PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Testosterone therapy in hypogonadal men: a systematic review and network meta-analysis
AUTHORS	Elliott, Jesse; Kelly, Shannon; Millar, Adam; Peterson, Joan; Chen, Li; Johnston, Amy; Kotb, Ahmed; Skidmore, Becky; Bai, Zemin; Mamdani, Muhammad; Wells, George

VERSION 1 - REVIEW

REVIEWER	Alisa Stephens-Shields University of Pennsylvania
REVIEW RETURNED	22-Dec-2016

GENERAL COMMENTS	Major: 1. The results section of the abstract should contain quantitative data to support statements. What metric/scoring system was used to assess for risk of bias, and what was the respective value of that measure? What are the measures of effect, 95% CIs, and p-values supporting the statements about benefits and lack of harm? 2. Page 9, Line 10. The main text, outcomes section needs a lot more detail. a. The grouping of outcomes into benefits and harms is ambiguous in that it's not clear whether 'benefit' was a single outcome combining information across domains, or if each domain was treated as a separate outcome and just combined into the groups benefits or harms for presentation. The results section clarifies this in the presentation, but it should be clear from the description of the outcomes as well. b. Were outcomes continuous or binary, and what was the measure of effect of these outcomes that was being combined across studies? This should all be detailed. c. A complete list of scales grouped by domain (QOL, libido, etc) should be provided in an appendix along with a definition of how benefit was defined for each. I would assume this is based on established minimum clinically important differences. 3. Page 9, Line 32. Why was the number who received treatment used as the denominator for harms? Use of this denominator prevents the causal comparison of treated and untreated subjects. A reference is needed to justify this approach. 4. Page 2, Line 46. Why were these models used for the

	network metaanalysis but not the standard metaanalysis? As stated in #2, outcomes should be sufficiently described so that the reader knows which were categorical and which were continuous. 5. Page 2, Line 53. What were all of the treatments that defined nodes for the network meta? 6. Page 2, Line 57. What is the measure of effect targeted in the network meta analysis? 7. Page 10, Line 46. What were comparator treatments in studies that were not placebo-controlled? Relatedly, what were the treatments that defined the nodes in the network meta? 8. Page 11, line 19. How many different treatments, and which treatments, were combined into 'any TRT'? 9. Page 11, line 21. The direction of effect is hard to interpret. Were all scales uniform in direction or appropriately transformed so that lower scores represented better outcomes? It should be stated that lower scores represent better outcomes so that the negative measure of effect is interpreted as favoring testosterone treatment. 10. Page 11, line 31. Related to #8, explain more clearly in the methods section the difference in how treatments were handled in the standard metaanalysis of RCTs compared to the network meta analysis. Relative to the standard meta, 2 additional trials are considered, but 14 treatments are compared instead of just two levels of treatment (resulting from grouping) in the meta. 11. Page 11, line 33. N=2698 can't be correct for both the standard meta and the network meta. Correct appropriately. 12. Page 11, line 39. How is consistency assessed? This should be described in the methods with relevant references. 13. Pages 11-13. All discussion of results from the standard meta of RCTs should be clear to note that the effect is of 'any testosterone' rather than just 'testosterone' since 'testosterone' alone is ambiguous. 14. Page 13, line 41. This statement is ambiguous. Not clear here whether 'any harm' means that a single indicator of harm was created, or that across multiple endpoints that were evaluated separately none were significant. The methods section should state which benefits and harms outcomes were evaluated separately. 15. Page 16, line 37. The lack of long term follow up, in addition to short treatment duration, is still a limitation. This should be stated. Minor: 1. The acronym NRS should be defined along with RCR in the section on eligibility for selecting studies. 2. Figure 2. A and B are inconsistent in the axis labels. Please add a footnote explaining this or correct so that they match. Similarly, the column headings are inconsistent between A and B.

