This is a repository copy of *Evaluating the ability of economic models of diabetes to simulate new cardiovascular outcomes trials : a report on the Ninth Mount Hood Diabetes Challenge*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/164412/

Version: Accepted Version

Article:
Si, L., Willis, M.S., Asseburg, C. et al. (23 more authors) (2020) Evaluating the ability of economic models of diabetes to simulate new cardiovascular outcomes trials : a report on the Ninth Mount Hood Diabetes Challenge. Value in Health, 23 (9). pp. 1163-1170. ISSN 1098-3015

https://doi.org/10.1016/j.jval.2020.04.1832

Article available under the terms of the CC-BY-NC-ND licence (https://creativecommons.org/licenses/by-nc-nd/4.0/).

Reuse
This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can't change the article in any way or use it commercially. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Evaluating the ability of economic models of diabetes to simulate new cardiovascular outcomes trials: a report on the Ninth Mount Hood Diabetes Challenge

Abstract

Objectives: The cardiovascular outcomes challenge examined the predictive accuracy of 10 diabetes models in estimating hard outcomes in two recent cardiovascular outcomes trials (CVOTs) and whether recalibration can be used to improve replication.

Methods: Participating groups were asked to reproduce the results of the Empagliflozin Cardiovascular Outcome Event Trial in Type 2 Diabetes Mellitus Patients (EMPA-REG OUTCOME) and the Canagliflozin Cardiovascular Assessment Study (CANVAS) Program. Calibration was performed and additional analyses assessed model ability to replicate absolute event rates, hazard ratios (HRs), and the generalizability of calibration across CVOTs within a drug class.

Results: Ten groups submitted results. Models underestimated treatment effects (i.e., HRs) using uncalibrated models for both trials. Calibration to the placebo arm of EMPA-REG OUTCOME greatly improved the prediction of event rates in the placebo, but less so in the active comparator arm. Calibrating to both arms of EMPA-REG OUTCOME individually enabled replication of the observed outcomes. Using EMPA-REG OUTCOME-calibrated models to predict CANVAS Program outcomes was an improvement over uncalibrated models but failed to capture treatment effects adequately. Applying canagliflozin HRs directly provided the best fit.

Conclusions: The Ninth Mount Hood Diabetes Challenge demonstrated that commonly used risk equations were generally unable to capture recent CVOT treatment effects but that calibration of the risk equations can improve predictive accuracy. While calibration serves as a practical approach to improve predictive accuracy for CVOT outcomes, it does not extrapolate generally to other settings, time horizons, and comparators. New methods and/or new risk equations for capturing these CV benefits are needed.

Highlights

- Diabetes health economic models are commonly developed based on risk equations using classic risk factors such as glycated hemoglobin, systolic blood pressure, lipid level, and body mass index. However, existing models might not account for the entire cardioprotective effects of new treatments observed in recent cardiovascular outcomes trials (CVOTs).
- This paper shows that existing risk factor-based health economics models in diabetes have limitations in predicting the results of CVOTs. Calibration of these risk functions to the observed data only partially resolves these issues.
- Future models may require new methods and/or new risk equations that promote standardization while better extrapolating CV outcomes across settings, time, and comparators.

Background

Use of economic modelling is widespread and necessary (1, 2), particularly for chronic and progressive diseases like diabetes mellitus (DM) for which the decision-maker's time horizon (often lifetime) is longer than clinical trial durations. Health economic modelling provides a unique tool that combines the best available epidemiological data for disease progression and health outcomes with trial (often relatively short-term) data and enables the extrapolation of the health and cost consequences of health interventions over long time horizons. Economic modelling, moreover, facilitates economic evaluation between competing treatment interventions in the absence of head-to-head data.

Economic modelling has a long history of use in type 2 DM (T2DM). Modelling T2DM is challenging, as it affects multiple inter-related organ systems, and complications occur over long-time horizons during which event rates tend to increase. Comorbid conditions such as hypertension, dyslipidaemia, and obesity are common, and treatments for diabetes and comorbid conditions frequently work on the same set of biomarker risk factors and require intensification over time (3). Given the expense and intellectual capital required to construct health economic models of DM, most have been developed with the goal of supporting multiple applications covering different settings and comparisons.

Historically, economic modelling of T2DM treatment interventions has featured projection of differences in these biomarker risk factors into economically relevant outcomes (e.g., event rates, life expectancy, quality-adjusted life years and costs) over long time horizons using risk prediction equations. Risk prediction equations, by design, reflect treatment conditions prevailing during the follow-up period of the underlying data, which may not match current standards of care. The widely used United Kingdom Prospective Diabetes Study (UKPDS) risk equations (4, 5), for instance, have been shown to overpredict cardiovascular (CV) risk in a wide range of contemporary cohorts (6-10). Given that important risk equations are inevitably backward-looking, as they require outcomes over long time horizons and practice standards change over this time, calibration has been proposed as a way to improve the predictive accuracy of valuable risk equations (11, 12). When conducted transparently, calibrated and appropriately validated, health economic modelling can be a valuable decision-making aid.

