia	1	1	
Duration since onset of any depressive symptoms (months)	24 (10)*	36 (18)	.45
Previous Psychiatric Treatment			
Psychiatrist	8	4	.23
Antidepressant drugs	8	5	.43
Hospitalization for depression	2	1	.62

*Means and (SD).

Figure 8. Table 1 of Singh (1997). Pink marks the five categorical p values that neither the chi-square test nor Fisher's exact test reproduces from the reported counts.

We recalculated this test, and Fisher's exact test, from the contingency table cells reported in Table 1 (see Table 6). Four of the 10 p values reproduce under the uncorrected chi-square test, those for annual income, previous treatment by a psychiatrist, previous antidepressant drugs, and previous hospitalization. Neither test reproduces the other six. For gender (5 men and 12 women vs 7 men and 8 women), the two tests return .31 and .47 against a reported p value of .71. For residence (13 house and 4 apartment vs 12 and 3),

Table 6. Baseline *p*-values for the categorical characteristics in Table 1, reproduced from the reported counts.

Characteristic	Exercise (n = 17)				Control (n = 15)				Reported <i>p</i>	Reproduced <i>p</i>		
	I	II	III	IV	I	II	III	IV		Chi-squared	Fisher	Reproduces
Gender	5	12	-	-	7	8	-	-	.71	.31	.47	No
Marital status	9	3	2	2	7	5	2	1	.76	.79	.85	No
Income	4	7	6	-	3	6	6	-	.95	.95	1.00	Yes
Residence	13	4	-	-	12	3	-	-	.63	.81	1.00	No
Employment	3	4	10	-	0	4	11	-	.63	.23	.35	No
Smoking	13	4	-	-	11	4	-	-	.78	.84	1.00	No
DSM-IV diagnosis	7	7	2	1	10	2	2	1	.40	.36	.32	No
Psychiatrist	8	9	-	-	4	11	-	-	.23	.23	.29	Yes
Antidepressant drugs	8	9	-	-	5	10	-	-	.43	.43	.49	Yes
Hospitalization	2	15	-	-	1	14	-	-	.62	.62	1.00	Yes

Note. Counts as reported in Table 1 of the article. Roman numerals are the levels of each characteristic in the article's row order: gender, I male, II female; marital status, I married, II widowed, III divorced, IV never married; income, I below \$10,000, II \$10,000 to \$30,000, III above \$30,000; residence, I house, II apartment; employment, I full time, II part time, III unemployed; smoking, I lifelong nonsmoker, II past smoker (the empty current smoker row is omitted); DSM-IV diagnosis, I minor depression, II major depression mild, III major depression moderate, IV dysthymia; previous psychiatrist, antidepressant drugs, and hospitalization, I yes, II no (the article reports the yes counts). Chi-squared is Pearson's test without a continuity correction, the test that reproduces the article's own comparisons of proportions in the Results; Fisher is the Fisher-Freeman-Halton exact test for tables larger than 2 by 2. None of the six 2 by 2 tables returns its reported value under the Yates corrected chi-square test either. Reproduces means that at least one test returns the reported value to within a unit in the last digit.

they return .81 and 1.00 against a reported .63. Two variants of the chi-square test, the Yates continuity correction and the likelihood ratio test, do not change this. None of the six 2 by 2 tables returns its reported value under the Yates corrected chi-square test. The likelihood ratio test agrees with the uncorrected chi-square test to the reported precision. Before counting a *p* value as irreproducible, we also tried every way of collapsing the categories of each variable into fewer groups, under all four tests. This collapsing reproduces one variable, employment. Collapsing full-time and part-time work into a single employed category and applying the Yates correction returns .625, which rounds to the reported value. The article describes no such collapsing and uses a continuity correction nowhere else. No collapsing of the marital status or DSM-IV diagnosis categories reproduces the reported *p* values under any of the four tests.

