Article details: 2016-0058	
Title	Dipeptidyl peptidase-4 (DPP-4) inhibitors and the risk of heart failure: A systematic review and meta-analysis of randomized trials
Authors	Subodh Verma MD PhD, Ronald M. Goldenberg MD, Deepak L. Bhatt MD MPH, Michael E. Farkouh MD MSc, Adrian Quan MPhil, Hwee Teoh PhD, Kim A. Connelly MBBS PhD, Lawrence A. Leiter MD, Jan O. Friedrich MD DPhil
Reviewer 1	Dr. George Honos
Institution	CHUM-HD, Cardiology, Montreal
General comments (author response in bold)	Excellent and timely paper on the potential effect of DPP-4 inhibitors on HF and potential within class differences between agents with respect to this important endpoint. 1. Please add the p values to the abstract (p=0.03 and p=0.1 respectively) as cited in the results section of the paper. Response: The p values are included in the abstract as requested.
Reviewer 2	Dr. Lee Green
Institution	University of Alberta, Family Medicine
General comments (author response in bold)	2. The abstract background says nothing about why the study is needed. Response: The abstract background has been expanded to support the need of the study. 3. Conversely, the abstract methods is somewhat more detailed than needed. Response: The abstract methods have been shortened. 4. The introduction does establish the need for the present work. Response: Thank you for this comment. 5. The description of the literature search is minimal. Response: A more detailed description of the literature search is now included in the Methods section in the manuscript text in the revised version of the manuscript and a new Supplementary Table 1 includes the detailed search strategy results for each searched database. 6. The approach to modeling appears appropriate. Response: Thank you for this comment. 7. No structured assessment of study quality is reported, but Table 3 provides sufficient information about study quality. Response: Study quality assessment is now included in the Methods section in the manuscript text and the detailed study quality data continues to be provided in Table 3 as noted by the Reviewer. 8. The discrepancy in statistical significance (though not effect size) between the full set of studies and the secondary analysis limited to the 3 large RCTs is not surprising, given the quite small effect size. The authors draw the appropriate conclusion from this finding in the first paragraph of the interpretation. Response: Thank you for this comment. 9. Limitations are generally well addressed. Response: Thank you for this comment. 10. The Tables and figure are useful, not redundant with text, and clearly presented. Response: Thank you for this comment.
Reviewer 3	Lars G. Hemkens MD MPH, Senior Scientist
Institution	Basel Institute for Clinical Epidemiology and Biostatistics, University Hospital Basel, Switzerland
General comments (author response in bold)	In their article “Dipeptidyl peptidase-4 (DPP-4) inhibitors and the risk of heart failure: A systematic review and meta-analysis of randomized trials”, Verma et al. examine heart failure as a potential adverse effect of DPP-4 inhibitors as treatment for patients with type 2 diabetes compared to no treatment. The background is that there are three recent very large trials (SAVOR-TIMI 53, EXAMINE, TECOS) including 5400 – 16500 patients with type 2 diabetes which examined three different types of DPP-4 inhibitors. One of the trials found an increased risk for hospitalizations due to heart failure (RR 1.26), the other two found no such association (the RR of one was 1.18, and of the other 1.0; all data from Figure 2). There are numerous further studies that are smaller, don’t focus on cardiovascular outcomes, and, at least when seen in isolation, provide comparatively little information on this adverse effect. This evidence base is suboptimal for clinical decision making. The authors “felt it was important to conduct an updated comprehensive systematic review and meta-analysis to inform clinicians who are concerned about the potential increased HF signal by providing them with the totality of the available randomized controlled trial evidence in the field.” They conclude that “Despite pooled data from 77,358 patients, whether DPP-4 inhibitors increase HF overall, or exhibit within-class differences (which would make pooling between agents inappropriate), remains unresolved highlighting the importance of ongoing trials which will address the overall but not class difference question.” While this is a well written manuscript on a systematic review and meta-analysis, it would benefit from a clearer structure and more consistent reporting to address the main questions (which in my view need to be clearer stated).