In 2015, the Empagliflozin Cardiovascular Outcome Event Trial in Type 2 Diabetes Mellitus Patients (EMPA-REG OUTCOME) (13) found that the sodium-glucose co-transporter-2 (SGLT2) inhibitor empagliflozin was not only safe but that it had significantly lower rates of the primary composite outcome of death from cardiovascular causes, nonfatal myocardial infarction, or nonfatal stroke versus placebo. Since then, five more cardiovascular outcomes trials (CVOTs) in patients with established CV disease or greatly increased CV risk have reported cardioprotective effects for SGLT-2-inhibitors and glucagon-like peptide-1 (GLP-1) receptor agonists: the Liraglutide Effect and Action in Diabetes: Evaluation of Cardiovascular Outcome Results (LEADER) (14), Trial to Evaluate Cardiovascular and Other Long-term Outcomes with Semaglutide in Subjects with Type 2 Diabetes (SUSTAIN-6) (15), the Canagliflozin Cardiovascular Assessment Study (CANVAS Program) (16), Dapagliflozin Effect on CardiovascuLAR Events (DECLARE-TIMI 58) (17) and the Albiglutide and Cardiovascular Outcomes in Patients with Type 2 Diabetes and Cardiovascular Disease (Harmony Outcomes) (18). These cardioprotective effects reported cannot be fully explained by improvements in known biomarker risk factors (19). Applications with three independent diabetes simulation models have failed to replicate these treatment effects based on changes in surrogate biomarkers for most CV outcomes (20-22), highlighting an important challenge for economic decision makers.

Initiated in 2000 by Andrew Palmer and Jonathan Brown at Timberline Lodge, Mount Hood, Oregon, USA (23), the Mount Hood Diabetes Challenge is a biennial congress in which diabetes modelling groups have met to compare and contrast models, methods, and data in the context of simulating standardized treatment scenarios and discussing the results. So far, eight Mount Hood Diabetes Challenge meetings have been held, with the aims of improving performance (24, 25) and input transparency (26) of diabetes models. In light of the evidence that cardioprotective benefit from newer diabetes medications such as GLP-1 receptor agonists and SGLT2-inhibitors cannot be fully explained by traditional physiological biomarkers such as glycated haemoglobin (HbA1c), systolic blood pressure (SBP) and body mass index (BMI), the Ninth Mount Hood Diabetes Challenge was convened with, in part, the aims of:

1. examining the predictive accuracy of diabetes models on hard endpoints in two recent CVOTs and
2. examining whether recalibration can be used to better replicate CVOT results.

Methods

The Ninth Mount Hood Diabetes Challenge was advertised on the Mount Hood Diabetes Challenge web site (https://www.mthooddiabeteschallenge.com/) and was open to all interested health economic modelling groups. Diabetes modelling groups registered within Mount Hood Diabetes Challenge Network were also informed directly via email.

The scope and parameters of the challenge were proposed by the modelling groups and debated, and the final conference program featured three challenge exercises (instructions were provided, and results were expected to be sent back prior to the conference). The first day of the conference featured a CV Outcomes Challenge and the second day featured a

Quality of Life Challenge and a Diabetes versus Non-Diabetes Simulation Models Challenge. The details of each challenge can be found on the Mount Hood Diabetes Challenge web site.

The focus of this article is limited to the CV Outcomes Challenge (Day 1), which aimed to evaluate how well existing diabetes simulation models replicated the absolute event rates and treatment effects observed for key endpoints in recent CVOTs, in particular EMPA-REG OUTCOME trial (13, 27) and the CANVAS Program (16, 28) A standard set of instructions (Supplementary Appendix 1) was provided to participating modelling groups in each challenge exercise and is summarised below.

EMPA-REG OUTCOME Challenge

The groups were asked to replicate the EMPA-REG OUTCOME trial (13, 27), separately by empagliflozin (pooling 10 mg and 25 mg doses) and placebo arms, by loading their models with weighted average baseline patient characteristics and treatment effects on key biomarkers over 3 years and then simulating for a period of 3 years (trial mean follow-up was 3.0 years for the treatment group). Modelling groups were asked to use the data provided in the instructions or in the study publications listed in the instructions to the extent possible. The groups were asked to document any deviations from the instructions (e.g., additional assumptions or covariates required for the model to run the simulations) and submit them with the results. Three scenarios were simulated:

Scenario A: Each modelling group simulated EMPA-REG OUTCOME using their model without modifications (i.e., without calibration).

Scenario B: Each modelling group was instructed to calibrate their model to the results of the placebo arm in EMPA-REG OUTCOME, using calibration techniques appropriate for their model construction and documenting the methodology used. The models re-simulated EMPA-REG OUTCOME using the same calibration factors for both empagliflozin and placebo arms.

Scenario C: Each modelling group calibrated their model to the results of the empagliflozin arm in EMPA-REG OUTCOME. These new calibration factors were used to re-simulate the empagliflozin arm only (the results for the placebo arm were recycled from Scenario B).

CANVAS Program Challenge

To evaluate the ability of models calibrated to one CVOT (in this case, EMPA-REG OUTCOME) to duplicate the results of another CVOT (16, 28), that is generalizability, the modelling groups were asked to replicate the CANVAS Program trial as well (but using the EMPA-REG OUTCOME calibration factors). The models were loaded with baseline patient characteristics and treatment effects on key biomarkers and then simulated for 4 years (mean follow-up in the CANVAS Program was 3.6 years), separately for the placebo and canagliflozin study arms. Modelling groups were asked to use the data provided in the instructions or in the study publications listed in the instructions to the extent possible. The groups were asked to document any deviations from the instructions (e.g., additional assumptions or covariates required for the model to run the simulations) and submit them with the results. Four scenarios were simulated:

Scenario A: Each modelling group simulated their model without modification (i.e., without calibration).