	3. Page 59, line 6. This note appears to belong on the following page 61.

REVIEWER	Dohle GR Erasmus MC Rotterdam, The Netherlands
REVIEW RETURNED	11-May-2017

GENERAL COMMENTS	The meta-analyses performed by Elliott and coworkers is an interesting manuscript, shedding new light on the controversies around testosterone therapy. The authors should be congratulated for performing such an extensive meta-analyses. Some questions remain: - the authors conclude that there is a serious risk of Bias with many papers they have reviewed. Still they are quite firm in their conclusions about the results of this meta-analyses. As always, if you include poor quality science, even a meta-analyses will not overcome this problem. heterogeneity, poor description of methods and outcomes in publications and potential publication bias (industry sponsored or not)severely limit the outcomes of this meta-analyses. - The authors conclude that testosterone has no influence on sexual function. This is in contrast to recent randomized controlled trials, that clearly show a beneficial effect, both on libido and erectile function. See ref 38 (Snyder et al NEJM) and 58 (Hackett et al). Also a recent large RCT (Brock G, et al. Effect of Testosterone Solution 2% on Testosterone Concentration, Sex Drive and Energy in Hypogonadal Men: Results of a Placebo Controlled Study. J Urol, 2016. 195: 699.) showed a clear beneficial effect of testosterone gel on male sexual dysfunction. The authors should assign more value to the results of large RCTs in judging the effects of testosterone therapy. Just adding the results of all studies, including the poor studies will not bring a valuable answer. - The reporting of side effects of different testosterone products is probably the most important part of this work: only data from many patients treated can give answers to the important question of for instance potential cardiovascular complications due to testosterone therapy. RCT are usually to small to show the negative effects of any drug, especially if the complications have a low recurrence rate. How many patients were included in the analyses of the different side effects? page 12 and 13.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1
Major:
1. The results section of the abstract should contain quantitative data to support statements. What metric/scoring system was used to assess for risk of bias, and what was the respective value of that measure? What are the measures of effect, 95% CIs, and p-values supporting the statements about benefits and lack of harm?

Response: We have added effect estimates and 95% confidence intervals (CIs) for the comparisons of any testosterone product versus placebo for the outcomes libido, erectile function, quality of life,

and depression to the results section of the abstract. Because of the nature of network meta-analysis results, we were unable to add individual effect measures for each of the outcomes. Likewise, because of the word limits, we were unable to add the effect estimate and 95% CIs for each of the harms outcomes without substantially increasing the word count.

(Abstract) “Results: 87 RCTs and 51 NRS were included. Most were at high or unclear risk of bias. When compared as a class against placebo, TRT improved quality of life (standardized mean difference [SMD] -0.26, 95% confidence interval [CI] -0.41,-0.11), libido (SMD 0.33, 95%CI 0.16,0.50), depression (SMD -0.23, 95%CI -0.44,-0.01), and erectile function (SMD 0.25, 95%CI 0.10,0.41). Most individual TRTs were significantly better than placebo at improving libido (6/10), with few differences among the TRTs. Only one TRT was better than placebo at improving quality of life, and no individual TRTs significantly improved depression or erectile function. There was no increased risk of adverse events, with the exception of withdrawals due to adverse events with the use of some TRTs; however, most included trials were of short treatment duration and follow-up, and at high or unclear risk of bias.”

2. Page 9, Line 10. The main text, outcomes section needs a lot more detail.
a. The grouping of outcomes into benefits and harms is ambiguous in that it’s not clear whether ‘benefit’ was a single outcome combining information across domains, or if each domain was treated as a separate outcome and just combined into the groups benefits or harms for presentation. The results section clarifies this in the presentation, but it should be clear from the description of the outcomes as well.

Response: In the meta-analyses and network meta-analyses, each outcome was considered separately. We had grouped the outcomes as “benefits” and “harms” for presentation only. On this reviewer’s advice, we have removed the grouping from the abstract and main text.

(Page 8, Paragraph 3) “Outcomes: Outcomes of interest were quality of life, depression, libido, erectile function, activities of daily living, and total testosterone level (at 3 mo, 6 mo, end of study) (continuous outcomes), as well as cardiovascular death, myocardial infarction, stroke, prostate cancer, diabetes, heart disease, serious adverse events, withdrawals due to adverse events, and erythrocytosis (dichotomous outcomes).”

b. Were outcomes continuous or binary, and what was the measure of effect of these outcomes that was being combined across studies? This should all be detailed.

