We appreciate the opportunity to revise and improve the manuscript. Reviewers' comments are listed in italics, and our response is listed immediately thereafter.

Reviewer Comments:

Reviewer #1:

In this systematic review and meta-analysis, Bekiari et al evaluated, as only outcome, % of time that sensor glucose level was within the near normoglycaemic range. They only considered randomized studies; they performed sensitivity analysis in order to reduce bias of the studies considered. The coverage of literature was accurate and up-to-date. The overall take-home message is clear and well substantiated.

Authors' response: We thank the reviewer for his/her comments about the rigor, quality and clinical relevance of our work.

However, there are two minor points:

1. *It is not true that this is the first meta-analysis, as one meta-analysis on the same issue was published in 2011 (J Diabetes Sci Technol. 2011 Nov 1;5(6):1352-62). The authors should comment on what is new in their meta-analysis as opposed to the preceding meta-analysis.*

Authors' response: Indeed, an early meta-analysis on closed loop systems was published in 2011. However, this was actually a pooled analysis of only 4 trials that were conducted by a specific group, and not a meta-analysis based on evidence acquired through a systematic review approach. Moreover, all 4 trials in this analysis were conducted in an inpatient setting, hence address a different research question compared to our systematic review. Similarly, an overview published in 2015 summarised existing data from RCTs until September 2014. More interestingly, a recent systematic review was conducted by Weisman et al, and was published while our paper was under peer review. Notably, compared to the systematic review of Weisman et al, we incorporated a larger pool of studies (34 instead of 24 studies used by Weisman et al), identified through a comprehensive search both of electronic databases and of grey literature (including conference abstracts which often present negative findings(1)), to ensure our conclusions are not prone to publication bias. In addition, we used appropriate methodology for handling trials that reported median instead of mean values, while retaining consistency and clinical relevance regarding intervention and outcome definitions. We have updated our introduction and discussion to briefly comment on these analyses and the differences from our manuscript.

2. *The authors should look for consistency in their Forest Plots, as "favours closed" appears in some instances on the left and in other instances on the right*

Authors' response: We are grateful to the reviewer for bringing up this issue. In our forest plots, the label "favours closed loop" appears in some instances on the left and in other instances on the right side of the vertical axis (line of null effect), depending on the type of outcome. More specifically, for outcomes for which an increase in absolute value is considered favourable (for example % of time within normoglycaemic range), "favours closed loop" is on the right side of the forest plot. On the other hand, for outcomes for

which a decrease in the absolute value is considered favourable (for example % of time within hypoglycemic range), “favours closed loop” is on the left side of the plot. Reversing the forest plots in the latter outcomes to achieve consistency of labeling, would result in positive values appearing on the left and negative values on the right side of the horizontal axis for these outcomes, which we believe would prove even more challenging for readers to interpret.
Alternatively, consistency in side of labeling at the horizontal axis of all forest plots could be achieved by reversing the reporting order of comparators on a case-by-case basis (closed loop versus control for outcomes for which an increase in absolute value is considered favourable, and control versus closed loop for outcomes for which a decrease in absolute value is considered favourable). Nevertheless, we opted for consistency in reporting order of comparators (i.e. always using closed loop versus control comparisons) rather than side of labelling.

Reviewer #2:
This analysis has combined data from trials comparing closed-loop insulin therapy with other types of insulin therapy in populations with type 1 diabetes. The authors have combined data on percentage of time spend in the therapeutic range, time spent out of range and mean blood glucose as well as other outcomes. In all cases the closed-loop system outperformed conventional insulin therapy. This work was well conducted and methodologically sound.

Authors' response: We thank the reviewer for their meticulous approach on reviewing our manuscript.

1. *Please clarify whether any adverse event data reported (in particular hypoglycaemic episodes) in the trials and how the numbers compared between the randomised groups. If there are sufficient data then a meta-analysis of this data should be included.*

Authors' response: We do agree with the reviewer that data on hypoglycaemic episodes would be of interest to the readers. However, we had decided a-priori not to include incidence of any hypoglycaemia as an outcome in our meta-analysis, assuming there would be considerable inconsistency across individual trials regarding definition of hypoglycaemia. In fact, this issue has been acknowledged both by ADA and EASD, who have recently published recommendations to improve consistency in defining and reporting hypoglycaemia in diabetes trials(2).