	General comments: 1. This is the seventh meta-analysis on this topics (in particular PICO 1) since 2014 (according to the authors, Tab 4), and this would be the fourth published in 2016. All meta-analyses find risk increases for heart failure in a range of 11% to 19% (those including all the three large trials including TECOS have 11% to 13%), albeit the nominal statistical significance (alpha level of .05) was not reached in all of them (see below). Although this manuscript provides a lot of information, the incremental benefit is in my view unclear. The authors address some interesting issues, for example the potential effect modification by substance type, or potential implications of results from ongoing studies (in the online appendix). A clear and succinct description of what is known and what this study adds would be extremely helpful. For example, the manuscript contains no clearly stated clinical questions that are aimed to be answered (e.g. "In patients with type 2 diabetes, what are the effects of a treatment with DPP-4 inhibitors compared to no treatment (with DPP-4 inhibitors and no alternative treatment) on heart failure?") Overall, I am not sure if this review adds much novel information for clinicians or provides additional guidance to inform clinical decisions. Response: Although there have been other meta-analyses published in this field, ours includes more randomized controlled trials and with this larger number of included RCTs our pooled results are the only ones able to demonstrate statistically higher overall heart failure risk, unlike the other post-TECOS meta-analyses. Our meta-analysis is also the only one to statistically compare differences in HF outcomes between different agents via subgroup comparisons which the Reviewer indicates is useful. Furthermore, by summarizing and comparing the results of other meta-analyses on this topic in Table 4, it has allowed the Reviewer to quickly summarize the risk increases pre and post TECOS as he does in his first general comment which would otherwise not be so readily apparent. In response to the reviewer's suggestion in the revised version of the manuscript we have added clearly stated clinical questions at the end of the Introduction and a clear and succinct description of what this study adds to what is already known in the Discussion/conclusion sections. 2. The second part of the conclusion, the potential existence of a within-class difference, can be seen as effect-modification by type of DPP-4 inhibitor. Since head-to-head comparisons of DPP-4-inhibitors are excluded from the review by design, the review does not directly address the clinical question "In patients with type 2 diabetes, what are the effects of a treatment with one specific DPP-4 inhibitor compared to a treatment with another specific DPP-4 inhibitor on heart failure?" In my view, the exclusion of active comparators is not strength but a limitation of the review. Isn't in patients with diabetes, who need treatment to control blood glucose, the clinical question about the comparative effectiveness of treatments, i.e. the question which treatment should be selected? Please discuss this central issue for decision making. Nevertheless, I agree that a separate analysis excluding active comparators (DPP-4 inhibitors or not) makes sense to evaluate potential impact of the comparators, but the stated assumption that most other antiglycemic drugs affect heart failure risk needs more discussion supported by empirical evidence. Response: Unfortunately there are minimal RCTs comparing DPP-4 inhibitors to each other. In the revised version of the manuscript we specifically reference the 11 head-to-head trials comparing different DPP-4 inhibitors to each other identified by our systematic search, and highlight that all are small trials with short follow up, and that none report a single heart failure outcome event making it impossible to pool results. With respect to the exclusion of non-DPP-4 inhibitor active comparators, we state in the manuscript that we chose this approach because it allows us to measure the effect of DPP-4 inhibitor treatment to avoid the confounding effect of other medications, and are pleased that the reviewer agrees "that a separate analysis excluding active comparators (DPP-4 inhibitors or not) makes sense …." We agree that comparison to active controls is also an important question but would require a separate analysis and was beyond the scope of our current review. One of the reasons for making this decision is that RCT data comparing DPP-4 inhibitors to active controls are much more limited without any large completed cardiovascular safety trials meaning that it would be even more unlikely to reach definitive conclusions comparing DPP-4 inhibitors to other medications given that definitive conclusions were difficult to reach comparing DPP-4 inhibitors to placebo even with multiple large completed cardiovascular safety trials. This can be illustrated using the data in one of the previously published meta-analyses by Monami et al that did include active control trials for which we grouped RCTs comparing DPP-4 inhibitors to other classes of diabetic medications in the figure below. (For