Received: 19 January 2022 | Revised: 21 April 2022 | Accepted: 28 April 2022

DOI: 10.1111/dom.14735

ORIGINAL ARTICLE

WILEY

Non-inferiority and clinical superiority of glucagon-like peptide-1 receptor agonists and sodium-glucose co-transporter-2 inhibitors: Systematic analysis of cardiorenal outcome trials in type 2 diabetes

Francesco Zaccardi PhD[1,2,3] | David E. Kloecker MPhil[1,2] | Kamlesh Khunti PhD[1,2,3] | Melanie J. Davies MD[2,4]

[1]Leicester Real World Evidence Unit, University of Leicester, Leicester General Hospital, Leicester, UK

[2]Diabetes Research Centre, University of Leicester, Leicester General Hospital, Leicester, UK

[3]NIHR Collaboration for Leadership in Applied Health Research and Care-East Midlands, University of Leicester, Leicester, UK

[4]NIHR Leicester Biomedical Research Centre, University Hospitals of Leicester NHS Trust and University of Leicester, Leicester, UK

Correspondence
Dr Francesco Zaccardi, PhD, Leicester Real World Evidence Unit, Diabetes Research Centre, Leicester General Hospital, Gwendolen Rd, Leicester LE5 4PW, UK.
Email: frazac@fastwebnet.it
@LRWEUnit@LDC_tweets

Funding information
National Institute for Health and Care Research (NIHR); NIHR Leicester Biomedical Research Centre

Abstract

Aims: Most trials leading to the approval of glucagon-like peptide receptor agonists (GLP-1RAs) and sodium-glucose co-transporter-2 inhibitors (SGLT2is) were primarily designed to confirm their non-inferiority to placebo (commonly using an upper 95% confidence limit threshold of 1.3) and, if confirmed, superiority (threshold 1): this asymmetry of margins (1 vs. 1.3) favours the active intervention. We aimed to quantify the probability of clinical superiority of the active treatment by applying the same threshold used to claim non-inferiority.

Materials and Methods: We searched PubMed and Cochrane CENTRAL for cardiorenal outcome trials in subjects with type 2 diabetes published before 5 December 2021, to reconstruct from Kaplan-Meier plots individual-level data for the primary outcome or all-cause mortality. We calculated Bayesian posterior densities to obtain the probability for a treatment effect (hazard ratio) <0.769, which is symmetric to the 1.3 threshold (i.e. its reciprocal 1/1.3), emulating a scenario where the active treatment is placebo and placebo is the active treatment.

Results: We extracted data from 27 Kaplan-Meier plots (18 for the primary outcome, nine for mortality). Probabilities of clinical superiority to placebo varied significantly: for GLP-1RAs, from a minimum of 0% to a maximum of 69% for the primary outcome and from 0% to 8% for mortality; corresponding estimates for SGLT2is were 0% to 96% and 0% to 93%. Probabilities were on average greater for SGLT2is, particularly in trials investigating kidney or heart failure outcomes.

Conclusions: The probability of clinical superiority to placebo varies widely across trials previously reported as showing superiority of GLP-1RAs or SGLT2is compared with placebo. These results showed within- and between-class differences, highlight the drawbacks of a binary interpretation of the results, particularly in the context of the current designs of non-inferiority trials, and have implications for decision makers and future clinical recommendations.

KEYWORDS
cardiovascular disease, diabetic nephropathy, GLP-1 analogue, randomized trial, SGLT2 inhibitor

1 | INTRODUCTION

Randomized controlled trials showed that treatments with glucagon-like peptide 1 receptor agonists (GLP-1RA) or sodium-glucose co-transporter-2 inhibitors (SGLT2is) result in notable reductions in the rates of classical atherothrombotic cardiovascular events, such as myocardial infarction and stroke, as well as in renal and heart failure outcomes in subjects with type 2 diabetes.[1] Most of these trials, popularized as 'cardiovascular outcome trials' (CVOTs), were designed according to the 2008 United States Food and Drug Administration (FDA) recommendations for industry,[2] after concerns around the cardiovascular risks of glucose-lowering medications such as rosiglitazone.[3] Newer glucose-lowering medications now had to be shown not to result in an unacceptable increase in the risk of cardiovascular disease. To secure FDA approval, these large phase 3 trials had to show safety in the form of non-inferiority of the active compared with the control treatment, defined as an upper bound of the two-sided 95% confidence interval (CI) lower than 1.3 (more rarely 1.8) in the active versus control comparison.

Many CVOTs also tested the hypothesis of efficacy through superiority of the active treatment in a hierarchical way, i.e. upon demonstration first of non-inferiority. This approach, however, may result in an important limitation, which stems from the way in which the thresholds used to claim non-inferiority (1.3) and superiority (1) are applied,[4] which becomes apparent in certain scenarios where the roles of the active treatment and the control are being reversed (Figure 1). As an example, if the comparison active versus control results in a hazard ratio (HR) of 0.95 (95% CI: 0.85-1.06), the trial is labelled as showing non-inferiority of the active treatment to control. When inverting the roles and estimating the reciprocal of the HRs (1/HR), the control versus active treatment results in an HR of 1.05 (0.94-1.18), which indicates the non-inferiority of the control (and thus its potential approval) compared with the active treatment (Figure 1). In that hypothetical trial, each arm is non-inferior to the other and neither is superior, a conclusion that is logically sound. By contrast, in a trial where the comparison of active versus control results in an HR of 0.90 (0.85-0.95) – labelled as showing both non-inferiority and superiority of the active treatment, the inverse comparison (1.11; 1.05-1.18) still indicates the non-inferiority of control to active treatment. In other words, the active treatment here is superior to the control and, at the same time, the control non-inferior (based on the arbitrarily chosen non-inferiority margin of 1.3) to the active treatment (Figure 1).

