Cardiovascular effects of sodium-glucose co-transporter-2 inhibitors and glucagon-like peptide-1 receptor agonists: The P value and beyond

David E. Kloecker MPhil1,2 | Melanie J. Davies MD1,3 | Kamlesh Khunti PhD1,2 | Francesco Zaccardi PhD1,2

1Diabetes Research Centre, Leicester Diabetes Centre, University of Leicester, Leicester, UK
2Leicester Real World Evidence Unit, Leicester Diabetes Centre, University of Leicester, Leicester, UK
3NIHR Leicester Biomedical Research Centre, Leicester General Hospital, Leicester, UK

Correspondence
David E. Kloecker, Leicester Real World Evidence Unit, Leicester Diabetes Centre, Leicester General Hospital, Gwendolen Road, Leicester LE5 4PW, UK.
Email: dk261@leicester.ac.uk; davidekloecker@gmail.com

Funding information
National Institute for Health Research Applied Research Collaboration East Midlands;
National Institute for Health Research Leicester Biomedical Research Centre

Abstract
Despite growing awareness of the dangers of a dichotomous interpretation of trial results based on the ‘statistical significance’ of a treatment effect, the uptake of new approaches has been slow in diabetes medicine. We showcase a number of ways to interpret the evidence for a treatment effect applied to the cardiovascular outcome trials of glucagon-like peptide-1 receptor agonists (GLP-1RAs) and sodium-glucose co-transporter-2 inhibitors (SGLT-2is): the P value function (or confidence curves), which depicts the treatment effect across the whole spectrum of confidence levels; the counter null value, which is the hazard ratio (i.e. treatment effect size) supported by the same amount of evidence as the null value (i.e. no treatment effect); and the S value, which quantifies the strength of the evidence against the null hypothesis in terms of the number of coin tosses yielding the same side. We show how this approach identifies potential treatment effects, highlights similarities among trials straddling the threshold of statistical significance, and quantifies differences in the strength of the evidence from trials reporting statistically significant results. For example, while REWIND, CANVAS and CREDENCE failed to reach statistical significance at the .05 level for all-cause mortality, their counter null values indicate that reduced death rates by 19%, 24% and 31%, respectively, are supported by the same amount of evidence as that indicating no treatment effect. Moreover, similarities among results emerge in trials of GLP-1RAs (REWIND, EXSCEL and LEADER) lying closely around the threshold of ‘statistical significance’. Lastly, several S values, such as for the primary outcome in HARMONY Outcomes (S value 10.9) and all-cause death in EMPAREG-OUTCOME (S value 15.0), stand out compared with values for other outcomes and other trials, suggesting much larger differences in the evidence between these studies and several others that cluster around the .05 significance threshold.
effect size or importance of the result. In this vein, the scientific conclusions and business or policy decisions on whether the probability of the test hypothesis. Some of these problems may be instead lead one to consider the drugs as equivalent. Moreover, significance can divert attention away from salient differences in the data (at $P = 1$), with parameter values closer to this midpoint indicating increasing consistency with the data. Because $P$ values and confidence levels are related to each other (e.g. when one is 0, the other is 1), the tails of the confidence curves show not only $p$ for any variable value but also the limit for any confidence level. Of note, the confidence curve differs from related (confidence) distributions, including the cumulative distribution function (probability of the parameter taking a value that is less than or equal to a given value), the probability density function (probability that a continuous random parameter falls within a range of values) and the probability mass function (probability that a discrete parameter is equal to a particular value).

Second, the $P$ value function reveals the counternull value, which represents the treatment effect size supported by the same amount of evidence as the null value of no treatment effect.