this figure we have grouped all DPP-4 inhibitors together. Comparing them individually to each of the other diabetic medications would make the number of RCTs in each comparison group even smaller.) Monami et al did not include the largest placebo controlled cardiovascular safety trial, TECOS, but even without TECOS, the RCTs comparing DPP-4 to placebo have >95% of the overall weighting in such an analysis and for most of the comparisons of DPP-4 inhibitors to other classes of medications there are only single RCT comparisons. Even for the comparison to glipizide with the largest number of RCTs, there are only a total of 6 RCTs and 25 patients with events. Although we did not formally update this search vs active medications, from our search we know that there have not been any large RCTs powered for cardiovascular outcomes comparing DPP-4 inhibitors to active controls published to date. This lack of data comparing DPP-4 inhibitors to other diabetic medications makes it difficult to draw meaningful conclusions for comparisons to active controls. In response to the Reviewer's comments, we have added a sentence to the limitations section addressing the issue of blood glucose control decision making in trials in which investigators targeted minimizing differences in A1C between groups. We have also added specific references supporting our statement that other antiglycemic drugs affect heart failure risk. 3. I wonder about the impact of existing heart failure on the outcome. The authors conclude that "it seems prudent to follow the FDA's recent Drug Safety Communication and be cautious about prescribing saxagliptin and

	alogliptin in patients with established HF, or at high risk of developing HF (previous HF, low eGFR and/or elevated NT-proBNP), and consider discontinuing these medications in any patient who develops heart failure". This indicates that the development of heart failure is important in this situation. The outcome used for the review (hospital admission for heart failure) indicates most likely a clinically more progressed stage of heart failure. In the online appendix (p.63) the authors state as limitation "It is important to point out that only a minority of patients in various studies had a history of HF, and patients with severe HF were excluded" but in the three large trials there are 13-28% of patients with heart failure (Tab 2). In EXAMINE, there were substantial differences in the effect estimates between patients with (HR 1.00; 0.71–1.42) and without history of heart failure (HR 1.76; 1.07–2.90) - although there was no nominally significant interaction (p=0.07), what might be simply due to lack of power in this single study. Are there subgroup analyses from the other large studies? The manuscript would substantially benefit from discussing and, if possible, analyzing the impact of stage of heart failure (or no heart function) at baseline as potential effect modifier; otherwise this would be a limitation to be more clearly discussed. Response: SAVOR TIMI-53 is the only other RCT that provides data for patients with and without a previous history of heart failure. In response to the Reviewer's suggestion we have added a post hoc analysis comparing the effect of DPP-4 inhibitors in patients with and without a previous history of heart failure (new Figure 4). Interestingly, although the risk of heart failure requiring hospitalization is higher in patients with previous heart failure as one would expect, the relative increase was considerably higher in patients without (RR 1.42, 95%CI 1.15-1.74) rather than with (RR 1.08, 95%CI 0.89-1.31) a prior history of HF, with interaction p=0.06 (Figure 4). However, these results need to be interpreted cautiously given the limited available data. 4. The authors discuss a lot nominal significance, especially of different approaches to assess the effect on heart failure ("achieved / achieved not statistical significance"; e.g. p.11 l.26 & l. 55; p.12, l. 23; p.13 l.12 &l. 21) and with regard to other meta-analyses. This is in my view very overemphasized. The effect estimates are more or less the same in most analyses and the CIs change slightly including or not including 1. For example, adding all the small to the three large trials changes the effect estimate from 1.13 to 1.14 and the lower CI limit changes from 0.97 to 1.01 (Abstract). One could argue that it makes not much difference which approach is used. Response: We agree that the minimal changes in the effect estimate adding the results of the small to the three large trials suggests that the results are consistent regardless of the approach used. As we highlight in the Discussion, we think that inclusion of the smaller trials not significantly changing the overall effect estimate is important because the smaller trials likely did not monitor patients for heart failure as comprehensively as the larger cardiovascular safety trials, so it is actually reassuring that their inclusion does not significantly change the results. 