The incongruence stems from the asymmetry of the two thresholds, HR <1.3 for non-inferiority and HR <1 for superiority, which favours the active over the control treatment.[4] If an increase in the rate of the outcome up to 1.3 (30%) is accepted to claim non-inferiority, it seems reasonable and ethical to declare clinical rather than mere statistical superiority to the control when the reduction in the rate with the active treatment is also at least 30%, rather than any reduction greater than 0% (i.e. 95% CI upper bound lower than 1).[4] Consequently, when comparing the active versus control treatment the upper bound of the 95% CI should be <1/1.3, which equates to 0.769; such a threshold could help distinguish between a 'statistically'

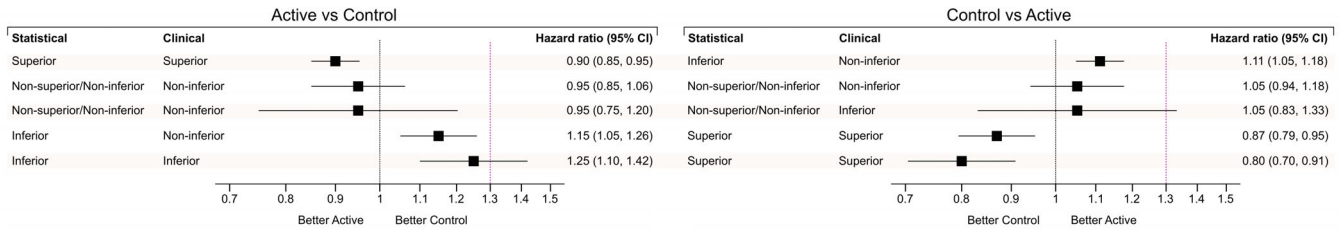

FIGURE 1 Statistical and clinical significance in non-inferiority trials. Five hypothetical trials showed the logical inconsistency between the criteria for non-inferiority and for superiority. Black dotted line shows the margin of 1 to determine statistical superiority, while the dotted magenta line indicates the margin of 1.3 commonly considered when investigating the non-inferiority of the active medication to control in cardiorenal outcome trials in type 2 diabetes. A trial comparing an active versus control intervention with an hazards ratio (HR) of 0.95 (0.85-1.06; second row) shows the non-inferiority (upper bound of 95% confidence interval lower than 1.3) of the active compared with the control intervention; when inverting the roles and considering control versus active intervention, the HR is 1.05 (0.94-1.18) and shows non-inferiority of the control compared with the active intervention; therefore, in this trial each arm is clinically non-inferior to the other and neither is clinically superior. A trial showing an HR of 0.90 (0.85-0.95; first row) comparing active with control intervention shows both clinical non-inferiority (upper bound of 95% confidence interval lower than 1.3) and statistical superiority (upper bound of 95% confidence interval lower than 1); however, this trial is considered to indicate clinical superiority as the threshold of 1 is also used to claim clinical superiority. This results in a paradox as the inverse comparison control versus active (1.11; 1.05-1.18) shows the clinical non-inferiority of the control to the active treatment; therefore, in the same trial the active treatment is superior to the control but, at the same time, the control is non-inferior to the active treatment. Note that not all possible results are shown in this active versus control example [i.e. 0.93 (0.60-1.45); 1.09 (0.95-1.25); 1.17 (0.95, 1.45)].

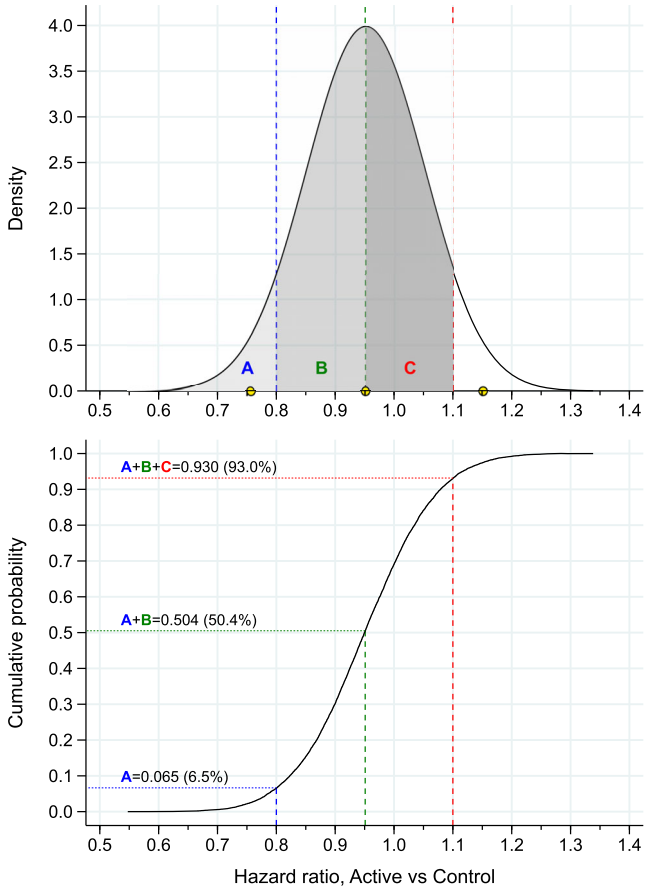

FIGURE 2 Posterior distribution and probabilities for a hypothetical trial. Top panel: Posterior hazard ratio distribution with mean 0.95 and 95% credible interval of 0.76-1.15 (yellow circles on the x-axis) in a simulated trial comparing active versus control treatment. The area under the curve A indicates the probability of values of hazard ratio ≤0.8: such probability (6.5%) can be determined in the bottom panel (cumulative density plot) as the value on the y-axis corresponding to the hazard ratio of 0.8. Probability of values of hazard ratio ≤1.1 is 93.0%, corresponding to the sum of the three areas A + B + C. This plot shows how, in a Bayesian context, it is possible to quantify the probability that the magnitude of a treatment effect is smaller (or larger) than a specific threshold.

(upper bound <1) and a 'clinically' (upper bound <0.769) significant reduction in the outcome rates.

Different thresholds may be defined to deem an active treatment clinically significant. It may be argued that 1.3 is high in some scenarios (e.g. the costs of the active and comparator treatment are similar) but not in others (e.g. the side effects in the current treatment are too severe or frequent), which would make the process of clinical interpretation more subjective. In this respect, a Bayesian estimation of the treatment effect is helpful, as it allows us to quantify the probability that the effect is lower (or greater) than any specific threshold (Figure 2).[5-9]

Within this context, we extracted individual level data from all published CVOTs including subjects with type 2 diabetes randomized to a GLP-1RA or SGLT2 inhibitor and reporting incident cardiovascular or renal (cardiorenal) or all-cause mortality events, to estimate their effects with two main goals: (a) quantify the probability that each treatment results in a reduction of the outcome at least equal to the 1.3 threshold used to declare non-inferiority to the control, i.e. the probability that the active treatment is clinically superior, and (b) estimate these probabilities for a range of other thresholds to aid the understanding of treatment effects across a continuum and facilitate comparisons within and between the two glucose-lowering classes.