Scenario B: Each modelling group was instructed to re-run their model with the calibration factors that were estimated using the placebo arm in EMPA-REG OUTCOME (i.e., calibration factors from EMPA-REG OUTCOME Challenge Scenario B above). The models were re-run using the same calibration factors for both canagliflozin and placebo arms.

Scenario C: To evaluate how well cardioprotection can be captured when the trial observed hazard ratios (HRs) are entered as inputs into the model, each modelling group was instructed to simulate the canagliflozin arm by using the EMPA-REG OUTCOME placebo-calibrated model together with the observed HRs for canagliflozin versus placebo published for the CANVAS Program. The models were re-run for the canagliflozin arm and the results for the placebo arm were recycled from Scenario B.

Scenario D: To evaluate how well calibration to the study arms in EMPA-REG OUTCOME individually captures cardioprotection in the CANVAS Program, each modelling group was instructed to simulate the canagliflozin arm using the calibration factors obtained in the EMPA-REG OUTCOME Challenge Scenario C (and without using direct HR inputs for canagliflozin). The models were re-run for the canagliflozin arm and the results for the placebo arm were recycled from Scenario B.

Results were presented and discussed by representatives from the modelling groups and other interested stakeholders. A recurring theme in the discussion was how to address the limitations of existing modelling approaches for considering cardioprotection, and paths to methodological improvement were debated. Representatives from each of the modelling groups were invited to participate in the development of meeting proceedings articles. Modelling groups were also contacted after the meeting to confirm their final submitted results, where correction of typographical errors was permitted but re-simulation was not.

Reporting of challenge results

A standard reporting format was provided to participating groups (29). Modelling groups were encouraged to submit results for each challenge. Groups were asked to document all their assumptions made in the challenge. Modelling groups were requested to report both cumulative incidence and event rates across 15 outcomes including death from any cause, death from CV cause, nonfatal myocardial infarction (MI), nonfatal or fatal MI, nonfatal stroke, nonfatal or fatal stroke, hospitalization for heart failure (HHF), hospitalization for angina, microalbuminuria, macroalbuminuria, end-stage renal disease (ESRD), major adverse cardiovascular events (MACE), coronary revascularization procedure, transient ischaemic attack (TIA) and amputations. Because few of the modelling groups submitted results for micro- and macroalbuminuria and it was unclear whether the results reflected new onset or overall prevalence, renal outcomes were excluded from analysis. In addition, for the sake of brevity, goodness-of-fit was reported only for event rates and for key macrovascular outcomes, namely CV death, nonfatal or fatal MI, nonfatal or fatal stroke, HHF and MACE. All data submitted by the modelling groups, however, are presented in Supplementary Appendix 2.

Mean absolute event rates for each outcome were computed separately by study arm based on individual rates submitted by the modelling groups. Treatment effects were calculated as HRs based on mean HRs reported by the modelling groups (using Microsoft Excel®). Concordance between the mean model predictions and the results of the CVOTs was measured using mean of absolute percentage errors (MAPEs) calculated by the modelling groups (30). Box and whisker plots were used to summarize the distribution of event rates for each of these endpoints, separately for each scenario using the ggplot2 package in R Project (31, 32).

Results

On October 6-7, 2018, 15 modelling groups gathered at the German Diabetes Center in Düsseldorf, Germany. Ten modelling groups participated in the CV Outcomes Challenge: BRAVO of diabetes model, Cardiff Model, CDC/RTI model, IQVIA-CDM, the Economic and Health Outcomes Model of Type 2 DM (ECHO-T2DM), Michigan Model for Diabetes (MMD), PROSIT diabetes modelling community, SPHR Type 2 Diabetes Treatment model, The Treatment Transition Model (TTM,) and UKPDS-OM2. The Cardiff Model submitted two sets of results, one using UKPDS-OM1 and one using UKPDS-OM2 risk equations.

Short biographies of participating models in the CV Outcome Challenge can be found in Supplementary Appendix 3. Not all groups submitted results for every endpoint and scenario. The outcomes reported by each modelling group are summarized in Table 1. Results for the full set of outcomes are presented in Supplementary Appendix 4.

<Table 1 should be inserted here>

EMPA-REG OUTCOME Challenge

The results for the EMPA-REG OUTCOME Challenge are presented in Table 2 and in Figure 1. Table 2 summarizes mean predicted event rates (by study arm), HRs and MAPE for key macrovascular outcomes (CV death, MI, stroke, HHF and MACE), together with the results observed in EMPA-REG OUTCOME. Panel A includes the uncalibrated set of results (Scenario A in the EMPA-REG OUTCOME challenge), Panel B the placebo-calibrated results (Scenario B in the EMPA-REG OUTCOME challenge), and Panel C the empagliflozin- and placebo-calibrated results (Scenario C in the EMPA-REG OUTCOME challenge).