5. In my view it is very valuable that the authors have identified and carefully extracted this enormous body of evidence from small studies. I wonder if the reports of these studies all clearly report that patients were systematically monitored for heart failure, I would expect that the reporting in the primary studies that are not focusing on such events is not always optimal. Please clarify and discuss any problems/biaste potentially arising from reporting quality of the primary studies with regard to adverse effects. Response: We are pleased that the Reviewer feels that it is very valuable that we have identified and carefully extracted this enormous body of evidence from small studies. With regards to monitoring for heart failure, the non-cardiovascular smaller outcome trials generally did not specify whether heart failure was a pre-specified adjudicated outcome and it seems likely that these trials did not systematically monitor for heart failure as compared to the larger cardiovascular outcome trials. The revised version of the manuscript discusses this issue in the limitations section. Given this limitation, it is encouraging that inclusion of the smaller RCTs does not significantly change the overall effect to also address the Reviewers' previous comment number 4. 6. While it is important to collect this information from the small studies, they also represent in the majority only studies with zero events in one or both groups. While excluding such studies would move the effect estimates away from the null, including the studies using continuity correction is probably not a much better approach as recent simulations showed (Kuss O; Stat Med 2015; 34;1097-1116). This and the potential advantage of the sensitivity analysis should be carefully discussed. Response: Using Review Manager software we were limited by the methods used within that software which excludes studies with zero events in both groups and is currently the most widely used approach. We augmented this analysis by conducting a sensitivity analysis including such trials using similar continuity corrections as used by Review Manager to include studies with zero events in one group and showed that the pooled results were unchanged. In the supplemental methods section in the revised version of the manuscript we now state that the recently published simulation study by Kuss et al suggests other approaches for including such studies. 7. The reporting is incomplete. The provided PRISMA checklist seems to be not correct for some items. For example, item 5: I found no details on a protocol or registration at the stated positions in the text or elsewhere; was there a protocol? Item 9: how was the screening done – was it done in duplicate? Item 10: how was the data extraction done – was this done in duplicate? Item 12: how was the bias assessment done – was this done in duplicate? Please peruse the manuscript and clearly address all reporting items from PRISMA. Please clearly report all objectives of the study in the introduction, including PICO questions since the aim is to inform clinicians. Response: The revised version of the manuscript addresses all reporting items from PRISMA. In particular, it now states that screening and data extraction including bias assessment were carried out in duplicate and that a review protocol was not published. We also now specifically report the objectives of the study including PICO questions at the end of the Introduction, as suggested by the Reviewer.

	8. The authors provide much information as online appendix. Response: In the revised version of the manuscript, most of the Methods have been moved from the online appendix to the manuscript text and redundant and duplicate material has been removed in response to the Reviewer's subsequent comments below. 9. Analyses: Some analyses and information on the main analysis are only described and reported in the appendix. I found this very complicated and very difficult to read. The supplementary methods are in my view not so extensive that they could not be reported in the main text. I found the duplication of text (e.g. l.15 – 21 on p. 56; l.6-18 on p. 57) difficult and cumbersome to read. I would suggest condensing and restructuring the text. Response: These sections have been moved to the main text and the duplication of text has been eliminated in the revised version of the manuscript. 10. Results: The description of the supplementary results could in my view be more focused and structured. I was not sure about the exact question the authors aim to address here and the underlying assumptions. The analyses for the "simulation" of results of ongoing studies should be clearly described in the Methods. With regard to the ongoing studies, I am not sure how they were obtained, and if there was a systematic search for ongoing studies (the clinicaltrials.gov search seems to filter ongoing studies out, Fig. 1). Therefore, the validity of these ancillary analyses is not clear. Response: In the revised version of the manuscript, we indicate in the online methods and results that we also searched clinicaltrials.gov for studies that had not yet completed to determine whether there were ongoing RCTs with cardiovascular primary outcomes since these trials dominate the pooled results. As these would not contribute to RCTs included in the pooled analysis of the completed RCTs they are not included in Figure 1. The presented results were simply meant to illustrate that the results from these RCTs, assuming baseline characteristics and effect sizes similar to the already completed large cardiovascular safety RCTs, could easily result in significant changes to the pooled result. In the revised version of the manuscript, we have added additional detail on how these studies were identified in the online methods and results sections. We have retained these results in the online appendix but would be agreeable to removing them if requested to do so by the Editors. 