Response: The analyses of the outcomes quality of life, libido, erectile function, depression, and total testosterone levels are based on continuous data, while the safety/harm outcomes were based on dichotomous data (occurrence of an event). We have clarified which outcomes are continuous and which are dichotomous in the main text, and we have indicated the measures of effect for each outcome.

(Page 8, Paragraph 3) “Outcomes: Outcomes of interest were quality of life, depression, libido, erectile function, activities of daily living, and total testosterone level (at 3 mo, 6 mo, end of study) (continuous outcomes), as well as cardiovascular death, myocardial infarction, stroke, prostate cancer, diabetes, heart disease, serious adverse events, withdrawals due to adverse events, and erythrocytosis (dichotomous outcomes).”

(Page 9, Paragraph 2) “Mean differences (MDs), standardized mean differences (SMDs) with standard deviations (SDs) or odds ratios (ORs) with 95% CrIs or CIs are reported for continuous or dichotomous outcomes as appropriate.”

c. A complete list of scales grouped by domain (QOL, libido, etc) should be provided in an appendix along with a definition of how benefit was defined for each. I would assume this is based on established minimum clinically important differences.

Response: We have added an appendix that includes a list of scales for each outcome assessed by use of a scale (libido, erectile function, depression, quality of life). Our analyses were based on the reported data for each outcome, without interpretation of "benefit". If the confidence interval or credible interval did not include zero for standardized mean differences and mean differences, we inferred that there was a statistically significant improvement in the outcome. We have clarified this in the main text:

(Page 8, Paragraph 3) We included data for quality of life, depression, erectile function, and libido that had been measured using a validated scale (eAppendix2), and the direction of each scale was standardized before analysis. A higher effect estimate (e.g., positive SMD or MD) indicates improvement in libido, erectile function and testosterone level, and a lower effect estimate indicates improvement in quality of life and depression.

3. Page 9, Line 32. Why was the number who received treatment used as the denominator for harms? Use of this denominator prevents the causal comparison of treated and untreated subjects. A reference is needed to justify this approach.

Response: Our analyses for adverse events (harms) involved the number who received the treatment to which they were randomized as the denominator in order to get a true picture of the harms associated with the treatment. In most of the included RCTs, the number who received treatment was the same as the number randomized. Overall, 99.7% of men randomized to placebo received the placebo treatment, while 99.5% of men randomized to a testosterone treatment received the treatment. As such, the use of the number randomized instead of the number treated as the denominator would not have a major impact on the results.

4. Page 2, Line 46. Why were these models used for the network metaanalysis but not the standard metaanalysis? As stated in #2, outcomes should be sufficiently described so that the reader knows which were categorical and which were continuous.

Response: The use of the binomial likelihood model for dichotomous outcomes and a normal likelihood model for continuous outcomes in the network meta-analysis allows the incorporation of multi-arm trials (e.g., placebo v. testosterone gel v. testosterone patch). In contrast, meta-analysis allows comparison of only two arms at a time. In the meta-analysis, for trials that involved more than one testosterone arm, we pooled the data for the comparison of testosterone v. placebo. We have added a description of which outcomes were continuous and which were dichotomous as described above.

5. Page 2, Line 53. What were all of the treatments that defined nodes for the network meta?

Response: The treatment that comprised the evidence network for each network meta-analysis are presented in Figure 3 (quality of life) and eAppendix5 (depression, erectile function, libido). We have added a call-out to the appendix to clarify this:

(Page 9, Paragraph 2) "Each dose of an individual testosterone product was included as a separate node (Figure 3, eAppendix6)."

6. Page 2, Line 57. What is the measure of effect targeted in the network meta analysis?

Response: The measures of effect are described in “data analysis” section of the main text as follows:

(Page 9, Paragraph 2) “Mean differences (MDs), standardized mean differences (SMDs) with standard deviations (SDs) or odds ratios (ORs) with 95% CrIs or CIs are reported for continuous or dichotomous outcomes as appropriate.”

7. Page 10, Line 46. What were comparator treatments in studies that were not placebo- controlled? Relatedly, what were the treatments that defined the nodes in the network meta?