<Table 2 should be inserted here>

<Figure 1 should be inserted here>

In the uncalibrated scenario (Panel A), MAPE for event rates of all outcomes combined for the treatment arm and placebo arm were 57.8% and 46.5% respectively. Of note, none of the modelling groups reproduced the increased but statistically nonsignificant increased risk for stroke that was observed in the EMPA-REG OUTCOME trial. In addition, treatment effects were generally underestimated, with MAPE ranging from 11.6% for the MI outcome to 55.0% for CV death (27.7% for all outcomes combined). The mean uncalibrated predictions varied between groups (Figure 1), and variation in predicted event rates across groups was largest for MACE. The EMPA-REG OUTCOME observed values fell within the interquartile ranges of the predicted results for 3 of the 5 endpoints for both study arms.

In the placebo-calibrated scenario (Panel B), MAPE for event rates overall and individually for the placebo arm was much improved when compared to Scenario A. The mean predicted event rate for the empagliflozin arm was also improved, but not to the same degree. Consistent with the improvement in MAPE for event rates in the placebo group, MAPE for HRs were also reduced. However, there was still a 23.7% MAPE between predicted and observed values for all outcomes after calibrating to the placebo group.

In the empagliflozin- and placebo-calibrated scenario (Panel C), MAPE for HRs was further reduced compared with the placebo-calibrated scenario as the result of the improvement in calibrating event rates in the treatment arm. The modelling groups were able on average to replicate the increased risk of stroke in the empagliflozin arm closely. All MAPEs for HR fell below 5% after calibration applied to both treatment and placebo groups.

CANVAS Program Challenge

The results for the CANVAS Program challenge are presented in Table 3 and Figure 2. Table 3 summarizes the mean predicted event rates by study arm and HRs for CV death, MI, stroke, HHF and MACE, together with the observed values and MAPE. Panel A includes the uncalibrated set of results (Scenario A in the CANVAS Program challenge), Panel B the placebo-calibrated results (Scenario B in the CANVAS Program challenge), Panel C the placebo-calibrated with CANVAS Program HR results (Scenario C in the CANVAS Program challenge), and Panel D the empagliflozin- and placebo-calibrated results (Scenario D in the CANVAS Program challenge).

<Table 3 should be inserted here>

<Figure 2 should be inserted here>

In the uncalibrated scenario (Panel A), modelling groups tended to overestimate event rates of macrovascular outcomes in both treatment and placebo arms. In addition, treatment effects predicted by modelling groups tended to be smaller than observed in the CANVAS Program. MAPEs of HRs were generally smaller compared to MAPE in the EMPA-REG OUTCOME challenge, even though the MAPE of HR for HHF remained high as modelling groups generally estimated little or no benefit in contrast to an observed 33% reduction in HHF in the CANVAS Program. Moreover, the CANVAS Program observed values fell within the interquartile ranges of the predicted results for all but one prediction (Figure 2).

In the placebo-calibrated scenario (Panel B), where models were calibrated to the event rates observed in the placebo arm of EMPA-REG OUTCOME trial, MAPEs for event rates in the placebo arm decreased for CV death, stroke, and MACE but increased for MI and HHF compared to scenario A. Not surprisingly, prediction of treatment effect did not improve compared to scenario A.

In the placebo-calibrated plus canagliflozin HRs scenario (Panel C), where the same calibration to the EMPA-REG OUTCOME placebo arm was combined with direct input of CANVAS Program observed HRs, mean predicted macrovascular event rates for the canagliflozin arm tended to improve compared to scenario A, except for MI and HHF. Notably, the predicted event rates for CV death and stroke were almost identical to those observed in the CANVAS Program. The MAPE of treatment effect for all outcomes in scenario C decreased to 8.7% from 23.0% in scenario A.

In the empagliflozin- and placebo-calibrated scenario (Panel D), MAPE for macrovascular event rates in the treatment arm were similar or larger (substantially so for stroke) than those for the placebo-calibrated with CANVAS Program HRs (Panel C). Not surprisingly, treatment effects for each macrovascular outcome further diverged from that in scenario C, HRs that exceeded 1.0 for stroke and HHF.

Discussion

This article summarizes the findings of the CV Outcomes Challenge in the Ninth Mount Hood Diabetes Challenge. Models tended to underestimate treatment effects using their existing risk prediction equations based on traditional biomarkers such as HbA1c, SBP and BMI. In the EMPA-REG OUTCOME challenge where calibration was conducted only using evidence from EMPA-REG OUTCOME trial, calibration considerably improved predictions and generally enabled the replication of outcomes that were observed in the EMPA-REG OUTCOME trial. In the CANVAS Program challenge, where calibration was conducted using evidence both from the EMPA-REG OUTCOME and CANVAS Program trials, calibration performed better using HRs from the CANVAS Program trial directly rather than relying on EMPA-REG OUTCOME calibration.

