PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Endogenous testosterone level and testosterone supplementation therapy in chronic obstructive pulmonary disease (COPD): a systematic review and meta-analysis
AUTHORS	Atlantis, Evan; Fahey, Paul; Cochrane, Belinda; Wittert, Gary; Smith, Sheree

VERSION 1 - REVIEW

REVIEWER	Johan Svartberg, M.D., Ph.D. Chief and Professor Section of Endocrinology Division of Internal Medicine University Hospital of North Norway 9038 Tromsø Norway
REVIEW RETURNED	08-May-2013

GENERAL COMMENTS	In this paper Atlantis and co-workers present a systematic review and a meta-analysis on endogenous testosterone levels and testosterone supplementation in COPD patients. The conclusion is that men with COPD have clinically relevant lower T levels and that testosterone therapy seems to improve exercise capacity (in small short-term studies). I have only one question to the authors and that is about the inclusion of the study by Sharma et al in the meta-analysis (even though they also perform separately analyses excluding this study). The study by Sharma et al was discontinued due to the lack of efficiency of nandrolone decanoate in improving any of the study outcomes and only 16 of the intended 44 patients were included. I would have excluded this study. Besides this the methods are solid and the results are presented in a sober way.

REVIEWER	Peter Watson Statistician MRC Cognition and Brain Sciences Unit 15 Chaucer Road Cambridge CB2 7EF I have no competing interests with the research described in this paper.
REVIEW RETURNED	09-May-2013

THE STUDY	I wasn't clear if Hedge's g was the effect size for the standardized mean difference (page 10) rather than, say, Cohen's d. I also wondered if the weightings used for the 'weighted mean difference' on the second line of page 10 uses inverse variance weights and if the DerSimonian estimate was used for any of the confidence intervals (third line of page 10) since at least some of these I assume must be derived from random effects meta-analyses (third last line of page 9). You could also remind us (in the third line on page 10) that the median-based estimates of means using medians rather than means are used with observed means to emphasise the pooling is based upon means (pseudo and actual). Like to see a one or two sentences inserted to mention the robustness of the median-based Hozo et al. (2005) estimates to distributions of responses within groups. On page 8 last paragraph there are at least three primary outcomes mentioned rather than one.
RESULTS & CONCLUSIONS	I found the sensitivity analyses rather clouded the picture as multiple results are presented (e.g. Tables 3-5 pages 30-32) for each primary outcome. Usually in the meta-analyses I have seen one chooses apriori the studies for inclusion and performs a meta-analysis on them or adjusts for a covariate whilst doing the meta-analysis if the studies differ on covariates. You can fit study covariate adjusted meta-analyses in R using the 'metafor' package so the results are adjustments for study differences and this has the advantage over subsamples of using all the studies in the one analysis.
GENERAL COMMENTS	Meta-analyses are performed yielding pooled estimates of standardised group comparisons to assess (page 6, first full paragraph) the influence of testosterone treatment on measures of exercise capacity (peak muscle strength and peak oxygen uptake) and health-related quality of life (HRQoL) on people with COPD in randomised control trials and differences in testosterone in those with chronic lung disease and controls. There are also some descriptive analyses of secondary outcomes (page 9, 'secondary outcomes' paragraph). I have a few minor comments below which the authors may like to address relating in the main to further clarification of the methods used and the usefulness of the sensitivity analyses described on pages 13-14 and in Tables 3 to 5 (pages 30-32). I also would welcome some reassurance on the plausibility of the apparent pooling of median-based estimates and means (page 10) since this implies that the groups they are obtained from have different distributions. The authors could reassure by adding into the text that Hozo et al. (2005) who they quote at the bottom of page 8 ran simulations to assess the closeness of these median-based estimates to means under a variety of distributions to assess their robustness and found the means were well reproduced albeit their variances less well so. I also found the sensitivity analyses confusing as more results were presented dropping various studies for the same outcome thus giving a range of results most of which suggest no, or little, benefit to using the testosterone therapy (forest plots of pages 36, 38 and 39 and results from last paragraph on page 13 to fourth line on page 15 which have 95% confidence intervals including, or near to, zero for treatment differences) and also use reduced samples. Perhaps my biggest concern is that possibly in light of these null results it is further suggested on page 18 that 'there is an absence of sufficient RCT evidence to draw firm conclusions' which would appear to rather limit the usefulness of this

