Responses to reviewers' comments on the manuscript submitted by Imamura et al., "Consumption of sugar-sweetened beverages, artificially sweetened beverages, and fruit juice and incidence of type 2 diabetes: a systematic review, meta-analysis, and estimation of population attributable fraction" (Manuscript ID BMJ.2014.023070-R1)

We appreciate all of the valuable comments from the reviewers of our work. We have revised our manuscript, according to the reviewers' comments, questions, and suggestions. We believe that the manuscript has been further improved. Here are the major revisions:

- Description of assessments of bias and overall quality of evidence, in response to Reviewer 1 and 3.
- The new supplementary figure (Figure S5) as to be found in our response to Reviewer #3-9.
- Along with Figure S5, we have also revised 95% confidence intervals (CI) for population attributable fraction (PAF). 95% CI have been derived from 2.5th and 97.5th percentiles of 1,000 iterations. In the original manuscript, 1,000 PAFs were based on two relative risk measures that were mutually independent. We have now used the same RR for the two risks in each repeat, as two risk measures are of the same individuals in each hypothetical condition; and substituted the updated 95% CI for the old ones in the current version of our manuscript (please see our response to #3-9).

Minor revisions, if not explained, are applied to the manuscript mostly for shorter manuscript and for minor modification along with revisions explained in this document. This document includes our responses to reviewers' comments:

Reviewer: 1

Comments:
Thank you again for giving me the opportunity to review this revised paper. I think that the authors have made substantial efforts to improve the paper and I agree with most of their responses; however, there are several points that I believe still need to be addressed.

1. GRADE system: I am pleased to see that the authors have adopted the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) system to rate the quality of evidence. However, I am not sure whether the authors have followed the principles of the GRADE system appropriately. According to the GRADE handbook (Schünemann H et al. GRADE handbook for grading quality of evidence and strength of recommendations. Updated October 2013. The GRADE Working Group, 2013. Available from www.guidelinedevelopment.org/handbook), the quality of evidence from observational studies is initially classified as low. Factors for downgrading the evidence include study limitations, indirectness of evidence, inconsistency of results (i.e., unexplained heterogeneity of results.), publication bias, and imprecision. Factors for upgrading the evidence include a dose-response relation, a large effect, or the existence of plausible confounders that would result in the underestimation of the true effects. Of note, the 5 factors for downgrading quality of evidence must be rated prior to the 3 factors for upgrading it and the decision to upgrade should only be made when serious limitations in any of the 5 factors reducing it are absent.
Please provide an explicit explanation for your evaluation. I also suggest that the authors specify the investigators who assessed the quality.

Response: We appreciate the suggestions. Accounting for the given suggestions, we have now revised the description and application of about the GRADE and used the website above as a citation.

For the evidence for sugar-sweetened beverages (SSB) and type 2 diabetes (T2D), we stated that the strength of evidence was moderate. The five factors that could limit evidence were considered as the following:

- **Publication bias**: Trim-and-fill analysis indicated possibility of publication bias, but its influence was identified to be too small to alter the conclusion (Table 1). For example, before and after

correction for publication bias, the estimate was 1.28 (95% confidence interval=1.12-1.46) and 1.27 (1.10 (1.10-1.46), respectively. Also, a contour plot for SSB (Figure 2) indicated that a missing study would not likely to alter the point estimate. Overall, there was little indication that publication bias would limit the confidence about the overall evidence.

- **Imprecision**: The main estimation for the SSB and T2D rejected the null hypothesis. The GRADE recommends confirming that the effect size approximates Optimal Information Size (OIS). We did not pre-specify OIS, which was indeed studied by assessing population attributable fraction. Given the PAF we reported, the effect size and its precision were sufficient not to downgrade the overall quality of evidence.
- **Indirectness of evidence**: We estimated a potential effect of SSB consumption on incidence of diabetes in a population at the risk, assuring temporality of the association. Except that we did not directly examine an intervention effect, indirectness of evidence was not likely to be a remarkable caveat. A population representativeness or directness to a target population would matter, because all prospective cohorts did not represent a contemporary general population. However, there was no plausible reason to suspect that this type of indirectness of evidence would lower overall quality of evidence noticeably.
- **Study limitations**. According to sensitivity meta-analyses concerning study limitations for the association of SSB with T2D (Tables S5 and S6), study limitations were not likely to limit quality of evidence substantially.
- **Heterogeneity**: We reported that crude I^2 as a measure of heterogeneity (Table 2). Of SSB, the crude value was >70%, indicating substantial heterogeneity. In meta-regression analysis, we identified no single factor that explained the heterogeneity. However, in multivariable meta-regression including demographics, body-mass index, and bias indicators together, I^2 went down to 23.4% (low level of heterogeneity); or values <50%, in analyses controlling for a few sets of covariates and after exclusion of bias-prone studies. We have now included this information in the manuscript (please see the end of this reply).
  In addition, we acknowledge a limitation of I^2. In general, meta-analysis including estimates with high precision tend to produce a high I^2 value together although estimates are not heterogeneous indeed. Given the overall information on heterogeneity, we considered that the limitation due to heterogeneity was not substantial.