Lastly, the $S$ value or surprisal value aims to overcome some of the problems associated with the interpretation of the $P$ value through a simple mathematical transformation, $s = -\log_2(p)$. Conceptually, an $S$ value translates an observed $P$ value $p$ into the probability of an outcome in a game of coin tossing, such that $p$ is the probability of the coin being unbiased given the result from a series of $s$ independent coin tosses that come out heads. Because the outcome of the coin tosses is always binary (i.e. heads or tails), the unit of the $S$ value is the binary digit (bit). As the probability of an unbiased coin coming out heads in a series of coin tosses is $.5^s$, the more heads we toss in a row (i.e. greater the $S$ value $s$), the smaller is the probability $p$ that the coin is unbiased, and, in the context of an analysis, the more bits of information we have against the test hypothesis. While the logic behind such a transformation goes back to the origin of information theory, its application to the interpretation of (medical) studies is more novel. The $S$ value helps to overcome at least two major problems with the $P$ value: first, the $P$ value lies on an inverse exponential scale, where absolute differences between points cannot be interpreted meaningfully, because differences between smaller $P$ values do not yield the same evidence as the same absolute

**KEYWORDS**
- cardiovascular disease, GLP-1, GLP-1 analogue, SGLT-2 inhibitor, type 2 diabetes

**INTRODUCTION**

The cardiovascular efficacy and safety of novel medications for individuals with type 2 diabetes (T2D) are often interpreted through the lens of non-inferiority and superiority to placebo measured at a single alpha level, generally .05, and the corresponding 95% confidence interval, as part of a randomized controlled trial (RCT) mandated by regulatory agencies such as the US Food and Drug Administration (FDA). Reporting $P$ values and interpreting them according to a single threshold for statistical significance has advantages.

However, the uncritical use of $P$ values comes with a range of well-known disadvantages. For one, it can lead to a binary understanding of a drug’s efficacy: either it is effective, or it is not. Such thinking may take place at the level of a consultation, the association creating a clinical guideline or that of a licensing body, and may have profound clinical implications. For example, when comparing two drugs, one which is narrowly ‘statistically significant’ and another which is narrowly not, one might dismiss the latter drug outright when its effects may be very similar, and other considerations, such as possible side effects or costs, may carry greater weight in clinical decision-making. Conversely, where two drugs have proven efficacy in lowering the risk of an outcome, a binary threshold for statistical significance can divert attention away from salient differences in the strength of their evidence bases or magnitude of their effects, and instead lead one to consider the drugs as equivalent. Moreover, $P$ values are also easily confused with other statistics, such as the probability of the test hypothesis. Some of these problems may be related to the pressure to eliminate uncertainty and publish ‘statistically significant’ results and, at a more technical level, to the statistical approach of null hypothesis significance testing.

Consequently, statisticians have called for more cautious use of the $P$ value (and indeed confidence intervals), counselling against basing scientific conclusions and business or policy decisions on whether a $P$ value passes a specific threshold and encouraging researchers to refrain from selective reporting of analyses, while interpreting $P$ values in the context of the model or hypothesis and measures of effect size or importance of the result. In this vein, the $P$ value may best be seen as a continuous measure of compatibility of data with the test model and hypothesis, and used alongside other measures to aid interpretation of research findings, several of which we will consider in the current study.

To move away from the danger of misinterpreting the observed $P$ value $p$, we can examine results across the whole spectrum of $P$ values (or confidence levels) via $P$ value functions. This can yield more nuanced conclusions about a drug’s efficacy. These so-called confidence curves have been described in the literature and their use has been promoted since at least the 1960s, but uptake by clinicians has been slow, hampered perhaps by a lack of convenient statistical packages. At their heart, $P$ value functions yield graphs depicting how compatible different parameter values are with the data behind the background assumptions. Because the degree of compatibility is expressed by the $P$ value, which ranges from 0 to 1, the peak of the confidence curve, which is also the middle of a symmetric curve, shows the parameter value most compatible with the data (at $P = 1$), while parameter values closer to this midpoint indicating increasing consistency with the data. Because $P$ values and confidence levels are related to each other (e.g. when one is 0, the other is 1), the tails of the confidence curves show not only $p$ for any variable value but also the limit for any confidence level. Of note, the confidence curve differs from related (confidence) distributions, including the cumulative distribution function (probability of the parameter taking a value that is less than or equal to a given value), the probability density function (probability that a continuous random parameter falls within a range of values) and the probability mass function (probability that a discrete parameter is equal to a particular value).