	study. Page 7. Was an inter-rater measure of agreement considered such as kappa to show the level of agreement between reviewers which could have been quoted in the 'data extraction' paragraph? Page 10. (top three lines of first full paragraph). I assume the standardised mean difference is actually Hedge's g which is commonly used in meta-analyses for comparing mean differences and that random effects (page 9, third last line) were incorporated using the DerSimonian and Laird estimate to produce 95% confidence intervals? Not sure even though it is referenced why you need to verify a random effects meta-analysis with a fixed one (second and third last line on page 9) since you are making the decision which to use by using I^2 (page 10, fourth to sixth line of the first full paragraph) to decide which to use and then using that. Doesn't using both fixed and random approaches invalidate using these decision criteria? Page 10. Third and fourth line from top. It might be less confusing here to say that it was only means that were pooled since (last two lines of age 8) the medians were transformed into means using the minimum and maximum values as suggested by Hozo et al. (2005) prior to pooling. I would say, however, that medians are usually quoted for skewed distributions whereas means assume more symmetric distributions within group so one would be pooling summary measures from different (shaped) distributions. It may be the distributions within groups differ due to different processes going on due to different sampling biases e.g. some groups may be based upon older people or mixed gender populations who contain more outlying observations than younger people or male only populations and so age and gender may be confounding variables for looking at group differences between studies. Do the authors, therefore, have any comments they could add to reassure on combining estimates based upon different distributions as implied by the use of medians and means. Medians also have different variances to means depending on sample size (Kenney and Keeping 1962, p. 211) and distribution which may influence pooling if weightings related to inverse variances are used (line two on page 10 mentions a pooled 'weighted' mean difference so these may have been used). Page 10. Line five of the second paragraph. Not sure here why a 40% threshold was used for deciding a high level of between study heterogeneity using I^2. This doesn't appear to correspond to any of the thresholds as suggested by Higgins et al. (2003) who suggest a value of 0% indicates no observed heterogeneity, 25%-49% is low heterogeneity, 50%-74% is moderate and 75% and above is large. Page 10. Third line from bottom of second full paragraph. Did you also consider in addition to funnel plots also using inference to assess the degree of publication bias using a statistical test? See for example Peters et al. (2010). Page 13. The last line has a point estimate of -0.31 which is outside the confidence interval. Page 18. The sensitivity analyses all seem to conclude that there is little or no benefit to using a testosterone therapy on primary

	outcomes (as for example shown in the confidence intervals on pages 34, 36, 38 and 39 and quoted from last paragraph on page 13 to first four lines on page 15 for the overall standardized mean difference containing or being close to zero) so I am not convinced the therapy improves several exercise capacity outcomes as stated on the second line of page 18 or that null results can make us hope that we have insufficient evidence to 'draw firm conclusions about the long-term benefits' of testosterone therapy as further stated on lines two to four of page 18. Page 30. Table 3. The first line in Table 3 quoted the use of a fixed effects model which by definition assumes no between study heterogeneity in effect sizes yet the test of heterogeneity in the right-most column in Table 3 appears to suggest that this is present ($p<0.001$) suggesting random effects may be more appropriate. I suspect the fixed model was used because the I^2 value was 'small' despite being statistically significant so the I^2 value could be inserted in this table to emphasise this as this is the value used to decide if fixed effects or random effect models are used (as intimated in line 5 of the second paragraph on page 10). Pages 30-32. The sensitivity analyses in Tables 3 to 5 (and described on pages 13-14) using subsets of studies can be interpreted as suggesting doubts about the usefulness of combining studies of, for example, different qualities in an overall analysis. I would have thought it simpler if the decision of what studies to include in a meta-analysis was made apriori and a meta-analysis only carried out on the studies deemed of appropriately high quality. The single confidence interval for these studies could then be quoted in the text and there would be no need for Tables 3 to 5. You might also wish to consider performing a single meta-analysis on each outcome that removes differences in study level covariates e.g. quality of study using, for example, the 'metafor' package in R although STATA may also do this. This would also help with multiplicity caused by repeated testing which might through chance yield confidence intervals which do not contain zero. I was also not sure from the third row of Table 3 (page 30) what 'unadjusted' and 'adjusted' related to? Is this adjustments for publication bias? Pages 36, 38 and 39. Judging from the overall 95% confidence intervals (using all the studies) of the standardized difference the effect of the testosterone therapy is either very small or not present. Pages 38 and 39. There appear to be no funnel plots corresponding to the standardized mean differences for outcomes in the forest plots in Figures 6 and 7 yet these are presented for the other forest plots on pages 34 and 36. Page 39. Figure 7 appears to be based upon only three studies (with one of these having a small weight < 10%) which seems a small number for obtaining pooled estimates. References Higgins, J.P., Thompson, S.G., Deeks J.J. and Altman, D.G. (2003) Measuring inconsistency in meta-analyses. BMJ 327 557-560. Kenney, J. F. and Keeping, E. S. "The Median," "Relation Between