Response: For each outcome analyzed by network meta-analysis, the included treatments and the connections between the trials (direct evidence) can be seen in the evidence networks (Figure 3, eAppendix 6). For example, in Figure 3 (evidence network for quality of life; below), we can see that IM TU (intramuscular testosterone undecanoate; 1000 mg/12 wk) has been compared to placebo in 6 RCTs, but that it has also been compared with Oral TU (160 mg/d) in 1 RCT, IM TC (testosterone cypionate 200 mg/4wk) in 1 RCT, and to IM Durateston (250 mg/4 wk) in 1 RCT. As well, the comparators for each RCT are listed in the table of characteristics for RCTs (eTable1).

8. Page 11, line 19. How many different treatments, and which treatments, were combined into ‘any TRT’?

Response: The number and type of treatments combined varies by outcome. Interested readers are referred to the meta-analysis figures (Figure 2, eFigures2-10) for information about which studies were included and to eTable1 for details of each study including the interventions and testosterone doses.

9. Page 11, line 21. The direction of effect is hard to interpret. Were all scales uniform in direction or appropriately transformed so that lower scores represented better outcomes? It should be stated that lower scores represent better outcomes so that the negative measure of effect is interpreted as favoring testosterone treatment.

Response: All scales were transformed to a common direction before analysis. We have added this information to the methods section, and we have indicated for which outcome a negative or positive difference indicates an improvement.

(Page 8, Paragraph 3) “We included data for quality of life, depression, erectile function, and libido that had been measured using a validated scale (eAppendix2), and the direction of each scale was standardized before analysis. A higher effect estimate (e.g., positive SMD or MD) indicates improvement in libido, erectile function and testosterone level, and a lower effect estimate indicates improvement in quality of life and depression.”

10. Page 11, line 31. Related to #8, explain more clearly in the methods section the difference in how treatments were handled in the standard metaanalysis of RCTs compared to the network meta analysis. Relative to the standard meta, 2 additional trials are considered, but 14 treatments are compared instead of just two levels of treatment (resulting from grouping) in the meta.

Response: We have clarified how the data were handled in the meta-analysis by adding a statement about pooling of data in multi-arm trials. For network meta-analysis, we used models that allowed for the inclusion of multi-arm trials.

(Page 8, Paragraph 4) “Data from RCTs with more than one testosterone arm were pooled before

meta-analysis."

(Page 9, Paragraph 2) "In the network meta-analysis, we used a binomial likelihood model for dichotomous and a normal likelihood model for continuous outcomes, allowing for the inclusion of multi-arm trials"

11. Page 11, line 33. N=2698 can't be correct for both the standard meta and the network meta. Correct appropriately.

Response: We have revised the text to better reflect the number of trials that informed the meta-analysis and network meta-analysis for each outcome. Differences between the two n values are because head to head trials of testosterone treatments that lacked a placebo control could not be included in the meta-analysis, resulting in a lower n for meta-analyses. For example:

(Page 10, Paragraph 4) "Quality of life: In total, 23 RCTs (21 placebo-controlled, 2 active-controlled) involving 3090 participants assessed quality of life. Compared with placebo, treatment with any TRT significantly improved quality of life (SMD –0.26, 95%CI –0.41,–0.11; n = 2834) with substantial heterogeneity (I2 = 71%; Figure 2A)."

12. Page 11, line 39. How is consistency assessed? This should be described in the methods with relevant references.

Response: The methods for evaluating consistency (inconsistency) between direct and indirect evidence in the network meta-analysis are described in the methods section and has been referenced:

(Page 9, Paragraph 2) "Inconsistency was assessed where possible by comparing the deviance, between-study variance, and deviance information criterion statistics of the consistency and inconsistency models.26 The posterior mean deviance of the individual data points in the inconsistency model was plotted against the posterior mean deviance in the consistency model.26

26. Dias S, Welton N, Sutton A, Caldwell D, Lu G, Ades A. Evidence synthesis for decision making 4: inconsistency in networks of evidence based on randomized controlled trials. Med Decis Mak. 2013;33(5):641-56

13. Pages 11-13. All discussion of results from the standard meta of RCTs should be clear to note that the effect is of 'any testosterone' rather than just 'testosterone' since 'testosterone' alone is ambiguous.

