PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Comparison of photoselective green light laser vaporization versus traditional transurethral resection for benign prostate hyperplasia: an updated systematic review and meta-analysis of randomized controlled trials and prospective studies
AUTHORS	Lai, Shicong; Peng, Panxin; Diao, Tongxiang; Hou, Huimin; Wang, Xuan; Zhang, Wei; Liu, Ming; Zhang, Yaoguang; Seery, Samuel; Wang, Jianye

VERSION 1 – REVIEW

REVIEWER	Jian Zhuo Department of Urology, Shanghai General Hospital Shanghai Jiaotong University China
REVIEW RETURNED	30-Jan-2019

GENERAL COMMENTS	The authors studied the most advanced treatment of BPH, compared it with the gold standard, and clarified the safety and effect of PVP technology. The results provide evidence-based medical for the further application of PVP in clinical practice. The article has a rigorous conception, detailed data and reliable statistical methods. Recommendations are given priority.

REVIEWER	C. Brunken Asklepios Westklinikum Hamburg, Dep. Urology, Germany
REVIEW RETURNED	31-Jan-2019

GENERAL COMMENTS	The labeling of all forest plots is missing. Every else is fine.

REVIEWER	Niraj Kumar VMMC and Safdarjung Hospital
REVIEW RETURNED	02-Feb-2019

GENERAL COMMENTS	Authors have made an effort to compare the current status of PVP with that of TURP. however, it needs some refinements: 1. Abstract: Result section line 3: due to insufficient data meta-analysis of data for 36 and 60 months follow up was not done, so, comment on 36.60 months outcome not justified. 2. Introduction, page 5 line 6: PVP is done with 532nm green light laser, the word "predominantly" not needed. 3. Results, section 1.1: line 2: IPSS difference at 12 months is not comparable

	4. Results, section 1.4: line 2: QOL difference at 6 months is not comparable 5. Results, section 1.5: line 2: IIEF at 24 months was not "slightly lower" but 'significantly lower' 6. Results, section 2.3: line 4: needs correction. 7. Results, section 3.1: line 3: the difference is significant 8. Discussion, paragraph 2 line 3: needs correction as QOL at 6 months and Qmax at 12 months are significant. 9. Page 15 line 5: needs correction, PVP done with saline as irrigant 10. 10 Forrest plots can be merged into 5. 11. Standard of English could have been better.

REVIEWER	Chris Jones Brighton and Sussex Medical School, UK
REVIEW RETURNED	05-Mar-2019

GENERAL COMMENTS	Generally the manuscript contains a good amount of detail and the results are reported reasonably clearly. However, there are many grammatical and spelling errors. The manuscript would benefit greatly careful proof reading and correction of the English. In particular, the manuscript needs to be proof read for spelling errors that won't be picked up by a spellchecker - such as "mouth" instead of month, "trails" instead of trials, and "sensitive" instead of sensitivity. Table 1 (and discussion): A column to indicate whether each study is a superiority design (or non-inferiority/equivalence) would be useful. I expect most (or all) of these studies were superiority designs - i.e. the null hypotheses are that PVP and TURP are the same with respect to each outcome. Lack of an apparent difference between them (i.e. a p-value >0.05) means we do not have enough evidence to reject the null hypothesis that PVP and TURP have the same effect - but it does not necessarily mean that they are the same (i.e. equivalent or non-inferior). It is therefore misleading to conclude that PVP is non-inferior to TURP if only superiority studies have been performed. The correct interpretation is that there is currently not enough evidence to conclude that they are different. Table 2: Sample size for IIEF in PVP group is 1.10? The tests for "overall effect" (difference in baseline outcome measure between intervention groups) here are invalid for randomised studies, as randomisation means that the null hypothesis of no difference is true - therefore there is no sense testing it. These tests should either be justified (if I have misunderstood their purpose) or removed. Tables in general: - Reporting (for example) "P<0.00001" is rather extreme, P<0.001 would suffice. It is good to see that all p-values in the manuscript are reported as their actual values (rather than just p<0.05 vs p>0.05). - Rounding of numbers (such as means and SDs) is inconsistent throughout the tables (some to 1 d.p., others to 2 d.p.) In the Forest plots, the weights of some of the studies for some outcomes at different time points vary wildly. The weights of the Tasci 2008 study for some of the outcomes at 3 and 12 months