The irreproducible categorical *p* values do not affect the conclusions of the study. The reported and recalculated values are all above .05, so under either set the groups do not differ significantly on these characteristics. We count them because they are incorrect. A baseline table that does not reproduce from its own contingency table cells also limits how far a reader can verify the article's other computations. The five *p* values that no test reproduces, from the reported cells or from any collapsing of them, are five instances of **Statistical Inconsistencies and Impossible Data Sets (RIV 2.2)**. Employment is described but not counted.

3.C Training intensity SDs that their ranges do not allow

The Results report the mean training intensity of the exercise group as "78% (*SD* = 24) of the 1RM (range 45–87%)". Excluding the one subject who had less than 50% compliance, they report "80% (*SD* = 10) of the 1RM (range 70–87%)". Both values are labeled *SD* in the text (the 1RM is the one-repetition maximum, the heaviest weight a subject can lift once, so intensity is the training load as a percentage of it). A set of *n* values that lies within a range of width *w* has an *SD* of at most $\frac{w}{2} \sqrt{n/(n-1)}$, reached when half

of the values sit at each end. For the 17 exercisers between 45% and 87%, the SD can be at most 22.2, even with each end of the range rounded outward by half a point. For the 16 exercisers between 70% and 87%, it can be at most 9.3. Neither reported SD is possible. Read as SEs, the implied SDs of 99 and 40 are larger still. No reading of these values is possible, so they are two instances of **Statistical Inconsistencies and Impossible Data Sets (RIV 2.2)**. The Discussion repeats the first value ("mean 78, SD = 24"), which is described but not counted.

3.D GRIM failure in a subgroup mean

The *granularity-related inconsistency of means* test (GRIM, Brown 2017) checks whether a reported mean can arise from integer scores, given the sample size. The BDI and HRSD are sums of integer item scores, so a group mean must be an integer divided by the group size. A mean reported to one decimal place can only fail the test when the group has fewer than 10 subjects. Of the 12 such means in Table 4, one fails. The baseline HRSD mean of 13.6 for the four control subjects with major depression would require an integer total between 54.2 and 54.6. No integer lies in that range (see Figure 9). The nearest achievable means are 13.5 and 13.75. The other 11 means are consistent.

SCREENSHOT | SINGH ET AL. (1997), TABLE 4

■ A reported mean that no set of four integer scores can produce

Table 4. Minor and Major Depression Outcomes by Diagnostic Group

Variable	Exercise	Control	Baseline <i>p</i> -value	Time <i>p</i> -value	Group × Time <i>p</i> -value
Minor Depression					
BDI					
Pre	16.7 (1.7)*	15.6 (1.0)	.13	<.0001	.001
Post	3.7 (1.3)	11.3 (1.9)			
HRSD					
Pre	9.3 (0.9)	10.6 (1.0)	.03	.0004	.19
Post	2.7 (0.8)	7.0 (1.3)			
Major Depression					
BDI					
Pre	24.6 (2.6)	21.5 (3.8)	.87	.002	.34
Post	14.9 (3.8)	16.3 (4.7)			
HRSD					
Pre	14.8 (0.8)	13.6 (2.9)	.34	.003	.04
Post	7.2 (2.2)	13.5 (2.5)			

Notes. Minor depression: 7 exercise and 10 control subjects; Major depression: 9 exercise and 4 control subjects. BDI = Beck Depression Inventory; HRSD = Hamilton Rating Scale of Depression.
*Means and (SD).

Figure 9. Table 4 of Singh (1997). Pink marks the baseline HRSD mean of the four control subjects with major depression, which no set of four integer scores can produce.

Rounding, the scale, and the sample size do not explain the failure. The notes to Table 4 give the subgroup size as 4, which matches the DSM-IV diagnosis counts in Table 1 (2 mild and 2 moderate major depression). The HRSD is scored in whole points. The test already allows for rounding. The Results state that the improvement in HRSD scores in this subgroup was twice that of controls, a comparison that depends on the impossible mean. This is one instance of **Statistical Inconsistencies and Impossible Data Sets (RIV 2.2)**.

