Vibration of effect in more than 16 000 pooled analyses of individual participant data from 12 randomised controlled trials comparing canagliflozin and placebo for type 2 diabetes mellitus: multiverse analysis

Henri Gouraud 1,2, Joshua D Wallach 3, Rémy Boussageon 4,5, Joseph S Ross 4,5, Florian Naudet 1,2,6

ABSTRACT

OBJECTIVE To evaluate the impact of conducting all possible pooled analyses across different combinations of randomised controlled trials and endpoints.

DESIGN Multiverse analysis, consisting of numerous pooled analyses of individual participant data.

SETTING Individual patient data from 12 randomised controlled trials comparing canagliflozin treatment with placebo, shared on the Yale University Open Data Access project (https://yoda.yale.edu/) platform, up to 16 April 2021.

PARTICIPANTS 15,094 people with type 2 diabetes mellitus.

MAIN OUTCOME MEASURES Pooled analyses estimated changes in serum glycated haemoglobin (HbA1c), major adverse cardiovascular events, and serious adverse events at weeks 12, 18, 26, and 52. The distribution of effect estimates was calculated for all possible combinations, and the direction and magnitude of the first and 99th centiles of effect estimates were compared.

RESULTS Across 16,332 distinct pooled analyses comparing canagliflozin with placebo for changes in HbA1c, standardised effect estimates were in favour of canagliflozin treatment at both the first centile (−0.75%) and 99th centile (−0.48%); 15,994 (97.93%) analyses showed significant results (P<0.05) in favour of canagliflozin. For major adverse cardiovascular events, estimated hazard ratios were 0.20 at the first centile and 0.90 at the 99th centile; 2,705 of 8,144 analyses (33.21%) were significant, all of which were in favour of canagliflozin treatment. For serious adverse events, estimated hazard ratios were 0.59 at the first centile and 1.14 at the 99th centile; 5,793 of 16,332 (35.47%) analyses were significant, with 5,754 in favour of canagliflozin and 39 in favour of placebo.

CONCLUSION Results from pooled analyses can be subject to vibration of effects and should be critically appraised, especially regarding the risk for selection and availability bias in individual participant data retrieved.

Introduction

Methods for synthesising evidence, such as systematic reviews and meta-analyses (SRMAs) and pooled analyses (ie, non-exhaustive quantitative syntheses that pool individual participant data (IPD) from several independent randomised controlled trials exploring similar research questions) are often considered to form the highest level of evidence to inform clinical practice and guidelines. However, in recent years, concerns have been raised about the numbers of overlapping and sometimes conflicting SRMAs/pooled analyses. In certain fields, the exponential increase in the numbers of SRMAs/pooled analyses has resulted in nearly one new SRMA/pooled analysis for every new randomised controlled trial. While SRMAs aim to comprehensively identify all studies in a specific area, specific concerns have been raised about pooled analyses, which—despite allowing for larger sample sizes, greater power, and more analytical options—have been described as a strategy used by pharmaceutical firms to hide negative study results by pooling their findings with those from positive studies.

Pooled analyses often explore subgroup effects or examine new questions across different combinations of all available studies. For instance, for the antidepressant duloxetine, at least 43 published pooled analyses have been conducted, compared with only 30 published randomised controlled trials...
for the treatment of major depression. Yet an a posteriori selection of trials to be pooled can result in p hacking (ie, conducting a large number of statistical tests until some significant findings arise). Different analytical decisions, such as the choice and timing of outcomes, can lead to vibration of effect (ie, the extent to which an effect estimate can change across multiple distinct analyses). In particular, evidence suggest that vibration of effect resulting from different methodological and analytical choices can lead to divergent and antagonistic conclusions in meta-analyses. Similar reproducibility issues could arise in pooled analyses, and this question could benefit from the increasing interest in shared IPD on data sharing platforms.

To gain a fuller understanding of the clinical implications of the different combinations of trials and repeated endpoint measures when conducting pooled analyses of IPD from randomised controlled trials, we explored the vibration of effect in pooled analyses in the field of type 2 diabetes mellitus. To date, the clinical value of drugs that reduce chronic hyperglycaemia, as measured by serum glycated haemoglobin (HbA1c), remains uncertain because of less clear effects on clinical outcomes, such as cardiovascular events. Canagliflozin is a drug used for glycaemic control among patients with type 2 diabetes mellitus and it has been consistently found to reduce HbA1c, a surrogate measure of diabetes control, and for which there is evidence for cardiovascular event reduction among patients at high cardiovascular risk. IPD from the canagliflozin company’s clinical development programme is shared on the Yale University Open Data Access (YODA) project (https://yoda.yale.edu/) platform. Using these shared IPD, our objective was to perform all possible pooled comparisons of canagliflozin with placebo in order to explore how three different outcomes change across the universe of all possible combinations of trials and endpoints. In particular, our pooled analyses estimated change in HbA1c as a surrogate measure of efficacy, in major adverse cardiovascular events as a clinical outcome measure of efficacy, and in serious adverse events as a measure of safety, at four time periods (weeks 12, 18, 26, and 52). Our hypotheses were that vibration of effect would not be observed for HbA1c, but would be observed for both major adverse cardiovascular events and serious adverse events.