11. Discussion: The supplementary discussion is also in some parts redundant and I am not sure how it contributes to the main interpretation of the study findings. The authors again discuss the three large trials, add selective information from further evidence (e.g. animal studies) and discuss potential pathomechanisms. I would suggest integrating the important information in the main text. In my view, all limitations of the main analysis should be clearly described in the main text. Response: In response to the reviewer we have eliminated the supplemental Discussion by removing redundant and non-essential components and moving all limitations to the main text. 12. Risk of bias assessment: some items of the Cochrane risk of bias tool are used, and the tool is cited, but not all items are used. Please explain the reasons. For example, blinding of outcome assessment and blinding of personnel is not assessed - but both might be a crucial issue since the decision to hospitalize for heart failure might be very subjective (and this is part of the discussion). Potential bias due to funding is also not addressed, please include or discuss. The potential source of bias due to "early trial stopping for efficacy before the planned enrollment was completed" is not clear and not part of the Cochrane tool. I agree that this might introduce bias; this could be categorized in the "other threads to validity" category of the Cochrane tool. Please provide a reference and explain why this is assessed. Response: The revised version of the manuscript now specifies in the Methods and a footnote to the Risk of Bias Table (changed from Study Quality Table as per the Reviewers' comment number 29 below) that patient, caregiver, and outcome assessor blinding was assessed. The revised version of the manuscript references the JAMA study supporting risk of bias for early stopping for efficacy (Montori et al Randomized trials stopped early for benefit. A systematic review. JAMA 2005; 294:2203-9) as a potential source of bias. We did not specifically assess for bias due to funding but now mention that the vast majority of studies were supported by pharmaceutical companies. 13. The search strategy in MEDLINE seems to use a rather specific search filter for RCTs and yielded only 492 hits. Please clearly state which filter was used. This could be a limitation. I would suggest to cross-check the findings with that of the other cited meta-analyses which used more sensitive search strategies to ensure the literature is complete. Please report the search strategy separately (e.g. in the webappendix), not included in the Flow Chart, and please clearly state the search dates, the search interface used, and the number of references after de-duplication. The "copy/paste" of the search result is not optimal. Response: In the revised version we have updated the search to August 2016 which increased the number of MEDLINE citations to 614. We have also added an EMBASE search for journals not indexed in MEDLINE. Most importantly, as discussed in our comparison to other meta-analyses, even prior to these changes, our search strategy identified more RCTs than any of the previous meta-analyses and did not miss any RCTs identified by other meta-analyses. In response to the reviewers' comments we have added the complete search strategies as a new Supplemental Table 1 in the revised version of the manuscript. As previously indicated in Figure 1 and now in the new Supplemental Table 1, a very general text word search of MEDLINE (and now also EMBASE) was performed (ie. using suffix .mp.) using the names of each DPP-4 inhibitor and this search was limited to "clinical trial, phase iii" or "clinical trial, phase iv" or "clinical trial" or "randomized controlled trial." The search date, including search periods within MEDLINE, and number of references at each stage were already shown in Figure 1 and have been updated (also now for EMBASE) in the Supplemental Table 1. No de-duplication was required within MEDLINE and the added EMBASE was for journals not indexed in MEDLINE. The search interface used