2 | MATERIALS AND METHODS

2.1 | Data sources and searches

We conducted this study in line with a pre-specified protocol and reported the results following current guidance for conducting and reporting systematic reviews (PRISMA checklist reported in the Appendix S1).[10] We searched PubMed and the Cochrane Central Register of Controlled Trials (CENTRAL) for randomized controlled trials published in English before 5 December 2021; the search strategy is reported in Figure S1.

2.2 | Study selection

We initially identified trials including adult patients with type 2 diabetes mellitus randomized to a GLP-1RA or SGLT2i and investigating the risk of cardiovascular outcomes or death. Records were included if the Kaplan-Meier plot was available for the primary outcome or all-cause mortality. In view of the prominent effect of the SGLT2is on chronic kidney disease and heart failure outcomes observed in earlier trials, more recent studies have primarily explored the effects of SGLT2is on these two outcomes in subjects with and without type 2 diabetes at randomization. We therefore identified also studies with outcomes related to chronic kidney disease or heart failure, which enrolled all or a subcohort of subjects with type 2 diabetes; in the latter case, studies were included if the Kaplan-Meier plot for the primary outcome or death was available for the group of subjects with type 2 diabetes. We excluded studies reporting subgroup analyses of the main trial (i.e. stratified Kaplan-Meier plots by the presence of chronic kidney disease or heart failure at randomization).

2.3 | Data extraction and quality assessment

We used a standardized, pre-defined form to extract trial data. In each study, information was retrieved for: first author name; trial acronym; ClinicalTrials.gov (NCT) number; PubMed ID (PMID) identifier number; GLP-1RA or SGLT2i agent; baseline characteristics of trial participants (number of randomized participants, age and sex); characteristics of the trial (inclusion criteria, follow-up duration, primary outcome definition, number of subjects with primary outcome

events and number of deaths); available Kaplan-Meier plots. Study quality was assessed with the Cochrane risk of bias tool.[11] Disagreement at any stage of the review process was solved by consensus or arbitration.

2.4 | Data synthesis and analysis

From each Kaplan-Meier curve, we extracted data on the number of patients at risk and, using Engauge Digitizer, on the time (x-axis) and survival probability (y-axis) coordinates. We then used the Stata *ipdfc* command to reconstruct individual level time to event data from the corresponding x-y values and the total number of events.[12] To estimate the treatment effect (the HR comparing the active treatment vs. placebo), individual-level data were analysed in R using the Bayesian survival analysis package *rstanarm*.[13] We modelled the log baseline hazard function with a B-spline (four degrees of freedom) and selected weakly informative normal priors for the intercept of the linear predictor and for each spline coefficient (mean ± standard deviation 0 ± 20) as well as for the treatment effect (mean ± standard deviation 0 ± 2.5):[13] these priors allow the results to be closer to the treatment effect calculated in the original trial with the Cox regression.[14] For each trial and outcome, we obtained the posterior density distribution with 10 000 Markov chain Monte Carlo iterations; convergence was evaluated by visual inspection of the trace and autocorrelation plots and using the $\widehat{R}$ statistic. The probabilities of a treatment effect lower than 0.769 and 1 were calculated as the proportions of values of HR lower than 0.769 and 1, respectively. As the value of the non-inferiority margin is based on a combination of clinical and statistical criteria, we estimated probabilities also for other values of HR.

We also estimated the HR with the frequentist Cox regression using the extracted data and compared it with the Bayesian and original, reported study estimate: this served to clarify whether potential differences between the Bayesian and reported estimate were because of the Bayesian modelling (Cox and reported estimate similar but Bayesian different) or to the quality of data extraction from the Kaplan-Meier curves (Cox and Bayesian estimate similar but different from the original report).

We used Engauge Digitizer (version 10.11) for Kaplan-Meier data extraction; Stata BE (version 17.0) for data manipulation, analyses and graphs; R (version 4.0.1) for the Bayesian analysis;[13] and Inkscape (version 0.92.3) for graph finalization.

3 | RESULTS

3.1 | Study characteristics

After excluding duplicates and screening by title or abstract, we assessed the full text of 44 articles: six reported the baseline characteristics of trial participants, trial protocol or comments on the results, while in 19 no Kaplan-Meier plots for the primary outcome or all-cause mortality were available, leaving 19 articles (reporting 27 Kaplan-Meier plots in 18 trials) for data extraction and quantitative analyses (Figure S1, Tables S1 and S2). The references of the included studies are reported in Appendix S1.

The characteristics of the included trials are summarized in Tables S1 and S2: Kaplan-Meier curves for the primary outcome were available in all trials (eight with GLP-1RAs and 10 with SGLT2is). The definition of the primary outcome was largely consistent across studies with GLP-1RAs [composite of a 3-point major adverse cardiac events (MACE): non-fatal myocardial infarction, non-fatal stroke or death from cardiovascular causes]; conversely, it varied across studies with SGLT2is, being 3-point MACE in five and a variable combination of hospitalization for heart failure and renal outcomes in the remaining five studies (Table S2). In all trials with a GLP-1RA and in seven with a SGLT2i, all subjects had type 2 diabetes at randomization; in DAPA-CKD and DAPA-HF, subjects with type 2 diabetes constituted a subcohort representing 67.5% and 45.1% of the overall participants, respectively; in EMPEROR Reduced, the proportion of subjects with type 1 or type 2 diabetes was 49.8%. Inclusion criteria also varied across trials: most studies with GLP-1RAs enrolled adult subjects with either established atherosclerotic cardiovascular disease or cardiovascular risk factors (both heterogeneously defined) while trials with SGLT2is included predominantly adult subjects with chronic kidney disease or heart failure (Table S2).

