Supplementary Appendix 2

AdViSHE Validation Assessment Tool:

The Economics and Health Outcomes Model of Type 2 Diabetes Mellitus (ECHO-T2DM)

Part A: Validation of the Conceptual Model

• **A1: Have experts been asked to judge the appropriateness of the conceptual model?**

Yes, ECHO-T2DM has been continuously evaluated for face validity since the first model version was completed in 1997, including feedback regarding design and programming from clinical experts at Janssen Inc. Additionally, the conceptual model has been peer-reviewed by scientific journal editors and independent reviewers in the course of two published model applications and one published model validity study. The model has also been presented and described in 39 posters and 8 podium presentations to the diabetes modeling community (see Reference List at the end of this document).

In a recent literature review of T2DM models, Charokopou and colleagues concluded that ECHO-T2DM fulfilled all criteria specified in their structured assessment of suitability for modeling long-term health-economic consequences in T2DM [1].

- “The thirteen newly-identified models were qualitatively assessed using a structural framework in order to assess which of these models are appropriate for projecting long-term health-economic consequences of chronic T2DM treatment. Based on this assessment, the CARDIFF, Sheffield T2DM, and ECHO-T2DM models fulfilled all criteria” (p. 216).

ECHO-T2DM has also been scrutinized by Health Technology Assessment (HTA) reviewers in a number of countries, including the National Institute for Health and Care Excellence (NICE) in the U.K. [2], Canadian Agency for Drugs and Technologies in Health (CADTH) [3], Scottish Medicines Consortium (SMC) [4], and National Centre for Pharmacoeconomics (NCPE) in Ireland [5]. These objective decision-makers in each country judged ECHO-T2DM to be suitable for estimating cost-effectiveness for the study problems considered, implicitly (and in some cases explicitly) approving the conceptual model of ECHO-T2DM. In particular:

- **NICE**
  - 2013: NICE Evidence Review Group (ERG) conducted by Southampton Health Technology Assessment Centre (SHTAC) independently investigated the model codes and concluded that: “The economic model captures all important aspects of the disease pathway and extrapolates intermediate outcomes to final outcomes in a robust and consistent manner, drawing upon standard sources from the literature.” [6] (p. 14-15). “Overall the ERG considers that the model is internally consistent […] and that the results of the model make intuitive sense” (p. 120). In addition, the ERG was able to independently run ECHO-T2DM and was able to replicate the results to within stochastic variability [6] (Table 67 and 68, p. 131).
  - 2014: “The Committee observed that the manufacturer’s ECHO-T2DM model followed the NICE reference case and methodology and noted the ERG’s view that the model had been well validated. Despite some uncertainty about the stability of the results, the Committee concluded that the manufacturer’s model was suitable for assessing the cost effectiveness of canagliflozin in combination therapy for treating type 2 diabetes” [7] (p. 45)
2016: “The committee acknowledged there were different advantages and disadvantages to the different modelling approaches, and it agreed with the AG that all models submitted were appropriate and of a reasonable quality” [2] (p.37)

- SMC
  - 2014: “As with other SMC submissions using diabetes models, there was some concern surrounding the appropriateness of using short term outcome measures to estimate long term treatment effects. However, the company has developed their economic analysis taking into account of relevant good practice guidance for economic modelling” [4] (p. 10)

- CADTH
  - 2014: The Common Drug Review (CDR) was also able to independently run ECHO-T2DM and ran additional analysis of their own (changing input values for e.g., HbA1c, SBP and BMI), “thus indicating the robustness of the manufacturer’s results” [7] (p. 20).

ECHO-T2DM has also been represented at the four most recent Mt. Hood Challenges, a biennial forum for computer modelers of diabetes to discuss and compare models, consider the appropriateness of conceptual model structure, and identify key areas of future development to advance the field [8-11]. Participants include representatives from pharmaceutical companies, academia, HTA bodies, and modelling groups. With the focal point of these challenges being to compare the structure and performance of diabetes simulation models, we refer the reader to Items A2, D2 and D4 for additional information. Continuous participation of ECHO-T2DM at the Mt. Hood Challenges and other fora however, has ensured a constant scrutiny of the conceptual model by leading experts in the field.