Second, the $P$ value function reveals the counternull value, which represents the treatment effect size supported by the same amount of evidence as the null value of no treatment effect. Lastly, the $S$ value or surprisal value aims to overcome some of the problems associated with the interpretation of the $P$ value through a simple mathematical transformation, $s = -\log_2(p)$. Conceptually, an $S$ value translates an observed $P$ value $p$ into the probability of an outcome in a game of coin tossing, such that $p$ is the probability of the coin being unbiased given the result from a series of $s$ independent coin tosses that come out heads. Because the outcome of the coin tosses is always binary (i.e. heads or tails), the unit of the $S$ value is the binary digit (bit). As the probability of an unbiased coin coming out heads in a series of coin tosses is $.5^s$, the more heads we toss in a row (i.e. greater the $S$ value $s$), the smaller is the probability $p$ that the coin is unbiased, and, in the context of an analysis, the more bits of information we have against the test hypothesis. While the logic behind such a transformation goes back to the origin of information theory, its application to the interpretation of (medical) studies is more novel. The $S$ value helps to overcome at least two major problems with the $P$ value: first, the $P$ value lies on an inverse exponential scale, where absolute differences between points cannot be interpreted meaningfully, because differences between smaller $P$ values do not yield the same evidence as the same absolute
differences between larger $P$ values; by contrast, the $S$ value lies on an equidistant scale, so that bits of information are additive and absolute differences between $S$ values carry the same meanings. This can make $S$ values useful for comparing the strength of the evidence for competing treatment effects across the spectrum of $P$ values. Second, we saw that $P$ values can be misinterpreted easily and are often conflated with Bayesian posterior probabilities; because $S$ values do not look like probabilities, they are less susceptible to such misuse.

Figure 1 illustrates these concepts using two example trials. $A$ depicts confidence curves and countrernull values; the cusp of each curve corresponds to the estimate of the hazard ratio (HR) comparing treatment and placebo. In Trial 1 (PIONEER-6, primary outcome), the HR with 95\% confidence interval is 0.79 (0.57, 1.10); the right arm of the confidence curve reaches 1 at an alpha level (i.e. $P$ value) of .166.\textsuperscript{24} The left arm has a corresponding estimate for this $P$ value, which is the countrernull value (0.62). For Trial 2 (EXSCEL, primary outcome), the HR is 0.91 (0.83, 1.00); the right arm of the confidence curve reaches 1 at a $P$ value of .047, and the countrernull value is 0.83.\textsuperscript{25} If we solely relied on a binary interpretation of the $P$ value to judge the efficacy of the two drugs, we would flatly dismiss the drug in Trial 1 and accept the drug in Trial 2. How then do our new measures help us to improve our interpretation? First, the broad shape of the confidence curve for Trial 1 suggests how imprecise the effect estimates are likely to be and, in light of the considerable area of the curve that falls below a HR of 1, cautious.

against a conclusion of no effect (compared with placebo). The other measures reinforce this idea. Because the countrernull and the null value lie at the same alpha/confidence levels, the countrernull value can be said to be the estimate that is supported by the same amount of evidence as the null value of no treatment effect. For Trial 1, we could say that there is as much evidence for a HR of 0.62 (a considerable treatment effect) as for a HR of 1 (no treatment effect relative to the comparator). $B$ depicts the curvilinear relationship between the $P$ value and the $S$ value: for Trial 1, $p = .166$ equates to $s = 2.6$; for Trial 2, $p = .047$ equates to $s = 4.4$. If we wanted to compare the trials, the inverse exponential scale of the $P$ value would make this task very difficult. By contrast, the equidistant scale of the $S$ value readily allows us to compare how many bits of information each study provides against the test hypothesis: Trial 2 gives almost two additional bits of information against the null (i.e. difference in $S$ values: 4.4–2.6 = 1.8) compared with Trial 1.