Quality of evidence was then judged to be moderate, for which we confirmed a linear dose-response relationship of SSB and incident T2D. The rating has been fit to the GRADE indication: "We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different" (GRADE handbook).

We did not uprate evidence based on the magnitude of a potential effect or based on residual confounding. The magnitude of a point estimate was not high or not greater than 2.0 which the GRADE system could anticipate. We could not rule out residual confounding enough to uprate quality of evidence, because unmeasured confounding might exist and cause bias in either direction toward false-positive or negative finding.

Revisions in text are made for more clarity including specific information:

- In Method section (Page 7, Line 14), we have added information on multivariable meta-regression: "In exploratory analysis using multiple variables of study-specific factors, I^2 was re-assessed as a magnitude of unexplained heterogeneity."

In Results:

- Original: "None of the study-specific factors evaluated could explain heterogeneity of results for SSB and ASB (P>0.1) (Table S5)."
- Revised: "None of the study-specific factors evaluated could explain heterogeneity of results for SSB and ASB (P>0.1) (Table S5). Exploratory meta-regression produced I^2 of 23.4% for SSB and of 67.8% for ASB, adjusting for population demographics (age, sex, country, incidence rate), BMI, follow-up duration, and measures of study quality." (Page 15, Line 18-20) (the next paragraph has the information on fruit juice, for which I^2 was observed to be zero in multivariate meta-regression.)

Text on the GRADE rating has now been revised:

- Original: "We rated 'moderate' quality for SSB, because the main findings were likely to be robust against different sources of bias, despite observational design."

- Revised: "We rated 'moderate' quality for SSB. The main finding rejected the null hypothesis and was likely to have a low degree of heterogeneity unexplained and have a dose-response relationship and robustness against potential bias or limitations including publication bias." (Page 11, Line 18-19)

Potential sources of bias and overall evidence were summarised by a single observer for group discussion and discussed among four authors to reach consensus: We have now added (Page 6, Line 11-13), "Results of bias assessment and quality of overall evidence were first summarised by FI and discussed among authors (FI, LOC, YZ, NGF) for consensus."

Reviewer: 2

Comments:

Thank you for providing me with an opportunity to re-read this manuscript. As initially noted this is a rigorously conducted meta-analysis, on an interesting topic which is gaining interest from both researcher and the public alike.

The manuscript now demonstrates much more clarity; the methods have greater transparency and the results are well presented. The additional information regarding confounding variables and acknowledgement of limitations provides a well-balanced and comprehensive article. The accuracy and attention to detail has been greatly improved, making the manuscript a pleasure to review. I believe the research will be of interest to the BMJ readership.

We appreciate the Reviewer's input to review the revised version and give this positive comment.

Reviewer: 3

Comments:

Statistical Review BMJ.2014.023070.R1

Consumption of sugar-sweetened beverages, artificially sweetened beverages and fruit juice and incidence of type 2 diabetes

The authors have submitted a revised manuscript.

The authors have now included a quality assessment process for the included studies which is based around the Cochrane ACROBAT-NSRI tool for assessing risk of bias in non-randomised intervention studies. I am not completely clear from the manuscript how they have undertaken this process or that the results of it have been appropriately presented.

Response: After we submitted the first manuscript, the BMJ's statistician team kindly recommended to us use of the ACROBAT-NSRI for assessment of bias in observational studies: "Our statistician team have advised they would encourage more documentation of confounders and how these were dealt with. They have advised to include these as part of the quality assessment (Cochrane has a new tool for assessing quality of non-randomised intervention studies which could help)."

To our knowledge, the ACROBAT-NRSI is the only one guide for bias assessment of non-randomised studies of interventions (NRSI), including prospective observational studies in which "allocation occurs during the course of usual treatment decisions or peoples' choices" (Sterne et al., A Cochrane Risk Of Bias Assessment Tool: for Non-Randomized Studies of Interventions (ACROBAT-NRSI), Version 1.0.0). Description has been revised and detailed in our response at #3-2.