	Mean, Median, and Mode," "Relative Merits of Mean, Median, and Mode," and "The Median." §3.2, 4.8-4.9, and 13.13 in Mathematics of Statistics, Pt. 1, 3rd ed. Princeton, NJ: Van Nostrand, pp. 32-35, 52-54, 211-212, 1962. Peters, J. L., Sutton, A. J., Jones, D. R. and Abrams K. R. (2010). Assessing publication bias in meta-analyses in the presence of between-study heterogeneity. Journal of the Royal Statistical Society A 173(3) 575-591.

VERSION 1 – AUTHOR RESPONSE

Reviewer: Johan Svartberg, M.D., Ph.D.
Chief and Professor
Section of Endocrinology
Division of Internal Medicine
University Hospital of North Norway
9038 Tromsø
Norway

In this paper Atlantis and co-workers present a systematic review and a meta-analysis on endogenous testosterone levels and testosterone supplementation in COPD patients. The conclusion is that men with COPD have clinically relevant lower T levels and that testosterone therapy seems to improve exercise capacity (in small short-term studies).

I have only one question to the authors and that is about the inclusion of the study by Sharma et al in the meta-analysis (even though they also perform separately analyses excluding this study). The study by Sharma et al was discontinued due to the lack of efficiency of nandrolone decanoate in improving any of the study outcomes and only 16 of the intended 44 patients were included. I would have excluded this study.
Besides this the methods are solid and the results are presented in a sober way.
>> This is a question of weighing bias from inclusion versus bias from exclusion of that study. While we agree that the inclusion of the study by Sharma et al, 2008 would have exerted some level of bias in our summary effect measures towards the null (showing no effect), it had to be included because it met our pre-specified inclusion criteria (please see pages 6 and 7 in the manuscript). Otherwise we would have biased our study selection process.

Second, despite the inclusion of the study by Sharma et al, 2008, the robustness of our meta-analysis findings (SDM was 0.31 [0.05,0.56]) presented in figure 4 was confirmed using sensitivity analysis. For instance, the sensitivity analysis presented in table 4 shows that the pooled SMD was almost identical after exclusion of that study (SMD was 0.31 [0.04,0.57]).

Finally, since the study by Sharma et al, 2008 had a small sample size; it contributed less than 7% and made no difference to the summary effect estimate.

Reviewer: Peter Watson
Statistician
MRC Cognition and Brain Sciences Unit
15 Chaucer Road
Cambridge
CB2 7EF

I have no competing interests with the research described in this paper.

I wasn't clear if Hedge's g was the effect size for the standardized mean difference (page 10) rather than, say, Cohen's d.
>> Thank you for identifying this omission. We have, top of page 9, inserted the sentence "Standardised mean differences were calculated using Glass's Delta method" as a follow-up to our existing comment that "Where necessary for RCTs, the post-treatment means were derived from the within group changes and the control group standard deviation carried forward from the baseline values".

I also wondered if the weightings used for the 'weighted mean difference' on the second line of page 10 uses inverse variance weights
>> Yes. We had now added the words 'inverse variance' at that location.

and if the DerSimonian estimate was used for any of the confidence intervals (third line of page 10) since at least some of these I assume must be derived from random effects meta-analyses (third last line of page 9).
>> Yes. We had now added the words 'DerSimonian and Laird' at that location

You could also remind us (in the third line on page 10) that the median-based estimates of means using medians rather than means are used with observed means to emphasise the pooling is based upon means (pseudo and actual).
>> Yes. We have replaced "Median and mean values were assumed to be equivalent estimates of central tendency..." with the more specific wording "Where papers presented medians and not means, we estimated the missing mean as being equal to the median..." We have also added this as a study limitation in our Discussion.