In total, 109 977 subjects with diabetes participated in the trials: 60 080 in studies with GLP-1RAs (range 3183-14 752) and 49 897 in those with SGLT2is (range 1222-10 584). At baseline, the weighted median (interquartile range) age was 64.3 (63.1-66.0) years and 64% were men; the median follow-up was 2.6 (1.6-3.5) years. The primary outcome, reported in all but one trial, occurred in 12 792 participants; the number of deaths was reported in all but one trial, totalling 8769; and the rate of primary outcome and death in the placebo arm ranged from 24 to 246 and 16 to 163 per 1000 person-years, respectively (Table S1).

Across all items and trials, the risk of bias was deemed low, high and unclear in 92%, 3% and 5% of cases, respectively (Table S3). The highest domain-specific bias was observed for 'incomplete outcome data' (three trials, 17%) followed by 'blinding of outcome assessment' (one trial, 6%).

3.2 | Treatment effects

Posterior medians and 95% credible intervals of the distributions of the HRs, for each trial and available outcome, are summarized in Figure 3; their densities and cumulative probabilities are shown in Figures S2 and S3, respectively.

In trials with GLP-1RAs, for the primary outcome the probability of resulting in any reduction in the event rate (i.e. upper limit of the 95% credible interval <1) ranged from 92% (PIONEER-6) to 100% (LEADER, HARMONY Outcomes and AMPLITUDE-O), with the exception of ELIXA (40%). The probability of a reduction greater than

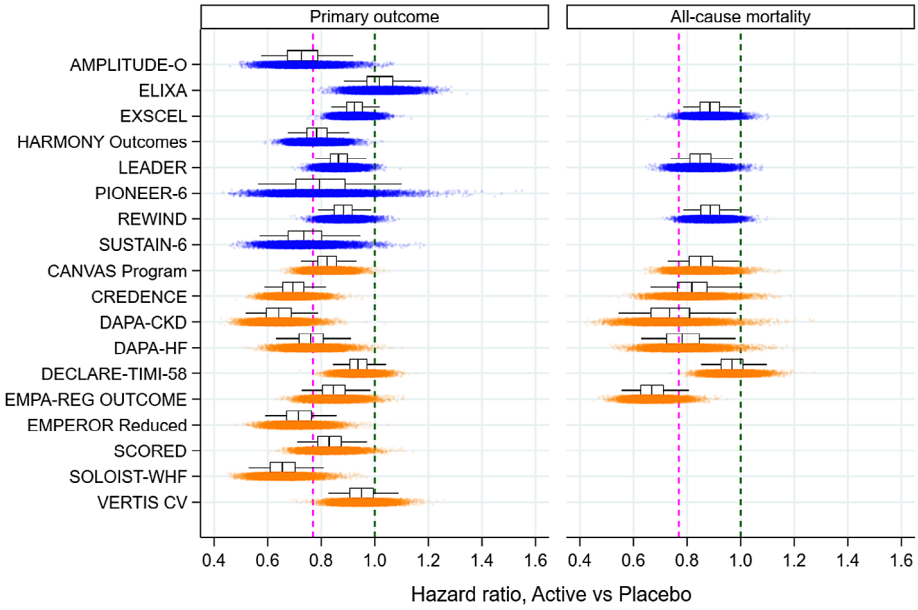

FIGURE 3 Bayesian treatment effect estimates. Each point corresponds to the estimate from a single iteration (10 000 for each trial and outcome); the box plot shows the median and interquartile range with spikes indicating the 95% credible interval; and the dotted magenta and green lines indicate the hazard ratio corresponding to the margin for clinical (1/1.3 = 0.769) and statistical (1) superiority, respectively, comparing active medication to placebo. Trials are sorted by drug class (blue, glucagon-like peptide-1 receptor agonists; orange, sodium-glucose co-transporter-2 inhibitors) and alphabetically.

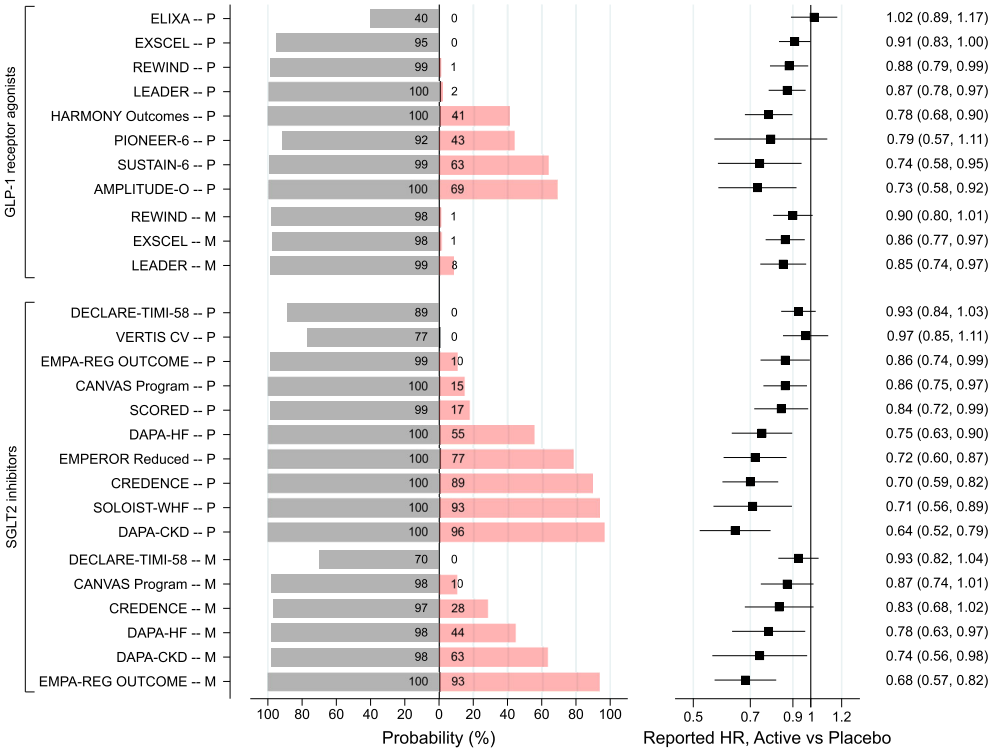

FIGURE 4 Statistical and clinical superiority of the active drug compared with placebo. Tornado plot showing the probability of statistical [grey; hazard ratio (HR) <1) and clinical (red; HR <0.769) superiority of the active drug compared with placebo; corresponding HRs reported in the original article are shown in the forest plot. Trials sorted by drug class, outcome (M, mortality; P, primary), and increasing probability of clinical superiority of the active drug. GLP-1, glucagon-like peptide-1; SGLT2i, sodium-glucose co-transporter-2.