	seem particularly odd as they are >90%, while at 6 and 24 months they are much lower (<40%). What is going on here? Even a weight of 40% is oddly high for one of the smaller studies. Tagu 2008 and Pereira-Correiaet 2012 also have extremely variable weights for different outcomes/time points (sometimes >90%) - why? The Forest plots are grouped into "80W", "120W", "180W" and "No mentioned" (should be "Not stated"), but no mention of this is made in the results. The figures would be simplified without the subgroups, so I would consider removing them unless there is a strong justification for showing them. If they are kept, I don't think the measures of heterogeneity for every subgroup add any useful information as the number of studies in each subgroup is so small. It would be better to only quantify heterogeneity over all of the studies. "reached a statistically significant difference" in the abstract should be changed to "at 12 month follow-up the difference was statistically significant, but was of no clinical significance" as "reached a statistically significant difference" implies the objective is to reach statistical significance. Units need to be given when stating mean differences/95% CIs in the results. I squared is also missing the % symbol. "less complications rates" should be "lower complication rates" in the abstract/discussion.

REVIEWER	Joanna IntHout, statistician Radboudumc, The Netherlands
REVIEW RETURNED	24-Mar-2019

GENERAL COMMENTS	General Thanks for the opportunity to review the paper. As I am a statistician, I focused on the methodology of the paper. However, one non-methodological remark: The English must be improved. I strongly advice that a person more experienced with the English language reads the manuscript. There are many errors in the language. The authors state it is an updated meta-analysis (strengths and limitations) but I wonder whether not an “up to date” meta-analysis was meant. Otherwise, they should refer to the previous meta-analysis. Page 6, add reference to PRISMA Page 6 / Table 1: not clear to me how the 22 publications are connected to the 19 clinical trials. I guess this must be clear from Table 1, but for me it is not clear. It is for example not clear why a study like Ruszat et al has many rows in this table. Explain in the legend of Table 1 which characteristics you show (e.g. mean plusminus SD), median (range)? Further the study of Tasci has a much higher prostate size than the other studies.

	Statistical analysis: Mouth = month You state that you used the standardized mean difference, but where did you use it? I guess you did not count the means and standard deviations but calculated them. I2 is not a test but a measure of heterogeneity. The chi-squared test is the test. You define in the paper when you used a random effects model and when a fixed effect model (if nonsignificant p-value and I2 <50%). However, the result is that for the same parameter at some timepoints a FE model and sometimes a RE model has been used. This affects the p-values (FE model results more easily in a significant treatment effect), and is very data-driven, so inconsistent with respect to the underlying assumptions whether heterogeneity between these studies is plausible or not. For low number of studies (<20) it is better to use the HKSJ (Hartung-Knapp-Sidik-Jonkman) approach than the DerSimonian Laird approach for the test of the pooled effect, as DL is way too optimistic. However Review Manager is not able to do this. It was not clear how you dealt with the different laser powers. Results General: Tasci et al have much smaller SD than the other studies. Is this correct? Or did they report the SE? 1. Functional outcomes Table 2. Make clear which results are MD and which are RR. Further for IIEF the values have been reported in the wrong columns. 1.1 pooled meta-analysis is a pleonasm. Pooled analysis or meta-analysis Results are very difficult to read, and the information is also present in the Figures. It might be more informative to describe the trend of the results instead of all those details.. The IPSS at the 12 month follow-up was statistically significant but comparable… 1.2, 1.3 You state that you did a sensitive analysis because of high heterogeneity. I guess a sensitivity analysis is meant. But it is not clear how you did this sensitivity analysis: What type of studies did you remove/select… 1.5 I guess the procedures itself have no sexual dysfunction but that it is due to the procedures. 1.6 trails must be trials. You cannot state that meta-analysis was not available 2 Perioperative parameters 2.1 operation time is 6 minutes less, but it is not reported whether this is from 12 to 6 minutes or from 2 hours to 1 hour 54 minutes, or what the variation in operation times is. In order to judge the relevance you should also report the group means (in the original units, with SDs, for all variables in this section 2). Further, how do the 6 minutes relate to the MD of 15.24? Really difficult to interpret these results. 2.2 Pooled analysis showed that the decreased Hb …. was lower. Is the word “decreased” correct? What unit is used? And CI is incorrect. 2.3 For which subgroup was the subgroup analysis performed? 3. Complications