In the current study, we use these measures to investigate the treatment effects of glucagon-like peptide-1 receptor agonists (GLP-1RAs) and sodium-glucose co-transporter-2 inhibitors (SGLT-2is) on cardiovascular outcomes in adult individuals with T2D; we focus on these two classes of medications because many of their members are currently recommended in subjects with T2D at high or very high cardiovascular disease risk for their proven benefit in reducing cardio-renal events. We aim to show that a focus on single thresholds of statistical significance fails to identify some potentially salient

---

**Figure 1** Interpretation of $P$ value function, countrernull value and $S$ value. $A$ shows the confidence curves and countrernull values for the primary outcome of two trials, Trial 1 (PIONEER-6, dark grey) and Trial 2 (EXSCEL, light grey). The hazard ratio is 0.79 for Trial 1 and 0.91 for Trial 2; these estimates correspond to the cusp of the $P$ value function graph. The dotted horizontal light grey line with arrows indicates the 90\% confidence interval (right y-axis) for Trial 2, which is (0.84, 0.98); therefore, from the curves, confidence intervals can be obtained at any alpha level (left y-axis). The countrernull value is shown as an empty dot and corresponds, in the $P$ value function graph, to the hazard ratio with the same $P$ value at hazard ratio 1 (full dot); in this example, for Trial 1, the $P$ value at hazard ratio 1 is .166 (full dark grey dot) and the corresponding hazard ratio for this $P$ value is 0.62 (empty dark grey dot); for Trial 2, the $P$ value is .047 and the countrernull value is 0.83 (empty light grey dot). Therefore, the countrernull value corresponds to the treatment effect size supported by the same amount of evidence as the null value (i.e. hazard ratio 1, no treatment effect relative to placebo). Note that the x-axis is on the log scale. $B$ shows the relationship between the $P$ value and the $S$ value [$S = -\log_2(P)$]. For Trial 1, the $P$ value at hazard ratio 1 is .166 (y-axis) and the corresponding $S$ value = $-\log_2(0.166) = 2.6$ (x-axis); for Trial 2, the $P$ value at hazard ratio 1 is .047 and the corresponding $S$ value is 4.4. Therefore, the strength of the evidence against the null hypothesis (i.e. hazard ratio 1, no treatment effect relative to placebo) can be expressed in terms of tossing a fair coin and recording 2.6 heads in a row for Trial 1 and 4.4 heads in a row for Trial 2. This indicates that Trial 2, compared with Trial 1, has 1.8 ($S$ value difference: 4.4–2.6 = 1.8) extra bits of information, corresponding to the information provided by 1.8 additional heads in a row. The dashed vertical line shows the $S$ value at 4.322, which corresponds to the canonical $p = .05$. 

---
treatment effects and overlooks important similarities and differences in the results of these trials.

2 | METHODS

2.1 | Study selection

We included RCTs in adults with T2D randomized either to treatment with GLP-1RAs or SGLT-2is or to placebo, with a primary outcome pertaining to major adverse cardiovascular events (MACE) and, when available, all-cause mortality. RCTs were searched using terms including the generic name of the pharmacological agent in PubMed and Cochrane CENTRAL until 17 October 2020, with scanning of reference lists from retrieved articles and additional hand searches. The PRISMA checklist is reported in the supporting information and the search strategy is detailed in Figure S1.

2.2 | Data extraction and analysis

For each RCT, data were extracted on: acronym of the trial; pharmacological agent; baseline characteristics of participants; duration; definition and numbers of primary outcome events and deaths; and the corresponding measures of treatment effect (i.e. HR with confidence intervals). Risk of bias was assessed using the Cochrane risk-of-bias tool and disagreements were resolved by consensus or arbitration. We used the R package concurve to obtain, for each RCT: (a) the confidence curve, plotted using estimates of the HR for 100,000 alpha levels between 0 and 1, and the corresponding confidence levels (1 - α x 100); (b) the counternull value; and (c) the S value.4,26 We also estimated the overlap of the areas under the confidence curves of two trials. Figures were created in Stata v. 16.0 and modified in Inkscape 0.92.3.