Response: We have revised the text to better distinguish between testosterone as a class (all testosterone products grouped together via meta-analysis) and the individual testosterone treatments compared via network meta-analysis. For example:

(Page 10, Paragraph 4) "Compared with placebo, treatment with any TRT significantly improved quality of life (SMD –0.26, 95%CI –0.41,–0.11; n = 2834) with substantial heterogeneity (I2 = 71%; Figure 2A)."

14. Page 13, line 41. This statement is ambiguous. Not clear here whether 'any harm' means that a single indicator of harm was created, or that across multiple endpoints that were evaluated separately none were significant. The methods section should state which benefits and harms outcomes were evaluated separately.

Response: We have revised the methods section to clarify that all outcomes were evaluated separately (removed the grouping of harms), and we have revised the “Other adverse events” section of the results to better reflect this.

(Page 13, paragraph 3) “Compared with placebo via meta-analysis, there were no increased odds of myocardial infarction, stroke, prostate cancer, heart disease, or erythrocytosis with the use of any TRT product (Table 2; eFigure 6-11). Withdrawals due to adverse events were significantly greater with the use of any TRT compared with placebo (OR 1.31, 95% CI 0.95, 1.73; I2 = 13%)(eFigure12). No RCTs reported incident diabetes during the treatment period.”

15. Page 16, line 37. The lack of long term follow up, in addition to short treatment duration, is still a limitation. This should be stated.

Response: We agree that the lack of long term follow-up in the RCTs is an important limitation, and we have added this to the limitation section as well as to the conclusion:

(Page 17, Paragraph 2) “Third, the duration of treatment and the length of follow-up may have been too short to see an effect of TRT for all outcomes, including adverse events."

(Page 17, Paragraph 3) “We found no increased risk of major harms; however, this must be viewed in light of the high risk of bias of the included studies, the rare nature of serious harms, and the short treatment duration and follow-up of most studies.”

Minor:
1. The acronym NRS should be defined along with RCR in the section on eligibility for selecting studies.

Response: We have expanded the acronym NRS in full on its first mention in the abstract and main text.

2. Figure 2. A and B are inconsistent in the axis labels. Please add a footnote explaining this or correct so that they match. Similarly, the column headings are inconsistent between A and B.

Response: We have updated the axis headings of all figures for consistency.

3. Page 59, line 6. This note appears to belong on the following page 61.

Response: We have removed the note on page 59, and have inserted the information contained within the note (random-effects model, how to interpret the colour in the network meta-analysis results table) as footnotes in each relevant table.

Reviewer: 2
- the authors conclude that there is a serious risk of Bias with many papers they have reviewed. Still they are quite firm in their conclusions about the results of this meta-analyses. As always, if you include poor quality science, even a meta-analyses will not overcome this problem. heterogeneity, poor description of methods and outcomes in publications and potential publication bias (industry sponsored or not)severely limit the outcomes of this meta-analyses.

Response: We agree with the reviewer that the risk of bias within the included studies is an important limitation. We have expanded the limitation section of our manuscript to further emphasize this point. We have concluded, on the basis of all available evidence, that testosterone improves quality of life,

depression, libido and erectile function; however, given the moderate to high risk of bias assessed for a large number of included studies, this review highlights the need for RCTs to evaluate the effectiveness of TRT in men with androgen deficiency. Future studies need to be rigorous in design and delivery, and include comprehensive descriptions of all aspects of methodology to enable appraisal and interpretation of results.

We have revised the conclusion of our manuscript to highlight this important consideration:

(Page 17, Paragraph 3) "To the best of our knowledge, this is the first study to compare the benefits and harms of individual testosterone products among hypogonadal men. Our study builds on previous meta-analyses by comparing the relative effects of individual testosterone treatments, most of which have never been compared in head-to-head trials. When considered as a class (any TRT compared to placebo), TRT improved quality of life, depression, erectile function, and libido; however, when the individual products were compared head to head, there were few differences between the treatments. We found no increased risk of major harms; however, this must be viewed in light of the high risk of bias of the included studies, the rare nature of serious harms, and the short treatment duration and follow-up of most studies. Future studies need to be rigorous in design and delivery, and include comprehensive descriptions of all aspects of methodology to enable appraisal and interpretation of results."