The design of these and other recent CVOTs reflects the guidance issued by the US Food and Drug Administration in December 2008, which responded to previous safety concerns by mandating long-term CVOTs for safety as a prerequisite to obtaining approval for antidiabetes drugs in type 2 diabetes (33). The resulting trials tend to have large sample sizes and long study durations, aim to maintain “glycaemic equipoise” rather than a glucose-lowering trial design, and measure hard outcomes like MACE directly (19). While CVOT durations are longer than most short-term glucose lowering trials, extrapolation using risk prediction equations is still required to fully capture the long-term costs and benefits (3). Adequately capturing all CV effects in CVOTs with current economic modelling methods poses several challenges. First, some drug classes, notably SGLT-2 inhibitors and GLP-1 receptor agonists, have reported treatment effects that cannot be explained only by improvements in known biomarker risk factors, and hence models including these risk factors in their risk predictions cannot be expected to fully capture the reported outcomes, at least until the mechanisms of action are better understood (20-22). The Ninth Mount Hood Diabetes Challenge has confirmed this finding. Second, widely used risk equations such as the UKPDS 68 and the UKPDS 82 reflect cohorts initially recruited in an earlier "therapeutic era", since when improved clinical care and many other factors have contributed to secular declines in morbidity and mortality in type 2 diabetes and in cardiovascular disease generally (34). The UKPDS study participants were also newly diagnosed with diabetes at recruitment (35), whereas recent CVOTs have typically recruited patients with long diabetes duration and established (sometimes quite severe) cardiovascular disease, although it should be noted that the UKPDS-OM2 was estimated using patient data with a median of 17.6 years follow up, during which many complications occurred. Third, the glycaemic equipoise design may encourage the adoption of more intensive and often multi-therapy treatment in patients in the placebo arm to reach similar biomarker values across study arms. This would be captured in a trial-based economic analysis as additional treatment costs in the comparator arm, and may not be unrealistic insofar as treating to target plays an increasing role in many standard treatment guidelines. There is also substantial heterogeneity within and across CVOTs, importantly including baseline patient characteristics such as the presence or absence of prior cardiovascular disease (13, 16). Although reported in aggregate results, such heterogeneity may be hard to model without access to individual patient data. Lastly, CVOTs are placebo-controlled, so to inform policy, evidence of differences with active comparators must be established indirectly, which requires knowledge of their impact on biomarkers and cardiovascular outcomes and is hampered by substantial CVOT heterogeneity.

Considering the above challenges, diabetes health economics modellers, are encouraged to adapt their methods and risk equations to reflect the cardioprotection observed in CVOTs. Calibration to individual trials is not a panacea, however, and some key limitations of the approach must be acknowledged. First, the successful replication in the EMPA-REG OUTCOME Challenge were limited to the short time horizon of the trial and it does not shed light on how well it applies to the longer time horizons that are often required in the economic modelling of DM. Nor did the exercise shed light on the robustness of extrapolation to different treatment settings, a frequent goal of health economic modelling. Second, the CANVAS Program Challenge demonstrated unsurprisingly that the calibrations from one CVOT extrapolate poorly to another CVOT. The differential features of the CVOTs, thus, dictate careful consideration and individual tailoring (and perhaps even separate models) to adequately model the cardioprotective effect observed in the CVOTs, though with careful documentation and explanation to ensure stakeholder face validity. Third, when the goal is to compare the results of two or more CVOTs (and not merely compare an active agent versus the placebo arm within a CVOT), it is unclear that calibrating to the separate trials individually would be viewed as credible. Moreover, comparisons of interventions using data from separate trials may be required and it remains unclear how calibration may be performed given substantial trial heterogeneity and a limited number of CVOTs. Alternative approaches that are straightforward and practical for predicting outcomes of CVOTs should therefore be explored. Two approaches may reward further exploration. First, it may be possible to modify existing models so that the effects of observed risk factor change on outcomes continue to be simulated, but in addition the residual “unexplained” treatment effects on outcomes are modelled directly as changes in the outcome probabilities for treated patients. Second, modellers need to closely follow and ideally participate in efforts to better understand mechanisms of action using data from existing and future SGL-2 and GLP-1 trials and include novel biomarkers being identified that could then be incorporated in future risk equations and models.

The CV Outcomes Challenge itself also had several limitations. First, it was impossible to provide instructions in sufficient detail to ensure that all modelling groups would apply exactly the same methods, given that model structures are so different, so some divergence in implementation was inevitable. Calibration methods in particular were left to each modelling group's discretion, which may flavour the results (the methods used by each model are reported in Supplementary Appendix 5). Moreover, most modelling groups generally limited calibration to the subset of endpoints in their models that were informed by risk prediction equations directly (e.g., models based on UKPDS Outcomes Model generally include all MI, but some of the models reported nonfatal MI as well), which can be observed in the modest, but non-zero MAPE values in Panel C of the EMPA-REG OUTCOME Challenge. Second, not all model groups participated in all parts of the challenge, which complicates comparison across scenarios. Third, some of the CVOT outcome definitions may differ from the corresponding outcomes in some of the models. For example, HHF in the CVOTs may match poorly to HF in models that use the UKPDS Outcomes Model.

Conclusions

The Ninth Mount Hood Diabetes Challenge provides evidence that existing risk factor-driven diabetes models have limitations when estimating the entire treatment effects observed in recent CVOTs. While calibration to the current diabetes models serves as a practical approach to improve the accuracy of predicting the reported CVOTs outcomes, it has several limitations regarding extrapolation to new settings and longer time horizons. Non-placebo comparisons also require special attention, and so new and better methods are needed: such as hybrid risk factor/treatment effect modelling, and the exploration of new biomarkers for newer drug classes.