the threshold used to declare the non-inferiority to placebo – corresponding to the probability for the active treatment of being clinically superior to placebo – varied significantly across trials (Figure 4): it ranged from 0% to 2% in ELIXA, EXSCEL, REWIND and LEADER; and from 41% to 69% in PIONEER-6, SUSTAIN-6 and AMPLITUDE-O. In the original reports, all trials showed superiority to placebo, except ELIXA and PIONEER-6 (Table 1 and Figure 4). Among the three trials with data on all-cause death, REWIND and EXSCEL showed a probability of an HR <1 of 98%, and LEADER of 99%; the corresponding probabilities of clinical superiority to placebo were 1%, 1% and 8%, respectively (Figure 4). REWIND and LEADER also showed superiority to placebo in the original reports (Table 1 and Figure 4).

For SGLT2is, the probability of any reduction in the rates of primary outcomes was lowest in VERTIS-CV (77%) followed by DECLARE-TIMI-58 (89%); for all other trials, it was ≥99%. Probabilities of clinical superiority varied more than those estimated for GLP-1RAs, ranging from 0% to 17% in DECLARE-TIMI-58, VERTIS-CV, EMPA-REG OUTCOME, CANVAS Program and SCORED; 55% to 77% in DAPA-HF and EMPEROR Reduced,

TABLE 1 Reported and estimated hazard ratios

Class, outcome	Randomized controlled trial	Cox PH regression	Bayesian modelling	Original report
GLP-1RAs				
Primary outcome	AMPLITUDE-O	0.73 (0.58, 0.91)	0.73 (0.58, 0.92)	0.73 (0.58, 0.92)
	ELIXA	1.02 (0.88, 1.17)	1.02 (0.89, 1.17)	1.02 (0.89, 1.17)
	EXSCEL	0.92 (0.84, 1.02)	0.92 (0.84, 1.02)	0.91 (0.83, 1.00)
	HARMONY outcomes	0.78 (0.68, 0.90)	0.78 (0.68, 0.90)	0.78 (0.68, 0.90)
	LEADER	0.86 (0.78, 0.96)	0.87 (0.78, 0.97)	0.87 (0.78, 0.97)
	PIONEER-6	0.79 (0.57, 1.11)	0.80 (0.56, 1.10)	0.79 (0.57, 1.11)
	REWIND	0.88 (0.79, 0.99)	0.88 (0.79, 0.99)	0.88 (0.79, 0.99)
	SUSTAIN-6	0.74 (0.57, 0.95)	0.74 (0.57, 0.94)	0.74 (0.58, 0.95)
All-cause mortality	EXSCEL	0.88 (0.78, 1.00)	0.89 (0.79, 1.00)	0.86 (0.77, 0.97)
	LEADER	0.85 (0.74, 0.97)	0.85 (0.74, 0.97)	0.85 (0.74, 0.97)
	REWIND	0.88 (0.79, 1.00)	0.89 (0.79, 0.99)	0.90 (0.80, 1.01)
SGLT2is				
Primary outcome	CANVAS program	0.82 (0.72, 0.92)	0.82 (0.73, 0.93)	0.86 (0.75, 0.97)
	CREDENCE	0.69 (0.59, 0.82)	0.70 (0.59, 0.82)	0.70 (0.59, 0.82)
	DAPA-CKD	0.64 (0.52, 0.79)	0.64 (0.52, 0.79)	0.64 (0.52, 0.79)
	DAPA-HF	0.76 (0.64, 0.91)	0.76 (0.63, 0.91)	0.75 (0.63, 0.90)
	DECLARE-TIMI-58	0.94 (0.85, 1.04)	0.94 (0.84, 1.04)	0.93 (0.84, 1.03)
	EMPA-REG OUTCOME	0.85 (0.73, 0.98)	0.85 (0.73, 0.98)	0.86 (0.74, 0.99)
	EMPEROR reduced	0.72 (0.59, 0.86)	0.72 (0.59, 0.86)	0.72 (0.60, 0.87)
	SCORED	0.83 (0.71, 0.97)	0.83 (0.71, 0.97)	0.84 (0.72, 0.99)
	SOLOIST-WHF	0.66 (0.53, 0.81)	0.66 (0.53, 0.81)	0.71 (0.56, 0.89)
	VERTIS CV	0.95 (0.83, 1.08)	0.95 (0.83, 1.09)	0.97 (0.85, 1.11)
All-cause mortality	CANVAS program	0.85 (0.73, 0.99)	0.85 (0.73, 0.99)	0.87 (0.74, 1.01)
	CREDENCE	0.82 (0.67, 1.00)	0.82 (0.66, 1.01)	0.83 (0.68, 1.02)
	DAPA-CKD	0.74 (0.55, 0.98)	0.74 (0.54, 0.98)	0.74 (0.56, 0.98)
	DAPA-HF	0.78 (0.63, 0.98)	0.79 (0.63, 0.98)	0.78 (0.63, 0.97)
	DECLARE-TIMI-58	0.97 (0.86, 1.10)	0.97 (0.85, 1.10)	0.93 (0.82, 1.04)
	EMPA-REG OUTCOME	0.67 (0.55, 0.80)	0.67 (0.56, 0.81)	0.68 (0.57, 0.82)

Note: Within each drug class, trials are sorted by outcome and alphabetically.
Cox PH and Bayesian modelling based on individual-level data extracted from the original study publication (Kaplan-Meier curve). The Bayesian estimation is mean and 95% credible interval of the posterior distribution.
GLP-1RA, glucagon-like peptide receptor agonists; PH, proportional hazards; SGLT2is, sodium-glucose co-transporter-2 inhibitors.

respectively; and ≥89% in CREDENCE, SOLOIST-WHF and DAPA-CKD (Figure 4). All trials, except VERTIS-CV, showed superiority to placebo in the original reports (Table 1 and Figure 4). Six trials yielded data on all-cause mortality; the probability of an HR <1 was 70% in DECLARE-TIMI-58 and ≥97% in CREDENCE, CANVAS Program, DAPA-CKD, DAPA-HF and EMPA-REG OUTCOME; the corresponding probabilities of clinical superiority were: 0% in DECLARE-TIMI-58; 10% in CANVAS Program; 28% in CREDENCE; 44% in DAPA-HF; 63% in DAPA-CKD; and 93% in EMPA-REG OUTCOME (Figure 4). Superiority to placebo was shown in the original report of DAPA-CKD, DAPA-HF and EMPA-REG OUTCOME (Table 1 and Figure 4).