Like to see a one or two sentences inserted to mention the robustness of the median-based Hozo et al. (2005) estimates to distributions of responses within groups.
>> We have added: "using methods which have been shown to be reasonably robust in non-extreme circumstances" to the citation at the bottom of page 8.

On page 8 last paragraph there are at least three primary outcomes mentioned rather than one.
>> Several primary outcomes are listed because our systematic review had several, equally important aims. (Secondary outcomes are listed on page 9).

I found the sensitivity analyses rather clouded the picture as multiple results are presented (e.g. Tables 3-5 pages 30-32) for each primary outcome. Usually in the meta-analyses I have seen one chooses apriori the studies for inclusion and performs a meta-analysis on them or adjusts for a

covariate whilst doing the meta-analysis if the studies differ on covariates. You can fit study covariate adjusted meta-analyses in R using the 'metafor' package so the results are adjustments for study differences and this has the advantage over subsamples of using all the studies in the one analysis.
>> Each sensitivity analysis corresponds to slightly different inclusion criteria and hence each addresses a slightly different research question. One way of viewing the sensitivity analysis is as a check of how robust the results are to slight variations (tightening) in the research question. Meta-analysis could be used to quantify the relative impact of various possible predictors: a different objective. However, with just three to nine studies in any of our meta-analyses, we do not believe we have sufficient information to derive reasonable estimate of these effects.

Meta-analyses are performed yielding pooled estimates of standardised group comparisons to assess (page 6, first full paragraph) the influence of testosterone treatment on measures of exercise capacity (peak muscle strength and peak oxygen uptake) and health-related quality of life (HRQoL) on people with COPD in randomised control trials and differences in testosterone in those with chronic lung disease and controls. There are also some descriptive analyses of secondary outcomes (page 9, 'secondary outcomes' paragraph).

I have a few minor comments below which the authors may like to address relating in the main to further clarification of the methods used and the usefulness of the sensitivity analyses described on pages 13-14 and in Tables 3 to 5 (pages 30-32).

I also would welcome some reassurance on the plausibility of the apparent pooling of median-based estimates and means (page 10) since this implies that the groups they are obtained from have different distributions. The authors could reassure by adding into the text that Hozo et al. (2005) who they quote at the bottom of page 8 ran simulations to assess the closeness of these median-based estimates to means under a variety of distributions to assess their robustness and found the means were well reproduced albeit their variances less well so.
>> (from above) We have added: "using methods which have been shown to be reasonably robust in non-extreme circumstances" to the citation at the bottom of page 8.

I also found the sensitivity analyses confusing as more results were presented dropping various studies for the same outcome thus giving a range of results most of which suggest no, or little, benefit to using the testosterone therapy (forest plots of pages 36, 38 and 39 and results from last paragraph on page 13 to fourth line on page 15 which have 95% confidence intervals including, or near to, zero for treatment differences) and also use reduced samples. Perhaps my biggest concern is that possibly in light of these null results it is further suggested on page 18 that 'there is an absence of sufficient RCT evidence to draw firm conclusions' which would appear to rather limit the usefulness of this study.
>> The main analyses establish the clinical and statistical significance of the effects, or lack thereof. The role of the sensitivity analyses is to confirm that the observed clinical effect is stable and consistent across various scenarios (i.e. is not an artefact of one or two 'unusual' studies). Whether statistical significance changes is of little interest: we would generally expect fewer studies to deliver lesser statistical power.

Page 7. Was an inter-rater measure of agreement considered such as kappa to show the level of agreement between reviewers which could have been quoted in the 'data extraction' paragraph?
>> No such measures were collected. Each reviewer independently extracted more than a dozen items of information from each paper. We cannot identify any one piece of information as of greater important than of the others. Further we have no mechanism to quantify the severity of disagreement within any individual data item. Without information on the relative importance of various discrepancies, it would be very difficult to interpret any reliability statistics.

Page 10. (top three lines of first full paragraph). I assume the standardised mean difference is actually Hedge's g which is commonly used in meta-analyses
for comparing mean differences and that random effects (page 9, third last line) were incorporated using the DerSimonian and Laird estimate to produce 95% confidence intervals?
>> (from above): Thank you for identifying this omission. We have, top of page 9, inserted the sentence "Standardised mean differences were calculated using Glass's Delta method" as a follow-up to our existing comment that "Where necessary for RCTs, the post-treatment means were derived from the within group changes and the control group standard deviation carried forward from the baseline values".