3 | RESULTS

The search identified 7386 results, leading to the inclusion of 12 RCTs (seven trials with GLP-1RAs and five with SGLT-2is; Figure S1). Baseline characteristics, risk of bias, definitions of MACE and references from the included studies are reported in Table S1; the risk of bias was deemed to be low for all RCTs.

Confidence curves for each drug class and outcome are shown in Figure S2. The HR of the treatment effect (the cusp of the confidence curve) is shown alongside its lower and upper confidence interval for

![FIGURE 2](image-url) Hazard ratios (HRs), P values, S values and counternull values in the included trials. For a P value of .05, the corresponding S value is 4.322 (dashed vertical line); HRs and 95% confidence intervals (CIs) are taken from the published data for each trial. GLP-1RA, glucagon-like peptide-1 receptor agonist; SGLT2i, sodium-glucose co-transporter-2 inhibitor
various confidence levels and $P$ values. Complementary estimates at selected $P$ values are shown in Table S2, and measures of the degree of overlap between curves in Table S3. HRs with 95% confidence intervals, counternull and $S$ values are reported in Figure 2.

As an example, we consider the results on all-cause death in REWIND: the HR for the active treatment on the risk of death was $0.90$ (95% confidence interval: $0.80$, $1.01$; $p = .076$), as shown in Figure 2. Table S2 displays selected confidence intervals for different $P$ values to better illustrate what the confidence curve in Figure S2 shows; for example, the confidence interval at $p = .1$ is $0.82–0.99$ and thus falls below the HR of null effect relative to the placebo, emphasizing that the trial only narrowly failed to reach the conventional significance threshold. Moreover, the corresponding counternull for the outcome is HR $0.81$ (Figure 2); therefore, the two statements of no treatment effect compared with placebo and a 19% reduction in the hazard of death are supported by the same amount of evidence, that is, stating that there is no evidence of a reduced hazard of mortality is no more correct than stating a 19% reduction in the hazard of death. Lastly, Figure S2 and Table S3 show the overlap between confidence curves of REWIND and the other GLP-1 RAs with data for all-cause death: there is considerable overlap of 41.9% and 37.4% with EXSCEL and LEADER, respectively, suggesting some similarity in their results. This can be quantified further by looking at $S$ values, which range from around four bits of information against the null hypothesis for REWIND to nearly seven bits for EXSCEL (Figure 2).

4 | DISCUSSION

In this study, we analysed the effects of SGLT-2is and GLP-1RAs using readily obtained confidence curves, counternull values and $S$ values. The results highlight and help to resolve three important criticisms of the use of single significance thresholds to determine the efficacy (or safety) of a medication in a patient population and qualify the strength of the evidence for these agents.

First, some drugs may have clinically important effects in reducing the risk of an outcome that judgement on the basis of a single significance level at $.05$ can overlook. Reporting a confidence curve and examining counternull values can help mitigate this problem. In our analysis, estimates for all-cause death in REWIND, CANVAS and CREDENCE all reached statistical significance at the $.1$ level. Because their counternull values were $0.81$, $0.76$ and $0.69$, respectively, there was as much evidence for a null effect relative to the placebo as for reduced death rates by 19%, 24% and 31%, respectively.