**Methods**

**Protocol, registration, and reporting**

Before starting the study or receiving the data, our protocol was registered on the Open Science Framework (https://osf.io/ac9w7/; 4 August 2020). The results are presented according to the PRISMA (preferred reporting items for systematic reviews and meta-analyses) checklist. This study used data obtained from the YODA project, which has an agreement with Janssen Research and Development (YODA project 2020-4417).

**Eligibility criteria**

We included randomised controlled trials including people (aged ≥18 years) with a diagnosis of type 2 diabetes mellitus without distinction in terms of baseline body mass index or baseline HbA1c concentrations. Studies were eligible if they focused on the comparison of canagliflozin at a dose between 50 and 300 mg, with or without another active treatment, versus a placebo. Eligible trials assessed at least one outcome among the following: HbA1c, major adverse cardiovascular events over the entire study period (which could be retrieved from the serious adverse events), and serious adverse events in the double blind study period.

**Search strategy, study selection process, and data retrieval**

Two reviewers (HG and FN) identified and selected the studies among the randomised controlled trials for which data were available on the YODA project platform. Data access was granted on 16 April 2021. One investigator (HG) conducted data management and collection, and gathered the following information for each patient in each study from the available IPD, in accordance with a statistical analysis plan registered on the Open Science Framework on 18 May 2021 (https://osf.io/mbeu7/): patient identification number, treatment and dose received, study in which the patient was included, relative end date of follow-up (all original dates relative to individual patients have been removed from the dataset, only relative days to the inclusion are provided), any deviation from trial protocol and, study outcomes listed below.

**Risk-of-bias assessment**

Two independent reviewers (HG and FN) evaluated the risk of bias in each individual study and used the Cochrane Collaboration ROB2 tool for assessing risk of bias.

**Study outcomes**

We explored the vibration of effect for three different outcomes: HbA1c difference from baseline (data were extracted at baseline and at weeks 12, 18, 26, and 52), time to occurrence of the first major adverse cardiovascular event, and time to occurrence of the first serious adverse event.

As defined by the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, serious adverse events are any untoward medical occurrence that results in death at any dose, is life threatening (ie, an event in which the patient was at risk of death at the time of the event; it does not refer to an event that hypothetically might have caused death if it were more severe), requires hospital admission or a prolonged existing hospital stay, results in persistent or clinically significant disability or
incapacity, or is a congenital anomaly or birth defect. All studies included used this definition of serious adverse events.

In most studies on diabetes, in accordance with the US Food and Drug Administration guidelines, major adverse cardiovascular events are defined as a combined endpoint of cardiovascular death, non-fatal infarction, and non-fatal stroke. These events were identified from the serious adverse events using a list detailed in our statistical analysis plan (online supplemental appendix 1).

### Assessment of vibration of effect

We computed all possible combinations of all combinations of the randomised controlled trials included, by using the formula below. HbA1c measures, major adverse cardiovascular events, and serious adverse events were analysed separately. All computations were performed for four different time points: at week 12 (or closest date), 18 (or closest date), 26 (or closest date), and 52 (or closest date). For HbA1c data, if an observation was missing at a time point (±3 weeks; except for baseline, when only measures at –3 weeks were considered), it was replaced using the last observation carried forward method.\(^{20}\)

\[
C_{n,k} = \frac{n!}{k!(n-k)!}
\]

(Formula to determine possible trial combinations shows samples of k trials in a set of n trials.)

For each pooled analysis (defined by a given combination of individual studies and a given time point), data were pooled using a two stage IPD meta-analysis approach.\(^{21}\) We used a random effects model estimated via restricted maximum likelihood estimation, and derived confidence intervals using the Hartung Knapp approach. Heterogeneity was estimated using \(\chi^2\).

Effect estimates were expressed in terms of mean differences for changes in HbA1c levels and hazard ratios for major adverse cardiovascular events and serious adverse events. In case of sparse events, we used the adaptation of Firth’s correction\(^{22}\) to compute hazard ratios. We computed the distribution of these effect estimates and their corresponding P values in all analytical scenarios. Pooled analyses were considered to be nominally significant if the effect estimate had a P value <0.05. The presence of a Janus effect was investigated by calculating the first and 99th centiles of the distribution of the effect estimates.\(^{7}\) A Janus effect occurs when the first and the 99th centiles of the effect estimates of pooled analyses are in the opposite direction, illustrating the presence of substantial vibration of effect.\(^{7}\)

All analyses were performed using R (version 3.6.3). Two stage analyses were undertaken by use of the meta package\(^{23}\) in R (version 4.15-1). Adaptation of Firth’s correction was implemented using the coxphf package.\(^{25}\) All the code necessary to reproduce the analyses are available on the Open Science Framework (https://osf.io/e9cfcf/).

### Additional analyses and changes from the initial protocol

We performed a post hoc sensitivity analysis for serious adverse events by excluding all major adverse cardiovascular events from the definition of serious adverse events. We did this analysis by focusing on the serious adverse events that were not related to the outcome of major adverse cardiovascular events, which had already been explored in our analysis and which could have reflected a potential benefit of canagliflozin.