	3.1 incidence of TUR syndrome, capsilar perforations, etc: should these not be RRs instead of MDs? 3.2 As before, you could add more details in original units. Conclusion It is somewhat contradictory to state that we can “safely” conclude that PVP can be offered …, but that the findings of this study should be confirmed by more large-sample RCTs.

VERSION 1 – AUTHOR RESPONSE

Responses to 1st reviewer’s comments
Comment 1: Please state any competing interests or state ‘None declared’: None declared
Response: This has been rectified.

Responses to 2nd reviewer’s comments
Comment 1: Please state any competing interests or state ‘None declared’: None declared
Response: This has been rectified.
Comment 2: The labelling of all forest plots is missing. Every else is fine.
Response 2: All labels have since been provided for each and every table and forest plot.
Responses to 3rd reviewer’s comments
Comment 1: Abstract: Result section line 3: due to insufficient data meta-analysis of data for 36 and 60 months follow up was not done, so, comment on 36.60 months outcome not justified.
Response: This sentence (and the associated sentence mentioned) has been completely removed due to inaccuracies.

Comment 2: Introduction, page 5 line 6: PVP is done with 532nm green light laser, the word "predominantly" not needed.
Response: The word “predominantly” has been completely removed. The sentence now reads as follows: “This technique is generally performed with a 532-nm green laser generated using potassium-titanyl-phosphate (KTP) or lithium triborate crystals”

Comment 3: Results, section 1.1: line 2: IPSS difference at 12 months is not comparable.
Response: We have adjusted this as follows: “IPSS data at the 12 month follow-up stage reached a statistically significant difference with a MD = -0.10 ($p < 0.01$)”

Comment 4: Results, section 1.4: line 2: QOL difference at 6 months is not comparable.
Response: Again, we have adjusted as follows: “Quality of Life at the 6-month follow-up point reached a statistically significant difference (MD = -0.08),”

Comment 5: Results, section 1.5: line 2: IIEF at 24 months was not "slightly lower" but 'significantly lower'
Response: We agree with this although we also felt the need to elaborate on this point. As such, we have adjusted the sentence which now reads, “Pooled analysis suggests IIEF at the 24 month follow-up was lower in the PVP group compared to the TURP group with a MD = -0.68, which can be considered statistically significant but again borderline (95%CI= -1.20 to -0.15, $p = 0.01$)”.

Comment 6: Results, section 2.3: line 4: needs correction.

Response: The previous sentence was incorrect as was pointed out. This sentence now reads as follows: “However, again the level of heterogeneity across this sample was extreme (I2 = 98%) therefore sensitivity analysis (Fig.3d) was again performed although this had a negligible impact on the results (MD = -1.83 days, 95% CI -2.25 to -1.40, p <0.01).” (Please see the result section in detail)

Comment 7: Results, section 3.1: line 3: the difference is significant
Response: Indeed, this issue needed to be clarified and now reads “ According to this meta-analysis, PVP was found to have significantly lower incidence of transfusion with an RR=0.14 (p <0.01), and clot retention (RR=0.14, p <0.01). There was also a small, but significant difference in the occurrence of TUR syndrome (RR=0.19, p <0.01) and capsular perforations (RR=0.09, p <0.01).”

Comment 8: Discussion, paragraph 2 line 3: needs correction as QOL at 6 months and Qmax at 12 months are significant.
Response: This report has been entirely revamped. As such, the sentence of concern is no longer in that place but toward the end of that paragraph. It now read, “Qmax at 6 months and QoL at 12 months highlighted a statistically significant difference, although the differences was not substantial.”

Comment 9: Page 15 line 5: needs correction, PVP done with saline as irrigant
Response: This has been rectified.