	(OVID) has been added. 14. I wonder about the data for EXAMINE that are used for the main meta-analysis: the events in Fig 2 are 106/2701 and 89/2679 (RR 1.18, 95%CI 0.9-1.56). The original paper reports these data in one table (cited Ref 9; Lancet. 2015;385:2067-76; Tab. 4) but in the abstract and Tab 3 provides different data, to my understanding the latter counting only the first occurrence of an event: "Hospital admission for heart failure was the first event in 85 (3·1%) patients taking alogliptin compared with 79 (2·9%) taking placebo (HR 1·07, 95% CI 0·79–1·46)". This information from the paper is also used in the two other recent meta-analyses (Fig. 3 "Risk of hospital admission for heart failure…" in Li et al. BMJ. 2016;352:i610) and (Fig. 2 "e/ Hospitalization for heart failure" in Abbas et al. Diabetes Obes Metab. 2016;18:295-9. Please discuss more the potential implications of using these different data (this is only very briefly mentioned on p.14, l.17) because the difference between both resulting effect estimates for EXAMINE might impact the meta-analyses. Response: The reviewer is correct that for EXAMINE, the original paper reports only first event heart failures in the abstract giving a lower hazard ratio of 1.07 which is that used in the Li et al and Abbas et al meta-analyses rather than all heart failures reported in the main paper which gives a higher hazard ratio of 1.18. In addition to mentioning this in the main text, we also highlight these issues in the footnote to Table 4 where we compare our results to the other meta-analyses. Li et al and Abbas et al used this lower number of events but did not provide an explanation or highlight the differences making one wonder whether they may have inadvertently missed the total event number. Since our focus was on heart failure we included all heart failure events for all studies rather than the arbitrarily chosen first event only rate reported in the abstract for EXAMINE paper, similar to the other meta-analyses. Minor issues: 15. Introduction: the description of the three large trials is rather long and anticipates some of the results. Please condense and avoid redundancy. Response: This has been condensed to avoid redundancy as requested by the Reviewer. 16. Introduction l.23: at this point the reader does not know what the primary outcomes are, please clarify. Response: The primary outcome has been added at the end of the Introduction. 17. Introduction, l.10: it would be informative to provide information also for more established treatments such as insulin, metformin etc (see p.14 l. 8 where this is mentioned as strength of the approach) Response: The Introduction provides information on empagliflozin which was recently shown to reduce heart failure in patients with Type 2 diabetes. To our knowledge, most other established treatments do not have RCT data demonstrating improved or worsening heart failure. Those that do are now specifically referenced in the Discussion as per the Reviewer's previous request in comment number 2. 18. Data analysis: l.23: please explain the rationale for the secondary analysis. It is not clear why the focus was on studies with cardiovascular outcomes as primary outcome. This approach would exclude studies with heart failure as a prespecified secondary outcome, but primarily looking at e.g. HbA1c. Please explain. Response: We initially chose studies with cardiovascular outcomes as the primary outcome for the secondary analysis because we thought these RCTs would include patients at higher cardiovascular risk but agree that our approach excluded studies not primarily examining cardiovascular outcomes that still had heart failure as a prespecified and adjudicated secondary outcome, and one trial (the VIVIDD trial) fell into this category. To address this, we also provided results for a modified secondary analysis which included this trial in the Results. 19. Results p.9, l.15: "did not specify… who developed HF". This is not entirely clear; does this mean these studies report cardiovascular events combined, but not heart failure separately? Please clarify. Response: In the revised version of the manuscript we clarified by changing this to "did not provide HF data." 20. Results: The description of the three large trials is long. Since much of the information is already in the table, I would suggest focusing on important differences (between the three large trails, and between the large and small trials). Response: The description of the three large trials has been shortened. 21. Discussion, p. 14 l.45: please provide time period corresponding to the NNH Response: the time period corresponding to the NNH has been added. 22. Discussion: please discuss the impact of strengths and limitations on the findings in more detail (see comment 11 above). Response: As suggested by the Reviewer's comment 11 above, all the limitations have been moved to the main text in a re-written limitations section which discusses the impact of strengths and limitations on the findings in detail. 23. Discussion, p.15: "highlighting the importance of ongoing trials which will address the overall but not class difference question". Are there any ongoing/planned head-to-head trials? This might be valuable to be reported. Response: In the revised version of the Discussion we have added references for the 11 head-to-head trials comparing different DPP-4 inhibitors to each other identified by our systematic search, and highlight that all completed trials are small with short follow up, and that none report a single heart failure outcome event. In the supplementary Results we provide the references for the ongoing cardiovascular-safety trials, none of which directly compare different DPP-4 inhibitors to each other.