We made several minor changes to our initial protocol after receiving the data. We included patients receiving a lower dose (50 mg) than initially planned (100-300 mg) because all doses were found to be efficacious in lowering HbA1c and were in use,\(^{26,27}\) excluded one study that did not match our selection criteria, analysed only intention-to-treat randomised controlled trials (no per protocol analysis was reported in the study reports or in the publications), and to simplify the analysis, decided during the peer review process to rely on a two stage approach for all meta-analyses.

### Patient and public involvement

Patients and the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research. This is a methodological study and we had no established contacts with specific patient groups who might be involved in this project.

### Results

#### Study selection and study characteristics

Among the 17 canagliflozin randomised controlled trials with IPD listed as available on the YODA project, 13 were judged eligible and were included in our protocol. Of these, we excluded one study after the data were received because it did not include a placebo comparator arm, which resulted in 12 eligible trials.\(^{12-15,26-32}\) A flowchart detailing the trial selection process is shown in online supplemental appendix 2.

Among the 15 525 patients randomised in these 12 studies, we excluded 431 patients from two trials (n=65 in NCT00642278\(^{26}\) and n=366 in NCT01106677\(^{12}\)) who were randomised to sitagliptin, resulting in the inclusion of 15 094 patients for analysis. For all studies, the numbers of patients randomised were consistent with the journal article except for one study (NCT01989754), which reported 5812 patients in the journal article, while 5813 randomised patients were identified in the IPD (the discrepancy in this study arose because one patient was included twice in two different sites, and the investigator retained only the first inclusion of this individual, and the two identification numbers were not available). Details of the
12 studies included are presented in table 1. The risk-of-bias assessment is presented in figure 1. The number of missing data per variable and the description of protocol deviations are detailed in the data management report registered on the Open Science Framework (https://osf.io/5kd89/) and missing HbA1c data are detailed in online supplemental appendices 3 and 4.

Vibration of effect
Overall, the 12 studies resulted in 4083 different combinations of randomised controlled trials at each time point (online supplemental appendix 5) resulting in 16 322 (4×4083) different pooled analyses for HbA1c and 16 332 for serious adverse events. For major adverse cardiovascular events, we excluded one study because no events occurred, resulting in 8144 (4×2036) pooled analyses. Detailed numerical results of these analyses are presented in table 2 and a visual representation is shown in figure 2 and figure 3.

For the difference in HbA1c, the mean difference was −0.59% (95% confidence interval −0.71% to −0.46%) at 52 weeks figure 2. The distribution of the mean differences estimated ranged from −0.97% to −0.37% (range of 0.60%), with a median of −0.60% (interquartile range −0.64% to −0.57%). The effect estimates were −0.75% for the first centile and −0.48% for the 99th centile, indicating the absence of a Janus effect. In 15 994 (97.93%) pooled analyses, we observed a statistically significant superiority of canagliflozin over placebo.

For major adverse cardiovascular events, the hazard ratio was 0.75 (95% confidence interval 0.61 to 0.92) at 52 weeks figure 2. The distribution of the hazard ratios estimated ranged from 0.05 to 2.76 (range of 2.71), with a median of 0.62 (interquartile range 0.50-0.73). Hazard ratios were 0.20 for the first centile and 0.89 for the 99th centile. Of 8144 meta-analyses, 8100 (99.46%) had hazard ratios in favour of canagliflozin (4.01% (433/10 802) were significant) and 1436 (11.73%) had hazard ratios in favour of placebo (3.69% (53/1436) were significant). The numbers of major adverse cardiovascular events and serious adverse events per randomised controlled trial are presented in online supplemental appendix 7.

Post hoc sensitivity analysis
The post hoc sensitivity analysis excluded major adverse cardiovascular events from the definition of serious adverse events (table 2). At weeks 12 and 18, one study was excluded because no events were observed, resulting in 2036 possible combination for these two time points and 12 238 (2×2036 + 2×4083) pooled analyses. The hazard ratio was 0.91 (95% confidence interval 0.83 to 1.01) at 52 weeks (figure 3). The distribution of the hazard ratios estimated ranged from 0.40 to 4.28 (range of 3.88), with a median of 0.91 (interquartile range 0.87-0.94). The hazard ratio was 0.66 for the first centile and 1.55 for the 99th centile, indicating the presence of a Janus effect. Of 12 238 meta-analyses, 10 802 (88.27%) had hazard ratios in favour of canagliflozin (4.01% (433/10 802) were significant) and 1436 (11.73%) had hazard ratios in favour of placebo (3.69% (53/1436) were significant). The numbers of major adverse cardiovascular events and serious adverse events per randomised controlled trial are presented in online supplemental appendix 7.