Comment 10: 10 Forest plots can be merged into 5.
Response: All the forest plots have been merged into 2 figures. Please see Figure 2 and Figure 3 for further detail.

Comment 11: Standard of English could have been better.
Response: We have revised this article entirely. There should no longer be a need to standardize the English.

Responses to 4th reviewer’s comments
Comment 1: The manuscript would benefit greatly careful proof reading and correction of the English. In particular, the manuscript needs to be proof read for spelling errors that won't be picked up by a spellchecker - such as "mouth" instead of month, "trails" instead of trials, and "sensitive" instead of sensitivity.
Response: The article has been entirely reworked and these problems should no longer be evidence.

Comment 2: Table 1 (and discussion): A column to indicate whether each study is a superiority design (or non-inferiority/equivalence) would be useful. I expect most (or all) of these studies were superiority designs - i.e. the null hypotheses are that PVP and TURP are the same with respect to each outcome. Lack of an apparent difference between them (i.e. a p-value >0.05) means we do not have enough evidence to reject the null hypothesis that PVP and TURP have the same effect - but it does not necessarily mean that they are the same (i.e. equivalent or non-inferior). It is therefore misleading to conclude that PVP is non-inferior to TURP if only superiority studies have been performed. The correct interpretation is that there is currently not enough evidence to conclude that they are different.
Response: The studies which were selected according to our method were all superiority designs, as you suggested. We were unable to find compelling evidence for either PVP or TURP when analyzing this dataset, in some instances, however; there were a few instances where we identified a significant and substantial effect. We have tried to report this tentatively using the term borderline in reference to the upper/lower confidence interval sitting relatively close to the null hypothesis. In other instances, we have either described the lack of evidence along with our interpretation hoping to elaborate on the issue within this field but also to provide caution for absolute conclusions.

Comment 3: Table 2: Sample size for IIEF in PVP group is 1.10?
Response: We have corrected the mistake, sample size for IIEF in PVP group is 351. Detail please see the Table 2.

Comment 4: The tests for "overall effect" (difference in baseline outcome measure between intervention groups) here are invalid for randomised studies, as randomisation means that the null hypothesis of no difference is true - therefore there is no sense testing it. These tests should either be justified (if I have misunderstood their purpose) or removed.
Response: As far as we are aware randomization is embedded to reduce systematic bias, specifically selection bias which may intentionally or unintentionally influence data. Perhaps, we misunderstand but we do not think randomization in itself pertains to testing the null hypothesis. You can still (and should) test hypotheses under random allocation of participants, except for those instances where it may not be possible such as end of life care interventions etc. However, we have adjusted the phrase “overall effect” to be a little more tentative when describing and interpreting statistical outputs.

Comment 5: Tables in general: Reporting (for example) "P<0.00001" is rather extreme, P<0.001 would suffice. It is good to see that all p-values in the manuscript are reported as their actual values (rather than just p<0.05 vs p>0.05). Rounding of numbers (such as means and SDs) is inconsistent throughout the tables (some to 1 d.p., others to 2 d.p.)
Response: Both issues raised have been corrected. We no longer report extreme p values but maintain a maximum of 3 decimal places.

Comment 6: In the Forest plots, the weights of some of the studies for some outcomes at different time points vary wildly. The weights of the Tasci 2008 study for some of the outcomes at 3 and 12 months seem particularly odd as they are >90%, while at 6 and 24 months they are much lower (<40%). What is going on here? Even a weight of 40% is oddly high for one of the smaller studies. Tagu 2008 and Pereira-Correiaet 2012 also have extremely variable weights for different outcomes/time points (sometimes >90%) - why?
Response: Yes, we felt this was also an issue so we initially re-analyzed the data using the Stata software though this had very little impact on the data. We also tried every effort to solve this problem including emailing to the RevMan, but there was no definitive answer. Due to the fact that more weight is allocated to effect sizes with narrower confidence intervals we suspected the issue was with the trial itself rather than a technical glitch so we contacted the authors of this study but have had no reply. As a final resort, we removed this study from the analysis because we felt this is extreme and may have skewed our data. However, removing this study did not have a significant impact on the results, therefore we have provided the more complete forest plot.