The other values in the article that can be checked are consistent. The three correlations reported in the Results for the exercise group ($r = -0.786, -0.589, \text{ and } 0.500$ with $n = 17$) give p values of .0002, .013, and .04, against reported values of .0002, .013, and .04. The diagnostic subgroup sizes sum to the group sizes once the dysthymic subject in each group is included. The available evidence points to problems in how the statistics are labeled and reported. It gives no reason to doubt that the trial took place.

4 | Conclusion

The audited article describes a small RCT whose tables misstate what they contain. Every table of continuous outcomes labels the values in parentheses as standard deviations, yet the Methods state that they are standard errors. The reported test statistics support the Methods (i.e., the values in parentheses are standard errors). For 25 variables across Tables 1 through 4, 6, and 7, the reported p value can not be reproduced with the values read as SDs and can be with the values read as SEs. Additionally, the reported age ranges and the Table 4 subgroup means are impossible under the labeled SDs. This warrants 86 instances of **Data Miscalculation or Erroneous Reporting (RIV 2.4)**, one for each of the values involved.

The effect of this error on meta-analyses is the central finding of this report. The largest post-test SMD the article can support is -0.72 (-0.44 for the BDI). Of the 24 meta-analyses we located that report an effect size for this trial, 19 assign Singh (1997) a value beyond anything a correct extraction can produce. The median assigned value is -2.74 . In two syntheses this 32-subject trial is the largest effect in the analysis by a factor of two or more. Four meta-analyses, including the Cochrane review, report exactly -1.75 , the value that results from meta-analysts treating the standard errors of the post-intervention BDI scores as standard deviations. For nearly 30 years, the evidence base on exercise and depression has carried a moderate effect as one of its largest.

The columns labeled “Baseline p -value” in the outcome tables do not report baseline comparisons. In the one table where the SDs are unambiguous, 7 of the 8 reported p values can not be reproduced as a test of the baseline scores. Under the correct reading of the depression outcomes, the groups differed at baseline on the somatic subscale of the BDI, the outcome with the strongest reported interaction. Yet the Results state that no variable differed at baseline. The 13 reported values that no baseline test can return are 13 instances of **Statistical Inconsistencies and Impossible Data Sets (RIV 2.2)**. Furthermore, three unpaired and five categorical p values in Tables 1 and 2 can not be reproduced under any reading of the reported summaries, which adds 8 instances of **Statistical Inconsistencies and Impossible Data Sets (RIV 2.2)**. Two training intensity SDs in the Results are larger than their ranges allow under either reading, which adds two more. One subgroup mean in Table 4 is impossible given its sample size, which adds one.

Because the misreporting in Singh (1997) has contaminated the evidence syntheses in this field, we recommend *retraction*. Every table labels the standard errors as standard deviations. Of the 24 meta-analyses we located, 19 assign Singh (1997) an effect size that no correct extraction can produce. Four of these meta-analyses, including the Cochrane review, report the value that results from treating the standard errors as standard deviations. A reader can only use the published tables as evidence after first correcting them. For nearly 30 years, most meta-analysts have not. In addition, the outcome tables report p values under the label of a baseline comparison that no test of the baseline scores can return. Furthermore, 8 p values in Tables 1 and 2 can not be reproduced under any reading. One subgroup mean in Table 4 is impossible given its sample size. As published, the article’s results have no evidentiary value. This recommendation does not rest on any evidence that the trial was not conducted.

The retraction notice should give the reasons. These are that (a) the values in parentheses in Tables 1 through 4, 6, and 7 and in the abstract are standard errors, (b) the first p value column of Tables 3 through 7 does not report a comparison of the groups at baseline, (c) the groups differed at baseline on the somatic subscale of the BDI, (d) 8 p values in Tables 1 and 2 can not be reproduced from the reported summaries, (e) the subgroup mean in Table 4 is impossible given its sample size, and (f) the two training intensity SDs in the Results are impossible given their ranges.