Discussion
Summary of evidence
In this case study, we explored vibration of effect in more than 16 000 pooled analyses of IPD data from 12 randomised controlled trials comparing canagliflozin with placebo for type 2 diabetes mellitus. We observed no Janus effect for the mean difference in HbA1c, suggesting that the vibration of effect did not have an impact on the direction of the effect observed, although the point estimate varied considerably. Almost all the pooled analyses on this endpoint showed significant differences, which indicates that the uncertainty related to vibration of effect concerned the magnitude of the change. Nevertheless, a difference in HbA1c is controversial as a valid surrogate for macrovascular and microvascular complications in type 2 diabetes mellitus, and it is difficult to interpret in terms of clinical relevance.11 33

With respect to our analyses examining the vibration of effect on major adverse cardiovascular events, a clinically relevant outcome, we observed no Janus effect. However, the vibration of effect had an impact on the detection of canagliflozin efficacy on major adverse cardiovascular events. Regarding individual studies, efficacy on these events was identified in only one of the 12 trials included, perhaps as a consequence of lack of power, or because it was only observed in certain populations, such as patients at high cardiovascular risk. Notably, two studies that included patients with high cardiovascular risk (NCT01032629 and NCT01989754) and reported major adverse cardiovascular events as their primary outcome were pooled into one paper.15
Table 1 | Summary of studies included. Data are number (%) of participants unless stated otherwise

<table>
<thead>
<tr>
<th>Trial No</th>
<th>Publication</th>
<th>No of participants</th>
<th>Placebo</th>
<th>Canagliflozin</th>
<th>Mean (SD) age (years)</th>
<th>Female sex</th>
<th>Duration of placebo comparison (months)</th>
<th>Primary outcome</th>
<th>Specific criteria</th>
<th>Glycaemic control</th>
<th>Previous treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCT00642278*</td>
<td>Rosenstock et al 2012</td>
<td>386</td>
<td>65 (16.8)</td>
<td>321 (83.2)</td>
<td>53 (8)</td>
<td>188 (48.7)</td>
<td>12</td>
<td>Change in HbA1c from baseline</td>
<td>—</td>
<td>Inadequate glycaemic control</td>
<td>Metformin monotherapy</td>
</tr>
<tr>
<td>NCT01340664</td>
<td>Qu et al 2014</td>
<td>279</td>
<td>93 (33.3)</td>
<td>186 (66.7)</td>
<td>56 (9.5)</td>
<td>149 (53.4)</td>
<td>18</td>
<td>Change in HbA1c from baseline</td>
<td>—</td>
<td>Inadequate glycaemic control</td>
<td>Metformin monotherapy</td>
</tr>
<tr>
<td>NCT01106625</td>
<td>Wilding et al 2013</td>
<td>469</td>
<td>156 (33.3)</td>
<td>313 (66.7)</td>
<td>57 (9.3)</td>
<td>230 (6.9)</td>
<td>52</td>
<td>Change in HbA1c from baseline</td>
<td>—</td>
<td>Inadequate glycaemic control</td>
<td>Metformin plus sulphonylurea</td>
</tr>
<tr>
<td>NCT01064414</td>
<td>Yale et al 2013</td>
<td>269</td>
<td>90 (33.5)</td>
<td>179 (66.5)</td>
<td>68 (8.2)</td>
<td>106 (39.4)</td>
<td>52</td>
<td>Change in HbA1c from baseline</td>
<td>—</td>
<td>Inadequate glycaemic control</td>
<td>Metformin monotherapy</td>
</tr>
<tr>
<td>NCT01081834</td>
<td>Stenlöf et al 2013</td>
<td>678</td>
<td>194 (28.6)</td>
<td>484 (71.4)</td>
<td>54 (10.8)</td>
<td>376 (55.5)</td>
<td>26</td>
<td>Change in HbA1c from baseline</td>
<td>—</td>
<td>Inadequate glycaemic control</td>
<td>Diet and exercise</td>
</tr>
<tr>
<td>NCT01106667†</td>
<td>Lavalle-González et al 2013</td>
<td>918</td>
<td>183 (19.9)</td>
<td>735 (80.1)</td>
<td>55 (9.4)</td>
<td>485 (52.8)</td>
<td>26</td>
<td>Change in HbA1c from baseline</td>
<td>—</td>
<td>Inadequate glycaemic control</td>
<td>Metformin monotherapy</td>
</tr>
<tr>
<td>NCT01032629</td>
<td>Neal et al 2013</td>
<td>4330</td>
<td>1442 (33.3)</td>
<td>2888 (66.7)</td>
<td>61 (8.1)</td>
<td>1669 (39.9)</td>
<td>&gt;52</td>
<td>Onset of MACES</td>
<td>High cardiovascular risk</td>
<td>HbA1c &gt; 7.0% and ≤ 10.5%</td>
<td>—</td>
</tr>
<tr>
<td>NCT01106651</td>
<td>Bode et al 2013</td>
<td>716</td>
<td>239 (33.4)</td>
<td>477 (66.6)</td>
<td>64 (6.3)</td>
<td>320 (44.7)</td>
<td>26</td>
<td>Change in HbA1c from baseline</td>
<td>Older patients (aged 55-80 years)</td>
<td>Inadequate glycaemic control</td>
<td>Metformin, sulfonylurea, dipeptidyl peptidase 4 inhibitor, α-glucosidase inhibitor, glucagon-like peptide 1 agonist, or insulin or pioglitazone</td>
</tr>
<tr>
<td>NCT01106690</td>
<td>Forst et al 2014</td>
<td>342</td>
<td>115 (33.6)</td>
<td>227 (66.4)</td>
<td>57 (10)</td>
<td>126 (36.8)</td>
<td>26</td>
<td>Change in HbA1c from baseline</td>
<td>—</td>
<td>Inadequate glycaemic control</td>
<td>Metformin and pioglitazone</td>
</tr>
<tr>
<td>NCT01381900</td>
<td>Ji et al 2015</td>
<td>676</td>
<td>226 (33.4)</td>
<td>450 (66.6)</td>
<td>49 (10.1)</td>
<td>314 (46.4)</td>
<td>18</td>
<td>Change in HbA1c from baseline</td>
<td>Asian patients with and without risk factors for CVD</td>
<td>Inadequate glycaemic control</td>
<td>Metformin or metformin+sulphonylurea</td>
</tr>
<tr>
<td>NCT01989754</td>
<td>Neal et al 2017</td>
<td>5813</td>
<td>2906 (50)</td>
<td>2907 (50)</td>
<td>62 (8.6)</td>
<td>2164 (37.2)</td>
<td>&gt;52</td>
<td>Onset of MACE</td>
<td>High cardiovascular risk</td>
<td>HbA1c &gt; 7.0% and ≤ 10.5%</td>
<td>—</td>
</tr>
<tr>
<td>NCT02025907</td>
<td>Rodbard et al 2015</td>
<td>218</td>
<td>109 (50)</td>
<td>109 (50)</td>
<td>57 (9.7)</td>
<td>92 (42.2)</td>
<td>26</td>
<td>Change in HbA1c from baseline</td>
<td>—</td>
<td>Inadequate glycaemic control</td>
<td>Metformin and sitagliptin</td>
</tr>
</tbody>
</table>