Indeed, across included studies, definitions for hypoglycaemia included symptomatic hypoglycaemic episodes, carbohydrate-treated events, average number of meter glucose values <70 mg/dL or different thresholds, hypoglycaemia requiring treatment, number of participants who experienced at least one hypoglycaemic event or other definitions. Moreover, all eligible studies also considered metrics based on continuous glucose monitoring, such % of time spent below a specific glucose value threshold, as more reliable and objective outcome measures, in accordance to a consensus report on optimal outcome measures for artificial pancreas clinical trials(3).

Interestingly, recently published international consensus on continuous glucose monitoring systems advocate standardization of definitions of terms and ways of reporting

hypoglycaemia. In particular, they support refining the classification of hypoglycaemia to include three main categories: severe events requiring assistance, clinically important events with values lower than 54 mg/dL (3.0 mmol/L), and events with values between 54 and 70 mg/dL (3.0 and 3.9 mmol/L), which may be regarded as warning signals for more serious events(4, 5).

On this ground, following reviewer's suggestion, we added incidence of severe hypoglycaemia (defined as hypoglycaemic episodes requiring assistance) as an outcome in our Methods (Outcomes section), and present relevant findings in the Results (Secondary outcomes section).

2. *Please include a more comprehensive explanation of low blood glucose index and interpretation of the implications of a lower LBGI.*

Authors' response: We agree with the reviewer that readers may not be familiar with LBGI as an outcome for measuring hypoglycaemia. Therefore, we have added a sentence in our methods section (Outcomes) reading: "We also used overnight low blood glucose index as an additional outcome for assessing hypoglycaemia. Low blood glucose index is a weighted average of the number of hypoglycaemic readings with progressively increasing weights as glucose levels decrease and is associated with risk for hypoglycaemia and prediction of severe hypoglycaemic episodes."(6)

3. *The numbers in Figure 1 are not quite consistent, please check.*

Authors' response: We thank the reviewer for bringing this oversight to our attention. Number of duplicate records is now correctly reported as 2355.

4. *There are 34 included studies reported in the text and Figure 1, but there are 37 studies listed in Table 1.*

Authors' response: We thank the reviewer for pointing this out. Three studies (Haidar 2015, Haidar 2016 and Haidar 2017 – references 30, 31 and 32) are reported twice in Table 1, because they compared both a single-hormone CL and a dual-hormone CL with control (three-way cross-over trials). Therefore, total number of studies is 34, but number of comparisons is 37.

We have now changed the labeling in Table 1 to clarify that it refers to comparisons rather than studies. This is also indicated in the flow diagram (figure 1) and in the relevant text in the manuscript (Characteristics of included studies) which now reads: "After juxtaposing different reports that referred to the same study, 32 publications describing 34 trials (792 participants with data for 37 comparisons) were used to inform our systematic review".

Moreover, we now present these three-way cross-over studies separately in the text, by adding the phrase: "Additionally, three studies evaluated both a single-hormone and a dual-hormone system against control treatment (three-way cross-over trials)."

5. *Some of the numbers in the text on page 5 do not add up to 34, for example 29 trials of single hormone and 8 trials of dual hormone exceeds the 34 included trials.*

Authors' response: As in the previous comment, we agree that this point requires clarification. Therefore, we have rephrased relevant text in the manuscript to read: "Twenty-five trials compared a single-hormone closed-loop system (mostly with unblinded SAP therapy), while six trials assessed dual-hormone closed-loop systems in comparison mainly to insulin pump therapy (consisting of CSII combined with a blinded CGM system). Additionally, three studies evaluated both a single-hormone and a dual-hormone system against control treatment (three-way cross-over trials). Of note, in four studies assessing SAP therapy, the control comprised a SAP combined with an LGS feature. ".

6. *Risk of bias: "Most studies were deemed at high risk for bias due to incomplete outcome data.." This is misleading and should be revised; should it read something like "Of those studies at high risk for bias due to incomplete outcome data, …." ?*

Authors' response: In accordance to reviewer's comment, we have removed the phrase "due to incomplete outcome data" from the sentence, while readers are referred to the appendix for a detailed presentation of risk of bias assessment across individual domains.