>> (from above) Yes. We had now added the words 'DerSimonian and Laird' at that location.

Not sure even though it is referenced why you need to verify a random effects meta-analysis with a fixed one (second and third last line on page 9) since you are making the decision which to use by using I^2 (page 10, fourth to sixth line of the first full paragraph) to decide which to use and then using that. Doesn't using both fixed and random approaches invalidate using these decision criteria?
>> Again, the focus of the sensitivity analysis is on the clinical effect size more than the statistical characteristics of the analysis. We would not expect the estimated effect size to change greatly between fixed and random effects model. If a change in estimated clinical effect was observed, we would be encouraged to explore further.

Page 10. Third and fourth line from top. It might be less confusing here to say that it was only means that were pooled since (last two lines of age 8) the medians were transformed into means using the minimum and maximum values as suggested by Hozo et al. (2005) prior to pooling.
>> (from above) Yes. We have replaced "Median and mean values were assumed to be equivalent estimates of central tendency..." with the more specific wording "Where papers presented medians and not means, we estimated the missing mean as being equal to the median..." We have also added this as a study limitation in our Discussion.

I would say, however, that medians are usually quoted for skewed distributions whereas means assume more symmetric distributions within group so one would be pooling summary measures from different (shaped) distributions. It may be the distributions within groups differ due to different processes going on due to different sampling biases e.g. some groups may be based upon older people or mixed gender populations who contain more outlying observations than younger people or male only populations and so age and gender may be confounding variables for looking at group differences between studies. Do the authors, therefore, have any comments they could add to reassure on combining estimates based upon different distributions as implied by the use of medians and means. Medians also have different variances to means depending on sample size (Kenney and Keeping 1962, p. 211) and distribution which may influence pooling if weightings related to inverse variances are used (line two on page 10 mentions a pooled 'weighted' mean difference so these may have been used).
>> We have added this as a study limitation in our Discussion. "we have replaced missing data points with estimates in some instances, introducing further uncertainty. This has included both estimating the mean from the median and range and carrying forward the pre-intervention standard deviation of control groups where the post-intervention statistic was not available"

Page 10. Line five of the second paragraph. Not sure here why a 40% threshold was used for deciding a high level of between study heterogeneity using I^2. This doesn't appear to correspond to any of the thresholds as suggested by Higgins et al. (2003) who suggest a value of 0% indicates no observed heterogeneity, 25%-49% is low heterogeneity, 50%-74% is moderate and 75% and above is large.
>> We have replaced "(I-squared values >40%)" with "(moderate being < 50% [26])" where [26] is

Higgins, J.P., Thompson, S.G., Deeks J.J. and Altman, D.G. (2003) Measuring inconsistency in meta-analyses. BMJ 327 557-560.

Page 10. Third line from bottom of second full paragraph. Did you also consider in addition to funnel plots also using inference to assess the degree of publication bias using a statistical test? See for example Peters et al. (2010).
>> With just 3 to 9 papers available for analysis, it is highly unlikely that a statistical test would return a positive result. Reporting a negative finding may encourage readers to form the wrong conclusion. We believe a plot alone to be more informative in this particular instance.

Page 13. The last line has a point estimate of -0.31 which is outside the confidence interval.
>> Thank you. The erroneous negative sign has been removed.

Page 18. The sensitivity analyses all seem to conclude that there is little or no benefit to using a testosterone therapy on primary outcomes (as for example shown in the confidence intervals on pages 34, 36, 38 and 39 and quoted from last paragraph on page 13 to first four lines on page 15 for the overall standardized mean difference containing or being close to zero) so I am not convinced the therapy improves several exercise capacity outcomes as stated on the second line of page 18 or that null results can make us hope that we have insufficient evidence to 'draw firm conclusions about the long-term benefits' of testosterone therapy as further stated on lines two to four of page 18.
>> (from above) The main analyses establish the clinical and statistical significance of the effects, or lack thereof. The role of the sensitivity analyses is to confirm that the observed clinical effect is stable and consistent across various scenarios (i.e. is not an artefact of one or two 'unusual' studies). Whether statistical significance changes is of little interest: we would generally expect fewer studies to deliver lesser statistical power.