Second, significance thresholds can inflate differences in the results of RCTs failing either side of the threshold, which are better quantified by the overlap among confidence curves and the $S$ values because of the rescaling of evidence against the null hypothesis from an inverse exponential scale ($P$ values) to an equidistant scale ($S$ values). The confidence curves of several GLP-1RAs (EXSCEL, LEADER, REWIND) showed considerable similarity for both MACE and all-cause death. Interestingly, the overlap between REWIND and EXSCEL was $41.9\%$ for all-cause death, but only EXSCEL and LEADER showed a ‘statistically significant’ reduction of mortality at the alpha level of $.05$, while EXSCEL more narrowly passed this threshold for MACE compared with REWIND (overlap $48.8\%$). Likewise, the $S$ values of these trials fell into a narrow range: EXSCEL, LEADER and REWIND had $S$ values between $4.4$ and $6.3$ against the null hypothesis of no benefit for MACE, and between $3.7$ and $6.6$ for all-cause death.

Third, differences in the strength of evidence among trials with ‘statistically significant’ effects are blurred by significance thresholds; as we have shown, $S$ values identify such differences more clearly. For example, for MACE, the results of EXSCEL ($S$ value $4.4$) were much closer to the non-significant finding from PIONEER 6 ($S$ value $2.6$) than to HARMONY Outcomes ($S$ value $10.9$), although both EXSCEL and HARMONY Outcomes yielded ‘statistically significant’ results at the $.05$ level while PIONEER 6 did not. Similarly, among SGLT-2is, the evidence for a reduction in all-cause death was much stronger for EMPAREG-OUTCOME ($S$ value $15.0$) than for the other trials ($S$ value range: $1.5$–$3.8$).

How should such results be interpreted and utilized? By themselves they are not sufficient for a comprehensive comparison of competing medications, which must consider all pertinent differences among trials. Instead, they shed light on the process by which evidence for efficacy might be evaluated in clinical guidelines and practice in place of single measures of the significance of treatment effects. This is especially salient in the case of the cardiovascular outcome trials for two reasons. First, these trials were largely designed and powered to show first non-inferiority (i.e., cardiovascular safety) and, in some cases, hierarchically superiority, creating an inherent problem of a ‘superiority’ interpretation of trials primarily designed for non-inferiority. Second, the FDA and current clinical guidelines on glucose-lowering agents in patients with T2D create label indications and clinical recommendations for the use of SGLT-2is or GLP-1RAs with proven cardiovascular benefits in line with a dichotomous interpretation of efficacy. To be sure, this process also includes an assessment of other relevant factors: for example, FDA indications are restricted to the outcome-specific effects underlying efficacy for a primary composite outcome, as in the case of empagliflozin, which showed a reduction in the risk of cardiovascular death, but not of non-fatal cardiovascular events. The use of the threshold and the associated $P$ value, however, represent the salient benchmark for such decision-making, preventing some potentially useful medications from receiving a label indication or clinical recommendation, while blurring differences among others.

By contrast, confidence curves alongside additional measures of the treatment effect give a more comprehensive and nuanced picture of the efficacy of a treatment. While applicable to any clinical area, in the case of cardiovascular outcome trials such an approach identifies modest evidence for efficacy for most medications and excellent evidence in a few cases (although often with small absolute effects).

$P$ value functions, counternull values and $S$ values are subject to some of the same limitations as $P$ values. For one, all these measures are generated from data and models that may be subject to bias or in violation of statistical assumptions; scrutiny of trial protocols and risk
of bias is therefore important and unconditional interpretation of results may be warranted. The same point applies to the fact that the strength of evidence depends also on sample size, and that meaningful clinical effects are not necessarily present, despite statistical evidence against a null effect; indeed, it has been argued that sample size along with the minimum meaningful effect size are needed to help make the P value informative about the effect of interest or null hypothesis. In the case of the large RCTs analysed here, sample size and expected effect sizes are explicitly accounted for in trial design alongside other important criteria, such as trial duration and, as discussed above, the design of non-inferiority trials, which may inherently limit our ability to understand the strength of the evidence for, and assess the magnitude of, the ‘true effects’ of the drugs under study. This makes it even more important to exercise caution in concluding that, where a large trial fails to show a statistically significant effect as judged by a significance threshold, this absence of evidence really constitutes evidence for the absence of a clinically meaningful treatment effect. Furthermore, there have been forceful arguments that Bayesian statistics offer the kind of intuitive approach to analysing clinical data, which the P value and measures derived from it lack. While noting that P value functions should not be interpreted as graphs of Bayesian posterior intervals, there remains no universally accepted guide to choosing the ‘prior’. Thus, no statistical approach is immune to misinterpretation or equally useful for all research settings, necessitating careful use of multiple statistics according to the research question, data at hand and previous subject matter knowledge.