Comment 7: The Forest plots are grouped into "80W", "120W", "180W" and "No mentioned" (should be "Not stated"), but no mention of this is made in the results. The figures would be simplified without the subgroups, so I would consider removing them unless there is a strong justification for showing them. If they are kept, I don't think the measures of heterogeneity for every subgroup add any useful information as the number of studies in each subgroup is so small. It would be better to only quantify heterogeneity over all of the studies.
Response: Actually upon reflection we found this to be really good advice. We had insufficient data in terms of available studies and therefore heterogeneity is not so useful. This also impacted on our reading of heterogeneity and therefore we recombined the data as advised and conducted sensitivity analysis in an effort to reduce some extreme heterogeneity while enabling us to generalize more concisely.

Comment 8: "reached a statistically significant difference" in the abstract should be changed to "at 12 month follow-up the difference was statistically significant, but was of no clinical significance" as "reached a statistically significant difference" implies the objective is to reach statistical significance.

Response: Yes, we have adjusted this not only in the abstract but later within the main article. It now reads, “There was a significant difference in Qmax at 6 months, and IPSS and QoL at 12 months although these differences were not clinically significant.”

Comment 9: Units need to be given when stating mean differences/95% CIs in the results. I squared is also missing the % symbol.
Response: This has been rectified.

Comment 10: "less complications rates" should be "lower complication rates" in the abstract/discussion.
Response: Again yes, this has been corrected and now reads, “PVP not only has an equivalent long-term efficacy in relation to IPSS, Qmax, QoL, PVR and IIEF, but is associated with fewer complications.”

Responses to 5th reviewer’s comments
Comment 1: The authors state it is an updated meta-analysis (strengths and limitations) but I wonder whether not an “up to date” meta-analysis was meant. Otherwise, they should refer to the previous metaanalysis.
Response: We have adjusted this within the main article and have adjusted the title to be more concise. We hope this will enable systematic researchers to source and select our article when conducting reviews of this kind but also for practitioners looking for the most up to date evidence.

Comment 10: Page 6, add reference to PRISMA
Response: We have added the reference for clarification as advised. Reference 18 now refers readers to PRISMA which we used as the format for our search and selection strategy.

Comment 10: Page 6 / Table 1: not clear to me how the 22 publications are connected to the 19 clinical trials. I guess this must be clear from Table 1, but for me it is not clear. It is for example not clear why a study like Ruszat et al has many rows in this table. Explain in the legend of Table 1 which characteristics you show (e.g. mean plusminus SD), median (range)? Further the study of Tasci has a much higher prostate size than the other studies.
Response: Yes we agree and have tried to clarify this for readers. The predetermined search and selection criteria yielded 22 publications [2, 7-11, 24-39], reporting 19 separate clinical studies. Three studies (i.e., Bachman et al., 2014[10], 2015[29] and Thomas et al. 2016[30]) refer to an identical study, and two studies (Kumar et al. 2013[24] and 2016[31]) were from the same trials in different period. We have clarified this at the beginning of results before points the reader to the tables which contain this information. (Please see the result section and table 1 in detail)

Comment 10: You state that you used the standardized mean difference, but where did you use it?
Response: We did not use standardized mean differences but rather simple mean differences. We have corrected this within the article and removed the phrase as you rightly pointed out.

Comment 10: I guess you did not count the means and standard deviations but calculated them.
Response: This has been rectified and all the data were extracted from the published literature.

Comment 10: I2 is not a test but a measure of heterogeneity. The chi-squared test is the test.
Response: Yes, we have also corrected this throughout the article.

Comment 10: You define in the paper when you used a random effects model and when a fixed effect model (if nonsignificant p-value and I2 <50%). However, the result is that for the same parameter at some timepoints a FE model and sometimes a RE model has been used. This affects the p-values

(FE model results more easily in a significant treatment effect), and is very data-driven, so inconsistent with respect to the underlying assumptions whether heterogeneity between these studies is plausible or not. For low number of studies (<20) it is better to use the HKSJ (Hartung-Knapp-Sidik-Jonkman) approach than the DerSimonian Laird approach for the test of the pooled effect, as DL is way too optimistic. However Review Manager is not able to do this.
Response: You are indeed correct, however; we designed our protocol to ensure this study was rigorously conducted. We also note that many researchers use this method and therefore we have not changed software for this study but will do so for future studies.