	24. Table 1: Instead of providing the results in this table, the prevalence of heart failure would be important (last column, patients with heart failure). Response: Overall heart failure rates for all participants including both the intervention and control groups can be calculated from the heart failure rates in the separate groups. Unfortunately, essentially none of the smaller studies provided the number of patients with a history of heart failure, except the VIVIDD trial which we highlight enrolled only patients with preexisting heart failure. The three cardiovascular safety trials did provide the proportions of patients with a history of heart failure and these proportions are provided in Table 2. 25. Figure 1: please follow PRISMA flow chart more closely. Response: We feel that Figure 1 follows the PRISMA flow diagram guidelines: "Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram." However, we would be agreeable to making any specific changes recommended by the Reviewer or Editor where it could follow the PRISMA flow chart more closely (we assume that the subsequent comments 26 and 27 by the Reviewer, which we address below, indicate where he feels specific changes should be made). 26. Figure 1: "partial duplicate"? What does this mean – why is the other part not relevant – please clarify. Response: "Partial duplicate" means that the results of one RCT were published in more than one publication. The revised version of the figure clarifies that the publication with longest follow up was included in the analysis. 27. Figure 1: the last box is very large and includes a lot of details and in my view rather distracting information at this place. This might be condensed. I assume "surrogate physiological outcomes" are the primary outcomes of these studies? Please use consistent descriptions. Response: The last box provides the number of identified RCTs by agent which we feel is useful information since evaluation of heart failure effect by agent is an objective of our meta-analysis. However, we would be agreeable to removing this information and reducing the size of the last box if requested to do so by the Editors. We have changed to "primary cardiovascular outcomes" and "primary physiological outcomes" to make the descriptions more consistent as suggested by the Reviewer. As per the Reviewers' request in comment 13 above, we have moved the search strategies from the first box in the PRISMA flow chart to a separate new Supplementary Table 1. 28. Figure 2: the numerator for VIVIDD is slightly different in Li et al. BMJ. 2016;352:i610 (n=124 vs 125), please double-check extractions. Response: We abstracted the numerators and denominators for VIVIDD from Table 3 in the published e-poster (the most recent publication) for this RCT which listed heart failure rates requiring hospitalization as 13/128 in the vildagliptin group and 10/125 in the placebo group (Krum H, Lukashevich V, Bolli GB, Kozlovski P, Kothny W, Ponikowski P. No significant difference in risk of heart failure hospitalization with vildagliptin in diabetic patients with systolic chronic heart failure: VIVIDD study. Diabetes. 2014; 63:A265; available at https://ada.scientificposters.com/epsView.cfm?3p2FCsEYNWbDTo3F%2FeHb2yebkmUUb%2BzjwMzz1p8yNw0mLnTrCjE0CMF13rT0bNIz). It is unclear why Li et al used a different denominator for the placebo group. 29. General: "Study quality" should not be used when risk of bias is addressed (many high quality observational studies have a high risk of bias, some high quality trials carry a high risk of bias, e.g. due to impossible blinding) Response: "Study quality" has been changed to "risk of bias" where indicated throughout the revised version of the manuscript, as suggested by the Reviewer. 30. General: I would suggest to be precise with the language and differentiate between risk "of heart failure" (what might some reader understand as developing heart failure, or incidence of heart failure) and "risk of heart failure leading to hospitalization" (what might rather indicate an exacerbation of preexisting heart failure). Please see comment 3. Response: "Risk of heart failure leading to hospitalization" might indicate an exacerbation of preexisting heart failure as the Reviewer suggests but could also indicate new development of heart failure serious enough to require hospitalization at the time of diagnosis. In the post hoc analysis comparing patients with vs without a previous history of heart failure added at the request of the Reviewer in comment 3, we have been precise with the language and used only "heart failure requiring hospitalization" since this analysis included patients who already had a previous history of heart failure.