In making decisions among competing glucose-lowering medications, healthcare professionals, regulatory authorities and clinical guidelines should therefore resist the temptation of a binary view of efficacy in favour of a clinical, and statistical, heuristic that better reflects such complexities.

ACKNOWLEDGEMENTS
FZ, MJD and KK acknowledge the National Institute for Health Research (NIHR) - Collaboration for Leadership in Applied Health Research and Care (CLAHRC) 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
DEK: no relevant conflict of interest. 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. 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. FZ has received honoraria for speaking at meetings from NAPP Pharmaceuticals and Boehringer Ingelheim.

AUTHOR CONTRIBUTIONS
Study idea and design: FZ; literature search: DEK and FZ; data preparation: DEK; data analysis: DEK; first draft: DEK; study critical revision and manuscript draft: all authors. All the authors provided final approval of the version to publish. The corresponding author (DEK) had full access to all the data in the study and had final responsibility for the decision to submit it for publication.

DATA AVAILABILITY
Data and statistical code are available in the supplementary material; for any questions please contact the corresponding author (DEK; dk261@leicester.ac.uk or davidekloecker@gmail.com).

COMPLIANCE WITH ETHICAL STANDARDS
This article does not contain any studies with human participants performed by any of the authors.

PEER REVIEW
The peer review history for this article is available at https://publons.com/publon/10.1111/dom.14384.

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

REFERENCES
1. Wasserstein RL, Lazar NA. The ASA statement on P values: context, process, and purpose. Am Stat. 2016;70(2):129-133.
2. McShane BB, Gal D. Statistical significance and the dichotomization of evidence. J Am Stat Assoc. 2017;112(519):885-895.
3. Gelman A, Stern H. The difference between “significant” and “not significant” is not itself statistically significant. Am Stat. 2006;60(4):328-331.
4. Rafi Z, Greenland S. Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprisal. BMC Med Res Methodol. 2020;20(1):244.
5. Gelman A. The problems with P values are not just with P values. Am Stat. 2016;70:1-2.
6. Hubbard R, Armstrong JS. Why we don’t really know what statistical significance means: implications for educators. J Market Edu. 2006;28(2):114-120.
7. Hubbard R, Bayarri MJ. Confusion over measures of evidence (p’s) versus errors (α’s) in classical statistical testing. Am Stat. 2003;57(3):171-178.
8. Perezgonzalez JD. Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. Front Psychol. 2015;6:223.
9. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, \( P \) values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol.* 2016;31(4):337-350.

10. Stark PB, Saltelli A. Cargo-cult statistics and scientific crisis. *Significance.* 2018;15(4):40-43.

11. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond \( "p < 0.05" \). *Am Stat.* 2019;73(supp. 1):1-19.

12. Infanger D, Schmidt-Trucksass A. \( P \) value functions: an underused method to present research results and to promote quantitative reasoning. *Stat Med.* 2019;38(21):4189-4197.

13. Shakespeare TP, Gebski VJ, Veness MJ, Simes J. Improving interpretation of clinical studies by use of confidence levels, clinical significance curves, and risk-benefit contours. *Lancet.* 2001;357(9265):1349-1353.

14. Birnbaum A. A unified theory of estimation. *Ann Math Stat.* 1961;32(1):112-135.

15. Fraser DAS. \( P \) values: the insight to modern statistical inference. *Ann Rev Stat Appl.* 2017;4(1):1-14.

16. Xie M-G, Singh K. Confidence distribution, the Frequentist distribution estimator of a parameter: a review. *Int Stat Rev.* 2013;81(1):3-39.