The probability of reducing the rates of the primary outcome and mortality, across all possible thresholds of HR, is shown in Figure S3.

3.3 | Quality of data extraction

HRs estimated with the Cox regression and the Bayesian modelling, both using data extracted from Kaplan-Meier curves, are shown in Table 1 and Figure S4. Their comparison indicated a virtually complete overlap in each trial, as expected using weakly informative priors. The comparison between the Cox and the original report showed a high quality of data extraction, as the two estimates were almost identical across trials: the smallest and largest ratios between the Cox and the reported HR were 0.93 (primary outcome in the SOLOIST-WHF trial) and 1.04 (all-cause mortality in the DECLARE-TIMI-58 trial; Figure S4); for the same studies, there were also the largest and smallest differences between the Cox and reported HR: −0.05 (0.66 vs. 0.71) and +0.04 (0.97 vs. 0.93), respectively (Table 1).

4 | DISCUSSION

Using data from trials investigating GLP-1RAs and SGLT2is, we estimated the treatment effect with a Bayesian survival model and quantified the probability for the active treatment of being not only statistically but also clinically superior to placebo, defined using the same threshold suggested by the FDA to claim non-inferiority. While most trials showed 'superiority' of the active treatment in the original published report, the probability of clinical superiority varied significantly: in studies with GLP-1RAs, from a minimum of 0% to a maximum of 69% for the primary outcome and from 0% to 8% for mortality; in those with SGLT2is, from 0% to 96% for the primary outcome and from 0% to 93% for mortality, although probabilities were on average greater for SGLT2is than for GLP-1RAs.

Virtually all trials investigating the efficacy or safety of medical products, devices, or strategies have been designed within a frequentist approach. Frequentist methods have several limitations, including the impossibility of interpreting the treatment effect in terms of probability statements and the complexities of testing more than one primary hypothesis (the null hypothesis of no difference vs. the alternative hypothesis of a difference).[8] Notwithstanding efforts to discriminate statistically from clinically significant results and avoid the equivalence between no evidence and evidence of no effect,[15] trial results are still commonly interpreted, defined, popularized and evoked in a dichotomous way: either 'positive' or 'negative' (the trial 'showed' or the trial 'failed to show' are other common definitions).[16] In most, if not all, cases, the criteria to separate these two worlds are a specific *P*-value (almost universally $P < 0.05$) or an upper bound of the confidence interval lower than 1. Besides the ease of recalling the results, there are no other advantages of such a binary interpretation while its negative scientific implications are well-known.[17] Particularly in the context of trials for drug approval, estimating a treatment effect is a matter of quantifying a magnitude rather than testing a hypothesis: the latter approach would indeed discard a hypothetical treatment showing an HR of 0.68 (95% CI: 0.46-1.01) in favour of another showing an HR of 0.95 (95% CI: 0.92-0.98), yet the magnitude of the effect is potentially much larger in the former study. Using a real example from our study, the reported HR for the primary outcome in PIONEER-6 was 0.79 (95% CI: 0.57-1.11), indicating non-inferiority but not superiority, while in HARMONY Outcomes it was 0.78 (95% CI: 0.68-0.90), indicating superiority; however, the probabilities of being clinically superior to the placebo were 43% and 41% and of resulting in any reduction (i.e. HR <1) 92% and 100%, respectively; these figures indicate a much greater similarity between the two treatment effects than what is suggested by the dichotomous interpretation of the results.

A Bayesian approach helps to circumvent some limitations of the frequentist, hypothesis-testing approach.[6-9] Rather than classifying the results in two alternatives, Bayesian estimates encompass the continuous spectrum of the magnitude of effect and permit a more nuanced and natural, probabilistic interpretation of the results. As an example, the 95% credible interval represents the probability within which lies, with a 95% probability, the treatment effect; such an interpretation is not possible when considering the frequentist 95% confidence interval. Estimating the probability of a treatment effect across varying thresholds gives deeper insights into the treatment effect and better aligns with a context-specific, personalized approach to glucose-lowering therapies, if compared with the results obtained within the frequentist framework. Thresholds, for example, may differ between patients and change in the same patient over time: instead of a 23% reduction (HR 0.77), a smaller reduction in the rate of MACE may be also considered clinically meaningful. Moreover, although several factors are simultaneously considered, label indications from regulatory agencies and clinical recommendations on glucose-lowering medications in subjects with type 2 diabetes are mainly based on the available evidence centred on dichotomous interpretations of the results.[18,19] The large differences in the clinical superiority observed within and between GLP-1RAs and SGLT2is across the investigated outcomes underline the limitations of the current decisional frameworks leading to their approval, call into question the within-class equivalence in most of the current clinical guidelines, and may better inform future indications, clinical recommendations and therapeutic decisions.

Notably, the estimated probabilities of clinical superiority for the primary outcomes were consistently greater in trials with SGLT2is investigating kidney or heart failure compared with atherothrombotic outcomes; this was true also for mortality, with the exception of the EMPA-REG OUTCOME trial, which showed the clinical superiority of empagliflozin. Probabilities of clinical superiority for GLP-1RAs were generally lower or much lower: except in HARMONY Outcomes, PIONEER-6, SUSTAIN-6 and AMPLITUDE-O, the highest probability in all remaining trials was only 8%, for both the primary outcome and mortality. While current guidance acknowledges the different target population and the divergent effects of the GLP-1RAs and SGLT2is,[18] reflecting their distinct pharmacodynamics,[1] our findings robustly show also clinically relevant differences among GLP-1RAs, with albiglutide, efpeglenatide and semaglutide more likely to confer clinically significant cardiovascular benefits than other GLP-1RAs. The heterogeneous results across trials, however, are related not only to the pharmacological differences between and within GLP-1RAs and SGLT2is but may also reflect differences in the designs and characteristics of the subjects included in the trials. This is particularly important when trial-specific results are interpreted on the absolute risk reduction (rather than HR) scale, as the risk of outcomes varied across trials. Moreover, while initial RCTs with SGLT2is mainly investigated MACE, more recent ones – based on the primary positive results – have expanded or changed the outcomes, focusing prevalently on kidney and heart failure events.