Reviewer 4	Dr. Edoardo Mannucci
Institution	University of Florence
General comments (author response in bold)	The paper reports a properly performed and well-written meta-analysis of RCTs on a very relevant clinical issue (i.e., is DPP4 inhibitor treatment associated with an increased risk of hospitalization for HF?). Response: Thank you for this comment. 1) The main concern with the paper is its originality. The authors appear to be concerned with this point as well, since they provide a table with available meta-analyses on DPP4 inhibitors and heart failure after the publication of SAVOR/EXAMINE. They correctly point out that their analysis collects a larger number of trials, including also results from VIVID, and that (thanks to this latter trial) the increase in HF risk reaches statistical significance - unlike what reported in other post-TECOS meta-analyses. The clinical relevance of a $p<0.05$ in this context remains obscure. I think that we all agree on the fact that there is a safety signal with respect to HF, with an increase of risk in the 1.15-1.20 range. What we would all like to know is whether the phenomenon is drug-specific (for saxagliptin and, probably, vildagliptin) or there is a class effect; we would also like to know if there are patient conditions associated with a potentially Greater increase in risk. Unfortunately, as correctly pointed out by the authors, available data are not sufficient to provide clear answers to those questions (and I am afraid that ongoing trials

	will not modify substantially this picture). In conclusion, the paper provides a methodologically sound and balanced synthesis of available evidence, without adding substantial knowledge. Response: We feel it is important to synthesize all the available evidence and, as the Reviewer indicates, our systematic review collects a larger number of trials and with this larger number of included RCTs our pooled results are able to demonstrate statistically higher overall heart failure risk, unlike other post-TECOS meta-analyses. We agree about not over interpreting a nominal $p<0.05$ and are pleased that the Reviewer feels we provide "a methodologically sound and balanced synthesis of the available evidence." We also agree that knowing whether there is a drug-specific or class effect is important and our meta-analysis is the only one to statistically compare differences in HF outcomes between different agents via subgroup comparisons. Unfortunately, given the available data it remains unclear whether there are significant within class differences given interaction p-values ranging from 0.07 to 0.13 comparing results between the two medications with the most extreme heart failure safety (sitagliptin) or harm (saxagliptin) results. Minor points: 2) One of the weaknesses of these analyses is that events in smaller trials were "assumed to require hospitalization"; this should be recognized. In addition, please specify in Methods if, in smaller trials, events of HF were counted only when reported as SAE, or all AEs were considered. Response: In response to the reviewer we have rephrased the primary and secondary analysis in the Methods to indicate that HF events were counted when reported as either AE or SAE (though HF was listed under SAE in all trials). In the limitations section we also address the issues that virtually all of the smaller RCTs did not specify and likely did not pre-specify or blindly adjudicate HF events. However, as we highlight, for the DPP-4 inhibitor analysis, including the results of the smaller RCTs does not significantly change the point estimate which only moves from RR 1.14 to 1.13, because the vast majority of the outcome events arise from the larger cardiovascular safety RCTs (at least 96% of the outcomes included in the current meta-analysis were specified to be adjudicated in a blinded manner and require hospitalization). 3) The protocol for review and analysis was apparently not previously published or registered. This should be specified. Response: The revised version of the manuscript indicates that a review protocol was not previously published or registered. 4) Page 15, 1st par.: it is true that previous meta-analyses also included active comparator-controlled trials, but some of them also provided separate analyses for placebo-controlled trials. Response: That is correct and in Table 4 we describe which previous meta-analyses also provided separate analyses for placebo-controlled trials but in these meta-analyses, these were secondary analyses. Our sentence on page 15, 1st paragraph states that ours was the only one to focus on placebo-controlled RCTs to avoid the confounding effect of other medications. 5) CVOTs, although placebo-controlled, are designed in such a way that concomitant drug therapies for diabetes can be modified by the investigators, with the goal of reaching and maintaining a fair glycemic control in all patients, thus minimizing between-group differences in A1c. As a consequence, patients in the plaebo groups receive a larger amount of non-DPP4 inhibitor drugs (insulin, SUs, etc.) than those in the active treatment group. It is possible that some of drugs interfere with the incidence of HF. This limitation of CVOTs should be recognized in the Discussion. Response: In response to the Reviewer's comment we have added the following sentence to the limitations paragraph: "Although we limited our analysis to placebo-controlled trials, in trials targeting A1C, placebo-treated patients would likely have received more non-DPP-4 inhibitor medications."