17. Poole C. Beyond the confidence interval. *Am J Public Health.* 1987;77(2):195-199.

18. Sullivan KM, Foster DA. Use of the confidence interval function. *Epidemiology.* 1990;1(1):39-42.

19. Schweder T, Hjort NL. Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions. Cambridge, UK: Cambridge University Press; 2016.

20. Rosenthal R, Rubin DB. The Counternull value of an effect size: a new statistic. *Psychol Sci.* 1994;5(6):329-334.

21. Shannon CE. A mathematical theory of communication. *Bell Syst Tech J.* 1948;27(3):379-423.

22. Greenland S. Valid \( P \) values behave exactly as they should: some misleading criticisms of \( P \) values and their resolution with \( S \)-values. *Am Stat.* 2019;73(supp. 1):106-114.

23. Cole SR, Edwards JK, Greenland S. Surprise! *Am J Epidemiol.* 2020;190(2):191-193.

24. Husain M, Birkenfeld AL, Domsmark M, et al. Oral Semaglutide and cardiovascular outcomes in patients with type 2 diabetes. *N Engl J Med.* 2019;381(9):841-851.

25. Holman RR, Bethel MA, Mentz RJ, et al. Effects of once-weekly Exenatide on cardiovascular outcomes in type 2 diabetes. *N Engl J Med.* 2017;377(13):1228-1239.

26. Rafi Z, Vigotsky A. concurve: Computes and Plots Compatibility (Confidence) Intervals, \( P \)-Values, \( S \)-Values, & Likelihood Intervals to Form Consonance, Surprisal, & Likelihood Functions. R package version 2.7.7. 2020. https://CRAN.R-project.org/package=concurve. Accessed 07 March 2021.

27. Ganju J, Rom D. Non-inferiority versus superiority drug claims: the (not so) subtle distinction. *Trials.* 2017;18(1):278.

28. Buse JB, Wexler DJ, Tsapas A, et al. 2019 update to: Management of hyperglycaemia in type 2 diabetes, 2018. A consensus report by the American Diabetes Association (ADA) and the European Association for the Study of diabetes (EASD). *Diabetologia.* 2020;63(2):221-228.

29. American Diabetes Association. 10. Cardiovascular disease and risk management: standards of medical Care in Diabetes-2020. *Diabetes Care.* 2020;43(Suppl 1):S111-S134.

30. U. S. Food and Drug Administration. Highlights of prescribing information for Jardiance. 2018. https://www.accessdata.fda.gov/drugsatfda_docs/label/2018/204629s018bld.pdf. Accessed 07 March 2021.

31. 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.

32. Davies MJ, Kloecker DE, Webb DR, Khunti K, Zaccardi F. Number needed to treat in cardiovascular outcome trials of glucagon-like peptide-1 receptor agonists: a systematic review with temporal analyses. *Diabetes Obes Metab.* 2020;22(9):1670-1677.

33. Greenland S, Rafi Z. To aid scientific inference, emphasize unconditional descriptions of statistics. ArXiv190908583 StatME. 2020. https://arxiv.org/abs/1909.08583. Accessed 07 March 2021.

34. Betensky RA. The \( P \) value requires context, not a threshold. *Am Stat.* 2019;73(supp. 1):115-117.

35. Gill CJ, Sabin L, Schmid CH. Why clinicians are natural Bayesians. *BMJ.* 2005;330(7499):1080-1083.

36. Spiegelhalter DJ, Abrams KR, Myles JP. Comparison of alternative approaches to inference. In: Senn SJ, Barnett V, Spiegelhalter DJ, Abrams KR, Myles JP, eds. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation.* Chichester, England: John Wiley & Sons Ltd.; 2003:121-137.

37. Rafi Z. Comparison to Bayesian Posterior Distributions. https://data. lesslikely.com/concurve/articles/bayes.html. Accessed 07 March 2021.

38. 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.

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** 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; 23:1685-1691. https://doi.org/10.1111/dom.14384