A Bayesian approach requires the inclusion in the statistical model of a prior probability for the treatment effect, which represents the belief about the distribution of its possible values before data collection; this information is then combined with the treatment effect observed in trial data (likelihood) to obtain the final results (posterior distributions).[14] Previous information may be completely subjective or derived from previous evidence, for example from single trials or meta-analyses. Incorporating prior information has been considered a

main limitation of the Bayesian methods, as the results may be largely influenced by the prior.[20] In line with the main goal of this analysis, we used weakly informative priors, which have a small impact compared with the trial data on the final HR estimate, as confirmed by the virtually identical Bayesian and reported HR estimates.

Differences between the HR calculated with the frequentist Cox regression or the Bayesian survival analysis from the extracted data and the HR reported in the original study were somewhat higher in two trials: SOLOIST-WHF and DECLARE-TIMI-58. As in both studies the Cox and Bayesian estimates were identical, the process of data extraction rather than the Bayesian statistical modelling may have contributed to the small discrepancies (further details are reported in Appendix S1).

To date, the uptake of Bayesian methods in clinical trials has been limited. This is partly related to the limited availability of software to conduct Bayesian analysis and partly to the apparently less applicable nature of the Bayesian results compared with the binary frequentist interpretation (evidence that an intervention 'improves' vs. 'does not improve') and the necessary nature of a health care label decision (yes/no).[17] However, a Bayesian approach to the analysis of clinical trials is not the solution to all (frequentist) problems and frequentist metrics may complement Bayesian estimates. One direction is to focus on the absolute effect, such as comparison of absolute event rates in the competing arms. This provides simple but powerful information, as a very large reduction (e.g. HR 0.4) results in a very small absolute rate difference if the outcome is rare. A further metric is the restricted mean survival time, which is the time of the outcome postponement because of the treatment during a specified time interval:[21] for non-inferiority trials with low event rates, short duration, or large non-inferiority margin, restricted mean survival time has better power than the common proportional hazard method.[22,23] Another direction is to move beyond single estimates or to derive metrics from the *p*-value to enhance the interpretation of results and enable comparisons among trials.[24] In this vein, the *p*-value function yields the magnitude of the effect across different *p*-values,[25] and the resulting curve looks similar to, but is conceptually very different from, the Bayesian posterior distribution.[26] Useful metrics derived from the *p*-value include the S-value[25] and the counter-null value.[27] All these metrics have been retrospectively estimated in CVOTs with GLP-1RAs and SGLT2is,[21,24,28] showing, for example, how small treatment effects from these trials are likely to be on the absolute scale (with most drugs postponing the outcome by only a few days over the course of the trial period), and how similar some of the treatment effects are that straddle the threshold for statistical significance. While such studies caution against the over-reliance on single metrics to estimate or interpret treatment effects, future trials should proactively report this complementary information and more frequently consider Bayesian approaches to treatment effect estimate.[5,6,9]

The results of randomized controlled trials investigating the safety and efficacy of GLP-1RAs and SGLT2is have led to the approval of several therapies and to an unprecedented growth in the number of available treatments for type 2 diabetes; directly or indirectly, the 2008 FDA regulatory requirement has been a strong influence on the design and implementation of such trials. Further consideration, however, is required if the goal is to claim not only non-inferiority of these medications but also their clinical superiority, as the recommended characteristics and current interpretation of the non-inferiority trials bias the comparison in favour of the medication under investigation for approval. Using the margin of 1.3 suggested by the FDA to declare non-inferiority, we showed that the evidence for clinical superiority is modest or null for some treatments and more robust for others, with remarkable differences across outcomes as well as between and within GLP-1RAs and SGLT2is. In view of the drawbacks that have so far characterized the design and interpretation of non-inferiority trials, changes in the recommendations on their design and complementary analytical approaches may help generate more informative and actionable evidence for health care professionals and patients.[29-31]

AUTHOR CONTRIBUTIONS

Study idea and design was performed by FZ. DEK and FZ performed the literature search. DEK prepared the data. FZ analysed the data and wrote the first draft. All the authors critically revised the study and the manuscript draft.

ACKNOWLEDGMENTS

FZ, MJD and KK acknowledge the National Institute for Health and Care Research (NIHR) - Collaboration for Leadership in Applied Health Research and Care (ARC) East Midlands and the NIHR Leicester Biomedical Research Centre. The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

CONFLICT OF INTEREST

FZ has received honoraria for speaking at meetings from NAPP Pharmaceuticals and Boehringer Ingelheim. DEK declares that he has no competing interests. KK has acted as a consultant and speaker for Novartis, Novo Nordisk, Sanofi-Aventis, Lilly and Merck Sharp & Dohme; has received grants in support of investigator and investigator-initiated trials from Novartis, Novo Nordisk, Sanofi-Aventis, Lilly, Pfizer, Boehringer Ingelheim and Merck Sharp & Dohme; and has received funds for research and served on advisory boards for Lilly, Sanofi-Aventis, Merck Sharp & Dohme and Novo Nordisk. MJD has acted as consultant, advisory board member and speaker for Novo Nordisk, Sanofi-Aventis, Lilly, Merck Sharp & Dohme, Boehringer Ingelheim, AstraZeneca and Janssen, as advisory board member for Servier and as a speaker for Mitsubishi Tanabe Pharma Corporation and Takeda Pharmaceuticals International Inc. She has received grants in support of investigator and investigator-initiated trials from Novo Nordisk, Sanofi-Aventis, Lilly, Boehringer Ingelheim and Janssen.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/dom.14735.

DATA AVAILABILITY STATEMENT

Kaplan-Meier plots to extract individual level data are available in the trial publication. The analytical codes are available on request from the corresponding author (FZ).