Response:

3-1. Page 6 – they do not describe whether the ACROBAT-NSRI assessment was undertaken by a single observer, checked by a second or done independently in duplicate.

Response: Potential sources of bias and overall evidence were summarised by a single observer for group discussion, and discussed among four authors to reach consensus. We have added the following information on Page 6, Line 11-13: "Results of bias assessment and quality of overall evidence were first summarised by FI and discussed among authors (FI, LOC, YZ, NGF) for consensus."

3-2. Page 6 -it also state here that sensitivity analysis was undertaken for each of the seven quality domains in the ACROBAT-NSRI tool, but no results of these are mentioned in the text or presented in any table.

Response: We apologise for our lack of clarity. We have now revised as the following:

- Original: "We tested influence of these sources of bias in sensitivity analyses. Bias related to exposure and outcome measures were incorporated quantitatively to meta-analysis (see below and Supplementary Text)."
- Revised: "Sources of bias were evaluated by using meta-regression for each as a potential source of heterogeneity, meta-analysis excluding studies with a certain type of bias, or meta-analysis incorporating quantitative measures of bias (see below and Supplementary Text)." (Page 6, Line 7-8)

Similar information has now been included in the Supplementary Materials: "Influence of potential sources of bias was examined in sensitivity analyses by testing heterogeneity due to presence or absence of bias, by performing meta-analysis after excluding studies with a certain type of bias (Table S5, Table S6), or by incorporating quantitative bias in meta-analysis (see below)."

We would agree that bias assessment involved subjectivity. We have noted it as a limitation, "Finally, assessment of bias and evidence quality involved subjectivity, though we objectively examined influence of potential bias in sensitivity analysis." (Page 13, Line 23-25)

Sources of bias have been evaluated in sensitivity analysis, along with several other sensitivity analyses. Results from analysis addressing concerns of bias have been presented in multiple tables including a table footnote; not in a single table. Briefly, the information for each bias domain is summarised here:

- Confounding: assessed in details in Table S3. Footnote includes additional information (eg results from secondary analysis in a specific study confirming little impact of potential confounders not adjusted for).
- Measurement errors (quality of dietary assessment): Analyses excluding studies with potential bias in exposure assessment are presented in Table S6. A new row of Table S6 has now been added to show results from analyses excluding studies that did not conduct internal validation study. We did not find substantial influence of the potential bias.
- Misclassification of exposure (exposure information during follow-up): Use of repeated measures was assessed. Influence of using repeated dietary measures (yes/no) was assessed in sensitivity analysis and is presented in Table S5.
- Diagnosis of type 2 diabetes: Use of objective records or of self-reported incidence of diabetes was assessed. Influence of using objective records to ascertain diabetes was assessed in sensitivity analysis and presented in Table S5.
- Selective report: Table S5 footnote has now included information that observed associations did not vary by selective reporting.
- Selection: We identified no study that was likely to have substantial selection bias and produce invalid estimates. Of note, selection bias has been distinguished from missing data on outcomes, as guided by the ACROBAT-NRSI.
- Missing data: Studies with missing information on outcome ascertainment were considered as having 'high' risk of overall bias. Analyses excluding 'high' risk of bias are presented in Table S6.
- Overall bias: Please see the next response.

Findings for fruit juice were identified to be sensitive to use of repeated measures of dietary assessment ($p=0.068$) and use of objective records for diabetes ascertainment ($p=0.008$) (Table S6). The other potential sources of bias were not identified to influence the main results significantly ($p>0.1$). Footnote of Table S4 has now included "Influence of sources of bias were examined in sensitivity analyses (Table S5 and S6)".

3-3. Page 9 – a key aspect of the ACROBAT-NSRI tool is documenting which confounders are balanced/matched/adjusted for in each study. There is no list of the confounders reported – the best

description is given in the middle of page 9 as "socio-demographic variables, clinical factors (family history of diabetes or prevalent diseases) and lifestyle factors including a diet". However, the adjustments used in Table 2 are noted in the footer as being only for "demographic and lifestyle covariates" which isn't the same. The legend for Figure 1 doesn't mention adjustment for these factors at all, which I presume is an oversight. I would have expected, particularly given the extension supplementary material provided, to have a tabulation of the actual adjustments made study-by-study. Table S4 gives the results of unadjusted and adjusted analyses and does not fully state what was adjusted for.