Page 30. Table 3. The first line in Table 3 quoted the use of a fixed effects model which by definition assumes no between study heterogeneity in effect sizes yet the test of heterogeneity in the right-most column in Table 3 appears to suggest that this is present ($p<0.001$) suggesting random effects may be more appropriate. I suspect the fixed model was used because the I^2 value was 'small' despite being statistically significant so the I^2 value could be inserted in this table to emphasise this as this is the value used to decide if fixed effects or random effect models are used (as intimated in line 5 of the second paragraph on page 10).
>> (from above) Again, the focus of the sensitivity analysis on the clinical effect size more than the statistical characteristics of the analysis. We would not expect the estimated effect size to change greatly between fixed and random effects model. If a change in estimated clinical effect was observed, we would be encouraged to explore further.

Pages 30-32. The sensitivity analyses in Tables 3 to 5 (and described on pages 13-14) using subsets of studies can be interpreted as suggesting doubts about the usefulness of combining studies of, for example, different qualities in an overall analysis. I would have thought it simpler if the decision of what studies to include in a meta-analysis was made apriori and a meta-analysis only carried out on the studies deemed of appropriately high quality. The single confidence interval for these studies could then be quoted in the text and there would be no need for Tables 3 to 5. You might also wish to consider performing a single meta-analysis on each outcome that removes differences in study level covariates e.g. quality of study using, for example, the 'metafor' package in R although STATA may also do this. This would also help with multiplicity caused by repeated testing which might through chance yield confidence intervals which do not contain zero.
>> (from above) Each sensitivity analysis corresponds to slightly different inclusion criteria and hence each addresses a slightly different research question. One way of viewing the sensitivity analysis is as a check of how robust the results are to slight variations (tightening) in the research question. Meta-analysis could be used to quantify the relative impact of various possible predictors: a different

objective. However, with just four to nine studies in any of our meta-analyses, we do not believe we have sufficient information to derive reasonable estimate of these effects.
>> (additionally) The quality of the findings is dependent on the quality of the data on which analysis is based. Removing various papers from the analyses provides one form of 'quality' check.

I was also not sure from the third row of Table 3 (page 30) what 'unadjusted' and 'adjusted' related to? Is this adjustments for publication bias?
>> This issue was explained in the text as "and in a model using unadjusted rather than adjusted values in one study[33]". To add further clarity we have now expanded this to read "Finally, for the one study[34] which provided both unadjusted mean differences and mean differences adjusted for age, waist circumference and smoking status, a model using unadjusted rather than adjusted values decreased the pooled WMD to -2.95 (-4.63, -1.27)."

Pages 36, 38 and 39. Judging from the overall 95% confidence intervals (using all the studies) of the standardized difference the effect of the testosterone therapy is either very small or not present.
>> We have provided the appropriate interpretations in our paper.

Pages 38 and 39. There appear to be no funnel plots corresponding to the standardized mean differences for outcomes in the forest plots in Figures 6 and 7 yet these are presented for the other forest plots on pages 34 and 36.
>> The first two graphs display statistically significant effects, the second two do not. Publication bias could reasonably be suspected of contributing to the positive findings in Figures 2 and 4, but is much less likely to be an important contributor to the negative results in Figures 6 and 7.

Page 39. Figure 7 appears to be based upon only three studies (with one of these having a small weight < 10%) which seems a small number for obtaining pooled estimates.
>> We agree. We acknowledge this in 'Limitations' of the 'Discussion': "there is an absence of sufficient RCT evidence to draw firm conclusions".

References

Higgins, J.P., Thompson, S.G., Deeks J.J. and Altman, D.G. (2003) Measuring inconsistency in meta-analyses. BMJ 327 557-560.

Kenney, J. F. and Keeping, E. S. "The Median," "Relation Between Mean, Median, and Mode," "Relative Merits of Mean, Median, and Mode," and "The Median." §3.2, 4.8-4.9, and 13.13 in Mathematics of Statistics, Pt. 1, 3rd ed. Princeton, NJ: Van Nostrand, pp. 32-35, 52-54, 211-212, 1962.

Peters, J. L., Sutton, A. J., Jones, D. R. and Abrams K. R. (2010). Assessing publication bias in meta-analyses in the presence of between-study heterogeneity. Journal of the Royal Statistical Society A 173(3) 575-591.