- The authors conclude that testosterone has no influence on sexual function. This is in contrast to recent randomized controlled trials, that clearly show a beneficial effect, both on libido and erectile function. See ref 38 (Snyder et al NEJM) and 58 (Hackett et al). Also a recent large RCT (Brock G, et al. Effect of Testosterone Solution 2% on Testosterone Concentration, Sex Drive and Energy in Hypogonadal Men: Results of a Placebo Controlled Study. J Urol, 2016. 195: 699.) showed a clear beneficial effect of testosterone gel on male sexual dysfunction. The authors should assign more value to the results of large RCTs in judging the effects of testosterone therapy. Just adding the results of all studies, including the poor studies will not bring a valuable answer.

Response: In our updated review, the recent study by Brock et al (2016), as well as the Testosterone Trials (Snyder et al 2016), has been included. After updating our systematic review, our updated analysis shows a positive effect of testosterone on quality of life, depression, erectile function, and libido, which is in keeping with findings from the RCTs mentioned by the reviewer, as well as previous meta-analyses.

- The reporting of side effects of different testosterone products is probably the most important part of this work: only data from many patients treated can give answers to the important question of for instance potential cardiovascular complications due to testosterone therapy. RCT are usually to small to show the negative effects of any drug, especially if the complications have a low recurrence rate. How many patients were included in the analyses of the different side effects? page 12 and 13.

Response: We have added a table to the main text (Table 2), which summarizes the meta-analysis findings for any testosterone product compared with placebo. This table includes the number of patients as requested by the reviewers, as well as the duration of treatment for each outcome. We feel that adding this table the main text will aid in the readability and transparency of our manuscript.

VERSION 2 – REVIEW

REVIEWER	Gert Dohle department of urology, Erasmus MC Rotterdam, The Netherlands
REVIEW RETURNED	26-Jun-2017

GENERAL COMMENTS	The article has been improved substantially and the authors have included new references from RCTs published in 2016. Still, I am somewhat surprised by the seemingly contradiction in the results of this meta-analyses: it may be my lack of understanding high level statistics, but I don`t understand why no individual TRTs (RCTs) improved depression and erectile dysfunction. In both the Snyder T-trial (NEJM 2016) and in the RCT of Brock et al testosterone 2% gel was used against placebo. Both trials showed an increase in erectile function, although moderate. This should result in a different conclusion or should be explained more clearly if the conclusion of the authors is correct.

VERSION 2 – AUTHOR RESPONSE

Reviewer 2: The article has been improved substantially and the authors have included new references from RCTs published in 2016. Still, I am somewhat surprised by the seemingly contradiction in the results of this meta-analyses: it may be my lack of understanding high level statistics, but I don`t understand why no individual TRTs (RCTs) improved depression and erectile dysfunction. In both the Snyder T-trial (NEJM 2016) and in the RCT of Brock et al testosterone 2% gel was used against placebo. Both trials showed an increase in erectile function, although moderate. This should result in a different conclusion or should be explained more clearly if the conclusion of the authors is correct.

Author response: We agree with Reviewer 2 that some individual RCTs reported a positive effect of TRT on depression and erectile function. This is apparent discrepancy between direct evidence from RCTs and the results of the network meta-analysis has been noted previously and is a known limitation of network meta-analysis, and has been discussed at the Cochrane Colloquium (Guyatt G., 2015). Overall, this results in a more conservative estimate of effect. We have added a statement to the limitation section of our manuscript to capture this.

“Fourth, although some individual RCTs showed a positive effect of TRT on depression and erectile function compared with placebo, in the network meta-analysis no individual TRTs showed a positive effect compared with placebo. This phenomenon has been noted in previously and results in a more conservative estimate of effect.47”

Reference 47: Guyatt G. Problems with Bayesian random effects in network meta-analysis. Cochrane Colloq. 2015;Vienna(October 3-7)

VERSION 3 – REVIEW

REVIEWER	Dohle GR Erasmus MC Rotterdam, The Netherlands
REVIEW RETURNED	10-Aug-2017

GENERAL COMMENTS	No further comments