Whatever the journal decides, meta-analysts and guideline developers who have included this trial should check the values they extracted from it. A meta-analysis that recorded the values in parentheses as standard deviations has treated a moderate effect as a very large one. Such a meta-analysis should re-extract the study with those values multiplied by the square root of the group size, or exclude the study.

Table 7. Summary of Research Integrity Violations identified in the audited Singh, Clements, and Fiatarone (1997) trial (ACE-RIV v1.0)

RIV	Title	Instances	Section
RIV 2.2	Statistical Inconsistencies and Impossible Data Sets	24	3.A, 3.B, 3.C, 3.D
RIV 2.4	Data Miscalculation or Erroneous Reporting	86	2.C
Total		110	

References

- Singh, Clements, Fiatarone (1997) A randomized controlled trial of progressive resistance training in depressed elders. *The Journals of Gerontology Series A*. DOI: 10.1093/gerona/52A.1.M27.
- Gordon, McDowell, Hallgren, Meyer, Lyons, Herring (2018) Association of efficacy of resistance exercise training with depressive symptoms: Meta-analysis and meta-regression analysis of randomized clinical trials. *JAMA Psychiatry*. DOI: 10.1001/jamapsychiatry.2018.0572.
- Dai, Ting, Malau, Wu, Lin, Lin, Chen (2024) Motivational barriers and facilitators for older adults' engagement in exercise during depressive episodes: A network meta-analysis of randomized controlled trials. *Healthcare*. DOI: 10.3390/healthcare12151498.
- Li, Chen, He, Yan, Zhang, Li (2026) Can exercise truly alleviate mild-to-moderate and subthreshold depression? Evidence from randomized controlled trials. *Frontiers in Psychology*. DOI: 10.3389/fpsyg.2026.1815160.
- Lawlor, Hopker (2001) The effectiveness of exercise as an intervention in the management of depression: Systematic review and meta-regression analysis of randomised controlled trials. *BMJ*. DOI: 10.1136/bmj.322.7289.763.
- Liu, Latham (2011) Can progressive resistance strength training reduce physical disability in older adults? A meta-analysis study. *Disability and Rehabilitation*. DOI: 10.3109/09638288.2010.487145.
- Yu, Wong, Lei, Nie, Shi, Zou, Kong (2022) Comparative effectiveness of multiple exercise interventions in the treatment of mental health disorders: A systematic review and network meta-analysis. *Sports Medicine - Open*. DOI: 10.1186/s40798-022-00529-5.
- Cooney, Dwan, Greig, Lawlor, Rimer, Waugh, McMurdo, Mead (2013) Exercise for depression. *Cochrane Database of Systematic Reviews*. DOI: 10.1002/14651858.CD004366.pub6.
- Heinzel, Lawrence, Kallies, Rapp, Heissel (2015) Using exercise to fight depression in older adults: A systematic review and meta-analysis. *GeroPsych*. DOI: 10.1024/1662-9647/a000133.
- Rhyner, Watts (2016) Exercise and depressive symptoms in older adults: A systematic meta-analytic review. *Journal of Aging and Physical Activity*. DOI: 10.1123/japa.2015-0146.
- Kvam, Kleppe, Nordhus, Hovland (2016) Exercise as a treatment for depression: A meta-analysis. *Journal of Affective Disorders*. DOI: 10.1016/j.jad.2016.03.063.
- Krogh, Hjorthøj, Speyer, Gluud, Nordentoft (2017) Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential analysis. *BMJ Open*. DOI: 10.1136/bmjopen-2016-014820.
- Silveira, Moraes, Oliveira, Coutinho, Laks, Deslandes (2013) Physical exercise and clinically depressed patients: A systematic review and meta-analysis. *Neuropsychobiology*. DOI: 10.1159/000345160.
- Schuch, Vancampfort, Richards, Rosenbaum, Ward, Stubbs (2016a) Exercise as a treatment for depression: A meta-analysis adjusting for publication bias. *Journal of Psychiatric Research*. DOI: 10.1016/j.jpsy-chires.2016.02.023.
- Schuch, Vancampfort, Rosenbaum, Richards, Ward, Veronese, Solmi, Cadore, Stubbs (2016b) Exercise for depression in older adults: A meta-analysis of randomized controlled trials adjusting for publication bias.