Response: We apologise for the lack of information that was taken out to shorten the material in response to a suggestion from the reviewer #1. We have now incorporated the tabulated information about covariates adjusted for in each study in the Table S4. Some study-specific information has also now been written in the footnote of Table S4.

Table 2 footnote has now stated, "Multivariable-adjusted estimates were based on meta-analysis of estimates adjusted for demographic and lifestyle covariates (see Table S4 for details)". Figure 1 legend also has now included the information, "Covariates adjusted for in each study are summarised in Table S4".

3-4. It is also not clear what criteria were used to rate the risk of confounding. In Table S2 it is noted that only one study was rated as being at high risk of bias due to confounding. Table S4 lists four studies of omitting adjustment for "diet and clinical factors" so it is not clear why these are also not flagged as being at risk of bias from confounding. Other studies may omit other key variables, but given that no list of variables is presented we cannot tell.

Response: We have considered risk of bias due to confounding to be high if available information indicated substantial bias and biological plausibility.

As the reviewer noticed, four studies (ARIC, EPIC-InterAct, MESA, and SCHS) were identified as having lack of adjustment for a set of confounders (dietary or clinical variables). We did not identify possibility of bias strong enough to downgrade these four studies based on this reason, as explained next.

ARIC, EPIC-InterAct, and MESA did not adjust for dietary factors in the main analysis from which effect estimates contributed to our meta-analysis. EPIC-InterAct and MESA confirmed that further adjustment for dietary variables little altered their estimates in secondary analysis. The same findings were reported from NHS I, NHS II, and HPFS. ARIC adjusted for intakes of alcohol, fibre, and total calorie, and we considered that further adjustment for dietary factors was unlikely to alter their results given by other studies testing influence of dietary confounding. We have kept this information in the footnote of Table S4. "Dietary factors were not adjusted in main analyses in EPIC-InterAct and MESA. ARIC did dietary adjustment for intakes of alcohol, total calorie, and fibre only. EPIC-InterAct and MESA confirmed little influence of potential dietary confounders in secondary analyses. The lack of substantial influence was also confirmed in NHS I, NHS II, and HPFS."

SCHS did not statistically control for clinical factors including medication use or family history of diseases. SCHS excluded adults with prevalent cardiovascular diseases, and consumption of SSB was associated with BMI to a very small extent (eg difference by 0.1 kg/m^2 between the low and high consumption groups of SSB). Within the study, after adjustment for other covariates, we considered that additional adjustment for clinical variables would give little impact on their main estimates. This consideration was supported by the observations in NHS I, NHS II, HPFS, and EPIC-InterAct, in which adjustment for clinical variables did not alter the main estimates remarkably.

To summarise these observations, we have kept the statement, "None of these factors was identified as a single cause of confounding, according to studies assessing influence of potential confounding in different regression models.[11,41,48–50,53,58,60,62–64]", in the main manuscript (Page 9, Line 22-24).

One study rated as having 'high' risk of confounding was E3N, which we pointed out in Table S2 as "Adjustment for adiposity was likely to be biased.[43–46]" The substantial confounding and bias in their main results are, we judged, likely to be plausible, as pointed out in the cited papers [43-46].

3-5. There is no mention whether the analyses of the drink types are mutually adjusted for each other (for example, is the SSB analysis adjusted for ASB and fruit juice?) It is hard to think that consumption of each drink is independent.

Response: We appreciate the query. We have now included this information in the Table S4. Ten out of 17 studies did the mutual adjustment. This information has now been evaluated as a potential source of heterogeneity of associations and added to the manuscript (Page 7, Line 11), "Publication status (peer-reviewed or not), selective reporting (yes or no), and mutual adjustment for three beverage types were also evaluated *post hoc*".

The mutual adjustment appeared not to explain heterogeneity for the main associations examined (P>0.5), as now written in the footnote of Table S5, "Heterogeneity was not significant (*P*>0.1) for the other factors for any types of beverages: duration of follow-up, use of FFQ or other methods (Table S3), selective reporting (yes or no, Table S2), publication status (peer-reviewed or not), and mutual adjustment for three different beverages (yes or no, Table S4)."

In addition, three cohorts (NHS I, NHS II, and HPFS) confirmed that results before and after mutual adjustment for different beverages little differed. This information has now been included in Table S4: "NHS I, NHS II, and HPFS confirmed that mutual adjustment did not affect results."