References

1. Buxton MJ, Drummond MF, Van Hout BA, et al. Modelling in economic evaluation: an unavoidable fact of life. Health Econ. 1997; 6: 217-27.
2. Caro JJ, Briggs AH, Siebert U, et al. Modeling good research practices-overview: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-1. Value in health. 2012; 15: 796-803.
3. ADA. Guidelines for computer modeling of diabetes and its complications. Diabetes Care. 2004; 27: 2262-5.
4. Clarke PM, Gray AM, Briggs A, et al. A model to estimate the lifetime health outcomes of patients with type 2 diabetes: the United Kingdom Prospective Diabetes Study (UKPDS) Outcomes Model (UKPDS no. 68). Diabetologia. 2004; 47: 1747-59.
5. Hayes AJ, Leal J, Gray AM, et al. UKPDS outcomes model 2: a new version of a model to simulate lifetime health outcomes of patients with type 2 diabetes mellitus using data from the 30 year United Kingdom Prospective Diabetes Study: UKPDS 82. Diabetologia. 2013; 56: 1925-33.
6. Zomer E, Liew D, Owen A, et al. Cardiovascular risk prediction in a population with the metabolic syndrome: Framingham vs. UKPDS algorithms. Eur J Prev Cardiol. 2014; 21: 384-90.
7. Tao L, Wilson EC, Griffin SJ, et al. Performance of the UKPDS outcomes model for prediction of myocardial infarction and stroke in the ADDITION-Europe trial cohort. Value in Health. 2013; 16: 1074-80.
8. Pagano E, Gray A, Rosato R, et al. Prediction of mortality and macrovascular complications in type 2 diabetes: validation of the UKPDS Outcomes Model in the Casale Monferrato Survey, Italy. Diabetologia. 2013; 56: 1726-34.
9. McEwan P, Bennett H, Ward T, et al. Refitting of the UKPDS 68 risk equations to contemporary routine clinical practice data in the UK. Pharmacoeconomics. 2015; 33: 149-61.
10. McEwan P, Ward T, Bennett H, et al. Validation of the UKPDS 82 risk equations within the Cardiff Diabetes Model. Cost Eff Resour Alloc. 2015; 13: 12.
11. Ramos M, Foos V, Ustyugova A, et al. Cost-Effectiveness Analysis of Empagliflozin in Comparison to Sitagliptin and Saxagliptin Based on Cardiovascular Outcome Trials in Patients with Type 2 Diabetes and Established Cardiovascular Disease. Diabetes Ther. 2019; 10: 2153-67.
12. Karnon J, Vanni T. Calibrating models in economic evaluation: a comparison of alternative measures of goodness of fit, parameter search strategies and convergence criteria. Pharmacoeconomics. 2011; 29: 51-62.
13. Zinman B, Wanner C, Lachin JM, et al. Empagliflozin, Cardiovascular Outcomes, and Mortality in Type 2 Diabetes. N Engl J Med. 2015; 373: 2117-28.
14. Marso SP, Daniels GH, Brown-Frandsen K, et al. Liraglutide and Cardiovascular Outcomes in Type 2 Diabetes. N Engl J Med. 2016; 375: 311-22.
15. Marso SP, Bain SC, Consoli A, et al. Semaglutide and Cardiovascular Outcomes in Patients with Type 2 Diabetes. New England Journal of Medicine. 2016; 375: 1834-44.
16. Neal B, Perkovic V, Mahaffey KW, et al. Canagliflozin and Cardiovascular and Renal Events in Type 2 Diabetes. N Engl J Med. 2017; 377: 644-57.
17. Wiviott SD, Raz I, Bonaca MP, et al. Dapagliflozin and Cardiovascular Outcomes in Type 2 Diabetes. New England Journal of Medicine. 2018; 380: 347-57.
18. Hernandez AF, Green JB, Janmohamed S, et al. Albiglutide and cardiovascular outcomes in patients with type 2 diabetes and cardiovascular disease (Harmony Outcomes): a double-blind, randomised placebo-controlled trial. Lancet. 2018; 392: 1519-29.
19. Cefalu WT, Kaul S, Gerstein HC, et al. Cardiovascular Outcomes Trials in Type 2 Diabetes: Where Do We Go From Here? Reflections From a Diabetes Care Editors' Expert Forum. Diabetes Care. 2018; 41: 14-31.