Comment 10: It was not clear how you dealt with the different laser powers.
Response: Actually, this became a critical issues because there are few studies and therefore we recombined studies without sub grouping around laser power. We realise this is not ideal but given the low number of studies available we feel this was best.

Comment 10: General: Tasci et al have much smaller SD than the other studies. Is this correct? Or did they report the SE?
Response: Yes, we also felt this was a little peculiar so we contacted the author for confirmation. As yet, the authors have not replied and so we are assuming the authors reported standard deviation for two reasons. Firstly, Standard error is generally provided for larger samples. The study in question had only 81 participants. Secondly, SEs are generally used when kurtosis is evident, otherwise SD is normally reported. As such, we assumed normal distribution and suggest SD was more likely to have been reported, in this instance.

Comment 10: Table 2. Make clear which results are MD and which are RR. Further for IIEF the values have been reported in the wrong columns.
Response: We have adjusted this accordingly.

Comment 10: 1.1 pooled meta-analysis is a pleonasm. Pooled analysis or meta-analysis
Response: We have corrected this throughout.

Comment 10: Results are very difficult to read, and the information is also present in the Figures. It might be more informative to describe the trend of the results instead of all those details.
Response: Thank you for your helpful suggestions. We have tried to report trends more generally rather than myopically focusing on minor, less meaningful outputs.

Comment 10: The IPSS at the 12 month follow-up was statistically significant but comparable…
Response: Adjusted according to a previous comment by the 3rd reviewer.

Comment 10: 1.2, 1.3 You state that you did a sensitive analysis because of high heterogeneity. I guess a sensitivity analysis is meant. But it is not clear how you did this sensitivity analysis: What type of studies did you remove/select…
Response: Thank you for your helpful suggestions. We have corrected the mistake and describe our method of sensitivity analysis more concisely to ensure our readers understand our processes. (Please see the method section in detail)

Comment 10: 1.5 I guess the procedures itself have no sexual dysfunction but that it is due to the procedures.
Response: Corrected.

Comment 10: 1.6 trails must be trials.
Response: Corrected.

Comment 10: You cannot state that meta-analysis was not available
Response: Thank you for your helpful suggestions. This comment has been complete removed.

Comment 10: 2.1 operation time is 6 minutes less, but it is not reported whether this is from 12 to 6 minutes or from 2 hours to 1 hour 54 minutes, or what the variation in operation times is. In order to judge the relevance you should also report the group means (in the original units, with SDs, for all variables in this section 2). Further, how do the 6 minutes relate to the MD of 15.24? Really difficult to interpret
these results.
Response: Yes, we have also included all measures for each output.

Comment 10: 2.2 Pooled analysis showed that the decreased Hb …. was lower. Is the word "decreased" correct? What unit is used? And CI is incorrect.
Response: Decreased was not the correct word to use in this instance. This has been corrected.

Comment 10: 2.3 For which subgroup was the subgroup analysis performed?
Response: Subgroup analysis was not conducted due to the lack of studies. We hope as more research becomes available we will be able to subgroup around laser power although this was not possible, this time. (The 4th reviewer also suggested to remove the subgroup analysis)

Comment 10: 3.1 incidence of TUR syndrome, capsilar perforations, etc: should these not be RRs instead of MDs?
Response: Corrected.

Comment 10: 3.2 As before, you could add more details in original units.
Response: Yes, we have added unit measures for each output to enhance clarity.