SD=standard deviation; HbA1c=serum glycated haemoglobin; MACE=major adverse cardiovascular events.

*In this study, 65 additional patients were randomised in an active comparator arm (sitagliptin) but were not included in our analyses.

†In this study, 366 additional patients were randomised in an active comparator arm (sitagliptin) but were not included in our analyses.

‡Study group includes Chinese and other Asian patients; the study was conducted in China, Malaysia, and Vietnam.
with a significant result of the pooled analysis, but not significant for any of the individual trials. This pooled analysis was planned a priori and the efficacy of canagliflozin on major adverse cardiovascular events has now been robustly established in another independent study for patients with type 2 diabetes mellitus who have chronic kidney disease, a study for which data were not available at the time of our initial request through the YODA project.

We observed a Janus effect in our analyses examining the vibration of effect on serious adverse events. However, because these results could be confounded by the results observed for major adverse cardiovascular events, we excluded this subgroup from our definition of serious adverse events. This approach, excluding the beneficial effects of a treatment from the composite of serious adverse events, has been proposed in the field of psoriasis research. Results on serious adverse events were robust in this post hoc sensitivity analysis.

Vibration of effect has been suggested as a standardised method that can be used to systematically evaluate the breadth and divergence of study results, depending on the various methodological choices. In the context of meta-analyses, this approach is quite similar to the GOSH (graphical display of study heterogeneity) method, which was proposed for meta-analyses on aggregated data. We believe that a method of this type, exploring all possible subsets, makes even more sense in the context of pooled analyses, because these studies do not, by nature, exhaustively cover

<table>
<thead>
<tr>
<th>NCT</th>
<th>Publication</th>
<th>RP</th>
<th>DEV</th>
<th>MIS</th>
<th>MO</th>
<th>SRR</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCT00642278</td>
<td>Rosenstock 2012</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NCT01340664</td>
<td>Qiu 2014</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NCT01106625</td>
<td>Wilding 2013</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NCT01064414</td>
<td>Yale 2013</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NCT01081834*</td>
<td>Stenlöf 2013</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NCT01106677</td>
<td>Lavalle-Gonzalez 2013</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NCT01032629; NCT01989754*</td>
<td>Neal 2017</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NCT01106651</td>
<td>Bode 2013</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NCT01106690</td>
<td>Forst 2014</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NCT01381900</td>
<td>Ji 2015</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NCT02025907</td>
<td>Rodbard 2016</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 1 | Risk-of-bias assessment. RP=randomisation process; DEV=deviations from the intended interventions; MIS=missing outcome data; MO=measurement of the outcome; SRR=selection of the reported result. Green circles with + signs indicate low risk. *For trials NCT01081834, NCT01032629, and NCT01989754, a risk of selective outcome reporting in the publication was identified, but this risk was rated low in the present analysis because all this information was available in the individual participant data. For missing outcome data, see online supplemental appendix 4.

Table 2 | Vibration of effect for comparison of canagliflozin with placebo for treatment of type 2 diabetes mellitus, by time point. Table includes numerical results for pooled analysis using all 12 studies included and for the first and 99th centiles in terms of point estimates.