Brazilian Journal of Psychiatry. DOI: 10.1590/1516-4446-2016-1915.

Hooper (2018) The efficacy of multiple non-pharmacological treatment modalities for depressive disorders: A systematic review and network meta-analysis. *Temple University*. URL: <https://ezproxy.lib.uconn.edu/login?url=https://www.proquest.com/dissertations-theses/efficacy-multiple-non-pharmacological-treatment/docview/2100697816/se-2?accountid=14518>.

Rossi, Santos, Rossi, Stubbs, Schuch, Neves (2024) Strength training has antidepressant effects in people with depression or depressive symptoms but no other severe diseases: A systematic review with meta-analysis. *Psychiatry Research*. DOI: 10.1016/j.psychres.2024.115805.

Noetel, Sanders, Gallardo-Gómez, Taylor, Pozo Cruz, Hoek, Smith, Mahoney, Spathis, Moresi, Pagano, Pagano, Vasconcellos, Arnott, Varley, Parker, Biddle, Lonsdale (2024) Effect of exercise for depression: Systematic review and network meta-analysis of randomised controlled trials. *BMJ*. DOI: 10.1136/bmj-2023-075847.

Cunha, Werneck, Santos, Oliveira, Zou, Schuch, Cyrino (2024) Can resistance training improve mental health outcomes in older adults? A systematic review and meta-analysis of randomized controlled trials. *Psychiatry Research*. DOI: 10.1016/j.psychres.2024.115746.

Nebiker, Lichtenstein, Minghetti, Zahner, Gerber, Faude, Donath (2018) Moderating effects of exercise duration and intensity in neuromuscular vs. Endurance exercise interventions for the treatment of depression: A meta-analytical review. *Frontiers in Psychiatry*. DOI: 10.3389/fpsyt.2018.00305.

Brown, Heathers (2017) The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science*. DOI: 10.1177/1948550616673876.

Heissel, Heinen, Brokmeier, Skarabis, Kangas, Vancampfort, Stubbs, Firth, Ward, Rosenbaum (2023) Exercise as medicine for depressive symptoms? A systematic review and meta-analysis with meta-regression. *British Journal of Sports Medicine*. DOI: 10.1136/bjsports-2022-106282.

Rosenbaum, Tiedemann, Sherrington, Curtis, Ward (2014) Physical activity interventions for people with mental illness: A systematic review and meta-analysis. *The Journal of Clinical Psychiatry*. DOI: 10.4088/JCP.13r08765.

Rethorst, Wipfli, Landers (2009) The antidepressive effects of exercise: A meta-analysis of randomized trials. *Sports Medicine*. DOI: 10.2165/00007256-200939060-00004.

Dauwan, Begemann, Slot, Lee, Scheltens, Sommer (2021) Physical exercise improves quality of life, depressive symptoms, and cognition across chronic brain disorders: A transdiagnostic systematic review and meta-analysis of randomized controlled trials. *Journal of Neurology*. DOI: 10.1007/s00415-019-09493-9.

Stubbs, Vancampfort, Rosenbaum, Ward, Richards, Ussher, Schuch (2016) Challenges establishing the efficacy of exercise as an antidepressant treatment: A systematic review and meta-analysis of control group responses in exercise randomised controlled trials. *Sports Medicine*. DOI: 10.1007/s40279-015-0441-5.

Version history

Version	Date	DOI	Change
1	2026-10-12	10.5281/zenodo.23331179	Initial publication. 110 instances across two categories.

Any change to a finding, RIV count, or recommendation creates a new version with a new version DOI. Superseded versions remain permanently resolvable.