Comment 10: It is somewhat contradictory to state that we can "safely" conclude that PVP can be offered …, but that the findings of this study should be confirmed by more large-sample RCTs.
Response: We have adjusted this to be more tentative when providing conclusions/recommendations. We have also noted that more studies with larger samples are necessary. (Please see the discussion section in detail)

VERSION 2 – REVIEW

REVIEWER	Chris Jones Brighton and Sussex Medical School, UK
REVIEW RETURNED	03-May-2019

GENERAL COMMENTS	The authors have clearly put considerable effort into improving this manuscript and the presentation of results is now much clearer. The addition of sensitivity analyses is also useful. I have a couple of further comments: 1. "Pooled analysis does suggest IIEF at the 24 month follow-up was lower in the PVP group compared to the TURP group with a MD = -0.68, which can be statistically significant but again must be presented with caution due to the upper confidence interval being so close to the null (95%CI= -1.20 to -0.15,p = 0.01), see Fig. 2 e4"

	Would be better written as: "Pooled analysis does suggest IIEF at the 24 month follow-up was lower in the PVP group compared to the TURP group with a MD = -0.68, which is statistically significant but again must be interpreted with caution due to the upper confidence interval being so close to zero (95%CI= -1.20 to -0.15,p = 0.01), see Fig. 2 e4" 2. Operation time, blood loss, periods of hospitalisation and catherisation time analyses - these are all variables that are likely to be right skew. These sections need to either show that mean is appropriate for these variables, or deal with the skewness in some way. It may be that the high heterogeneity is a symptom of the variables distributions being inappropriately skewed.

REVIEWER	J IntHout Radboudumc Nijmegen the Netherlands
REVIEW RETURNED	16-Jun-2019

GENERAL COMMENTS	The paper is much better than the previous version. Well done. However, I have a few remarks left. 1. I think that your conclusions are a bit too strong, the differences between both treatments don't seem so relevant if I read the complete paper. 2. Further, I would not do fixed analysis if I2 <50% and random effects meta-analysis if I2 > 50%, but I would consider that the methods differ with regard to power, studies with regard to design etc. so I would recommend a random effects analysis for all analyses. But probably the results would not be so very different. 3. Further, the forest plots are very small and difficult to read, apparently the resolution is not high enough for this size. My remarks including some edits are in the attached paper. I would also recommend to have the English (even though it was much better than before) read by a critical reader. Success. The reviewer provided a marked copy with additional comments. Please contact the publisher for full details.

VERSION 2 – AUTHOR RESPONSE

Responses to reviewer's comments

4th reviewer

Comment 1: Please state any competing interests or state 'None declared': None declared

Response: Adjusted accordingly.

Comment 2: "Pooled analysis does suggest IIEF at the 24 month follow-up was lower in the PVP group compared to the TURP group with a MD = -0.68, which can be statistically significant but again must be presented with caution due to the upper confidence interval being so close to the null (95%CI= -1.20 to -0.15,p = 0.01), see Fig. 2 e4"

Would be better written as: "Pooled analysis does suggest IIEF at the 24 month follow-up was lower in the PVP group compared to the TURP group with a MD = -0.68, which is statistically significant but again must be interpreted with caution due to the upper confidence interval being so close to zero (95%CI= -1.20 to -0.15,p = 0.01), see Fig. 2 e4"

Response: Thank you for your comments. We have modified this section according to your advice.

Comment 3: Operation time, blood loss, periods of hospitalisation and catheterization time analyses - these are all variables that are likely to be right skew. These sections need to either show that mean is appropriate for these variables, or deal with the skewness in some way. It may be that the high heterogeneity is a symptom of the variables distributions being inappropriately skewed.

Response: Thank you for your comments. Indeed, this may be a symptom of variable distribution, however; under sensitivity analysis 77% of the heterogeneity identified for operation times could be attributed to lower quality studies (please see the Meta-analysis of perioperative parameters: Operative time, page 12).

Issues with blood loss, period of hospitalization and catheterization have been discussed in the latter stages of this report (please see the Discussion section, page 17). But actually, we felt it unnecessary to report heterogeneity for these particular outcomes given the potential for right-sided skewness, as you pointed out. If you would still like for us to add a measure of heterogeneity, we will certainly do that.

5th reviewer

Comment 1: Please state any competing interests or state 'None declared': None declared

Response: Adjusted as previously stated.

Comment 2: I think that your conclusions are a bit too strong, the differences between both treatments don't seem so relevant if I read the complete paper.

Response: Thank you for your comments. Yes, we have tried to be a little bit more tentative when providing conclusions. However, while the differences are only slight they are significant. We feel, as urologists, the findings are meaningful and therefore we have tried to provide balanced conclusions useful to clinical practice.