- A2: Has this model been compared to other conceptual models found in the literature or clinical textbooks?
  Yes. Comparing the structure and performance of diabetes simulation models is a focal point of the biennial Mt. Hood Challenges. There, the modeling groups are asked to attempt to replicate outcomes from key RCTs or observational data sets. Participants include representatives from pharmaceutical companies, academia, HTA bodies, and modelling groups. Comparison of conceptual models has been a key topic at all Mt. Hood Challenges, but the following activities in Mt. Hood Challenge 6 and 7 shed specific light on comparisons of the conceptual models:

  - Mt. Hood 6 Challenge (2012): Eight modelling groups participated and the focus of the conference was validation and uncertainty. The validation challenge consisted of predicting CVD and mortality patterns across a number different country settings and using naturalistic data. Kaiser Permanente Northwest [KPNW] in the US and the Swedish National Diabetes Registry [NDR] in Sweden) and an RCT (sub-groups in ADVANCE). Each modeling group presented their replication results, which were then compared and contrasted by all the modeling groups and public. The uncertainty challenge consisted of estimating mean and Monte Carlo error across a series of outcomes and pre-defined number of replicates in order to understand how many replicates are required for the participating models to limit Monte Carlo error to acceptable levels

  - Mt. Hood 7 Challenge (2014): 11 modelling groups participated and the focus was on validation. This conference consisted of three challenges: (1) replicate the Look AHEAD study, (2) replicate mortality after various CVD events from an observational study [12], and (3) investigate the impact on model outcomes due to assumptions about ethnicity. Each modeling group presented their replication results, which were then compared and contrasted by all the modeling groups and the public.

In addition, NICE in the UK conducted a mixed treatment appraisal of canagliflozin, empagliflozin, and dapagliflozin in 2016 [2], where the ECHO-T2DM model was used to generate evidence for the canagliflozin submission and concluded that:
“(...) there were different advantages and disadvantages to the different modelling approaches, and it agreed with the AG that all models submitted were appropriate and of a reasonable quality” (p. 37)

Part B: Input Data Validation

- **B1: Have experts been asked to judge the appropriateness of the input data?**

  Yes. Although it is unclear where to draw the line between the conceptual model and input data, experts have regularly weighed in on the appropriateness of inputs used in ECHO-T2DM. In a broad sense, any risk equation used in ECHO-T2DM is “input data”, and the decision on which risk equations to use was made in consensus with clinical experts from Janssen Inc. In a narrow sense, “input data” can be interpreted as the values that are entered into ECHO-T2DM by a user to parameterize a given application (e.g., mean age of the cohort at study baseline), which differs from programmed features of the model which cannot be modified by the user without changing the source code (e.g., the macrovascular risk equations supported in the model).

  The sources of risk inputs concerning the health states modelled in ECHO are presented in Supplementary Table 1 and Table 2 in the manuscript under submission. While some of the choices made in selecting risk equations may be unique to ECHO-T2DM, there is broad consensus among diabetes modelers about others, such as the UKPDS equations used to estimate risks of cardiovascular events and mortality. One of our secondary aims in the validation study described in this submission is to compare the replicability of empirical data across the different sets of macrovascular event risk equations.

  In the narrow sense of “input”, that is, the data required to populate the simulations for this manuscript (12 studies with a total of 17 study populations), not all inputs have been judged by independent experts. However, our goal with the external validity tests was to replicate trials as closely as possible, in order to facilitate accurate estimates of the outcomes. Most of the inputs required (e.g., baseline patient characteristics, treatment patterns, and biomarker evolution) were reported comprehensively in the various study publications and there is relatively little concern that the figures used are not correct. Where key covariates were missing, however, we sourced data from the most similar studies (e.g., most studies did not report values for eGFR, and heart rate). The details are described in the manuscript under submission.

- **B2: When input parameters are based on regression models, have statistical tests been performed?**

  Yes. Regression models as a source of risk predictions are hard-coded into model programing, but regressions models for QALY disutility values, such as the CODE-2 [13], can be customized by the model as “user inputs”. Statistical assessments of the validity and adequacy of these regression models are largely available in their respective publications. The UKPDS equations used in ECHO for the prediction of macrovascular risks and mortality are well-established in T2DM modeling and appropriate statistical tests were performed.