7. *Figure 2 subgroup labelling is "single hormone CL" and" dual hormone CL", whereas the text on page 6 reports findings from "Closed-loop overnight" and "throughout24 hours".*

Authors' response: We appreciate the reviewer for bringing this oversight of ours to our attention. Figure 2 has now been replaced with the correct forest plot, which includes a subgroup analysis based on time (overnight or 24-hour) rather than type of CL (single or dual).

8. *The protocol states that data for area under the curve of glucose<3.5 mmol/l will be reported but it has not been reported in the manuscript. Please explain the reasons for this.*

Authors' response: Indeed, we do not report area under the curve (AUC) of glucose<3.5 mmol/l in the manuscript, as it was not consistently reported in eligible studies. In addition, after completing data extraction for our systematic review, we felt that AUC as an outcome would not be easy to interpret and thus opted for more clinically relevant outcomes related to hypoglycaemia, such as % of time below 3.9 mmol/L and low blood glucose index. In addition, following the reviewer's suggestion for adding information on hypoglycaemic episodes (reviewer#2, comment 1), we have included incidence of severe hypoglycaemia as an outcome in our methods and results.

9. *The table of included studies would benefit from a column showing length of follow-up.*

Authors' response: We agree with the reviewer's suggestion and have added a column in Table 1 describing length of follow-up.

10. *Please report the mean time study participants were within therapeutic range at baseline if this data is available.*

Authors' response: Baseline values for mean time participants within therapeutic range could potentially be available only for patients already using a sensor outside the context of the trial. Nevertheless, such data are not available in any of the eligible studies.

11. *Discussion – comparison of the findings with the existing literature has not been included. Please add a section comparing this work with other research, in particular the previous systematic review published in 2014.*

Authors' response: We have identified an overview of randomized controlled trials on closed-loop systems, published in 2015 (Battelino 2015). We report this paper in the introduction, but it is not further discussed, as it solely a review summarizing RCTs up to 2014, rather than a rigorous systematic review or a meta-analysis.
Furthermore, we have also identified two previously published meta-analyses, which are mentioned in the introduction and commented in the discussion. The first (Kumareswaran 2011) is a 2011 pooled analysis of 4 RCTs of patients with type 1 diabetes in the inpatient setting, as opposed to our systematic review which is focused to the outpatient setting (see also answer to reviewer#1, comment 1). The second is a systematic review conducted by Weisman et al, which was published in 2017, while our paper was under peer review. Notably, compared to the systematic review of Weisman et al, we incorporated a larger pool of studies, identified through a comprehensive search both of electronic databases and of grey literature (including conference abstracts which often present negative findings(1)), to ensure our conclusions are not prone to publication bias. In addition, we used appropriate methodology for handling trials that reported median instead of mean values, while retaining consistency and clinical relevance regarding intervention and outcome definitions.
We have now updated our discussion (Strengths and limitations section), highlighting the main differences of our systematic review compared to the two aforementioned analyses.

Reviewer #3:

1. *The outcome is % of time that sensor glucose level was within the near normoglycaemic range. Secondary outcomes are also defined on sensor glucose levels. I can't see (Figure 1) any trials being excluded because the outcome was not available in the control arm. I lack expertise on different types of insulin therapy (well, I'm reviewing for a generalist journal, not a diabetes journal). I can guess that when the control arm is "Sensor augmented pump therapy" then the outcomes, that depend on sensors, are available in the control arm. What about when the control arm is "Insulin pump therapy" or "Low Glucose Suspend"? Do these therapy protocols also use continuous glucose sensors? Or was continuous monitoring added by the trialists, for the sake of comparability? If the latter, did you pre-specify that you would exclude trials that did not use sensors in the control arm - or was it just that such trials did not arise during review?*

Authors' response: We agree with the reviewer that, given that BMJ is a generalist journal, our description of intervention and comparator therapies warrants further clarification and explanation in the manuscript. Therefore, we have revised the first two paragraphs in the

Introduction to include descriptions and relevant references for different types of insulin pump therapy and related features, including continuous subcutaneous insulin infusion (insulin pump), continuous glucose monitoring, sensor augmented pump (SAP) therapy and low glucose suspend (LGS) pump therapy.