<table>
<thead>
<tr>
<th></th>
<th>Week 12 Mean (95% CI)</th>
<th>Week 18 Mean (95% CI)</th>
<th>Week 26 Mean (95% CI)</th>
<th>Week 52 Mean (95% CI)</th>
<th>All Mean (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HbA1c</td>
<td>-0.6 (-0.72 to -0.49)</td>
<td>-0.62 (-0.73 to -0.5)</td>
<td>-0.6 (-0.72 to -0.49)</td>
<td>-0.59 (-0.71 to -0.46)</td>
<td>-0.6 (-0.72 to -0.49)</td>
</tr>
<tr>
<td>τ²</td>
<td>0.02</td>
<td>0.03</td>
<td>0.02</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>First centile</td>
<td>-0.75</td>
<td>-0.77</td>
<td>-0.75</td>
<td>-0.74</td>
<td>-0.75</td>
</tr>
<tr>
<td>99th centile</td>
<td>-0.5</td>
<td>-0.51</td>
<td>-0.5</td>
<td>-0.46</td>
<td>-0.48</td>
</tr>
<tr>
<td>Median</td>
<td>-0.61</td>
<td>-0.62</td>
<td>-0.61</td>
<td>-0.59</td>
<td>-0.61</td>
</tr>
<tr>
<td>Interquartile range</td>
<td>0.06</td>
<td>0.09</td>
<td>0.06</td>
<td>0.08</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Major adverse cardiovascular events

<table>
<thead>
<tr>
<th></th>
<th>Week 12 Hazard ratio (95% CI)</th>
<th>Week 18 Hazard ratio (95% CI)</th>
<th>Week 26 Hazard ratio (95% CI)</th>
<th>Week 52 Hazard ratio (95% CI)</th>
<th>All Hazard ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>τ²</td>
<td>0.58 (0.4 to 0.85)</td>
<td>0.67 (0.5 to 0.9)</td>
<td>0.72 (0.55 to 0.93)</td>
<td>0.75 (0.61 to 0.93)</td>
<td>0.72 (0.55 to 0.93)</td>
</tr>
<tr>
<td>First centile</td>
<td>0</td>
<td>0.2</td>
<td>0.22</td>
<td>0.23</td>
<td>0.20</td>
</tr>
<tr>
<td>99th centile</td>
<td>0.96</td>
<td>0.89</td>
<td>0.95</td>
<td>0.81</td>
<td>0.89</td>
</tr>
<tr>
<td>Median</td>
<td>0.58</td>
<td>0.61</td>
<td>0.66</td>
<td>0.72</td>
<td>0.62</td>
</tr>
<tr>
<td>Interquartile range</td>
<td>0.16</td>
<td>0.2</td>
<td>0.21</td>
<td>0.23</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Serious adverse events

<table>
<thead>
<tr>
<th></th>
<th>Week 12 Hazard ratio (95% CI)</th>
<th>Week 18 Hazard ratio (95% CI)</th>
<th>Week 26 Hazard ratio (95% CI)</th>
<th>Week 52 Hazard ratio (95% CI)</th>
<th>All Hazard ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>τ²</td>
<td>0.78 (0.66 to 0.92)</td>
<td>0.87 (0.77 to 0.98)</td>
<td>0.87 (0.79 to 0.96)</td>
<td>0.91 (0.86 to 0.96)</td>
<td>0.91 (0.86 to 0.96)</td>
</tr>
<tr>
<td>First centile</td>
<td>0</td>
<td>0.65</td>
<td>0.67</td>
<td>0.66</td>
<td>0.59</td>
</tr>
<tr>
<td>99th centile</td>
<td>1.16</td>
<td>1.2</td>
<td>1.06</td>
<td>0.97</td>
<td>1.14</td>
</tr>
<tr>
<td>Median</td>
<td>0.78</td>
<td>0.87</td>
<td>0.87</td>
<td>0.9</td>
<td>0.87</td>
</tr>
<tr>
<td>Interquartile range</td>
<td>0.09</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Serious adverse events without major adverse cardiovascular events

<table>
<thead>
<tr>
<th></th>
<th>Week 12 Hazard ratio (95% CI)</th>
<th>Week 18 Hazard ratio (95% CI)</th>
<th>Week 26 Hazard ratio (95% CI)</th>
<th>Week 52 Hazard ratio (95% CI)</th>
<th>All Hazard ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>τ²</td>
<td>0.84 (0.66 to 1.05)</td>
<td>0.91 (0.79 to 1.11)</td>
<td>0.91 (0.81 to 1.01)</td>
<td>0.91 (0.83 to 1.01)</td>
<td>0.91 (0.83 to 1.01)</td>
</tr>
<tr>
<td>First centile</td>
<td>0</td>
<td>0.69</td>
<td>0.7</td>
<td>0.68</td>
<td>0.66</td>
</tr>
<tr>
<td>99th centile</td>
<td>1.81</td>
<td>1.61</td>
<td>1.27</td>
<td>1.25</td>
<td>1.55</td>
</tr>
<tr>
<td>Median</td>
<td>0.85</td>
<td>0.94</td>
<td>0.91</td>
<td>0.92</td>
<td>0.91</td>
</tr>
<tr>
<td>Interquartile range</td>
<td>0.11</td>
<td>0.08</td>
<td>0.07</td>
<td>0.03</td>
<td>0.07</td>
</tr>
</tbody>
</table>

CI=confidence interval; HbA1c=serum glycated haemoglobin (HbA1c), expressed as a percentage.
all existing studies. In addition, the use of IPD enabled us to explore and extract outcomes (i.e., major adverse cardiovascular events and serious adverse events) that would have been difficult to extract from aggregated data, because they would not have been measured or reported in the initial publications, thus increasing the relevance of our study beyond the classic GOSH approach.