20. Kuo S, Ye W, Duong J, et al. Are the favorable cardiovascular outcomes of empagliflozin treatment explained by its effects on multiple cardiometabolic risk factors? A simulation of the results of the EMPA-REG OUTCOME trial. Diabetes Res Clin Pract. 2018; 141: 181-89.
21. Willis M, Neslusan C, Johansen P, et al. The Importance of Considering the Evolving Evidence Base on Cardiovascular Effects of Anti-Hyperglycemic Agents on Estimates of 'Value for Money'. poster presentation at 77th ADA San Diego, USA, 2017.
22. Evans M, Johansen P, Vrazic H. Incorporating cardioprotective effects of once-weekly semaglutide in estimates of health benefits for patients with type 2 diabetes. ADA 78th Scientific Sessions June 22-26 Orlando, FL 2018.
23. Brown JB, Palmer AJ, Bisgaard P, et al. The Mt. Hood challenge: cross-testing two diabetes simulation models. Diabetes Research and Clinical Practice. 2000; 50: S57-S64.
24. Palmer AJ. Computer Modeling of Diabetes and Its Complications: A Report on the Fifth Mount Hood Challenge Meeting. Value in Health. 2013; 16: 670-85.
25. Computer modeling of diabetes and its complications: a report on the Fourth Mount Hood Challenge Meeting. Diabetes Care. 2007; 30: 1638-46.
26. Palmer AJ, Si L, Tew M, et al. Computer Modeling of Diabetes and Its Transparency: A Report on the Eighth Mount Hood Challenge. Value in Health. 2018; 21: 724-31.
27. Wanner C, Inzucchi SE, Lachin JM, et al. Empagliflozin and Progression of Kidney Disease in Type 2 Diabetes. N Engl J Med. 2016; 375: 323-34.
28. Mahaffey KW, Neal B, Perkovic V, et al. Canagliflozin for Primary and Secondary Prevention of Cardiovascular Events: Results From the CANVAS Program (Canagliflozin Cardiovascular Assessment Study). Circulation. 2018; 137: 323-34.
29. Network MHDC. Challenge Session Final instructions. 2018.
30. Armstrong JS, Collopy F. Error measures for generalizing about forecasting methods: Empirical comparisons. International Journal of Forecasting. 1992; 8: 69-80.
31. Team RC. R: A language and environment for statistical computing. 2013.
32. Wickham H. ggplot2: elegant graphics for data analysis. Springer, 2016.
33. Administration USFaD. Guidance for industry: diabetes mellitus—evaluating cardiovascular risk in new antidiabetic therapies to treat type 2 diabetes. In: Services USDoHaH, ed. Silver Spring, MD, 2008.
34. Rawshani A, Rawshani A, Franzén S, et al. Mortality and Cardiovascular Disease in Type 1 and Type 2 Diabetes. New England Journal of Medicine. 2017; 376: 1407-18.
35. UK Prospective Diabetes Study (UKPDS). VIII. Study design, progress and performance. Diabetologia. 1991; 34: 877-90.

Table 1: Outcomes submitted to the cardiovascular outcomes challenge, by modelling group

	BRAVO	CARDIFF		CDC	ECHO-T2DM	IQVIA CDM	MMD	PROSIT	SPHR	TTM	UKPDS-OM2
		UKPDS-OM1	UKPDS-OM2								
ACM	X	X	X	X	X	X	X	X	X	X	X
CV Death	X	X	X	X	X	X	X	X	--	X	X
Nonfatal MI	X	X	--	--	X	--	X	X	--	X	--
Nonfatal or Fatal MI	X	X	X	X	X	X	X	X	X	X	X
Nonfatal Stroke	X	X	--	--	X	--	X	X	--	X	--
Nonfatal or Fatal Stroke	X	X	X	X	X	X	X	X	X	X	X
HHF	X	X	X	--	X	X	X	--	X	X	X
HA	X	--	--	X	--	X	--	--	X	X	X
MACE	X	--	--	--	X	--	X	X	--	X	--
Coronary Revascularization	X	--	--	--	--	--	X	--	--	--	--
TIA	--	--	--	--	--	--	--	--	--	--	--
Amputation	--	X	X	--	X	X	--	X	X	X	X

ACM, all-cause mortality; CV, cardiovascular; MI, myocardial infarction; HHF, hospitalization for heart failure; HA, hospitalization for angina; MACE, major adverse cardiovascular events; TIA, transient ischemic attack

Table 2: Event rates (per 1,000 patient-years), hazard ratios (HRs), and mean absolute percentage error (MAPE) for the EMPA-REG OUTCOME Challenge

	n	Empagliflozin			Placebo			HR		
		EMPA-REG OUTCOME	Predicted	MAPE (%)	EMPA-REG OUTCOME	Predicted	MAPE (%)	EMPA-REG OUTCOME	Predicted	MAPE (%)
A. Uncalibrated										
CV Death	10	12.4	19.0	57.9	20.2	19.5	33.7	0.62	0.96	55.0
MI	11	16.8	17.1	21.1	19.3	17.8	19.4	0.87	0.97	11.6
Stroke	11	12.3	8.3	48.2	10.5	9.1	42.0	1.18	0.91	23.3
HHF	9	9.4	6.8	31.1	14.5	7.5	48.5	0.65	0.92	41.4
MACE	5	37.4	37.9	47.6	43.9	39.7	43.8	0.86	1.11	30.5
All Outcomes*	85	NA	NA	57.8	NA	NA	46.5	NA	NA	27.7
B. PBO-Calibrated										
CV Death	7	12.4	14.5	32.4	20.2	18.1	11.2	0.62	0.81	40.9
MI	7	16.8	18.2	8.4	19.3	19.3	4.3	0.87	0.94	8.4
Stroke	7	12.3	9.7	21.3	10.5	10.7	2.7	1.18	0.90	23.3
HHF	6	9.4	12.6	33.5	14.5	13.9	4.0	0.65	0.90	39.1
MACE	3	37.4	37.7	14.9	43.9	40.2	12.0	0.86	0.94	10.4
All Outcomes*	55	NA	NA	27.1	NA	NA	10.1	NA	NA	23.7
C. Empagliflozin-& PBO-Calibrated										
CV Death	7	12.4	11.3	11.0	20.2	18.1	11.2	0.62	0.62	1.8
MI	7	16.8	16.6	2.2	19.3	19.3	4.3	0.87	0.86	2.0
Stroke	7	12.3	12.5	3.8	10.5	10.7	2.7	1.18	1.17	4.9
HHF	6	9.4	9.4	2.6	14.5	13.9	4.0	0.65	0.68	4.4
MACE	3	37.4	35.1	9.1	43.9	40.2	12.0	0.86	0.88	4.4
All Outcomes*	55	NA	NA	9.2	NA	NA	10.1	NA	NA	4.1

NA, not applicable; EMPA-REG OUTCOME, the Empagliflozin Cardiovascular Outcome Event Trial in Type 2 Diabetes Mellitus Patients; CV, cardiovascular; MI, myocardial infarction; HHF, hospitalization for heart failure; MACE, major adverse cardiovascular event.