Furthermore, we did not pre-specify to exclude studies that did not use a continuous glucose monitoring (CGM) system. In particular, in our Methods we report that eligible comparators included multiple daily injections (MDIs), insulin pump therapy with or without CGM, or sensor-augmented insulin pump combined with or without an LGS feature. However, a blinded CGM system was added in all trials that assessed insulin pump as a control. We now report this in the Results (Characteristics of included studies): "Twenty-five trials compared a single-hormone closed-loop system (mostly with unblinded SAP therapy), while six trials assessed dual-hormone closed-loop systems in comparison mainly to insulin pump therapy (consisting of CSII combined with a blinded CGM system). Additionally, three studies evaluated both a single-hormone and a dual-hormone system against control treatment (three-way cross-over trials). Of note, in four studies assessing SAP therapy, the control comprised a SAP combined with an LGS feature. ".

2. *Related to this: did the trials ensure equivalence of outcome measurement in the intervention and control patients? Specifically, was the frequency of sensor monitoring the same in both the treatment and control arms, to prevent confounding between the outcomes and the intervention? I can't see any mention or exploration of this in the manuscript - apologies if I have overlooked it.*

Authors' response: We agree that information regarding equivalence of outcome measurement, which is related to exact type of CGM sensor used in the closed-loop and control arms, is important and thank the reviewer for highlighting this. On this account, we have added relevant text in the Results (Characteristics of included studies) which reads: "Seventeen closed-loop comparisons utilised the Dexcom G4® CGM sensor,[25 24 26 28-30 32 34 35 37 38 54 45 46] while an Enlite™ Sensor, a FreeStyle Navigator® or a Medtronic 4s sensor were used in the closed-loop systems in nine,[23 27 31 40 42-44 49] eight,[33 36 47 50-53] and one comparisons,[39] respectively. Type of CGM sensor wan not reported in two trials.[22 48] Of note, in 30 comparisons, type of CGM sensor was identical between closed-loop and control arms, one trial used a different sensor in the control arm,[39] and six trials did not report information for type of sensor used in the control arm.[22 25 26 38 41 48]"

3. *Given that this has been submitted to a generalist journal, rather than a diabetes journal, explanation of the different types of therapy would be welcome.*

Authors' response: We thank the reviewer for this useful suggestion. We have revised our introduction and updated relevant sections in the methods and results to incorporate better description of different types of insulin therapy (see reply to reviewer#3 comment 1).

4. *Please could you supply a reference for "counter-enhanced" funnel plots? Standard funnel plots are centered on the main effect; this one appears to be centred on zero. Visual inspection of the funnel plot is mentioned in the manuscript; are you aware that there is empirical evidence that visual inspection is too subjective to be useful? See Terrin, Schmid & Lau, 2005, J Clin Epi 58:894-901. Based on this it is more useful to include the p-value than "visual inspection". The results refer to this p-value as "significant publication bias", but it is more strictly correct to describe this as "evidence of small-study effects" than to assume that small-study effects are always due to publication bias.*

Authors' response: We added a reference for the counter-enhanced funnel plots(7). It is true that the counter-enhanced funnel plot is not centered at the model estimate, as is usually done when drawing funnel plots, but at zero indicating the value under the null hypothesis of no effect. Funnel plots enhanced by including contours that partition it into areas of statistical significance and non-significance, are more useful to differentiate asymmetry due to publication bias from that due to other factors. We agree that visual inspection of funnel plots is inherently subjective and should complement the use of statistical tests and the Egger's test was used as well. Moreover, we agree that it is more proper to describe this as "evidence of small-study effects" than "significant publication bias", therefore this has been corrected throughout the manuscript.

5. *The statistical heterogeneity is high, but that is not uncommon in evaluations of complex interventions. The authors have explored some candidate explanations for heterogeneity: single vs dual hormone therapy; overnight vs 24hr use of the system. Both of these are characteristics of the intervention. I don't wish to encourage the authors to explore too many hypotheses, on this number of studies, but is it worth also exploring whether characteristics of the comparator (SAP vs insulin pump) explain some of the heterogeneity between studies?*

Authors' response: We thank the reviewer for this suggestion. However, we do not believe that our conclusions would benefit from an additional subgroup/sensitivity analysis based on type of comparator (SAP vs insulin pump), since such an analysis would not substantially differ from our prespecified subgroup analysis based on type of intervention (single vs dual CL). The reason is that the vast majority of single CL trials used SAP as a comparator (26 of 27 studies), while, in a similar manner, almost all dual CL trials (7 out of 9) used insulin pump therapy as control. We specify this in our results (both in Characteristics of included studies, and Sensitivity and subgroup analysis sections).