Lastly, we used a definition of the Janus effect that is only contingent on point estimates, and not on statistical significance, as in previous work.\(^7\)\(^9\)\(^36\) When looking for statistical significance, our case study very rarely identified contradictory results. Observing changes in the direction of effect estimates and occasionally in significance is to be expected because of sampling variability only. Heterogeneity, bias in some of the initial studies, and the magnitude of the effect might also affect the existence of vibration of effect and the presence of a Janus effect. However, we believe that the bigger concern for pooled analyses is the presence of selection or availability bias in the IPD used in the meta-analysis.\(^38\)

**Strengths and limitations of the study**

The findings from our case study might not be generalisable across all fields. Firstly, we selected an example involving a large number (and therefore a large number of possible combinations) of randomised controlled trials. Identifying vibration of effect in fields with only a few trials could be more challenging. Secondly, we only considered two methodological choices—that is, study inclusion and timing of endpoints. Vibration of effect can be influenced by many other characteristics, including subgroup analyses, different definitions of outcomes (e.g., a different construction of major adverse cardiovascular events), different groupings of doses, and different analytical strategies (e.g., choice of one stage vs two stage IPD, model specification, or different handling of missing data). Therefore, our results could in fact...

---

**Figure 2 | Vibration of effects for the comparison of canagliflozin with placebo for treatment of type 2 diabetes mellitus, according to haemoglobin A1c (HbA1c) difference from baseline and time to occurrence of the first major adverse cardiovascular event. For HbA1c, a negative mean difference is in favour of canagliflozin. For major adverse cardiovascular events, a hazard ratio <1 is in favour of canagliflozin. In the graphs on the right, dots represent meta-analyses and colours represent densities of dots (red=high; blue=low)**
underestimate the vibration of effect that might have resulted from many other researchers' degrees of freedom in the analysis. However, our evaluation concerned 16,332 analyses and provides an idea of the impact of different trial combinations and repeated endpoint measures. Our choice to focus such combinations of studies was driven by the fact that many pooled analyses are run with the risk of manipulating the results, by selecting favourable combinations of studies. Nevertheless, subgroup analyses might be frequently conducted in pooled analyses (eg, for duloxetine), a consideration that deserves attention in future research.

Thirdly, we relied only on studies available on the YODA platform at the time of our request. All these studies were sponsored by Janssen, and we considered that this specific subset of studies was adequately represented a sample that a given sponsor would use when conducting a series of secondary analyses. Relying on IPD from such a homogeneous subset of studies allowed the quality of studies to be better assessed, and for analyses to be standardised. Therefore, selected studies for the analysis could have less variability and less potential for vibration of effect than in the present study. We did not conduct a systematic search for other studies, for example, those conducted in an academic context. Whether the authors of academic studies would have shared the trial IPD necessary to conduct our analyses for vibration of effect is uncertain, and including studies of this type could have added heterogeneity and vibration of effect.

Implications of the findings
Our findings have several implications. Especially when performing post hoc evaluations of published trials, pooled analyses focusing on a subset of all available studies cannot be simply assumed to be the preferred method. In particular, our findings suggest that results from pooled analyses should be critically appraised. Health authorities, for instance, should not rely exclusively on findings from pooled analyses when approving treatments. Evidence suggests that findings from pooled analyses have been used to guide approvals by the European Medicine Agency, including that for naloxone for alcohol

---

Figure 3 | Vibration of effects for the comparison of canagliflozin with placebo for treatment of type 2 diabetes mellitus, according to time to occurrence of the first serious adverse event (with and without major adverse cardiovascular events). A hazard ratio <1 is in favour of cangliflozin. In the graphs on the right, dots represent meta-analyses and colours represent densities (red=high; blue=low).
use disorders. To enhance the quality of pooled analyses and the evidence generated by them, we suggest that pooled analyses should be planned a priori, with detailed, pre-registered study protocols, as with prospective meta-analyses. This step would minimize any methodological changes during the analyses that could introduce vibration of effect. If pre-registration is not possible (eg, when the researchers conducting pooled analyses are not involved in the design or conduct of the original randomised controlled trial), analytical plans should be registered before data analysis, to maintain full transparency regarding any decision made during the conduct of the study, such as the selection of studies to be pooled in the analysis. Pooled analyses should rely on IPD from studies that are representative of the target population of interest and high quality, in order to best estimate the estimands of interest. These steps will continue to be important as data sharing increases in medicine and secondary uses of this type become more popular.