Comment 3: Further, I would not do fixed analysis if I2 <50% and random effects meta-analysis if I2 > 50%, but I would consider that the methods differ with regard to power, studies with regard to design etc. So, I would recommend a random effects analysis for all analyses. But probably the results would not be so very different.

Response: We did not select the fixed or random effect models based entirely on an I2 50% threshold. While this is common practice, we felt that the number of studies is still rather small therefore we went for a more descriptive analysis. That is to say, we were attempting to analyse common effect sizes rather than generalizing to the wider population, which would not have been possible given the relative simplicity of participant characteristics reporting. As such, we feel the fixed effect model is still more appropriate in this instance but the random effects model was implemented where high levels of heterogeneity was observed.

Comment 4: Further, the forest plots are very small and difficult to read, apparently the resolution is not high enough for this size. My remarks including some edits are in the attached paper. I would also recommend to have the English (even though it was much better than before) read by a critical reader.

Response: We have since increased the resolution of the forest plots for clarity. We have also meticulously reread the report, adding and adjusting critical commentary where necessary. We hope this will be met with your approval.

All additional feedback provided on the manuscript has been addressed and highlighted in yellow for your ease. Thanks once again, we feel these comments have really helped us in improving the report.

VERSION 3 – REVIEW

REVIEWER	Chris Jones Brighton and Sussex Medical School, UK
REVIEW RETURNED	03-Jul-2019

GENERAL COMMENTS	In general I am satisfied with the changes made to the manuscript, I only have a few very minor wording changes left: Table 1: Needs to say somewhere what numbers reported are (e.g. means and SDs). Table 2: Write out "MD" as "Mean Difference". Use of the word "juncture", e.g. in "At the 6 month juncture the MD = -0.17 " - I would just write "follow up" instead. "Juncture" is not wrong, but it's less clear and only used twice in the whole manuscript. In the change to "1.4. Qol at 3, 6, 12, and 24-month follow-up", change "...however; at six months there appears to be a statistically significant difference." to "...however; there was one statistically significant difference at six months". This is stated again a couple of sentences later, which could be removed/simplified.

	In "1.5. IIEF at 6, 12, and 24 month follow-up", "may be" has been added to "which may be statistically significant" - change "may be" to "is", otherwise the meaning of the sentence is ambiguous.

VERSION 3 – AUTHOR RESPONSE

Responses to reviewer's comments
4th reviewer： Chris Jones
Comment 1: Please state any competing interests or state 'None declared': None declared

Response: Adjusted accordingly.

Comment 2: Table 1: Needs to say somewhere what numbers reported are (e.g. means and SDs)
Response: Thank you for your comments. We have modified this section according to your kind advice. We have added the sentence "Continuous variables were expressed as (mean±standard deviation) ,mean (range)$ or median (interquartile range)% "(Detail please see Table 1)

Comment 3: Table 2: Write out "MD" as "Mean Difference".
Response: Thank you for your comments. We have modified this section according to your kind advice. (Detail please see Table 2)

Comment 4: Use of the word "juncture", e.g. in "At the 6 month juncture the MD = -0.17 " - I would just write "follow up" instead. "Juncture" is not wrong, but it's less clear and only used twice in the whole manuscript.
Response: Thank you for your comments. We have modified this section according to your kind advice.

Comment 5: In the change to "1.4. Qol at 3, 6, 12, and 24-month follow-up", change "...however; at six months there appears to be a statistically significant difference." to "...however; there was one statistically significant difference at six months". This is stated again a couple of sentences later, which could be removed/simplified.
Response: Thank you for your comments. We have modified this section according to your kind advice.

Comment 6: In "1.5. IIEF at 6, 12, and 24 month follow-up", "may be" has been added to "which may be statistically significant" - change "may be" to "is", otherwise the meaning of the sentence is ambiguous.
Response: Thank you for your comments. We have modified this section according to your kind advice.

All the feedback provided on the manuscript has been addressed and highlighted in red for your ease. Thanks once again, we feel these comments have really helped us in improving the report.