Reviewer #4:
Thank you for the chance to review this comprehensive and important paper.

1. *The topic is absolutely relevant for BMJ's readers. Everyone working with type 1 diabetes is looking for the size of overall treatment improvements, with closed loop therapy systems. All because we are looking for anything, which can diminish the development of diabetes late complications in our patients. Secondly, closed-loop insulin therapy most probably will impact positively the patient burden in decision making on insulin dosing; however, this was not the topic for the current review.*

Authors' response: We thank the reviewer for acknowledging the clinical relevance of our systematic review. Although we did not explore the effect of artificial pancreas on patient burden, we do agree that closed-loop systems can favourably affect their quality of life, by reducing insulin-related burden. On this account, we have added a sentence and relevant references in the discussion that reads: "Moreover, the impact of artificial pancreas on quality of life and its effect on reducing patient burden should be further explored, considering that patients with type 1 diabetes and their carers have demonstrated a positive attitude towards closed-loop systems".

2. *This systematic review and metaanalysis is done according to guidelines and best principles for such kind of research – and I cannot find anything which should have been done methodological differently.*

Authors' response: We appreciate the reviewer's comment on the robustness of our methods.

3. *Regarding the primary and secondary outcomes; comparisons of time in target (and above/below target), I agree that the differences between closed loop and control treatment are the main outcomes and should be reported in text and figures. However, data on the actual time spent in target during closed loop and control treatment in the different studies (and above/ below targets) would be of additional values for the reader, given in result text only. I suggest roughly statements on how these percentages vary between studies for closed loop and control treatments.*

Authors' response: We agree with the reviewer that in the case of parallel studies, where the intervention and control are applied to two different groups of patients, it would be reasonable to present outcome values separately for intervention and control, in addition to the main effect estimate (difference between intervention and control). However, almost all eligible studies (except three) had a cross-over design, therefore both closed-loop and control therapy were evaluated on the same patients, allowing for comparison at the individual rather than group level.

The Cochrane collaboration (Cochrane Handbook – Chapter 16.4, as well as relevant expert guidance(8)), advocate that ideally, cross-over trials should analyse and present data as within-participant differences, rather than within-group differences, given that each participant serves as their own control. These means that properly reported cross-over trials should provide within-patient differences between intervention and control (derived

through a paired t-test) and not present data for each group separately. Reporting actual values for intervention and control separately would treat data as if they arose from a parallel trial, ignoring the cross-over effect (i.e. the fact that the same patients appear in both arms of the study and so they are not independent of each other), thus violating recommended guidance.

In fact, the majority of eligible studies reported results as recommended (within-patient difference between intervention and control). However, some trials did deviate from this guidance and improperly reported results for intervention and control separately, estimating a within-group difference. In that case, we used appropriate methodology to impute within-patient differences(8).

4. *I suggest adding the actual values for sensor glucose and HbA1c in text as well (and not only changes).*

Authors' response: Please see our reply at previous comment.

5. *For the discussion I appreciate its precise and short form pointing on better reporting and broader patient selection in further closed loop studies.*

Authors' response: We thank the reviewer for his/her favourable view on our implications for future research on closed-loop systems.

6. *In Table 2, I will suggest for the readers less familiar with parameters as time in target etc. to emphasize in legends that here is given differences between the two treatments compared (and not actual values). Also in all figure legends, I suggest to add, mean difference between closed loop treatment and control treatment.*

Authors' response: We have revised Table 2 legend to clarify that values refer to weighted mean differences between closed-loop and comparator. Moreover, we added number of studies and heterogeneity values for all values in Table 2.

All figure legends also explicitly report that presented effect estimates refer to weighted mean differences between closed-loop and comparator.

7. *If space is a problem for the journal, I will suggest Fig 8 to be moved to appendix rather than in the main paper and results given in main text only.*

Authors' response: Following the reviewer's suggestion, we have moved Figure 8 to the appendix.