We think that the vibration-of-effect approach shows promise in exploring issues related with reproducibility, especially because overlapping meta-analyses with divergent conclusions are not rare in the literature. However, to recommend implementing the method in all IPD meta-analysis/pooled analyses would be immature. Therefore, we recommend that future research systematically explores vibration of effect in a large set of meta-analyses in order to give a better indication of its relevance. Such a study will also help to investigate associations between the vibration of effect and the Janus effect with many parameters such as heterogeneity, effect size, study quality, and random sampling.

Conclusion
In this case study, we explored the vibration of effect in more than 16 000 pooled analyses of IPD data from 12 randomised controlled trials comparing canagliflozin with placebo for treating patients with type 2 diabetes mellitus. We found substantial variations in the magnitude and, for serious adverse events, the direction of the effects estimated. These findings suggest that when conducting pooled analyses of IPD from randomised controlled trials, trial selection, analysis of subsets of all trials and their selection or availability of IPD could have considerable consequences on treatment effect estimation.

AUTHOR AFFILIATIONS
1Inserm, CIC 1444 (Centre d’Investigation Clinique de Rennes), Rennes 1 University, Rennes, France
2Inserm, U1087 (Institut de recherche en santé, environnement et travail), Rennes 1 University, Rennes, France
3Department of Environmental Health Sciences, Yale University School of Public Health, New Haven, CT, USA
4UCBL, CNRS, UMR 5558, BBBE, EMET, University Claude Bernard Lyon 1, Villeurbanne, France
5Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA
6Institut Universitaire de France, Paris, France

Twitter Joshua D Wallach @JoshuaDWallach and Florian Naudet @NaudetFlorian

Acknowledgements We thank Anne Hespel and Frederic Rimattei for processing the data requests, and Sarah Verdier Leyshon and Angela Swaine Verdier for revising the paper for English. The interpretation and reporting of research using this data are solely the responsibility of the authors and do not necessarily represent the official views of the YODA Project or Janssen Research and Development.

Contributors FN HG conceived and designed the experiments. HG, JDW, RB, JSR, and FN validated the protocol. HG extracted and analysed the data, and wrote the first draft of the manuscript. FN and HG interpreted the results. JDW, JSR, FN, and HG contributed to the writing of the manuscript. All authors agreed with the results and conclusions of the manuscript. FN was study guarantor. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. Transparency: The lead author (the guarantor) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Funding The project was funded by the Agence Nationale de la Recherche (reference ANR-17-CE-36-0010-01). The sponsor had no role concerning the preparation, review, or approval of the manuscript.

Competing interests All authors have completed the ICMJE uniform disclosure form at www.icmje.org/disclosure-of-interest/ and declare: support from the Agence Nationale de la Recherche for the submitted work. JDW currently receives research support from the US Food and

Figure 4 | Overlap between points estimates and 95% confidence intervals, for all pooled analyses and for the full meta-analysis (in black). Colours represent the densities of estimates (red=high; blue=low). for clarity, X axis limits were set at −1.5 to 0.5 for continuous outcomes and 0.05 to five for survival outcomes. Large central black dots and vertical lines correspond to the estimate and 95% confidence interval of the pooled analysis including the all 12 studies. Curved lines are composed of smaller black dots and vertical lines correspond to the 95% confidence interval of all the pooled analyses. Full figures including extreme values are presented in online supplemental appendix 6.
Drug Administration and the National Institute on Alcohol Abuse and Alcoholism of the National Institutes of Health under award K04Ao208258. JSR currently receives research support through Yale University from Johnson and Johnson to establish the YODA project and develop methods of clinical trial data sharing; from the Medical Device Innovation Consortium as part of the National Evaluation System for Health Technology, from the FDA for the Yale-Mayo Clinic Center for Excellence in Regulatory Science and Innovation program (U3UFD005938); from the Agency for Healthcare Research and Quality (R01HS22882); from the National Heart, Lung, and Blood Institute of the National Institutes of Health (R01HL144644); and from the Laura and John Arnold Foundation to establish the Good Pharma Scorecard at Bioethics International; JSR is an expert witness at the request of Relator’s attorneys, the Greene Law Firm, in a qui tam suit alleging violations of the False Claims Act and Anti-Kickback Statute against Biogen; JSR is an associate editor at BMJ and an editorial board member at BMJ Medicine; he was not involved in BMJ Medicine’s handling of this manuscript.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility upon this work non-commercially, and license their derivative works upon different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iDs Henri Gouraud http://orcid.org/0000-0003-3882-3160 Joshua D Wallach http://orcid.org/0000-0002-2816-6905 Rémy Boussageon http://orcid.org/0000-0002-8598-3186 Joseph S Ross http://orcid.org/0000-0002-9218-3320 Florian Naudet http://orcid.org/0000-0003-3760-3810

REFERENCES

2 loanidis JPA. The mass production of redundant, misleading, and Conflicted systematic reviews and meta-analyses. Milbank Q 2016;94:456–514. doi:10.1111/1468-0013.12120
8 loanidis JPA. Why most discovered true associations are inflated. Epidemiology 2008;19:60–8. doi:10.1097/EDE.0b013e31813f8577

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjmed-2022-000154).