* Based on all outcomes presented by participating modelling groups, not just the 5 macrovascular outcomes reported in Table 2 (see Supplementary Appendix 4).

Table 3: Event rates (per 1,000 patient-years), hazard ratios (HRs), and mean absolute percentage error (MAPE) for CANVAS Program Challenge

		Canagliflozin			Placebo			HR		
	n	CANVAS Program	Predicted	MAPE (%)	CANVAS Program	Predicted	MAPE (%)	CANVAS Program	Predicted	MAPE (%)
A. Uncalibrated	-									
CV Death	9	11.6	14.3	51.5	12.8	15.0	46.0	0.87	0.96	12.6
MI	10	11.2	15.2	53.8	12.6	16.0	49.1	0.89	0.98	10.7
Stroke	10	7.9	8.0	44.9	9.6	8.7	41.4	0.87	0.92	11.4
HHF	8	5.5	6.1	24.5	8.7	6.3	31.5	0.67	0.99	47.1
MACE	4	26.9	32.6	72.8	31.5	34.2	66.7	0.86	1.01	17.7
All Outcomes*	71	NA	NA	64.4	NA	NA	80.1	NA	NA	23.0
B. Placebo-calibrated										
CV Death	7	11.6	13.1	35.5	12.8	14.5	30.3	0.87	0.90	9.5
MI	7	11.2	19.0	77.5	12.6	20.9	66.2	0.89	0.89	8.6
Stroke	7	7.9	8.5	25.9	9.6	9.9	15.6	0.87	0.86	12.3
HHF	6	5.5	14.6	165.9	8.7	14.4	85.2	0.67	1.28	91.4
MACE	3	26.9	35.5	36.4	31.5	37.8	31.3	0.86	0.94	9.3
All Outcomes*	51	NA	NA	61.2	NA	NA	48.4	NA	NA	24.8
C. Placebo-calibrated & canagliflozin HRs										
CV Death	7	11.6	11.4	21.8	12.8	14.5	30.3	0.87	0.78	10.2
MI	7	11.2	17.3	55.4	12.6	20.9	66.2	0.89	0.83	6.3
Stroke	7	7.9	7.9	14.1	9.6	9.9	15.6	0.87	0.80	9.1
HHF	6	5.5	9.1	85.1	8.7	14.4	85.2	0.67	0.63	6.8
MACE	3	26.9	30.9	28.6	31.5	37.8	31.3	0.86	0.82	5.9
All Outcomes*	51	NA	NA	41.4	NA	NA	48.4	NA	NA	8.7
D. Empagliflozin-& placebo-calibrated										
CV Death	7	11.6	8.8	24.2	12.8	14.5	30.3	0.87	0.61	29.8
MI	7	11.2	16.9	59.1	12.6	20.9	66.2	0.89	0.81	9.4
Stroke	7	7.9	10.6	44.0	9.6	9.9	15.6	0.87	1.08	32.7
HHF	6	5.5	10.9	99.0	8.7	14.4	85.2	0.67	1.07	65.4
MACE	3	26.9	32.4	27.6	31.5	37.8	31.3	0.86	0.87	5.6
All Outcomes*	51	NA	NA	47.4	NA	NA	48.4	NA	NA	31.7

NA, not applicable; CANVAS, the Canagliflozin Cardiovascular Assessment Study; CV, cardiovascular; MI, myocardial infarction; HHF, hospitalization for heart failure; MACE, major adverse cardiovascular event.

* Based on all outcomes presented by participating modelling groups, not just the 5 macrovascular outcomes reported in Table 3 (see Supplementary Appendix 4).

Figure 1: Box-and-Whisker Plots of predicted event rates (per 1000 patient-years) for key macrovascular outcomes, by scenario and arm in the EMPA-REG OUTCOME Challenge. Event rates observed in EMPA-REG OUTCOME are depicted by stars.
EMPA-REG OUTCOME, the Empagliflozin Cardiovascular Outcome Event Trial in Type 2 Diabetes Mellitus Patients. PBO, placebo; EMPA, empagliflozin; SoC, standard of care; CV, cardiovascular; MI, myocardial infarction; HHF, hospitalization for heart failure; MACE, major adverse cardiovascular events

Figure 2: Box-and-Whisker plots of predicted event rates (per 1000 patient-years) for key macrovascular outcomes, by scenario and arm in the CANVAS Program Challenge. Event rates observed in the CANVAS Program are depicted by stars.
CANVAS, the Canagliflozin Cardiovascular Assessment Study; PBO, placebo; CANA, canagliflozin; HR, hazard ratio; EMPA, empagliflozin; SoC, standard of care; CV, cardiovascular; MI, myocardial infarction; HHF, hospitalization for heart failure; MACE, major adverse cardiovascular event.