Reviewer #5:

Thank you for the opportunity to review this interesting paper. The authors have clearly done a lot of work to identify, appraise and synthesise the existing evidence in this field. I have reviewed this from a statistical perspective, and have some comments for the authors to address going forward

Authors' response: We thank the reviewer for his comments about the rigor of our work and appreciate his meticulous and insightful statistical approach on our manuscript.

1. *I worry about the quality of the individual studies. Many of them have just 10-30 patients. I doubt randomisation could have balanced the groups in such a short sample. Did the individual studies have balance at baseline? Was adjustment for any imbalance undertaken? When taking results from published studies, it is hard to overcome issues of baseline imbalance without the IPD. Even then, adjustment for non-recorded baseline variables is not possible. Can the authors reassure the BMJ that there syntheses are meaningful? Was this issue accounted for in the risk of bias classification?*

Authors' response: We agree with the reviewer that imbalance between the two arms would be an important issue in case of small sample trials with a parallel design. Nevertheless, in cross-over trials, as is the case of all studies (except three) in our systematic review, to our knowledge there is no issue of imbalanced groups at baseline since each patient serves as their own control and there is no need for adjustment.

2. *I am also concerned about the primary outcome definition: % of time in a normal range. Is this meaningful clinically? Is 'normal' well defined. For example, for a value just outside the range (e.g. 10.01mmol/l), why should this be abnormal when a neighbouring (almost identical) value (e.g. 9.99mmol/l) that happens to fall inside the range is classed as normal. Of course, this may be how the primary studies recorded the outcome, but this classification has consequences for interpretation of the meta-analysis results. I find the outcome uncomfortable. Similarly, others like % of time > 10mmol/l, or below 3.9mmol/l.*

Authors' response: We agree with the reviewer that the choice of optimal outcome measures in artificial pancreas research is an important issue that has been acknowledged by relevant research groups and the wider artificial pancreas community(3). On this account, our choice of primary and secondary outcome measures was decided a-priori, based on guidance issued by the Juvenile Diabetes Research Foundation International (JDRF) Artificial Pancreas Project Consortium(3). Moreover, in addition to outcomes that are based on specific threshold values, we included clinical relevant outcomes, such as mean sensor glucose value, HbA1c and incidence of severe hypoglycaemia.
Of note, recently published international consensus on continuous glucose monitoring systems (CGMS) support standardization of definitions of terms and ways of outcome reporting. In particular, with regards to our primary outcome, current guidance advocates that measurement of time spent in glucose range 70–180 mg/dL [3.9–10 mmol/L] adds

valuable information to assess the level of current glycemic control in addition to what is known from HbA1c(4, 5). We now cite both articles in our discussion when referring to clinically important outcomes in CGMS trials.

3. *I am pleased that the authors use a random effects analysis. However, they do not say what estimation method was used.*

Authors' response: We now report in the manuscript that we applied an inverse-variance weighted random effects model using the DerSimonian and Laird estimation method.

4. *In relation to this, there is increasing evidence that the uncertainty in heterogeneity estimates should be accounted for in the derivation of CIs. See refs below [1, 2]. For example, the Hartung Knapp correction works generally quite well.*

Authors' response: Following reviewer's suggestion, we calculated CIs applying the Hartung Knapp correction. We now report this in the Methods (statistical analysis) and Results (Primary outcome) in the main text with the relevant reference(9), as well as in appendix 13 (for all outcomes).

5. *With large heterogeneity as observed here, 95% prediction intervals can be helpful to summarise the range of effects across settings better than the summary effect itself. i.e. the average effect is perhaps not so meaningful. The authors might consider an approximate prediction interval to address this. [3]*

Authors' response: We have now calculated 95% prediction intervals, as reported in the Methods (statistical analysis) and Results (Primary outcome) in the main text with the relevant reference(10), as well as in appendix 13 (for all outcomes). Furthermore, in the discussion (strengths and limitations section), we acknowledge that findings for most analyses should be interpreted witch caution, due to high heterogeneity. More specifically, this cautionary note was quantified by presenting 95% prediction intervals which were not statistically significant (i.e. they included value of 0) for most outcomes, except for overnight time in normoglycaemia, overnight time in hypoglycaemia and overnight low glucose blood index.