ORCID

Francesco Zaccardi https://orcid.org/0000-0002-2636-6487
David E. Kloecker https://orcid.org/0000-0002-8910-2091
Kamlesh Khunti https://orcid.org/0000-0003-2343-7099
Melanie J. Davies https://orcid.org/0000-0002-9987-9371

TWITTER

Francesco Zaccardi @LRWEUnit @LDC_tweets

REFERENCES

1. Brown E, Heerspink HJL, Cuthbertson DJ, Wilding JPH. SGLT2 inhibitors and GLP-1 receptor agonists: established and emerging indications. *Lancet*. 2021;398(10296):262-276.
2. Food and Drug Administration Center for Drug Evaluation and Research. Guidance for industry: diabetes mellitus - Evaluating cardiovascular risk in new antidiabetic therapies to treat type 2 diabetes. https://www.fda.gov/media/71297/download. Accessed on 11 Jan 2022.
3. Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *N Engl J Med*. 2007;356(24):2457-2471.
4. Ganju J, Rom D. Non-inferiority versus superiority drug claims: the (not so) subtle distinction. *Trials*. 2017;18(1):278.
5. Wijeysundera DN, Austin PC, Hux JE, Beattie WS, Laupacis A. Bayesian statistical inference enhances the interpretation of contemporary randomized controlled trials. *J Clin Epidemiol*. 2009;62(1):13-21 e15.
6. Zampieri FG, Casey JD, Shankar-Hari M, Harrell FE Jr, Harhay MO. Using Bayesian methods to augment the interpretation of critical care trials. An overview of theory and example reanalysis of the alveolar recruitment for acute respiratory distress syndrome trial. *Am J Respir Crit Care Med*. 2021;203(5):543-552.
7. Harrell FE. Why Bayes for Clinical Trials? http://hbiostat.org/doc/bayes/why.pdf. Accessed on 11 Jan 2022.
8. Harrell FE. Introduction to Bayes for Evaluating Treatments. https://hbiostat.org/doc/bayes/course.html#content. Accessed on 11 Jan 2022.
9. Yarnell CJ, Abrams D, Baldwin MR, et al. Clinical trials in critical care: can a Bayesian approach enhance clinical and scientific decision making? *Lancet Respir Med*. 2021;9(2):207-216.
10. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med*. 2009;6(7):e1000097.
11. Higgins JP, Altman DG, Gotzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*. 2011;343:d5928.
12. Wei Y, Royston P. Reconstructing time-to-event data from published Kaplan-Meier curves. *Stata J*. 2017;17(4):786-802.
13. Brilleman SL, Elci EM, Novik JB, Wolfe R. Bayesian survival analysis using the rstanarm R package. *arXiv preprint arXiv:200209633*. 2020.
14. van de Schoot R, Depaoli S, King R, et al. Bayesian statistics and modelling. *Nature Reviews Methods Primers*. 2021;1(1):1.
15. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ*. 1995;311(7003):485.
16. McShane BB, Gal D. Statistical significance and the dichotomization of evidence. *J Am Stat Assoc*. 2017;112(519):885-895.
17. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "p < 0.05". *The American Statistician*. 2019;73(sup1):1-19.
18. Buse JB, Wexler DJ, Tsapas A, et al. 2019 update to: Management of Hyperglycemia in type 2 diabetes, 2018. A consensus report by the American Diabetes Association (ADA) and the European Association for the Study of diabetes (EASD). *Diabetes Care*. 2020;43(2):487-493.
19. American DA. 10. Cardiovascular disease and risk management: standards of medical Care in Diabetes-2020. *Diabetes Care*. 2020;43-(Suppl 1):S111-S134.
20. Vallverdú J. *Bayesians versus frequentists: a philosophical debate on statistical reasoning*. 1st ed. Berlin, Heidelberg: Springer Berlin Heidelberg; 2016.
21. Kloecker DE, Davies MJ, Khunti K, Zaccardi F. Uses and limitations of the restricted mean survival time: illustrative examples from cardiovascular outcomes and mortality trials in type 2 diabetes. *Ann Intern Med*. 2020;172(8):541-552.
22. Freidlin B, Hu C, Korn EL. Are restricted mean survival time methods especially useful for noninferiority trials? *Clin Trials*. 2021;18(2):188-196.
23. Quartagno M, Morris TP, White IR. Why restricted mean survival time methods are especially useful for non-inferiority trials. *Clin Trials*. 2021;18(6):743-745.
24. Kloecker DE, Davies MJ, Khunti K, Zaccardi F. Cardiovascular effects of sodium-glucose co-transporter-2 inhibitors and glucagon-like peptide-1 receptor agonists: the P value and beyond. *Diabetes Obes Metab*. 2021;23(7):1685-1691.
25. Rafi Z, Greenland S. Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. *BMC Med Res Methodol*. 2020;20(1):244.
26. Rafi Z. Comparison to Bayesian Posterior Distributions. Available at: https://data.lesslikely.com/concurve/articles/bayes.html. Accessed on 19 Oct 2021.
27. Rosenthal R, Rubin DB. The Counternull value of an effect size - a new statistic. *Psychol Sci*. 1994;5(6):329-334.
28. Ferrannini E, Rosenstock J. Clinical translation of cardiovascular outcome trials in type 2 diabetes: is there more or is there less than meets the eye? *Diabetes Care*. 2021;44(3):641-646.
29. Sharma A, Pagidipati NJ, Califf RM, et al. Impact of regulatory guidance on evaluating cardiovascular risk of new glucose-lowering therapies to treat type 2 diabetes mellitus: lessons learned and future directions. *Circulation*. 2020;141(10):843-862.
30. Food and Drug Administration Center for Drug Evaluation and Research. Guidance for industry: Type 2 Diabetes Mellitus: Evaluating the Safety of New Drugs for Improving Glycemic Control. https://www.fda.gov/media/135936/download. Accessed on 11 Jan 2022.
31. Khunti K, Davies MJ, Marx N, Buse JB. Draft FDA guidance for assessing the safety of glucose lowering therapies: a missed opportunity? *Lancet Diabetes Endocrinol*. 2020;8(10):810-811.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Zaccardi F, Kloecker DE, Khunti K, Davies MJ. Non-inferiority and clinical superiority of glucagon-like peptide-1 receptor agonists and sodium-glucose co-transporter-2 inhibitors: Systematic analysis of cardiorenal outcome trials in type 2 diabetes. *Diabetes Obes Metab*. 2022;24(8):1598-1606. doi:10.1111/dom.14735