3-6. Table S2 and Supplementary material on page 17. Overall quality assessment has to make a leap from the ratings of the individual domains to obtaining an overall assessment of likelihood of bias. It is not clear what rule the authors used to achieve this. The text on page 17 does not describe a consistent system for doing this. For example, studies 25 and 48 are stated as being at high risk of bias because their classification of diet was wrong, but the other seven studies marked as having high risk of bias on the dietary measures domain are not classified as being at high risk of bias overall. The same problem appears across multiple domains in the tool.

Response: We apologise for the lack of information. We have primarily considered whether or not potential bias of seven domains influenced main results or lost credibility of results substantially and plausibly, by which study results might not approximate a true effect. Little influence of a specific type of bias was supported by analysing heterogeneity across studies, as we responded at #3-2 above. Therefore, we considered that overall risk of bias could be 'unknown', when some bias, even if present, was not likely to influence the main result substantially and plausibly.

We previously described the key study-specific information influencing overall quality of bias in the main text (Page 9). We have kept it as the following:
"We rated six cohorts having potential bias in quantitative results based on at least one of the following reasons: publication of a conference abstract only[62]; exclusion of participants lost during follow-up[57,63]; likelihood of substantial residual confounding[40]; and no separation between fruit juice and SSB (fruit drinks) or between SSB and ASB[57,64]. Selective reporting might exist in some studies[39,52,55,58,62,64], but was unlikely to cause bias, for example, reporting only non-quantitative results for SSB in a study mainly on ASB.[39] Other potential sources of bias were detected, but we considered them not substantially influential on overall bias in each study, partly based on sensitivity analyses."

For example, 'high' risk of bias was rated for several cohorts that did not assess quality of dietary measures within a study (the footnote of Table S2, Page 18 of the Supplementary Materials). Despite likelihood of presence of bias, we have considered that this potential bias would not necessarily influence point estimates of interest substantially. Thus, we have not assigned 'high' risk of bias to those cohorts (BWHS, MESA, EPIC-InterAct, Occup. Cohort, KIHDS, FMCHES). By contrast, ARIC and SCHS were considered as having 'high' overall risk of bias by accounting for potential bias in dietary variables: SSB and ASB were aggregated in ARIC, and fruit juice and SSB were aggregated in SCHS. Influence on their point estimates was biologically plausible, because two types of beverages are likely to have different biological effects. This biological plausibility was considered important in our bias assessment.

We have revised descriptions of bias assessment and clarified what considerations influenced overall risk of bias, on Page 17-18 of the Supplementary Materials. The description of the overall bias has now been written as the following:

"Overall bias: We acknowledged ACRBAT-NRSI's recommendation that an observational study is not likely to have 'low' risk of bias.[89] Then, we assigned 'high' or 'unknown' risk of bias to each study.[88] We considered whether or not multiple sources of bias would impact estimated effects and uncertainty. We did not assign 'high' risk of bias even when studies were likely to have domain-specific bias, if the sources of bias were not likely to impact study estimates or uncertainty substantially and plausibly. Thus, 'unknown' overall bias was assigned to several studies, although they might have bias in some domains[26,31,41,48,49,52–54], because there was no strong plausibility that bias caused substantial impact on effect estimates to be used in this meta-analysis. Studies rated to have 'high' risk of bias had <20% of weights in the main meta-analyses, and exclusion of these studies did not change results (see sensitivity analysis, below).

Here, we describe considerations of seven domains: confounding, selection, measurement errors, misclassification of exposure, missing data, outcome assessment, and selective reporting. Influence of a single source of bias on overall risk of bias is also described. A single source of bias was not necessarily considered influential on overall bias as described above, with regards to biological plausibility and sensitivity analyses."

This has been followed by descriptions of each of the seven bias domains.

3-7. Page 13 – the authors indicate that publication bias created a false positive effect for ASB. However, the degree of publication bias seems rather small (not really visible at all in the funnel plot) and adjustment for it did not substantially change the magnitude of the effect.

Response: We agree that possibility of publication bias was not definite. Nevertheless, we have thought that publication bias was practically important based on the observation that trim-and-fill analysis shifted point estimate of 1.29 to 1.22 and that the latter estimate failed to reject the null (95% CI=0.98-1.52). The shift from 1.29 to 1.22 means 22% change in the effect estimate. This magnitude is, we judged, substantial, or at least not negligible: Of note, 5% or 10% change in an effect estimate is generally considered substantial, when we consider bias (eg confounding).