6. *Publication bias assessments should be better explained as assessments of small study effects [4]*

Authors' response: We thank the reviewer for his useful suggestion. We rephrased "publication bias" to "small study effects", in accordance to reviewer's suggestion.

7. *In the abstract, please state how many trials were at low risk of bias*

Authors' response: We have added a sentence in the abstract reporting that only 9 studies were at low risk of bias. Number of studies at low ROB is now also reported in the main text (Results, Sensitivity and subgroup analyses).

8. *Referring back to the outcome, I see a range of secondary outcomes. But was the trend in glucose levels not summarised? E.g. in some studies, were the repeated measures of glucose not modelled, and could these have not been synthesised (e.g. mean trend)?*

Authors' response: We agree that synthesising mean trend data for glucose levels could enhance the clinical relevance of our systematic review. However, this was not possible, because trials reported glucose values at an aggregate mean level, rather than applying any models on repeated glucose measurements.

9. *STATA should be Stata*

Authors' response: Name corrected as suggested.

10. *An I-squared > 50% does not necessarily indicate high heterogeneity. Please see Rucker [5]. It depends on the size of the studies. An actual estimate of tau-squared (between-study variance) is preferable.*

Authors' response: We have updated the relevant phrase in the Methods (Statistical analysis) to include Tau^2 as an additional estimate of heterogeneity. Tau^2 values are also reported in the Results and appendix 13.

11. *The authors say they explored risk of bias using a funnel plot and Egger's test. But why does this relate to risk of bias? All studies could have high risk of bias but the funnel plot may be symmetrical. Perhaps the 'risk of bias' is misleading language. Do you mean risk of bias of the original studies or risk of bias in the summary effect due to small study effects? I think the latter perhaps? Anyway, there is some confusion with the risk of bias tool.*

Authors' response: We thank the reviewer for bringing this issue to our attention. Indeed, interpretation of funnel plot and Egger's test were not used for risk of bias assessment in individual studies, but for assessment of small study effects, as accurately stated by the

reviewer. Thus, we have now rephrased "publication bias" to "small study effects" throughout the manuscript. Additionally, in accordance to the PRISMA statement, we report risk of bias in individual studies and risk of bias across studies (small study effect) separately in the Methods section.

12. *Most studies were high risk of bias because "they reported median instead of mean values or reported results that required extensive use of imputation methods to be used in meta-analyses". Why is median is worse than mean. Indeed, if the % time in normal level is skewed, the median may be preferred. Also, if the imputation methods are not reliable, then why does it say 'appropriate formulas to calculate mean and variance" in the methods. This warrants further explanation please.*

Authors' response: We do agree that when data are skewed, reporting medians and IQR is preferred, as opposed to reporting means and standard deviations. In fact, that was the case in most eligible studies for continuous outcomes related to hypoglycaemia (for example % of time below 3.9 mmol/l). For these trials, rather than assuming that medians equal means (which would be reasonable only if data had a normal distribution), we imputed means and standard deviations using published methodology.

However, the main reason that several trials were deemed at high risk of bias was due to the fact that they improperly handled and reported data as parallel group studies (i.e. presenting within-group differences), despite having a cross-over design (see our reply to reviewer 4 comment 3). These studies required imputation methods to derive mean within-patient differences from within-group values(8).

Even though imputation methods utilised (for both cases mentioned above) are considered reliable, we opted for a conservative approach when assessing study quality and regarded the need for any imputation as a potential source of bias in our findings.

13. *Contour enhanced funnel plot is not mentioned in the methods*

Authors' response: We now mention contour enhanced funnel plot in the Methods (under Assessment of risk of bias across studies) and provide a relevant reference.

14. *In the results, when giving a summary result I suggest saying 'summary' or 'average' effect explicitly. Also, in the brackets please give the number of studies next to each m-a result. Some description of the amount of heterogeneity is also warranted in the primary and secondary outcome results sections. This is where a prediction interval may be warranted, to summarise the range of effects across settings.*

Authors' response: We have added a phrase in the Results section reading: "All meta-analysis results are presented as summary effect estimates for closed-loop versus control".

We have also added data on heterogeneity (I^2 AND Tau^2 values) and number of studies and calculated prediction intervals for all meta-analyses results for the primary and secondary outcomes (Appendix 13).