Besides, we have also considered long-term public interest and debate over the health effect of artificial sweeteners (De la Peña C. Artificial sweetener as a historical window to culturally situated health. *Ann NY Acad Sci*. 2010;1190:159–65; I Hellsten et al., Implicit media frames: Automated analysis of public debate on artificial sweeteners, *Public Understanding Sci*, 2009;19(5):590-608). Thus, we have thought that publication bias for ASB is plausible to cause a false-positive inference. We have discussed this in the Discussion (Page 14 Line 26-27).

Nonetheless, we agree that we should clarify further that 1) we detected indication of publication bias, but not evidence of a false-positive finding; and 2) overall, a false-positive finding for ASB would be plausible, but not determined. Accordingly, we have now revised the Discussion section as the following:

- Original: "Our analyses indicated a false-positive association of ASB with T2D because of possible publication bias. The bias would be expected by existing public interest over their health effects.[6,76],,
- Revised: "Our analysis indicated possibility of publication bias of the associations of ASB with T2D. The bias toward a false-positive finding would be plausible according to public interest over the health effects.[6,76] The finding at least underscores potential low quality of evidence and the need for cautious interpretation." (Page 14, the last two lines)

We have confirmed that, throughout the manuscript, we have not implied any definitive evidence of a false-positive finding. We have stated, "the findings were likely to involve bias" in the Abstract, for example. Our conclusion for ASB (and fruit juice) has been kept: "Nonetheless, both ASB and fruit juice were unlikely to be healthy alternatives to SSB for the prevention of T2D." (Abstract); "although ASB and fruit juice showed a positive association with incident T2D, potential bias and heterogeneity by study design limit quality of evidence." (Conclusion on Page 16)

3-8. Page 11 - The authors make GRADE assessment and place two outcomes (ASB and fruit juice) as being of low quality and one (SSB) as being of moderate quality. There is no strong argument why SSB is argued to be of moderate quality. It is hard to see why one outcome would differ from the others given that they are all reported in the same studies which were done using the same methods and adjusted for

the same confounders. Given that the estimates for both SSB and ASB shift considerably between the analyses adjusting for measurement error, confounders and publication bias, it is hard to attribute moderate or high credibility to any of them. The estimate for fruit juice seems to be close to a null effect in all analyses – the authors seem distracted by it moving either side of the null effect value, but it seems consistently close to it in all analyses.

We appreciate the comment on the GRADE assessment. Please see our response to the first reviewer about considerations for overall quality of evidence for SSB. Considering available information, we could be confident that SSB consumption was positively associated with incidence of type 2 diabetes, before and after adjustment for adiposity; and before and after adjustment for variability of measurement errors, for which publication bias was indicated, however the influence was likely to be little (Table 2).

The Reviewer mentions in the current comment that overall quality cannot be moderate when very different estimates were given by different models. We would like to state that estimates before and after adjustment for adiposity represent different biological effects of SSB: the effect confounded but partly mediated by adiposity; and other effects independent of adiposity. Therefore, both could be valid as different measures of effects. This is the same for correction for measurement errors. Different estimates did or did not assume within-individual variations in dietary assessments. In practice, there should be variability in dietary exposure, while effects under assumption of no errors are of biological interest. Therefore, different estimates from different models do not necessarily mean a magnitude of bias or lack of stable estimates. In the GRADE assessment, we have kept this view.

We have not incorporated these aspects in the manuscript for the sake of simplicity. Nevertheless, we would like to note here that different measures could be valid as they are; and overall evidence could be drawn for different contexts.

The comment here indicates that the same sets of studies would give the same quality of evidence overall. As this comment, we would like to note that overall quality of evidence could vary even if exactly the same methods were used in the same series of studies. We are confident, based on available evidence, that three types of beverages have different biological characteristics and consumption pattern in the real-life context; and have likelihood of producing different quality of evidence. We have written the example below (a paragraph before the comment of #3-9).

We think it is plausible and possible that quality of evidence could vary by different types of exposure, even if the same sets of cohorts did exactly the same prospective analysis. We have kept different levels of quality of evidence for the three types of beverages.

Nonetheless, we would agree that quality assessment involved subjectivity, as the GRADE group anticipates involvement of subjectivity. As we responded above (#3-2), we have noted the limitation of subjectivity, "Finally, assessments of bias and evidence quality involved subjectivity, which we attempted to avoid by objectively comparing estimates with and without potential bias." (Page 13, Line 22-23)

Here is the example: ASB is more prone to epidemiological bias than SSB and fruit juice, because a measure of validity was lower for ASB than for SSB or fruit juice (eg Salvini et al., Food-based validation of a dietary questionnaire: the effects of week-to-week variation in food consumption. Int J Epidemiol. 1989;18(4):858–67); and because ASB-disease association is likely to be confounded by health consciousness (Malik et al., Sweeteners and Risk of Obesity and Type 2 Diabetes: The Role of Sugar-Sweetened Beverages. *Curr Diabetes Rep*, 2012;12(2):195-203). The differences in biological and behavioural characteristics between types of beverages are likely to vary quality of epidemiological evidence.

3-9. The population attributable fraction computations are based on assumptions of causality, and on reducing consumption of the three types of drink to zero. Given that the estimates of effect vary considerably between the sensitivity analyses, they perhaps should investigate how much the estimates varies.

Thank you for this suggestion. We agree on the needs to understand how much PAF could vary under different assumptions. We have now performed iterative sensitivity analyses by varying relative risk

(RR) and calculating PAF, by which a reader can interpret PAF based on RR of interest. In the additional analyses, we accounted for I^2 in the process, by adding this variability to standard error (SE) of RR. I^2 was 23.4% in the multivariable meta-regression, but we acknowledge potential uncertainty of I^2 (Ioannidis and Nikolaos, *BMJ*, 2007;335:914) and also uncertainty of applying cohort-based RR to general population (Greenland, *Int J Epidemiol*, 2004;33:389-397). For the purpose of sensitivity analysis, we have conducted analyses using I^2=25%, 50%, and 75% separately. Confidence interval was derived from 2.5th and 97.5th percentiles of 1000 iterations (Greenland, *Int J Epidemiol*, 2004;33:389-397). The following figure has now been generated and inserted in the Supplementary Material as Figure S5 (Page 16); and brief description, in the text of the Supplementary Materials.

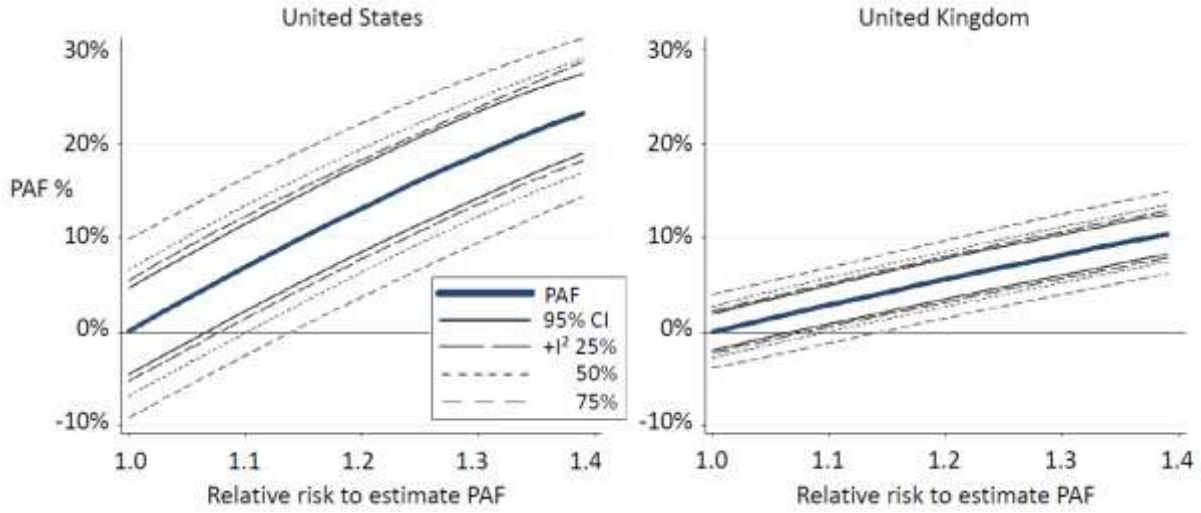

Figure S5. Population attributable fraction (PAF) for different degrees of associations between consumption of sugar sweetened beverages (SSB) and incidence of type 2 diabetes in the United States and the United Kingdom: Sensitivity analyses. Thick solid line is the best estimate of PAF; thin solid line, 95% confidence interval (CI) of the point estimate with a fixed standard error adjusted for adiposity; and dashed lines further incorporated measures of heterogeneity of potential effect across populations (I^2=25%, 50%, 75%). I^2 was 23.4% after controlling for measured characteristics of populations and identified studies (age, sex, absolute incidence, body-mass index, location of studies, methods for diabetes ascertainment, measures of validity and reproducibility of dietary assessment).