15. *Please include the results for low risk of bias in the main paper, not just appendix.*

Authors' response: We have added forest plots for low risk of bias sensitivity analyses in the main text (figures 8 and 9), as suggested.

16. *I like the discussion about the limitations of the existing studies, and what needs to improve in terms of reporting, included populations, sample sizes, and follow-up length. However, the abstract never mentions these limitations or recommendations, and this should be addressed to give a better-rounded summary of the evidence found. I would also include the outcome definition as a major limitation and something for new trials to address.*

Authors' response: Following the reviewer's recommendation, we have added a relevant phrase in the Abstract (Conclusions) reading: "The main limitations of current research evidence on closed-loop systems are related to inconsistency in outcome reporting, small sample size and short follow-up duration of individual trials." With regards to the issue of outcome definition in closed-loop research, we do include a relevant phrase in the Implications section which reads: "It is important for research groups to report a minimum set of agreed outcome measures and respective metrics."

17. *A similar comment applies for the 'what this study adds'.*

Authors' response: We added a similar phrase to "what this study adds", which reads "The main limitations of current research evidence on closed-loop systems are related to inconsistency in outcome reporting, small sample size and short follow-up duration of individual trials"..

18. *Table 2 – are these summary meta-analysis results? If so, we have no idea of the number of studies, amount of heterogeneity, etc. Also, if the aim is to compare single and dual therapy, then a meta-regression should have been conducted and the difference in the groups formally estimated for each outcome.*

Authors' response: We thank the reviewer for his suggestions on improving table 2. We have now revised the contents of table 2 accordingly, by adding values for number of studies and heterogeneity. In addition, at the table legend we clarify that values are summary results of subgroup meta-analyses based on type of closed-loop utilised.

Furthermore, we refrained from drawing any conclusions regarding differences between single and dual hormone systems based on a meta-regression analysis, since they have been mostly compared against different controls (single-hormone versus SAP and dual-hormone versus insulin pump therapy. Therefore, we believe that a subgroup analysis based on type of closed-loop (single vs dual hormone), rather than a meta regression analysis, would be more clinically relevant and interpretable to the readers.

I hope these comments are helpful the authors to revise their article and improve their work further.

Authors' response: We thank the reviewer for providing his statistical expertise and helping us indeed improve our paper.

References

1. Scherer RW, Langenberg P, von Elm E. Full publication of results initially presented in abstracts. The Cochrane database of systematic reviews. 2007(2):MR000005.
2. International Hypoglycaemia Study G. Glucose Concentrations of Less Than 3.0 mmol/L (54 mg/dL) Should Be Reported in Clinical Trials: A Joint Position Statement of the American Diabetes Association and the European Association for the Study of Diabetes. Diabetes care. 2017;40(1):155-7.
3. Maahs DM, Buckingham BA, Castle JR, Cinar A, Damiano ER, Dassau E, et al. Outcome measures for artificial pancreas clinical trials: a consensus report. Diabetes Care. 2016;39(7):1175-9.
4. Danne T, Nimri R, Battelino T, et al. International consensus on use of continuous glucose monitoring. Diabetes Care 2017;40:1631–1640.
5. Agiostratidou G, Anhalt H, Ball D, et al. Standardizing clinically meaningful outcome measures beyond HbA1c for type 1 diabetes: a consensus report of the American Association of Clinical Endocrinologists, the American Association of Diabetes Educators, the American Diabetes Association, the Endocrine Society, JDRF International, The Leona M. and Harry B. Helmsley Charitable Trust, the Pediatric Endocrine Society, and the T1D Exchange. Diabetes Care 2017;40:1622–1630.
6. Kovatchev BP, Cox DJ, Gonder-Frederick LA, Young-Hyman D, Schlundt D, Clarke W. Assessment of risk for severe hypoglycemia among adults with IDDM: validation of the low blood glucose index. Diabetes care. 1998;21(11):1870-5.
7. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. J Clin Epidemiol. 2008;61(10):991-6.
8. Elbourne DR, Altman DG, Higgins JP, Curtin F, Worthington HV, Vail A. Meta-analyses involving cross-over trials: methodological issues. International journal of epidemiology. 2002;31(1):140-9.
9. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol. 2014;14:25.
10. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. BMJ. 2011;342:d549.