The incorporation of uncertainty indicates a concern in the precision. We highlight it in the limitations and in the conclusion. We have now stated these further relevant points in the manuscript:

- "PAF had limitation in precision related to a sizable number of T2D cases." (Page 13 Line 28)
- "Future work should seek to improve precision of evidence and to characterise efficacy and effectiveness of policy interventions for different populations." (Page 14 Line 2-4)
- "Although causality has not been established and precision needs to be improved, this study informs potential efficacy of reducing SSB consumption in a contemporary population." (Page 16 Line 5-6)

The authors do state the assumptions behind this, and indicate that causality is a concern, but are keen to promote intense public health interventions based on this evidence. Would not trials of SSB reduction be justified now rather than public health interventions? Also I wonder whether the illustration would be more helpful if it were based on the sort of magnitude of reduction in SSB that was achievable by a public health intervention and not an unachievable reduction to zero.

We thank for this opportunity to clarify this point. We would like to note that we are not keen to promote intense public health interventions to reduce SSB. We have noted that we estimated efficacy and that effectiveness is to be estimated in future work. This means that planning an intervention requires additional research on effectiveness. To clarify this point, we have now noted: "future research should also include randomised trials examining people's health and behaviours and informing effectiveness." (Page 14 Line 7-8); "In future, our work on efficacy should be extended to work on effectiveness to identify needs for interventions. In addition to observational evidence, trial evidence should be available, accounting for effects on cardiometabolic health and lifestyle change associated with a possible intervention." (Page 15 Line 18-21)

In the first paragraph of the Conclusion, we have also revised:

- Original: "We estimated that two millions of T2D events in US and 80 thousands of T2D cases in UK over 10 years would be related to SSB consumption."
- Revised: "We provided efficacy estimates that two million T2D events in the US and 80 thousands of T2D cases in the UK over 10 years would be related to SSB consumption."

We have confirmed that the abstract includes no indication to promote a policy intervention. Neither does the first paragraph of the Discussion section. The last paragraph of the Discussion has been revised as the following to clarify that we produced efficacy estimates:

- Original: "Although causality has not been established, the available evidence justifies an intervention to reduce SSB consumption in a population level."
- Revised: "Although causality has not been established and precision needs to be improved, this study informs potential efficacy of reducing SSB consumption in a contemporary population." (Page 16 Line 5-6)

We thank the reviewer for the query that we should also discuss trials to implement. We have now made the following revisions:

- Original: "For future implementation of a policy-based intervention to reduce SSB consumption,[12,13] our estimate of efficacy should be extended to estimates of effectiveness of interventions of reducing SSB, accounting for practical issues in interventions and effects on obesity, T2D risk, and lifestyle change associated with reduction of SSB consumption.[8,77] Despite PAF of no more than 15%, estimates of efficacy and effectiveness are crucial, as 535 million adults are estimated to have T2D in 2035."
- Revised: "In future, our work on efficacy should be extended to work on effectiveness to identify needs for interventions. In addition to observational evidence, trial evidence should be available, accounting for effects on cardiometabolic health and lifestyle change associated with a possible intervention.[8,77] Despite PAF estimates of no more than 20% in the current work, effectiveness should be evaluated for different populations, as 592 million adults are estimated to have T2D in 2035 globally." (Page 15 Line 18-22) (of note, 15% was revised to 20%, accounting for the 95% CI of PAF. Additionally, '535 million' has now been corrected to '592 million' as presented in International Diabetes Federation Diabetes Atlas, 6th edition.)

We also kept the following statement in the end of the paragraph for limitations to highlight needs for trials: "To address limitations typical of observational research and also needs to conceive a policy intervention in different populations, future research should also include randomised trials examining people's health and behaviours and informing effectiveness."(Page 14, Line 6-8)

We understand that zero consumption of SSB would not be achievable. However, we believe our efficacy estimate is informative for health professionals to recognise the minimal risk to target. We would also like to argue that efficacy can tell important information: even if we achieved zero consumption of SSB, 80% of future cases would still occur, and we have to conceive multiple modifiable risk factors. To express this, we have noted: "the average PAF no more than 20% confirms importance of modifying multiple lifestyle risk factors, rather than a single dietary component, for the primary prevention of T2D." (Page 15, Line 21-24)

Reviewer: 4

Comments:

The authors have answered all queries appropriately. The manuscript has been improved. No more concerns on the manuscript. Congratulations to the authors.

We appreciate this positive comment.