Part C: Validation of the Computerized Model

- **C1: Has the computerized model been examined by modeling experts?**

  Yes. As part of HTA submissions, independent and objective HTA boards such as NICE (UK), SMC (Scotland), NCPE (Ireland), and CADTH (Canada) have all reviewed the structure of ECHO-T2DM and its cost-effectiveness applications (see the entry for A1)

  o NICE Evidence Review Group (ERG) conducted by Southampton Health Technology Assessment Centre (SHTAC) independently investigated the model codes and was able to replicate the results to within stochastic variability [6] (Table 67 and 68, p. 131).
The NICE review of canagliflozin, empagliflozin, and dapagliflozin [2], where ECHO-T2DM was submitted to support the canagliflozin application, and “The committee […] agreed with the AG that all models submitted were appropriate and of a reasonable quality” (p. 37)

SMC assessed a cost-effectiveness application of ECHO-T2DM in 2014, and concluded that “the economic case has been demonstrated” (p. 10) [4], thus implicitly asserting adequacy of the computerized model

CADTH assessed the cost-effectiveness of canagliflozin including evidence generated using ECHO-T2DM and ruled for reimbursement, thus implicitly accepting ECHO-T2DM as a credible source of health-economic evidence [3]

- **C2:** Has the model been run for specific, extreme sets of parameter values in order to detect any coding errors?

Yes, every time a programming change is made to the model, the implementation is tested systematically and debugged by both the programmer and other staff at the Swedish Institute for Health Economics.

As part of a CADTH 2014 submission, the Common Drug Review (CDR) was also able to independently run ECHO-T2DM and ran additional analysis of their own (changing input values for e.g., HbA1c, SBP and BMI), “thus indicating the robustness of the manufacturer’s results” [14] (p. 20).

Formal verification tests of ECHO-T2DM have been conducted twice. The first was conducted by external analysts and included a structured audit of ECHO-T2DM. 67 “stress tests” were specified and then conducted, and all results were consistent with expectations [15]. The results of a follow-up verification of the current model version are part of the manuscript under submission here.

- **C3:** Have patients been tracked through the model to determine whether its logic is correct?

Yes. While ECHO-T2DM does not ordinarily produce any output on individual patients, the code is regularly inspected and traces of health states and biomarkers are examined for individual patients to assess correct implementation in the computerized model and plausibility of assumptions. Because of space limitations in journal publications, and also due to the repetitive and iterative nature of this type of testing in multi-application models, these tests and their outcomes have not been reported.

- **C4:** Have individual sub-modules of the computerized model been tested?

Yes, see response to C2, above. Furthermore, when model improvements are made to any sub-module of ECHO-T2DM, the modified sub-module is inspected by running ECHO-T2DM with inputs that are specifically designed to isolate and investigate the effects of this sub-module on other parts of the model simulation. While these tests are being carried out regularly to verify that the ECHO-T2DM model calculations perform in line with expected outcomes, we do not follow a set protocol nor are these test results reported.

Additionally, prior to model execution, ECHO-T2DM evaluates the model inputs for nonsensical, impermissible values (e.g., negative ages or biomarker values), and the user is prompted to adjust any offending inputs to ensure consistency with model requirements.

Part D: Operational Validation

- **D1:** Have experts been asked to judge the appropriateness of the model outcomes?

Yes. Several HTA bodies have judged that the model outcomes are appropriate. NICE (UK), SMC (Scotland), NCPE (Ireland), and CADTH have all reviewed the model itself and applications of the model as part of reimbursement submissions.

NICE Evidence Review Group, The Southampton Health Technology Assessment Centre (SHTAC) [6] concluded that:
“The economic model captures all important aspects of the disease pathway and extrapolates intermediate outcomes to final outcomes in a robust and consistent manner, drawing upon standard sources from the literature. The model has been very well validated against external data” (p. 14-15)

“Overall the ERG considers that the model is internally consistent and very well-validated, and that the results of the model make intuitive sense” (p. 120)

Modeling guidelines were followed in the construction of ECHO-T2DM, including the decision of what outcomes to report. ECHO-T2DM reports a comprehensive set of outcomes that includes the standard set expected of DM models plus many more to enable a full examination of model results. By providing this level of transparency, we enable all concerned experts to assess appropriateness of the model for a broad range of outcomes, including for example,

- Summary measures of health economic endpoints, including, costs, LYS, survival, QALYs, ICERs, NNT, NMB, CEACs, scatterplots
- Cumulative incidence of individual micro- and macrovascular complications, survival
- Simulated biomarker evolution over time (e.g., HbA1c, SBP, BMI, cholesterol, eGFR, heart rate)
- Treatment use over time, average time until treatment failure
- Adverse event rates

Adequacy of outcomes reporting was a key topic also of the Mt. Hood Challenge 8 (2016). The conference consisted of two challenges:

1. Transparency: Consisted of replicating two studies (Baxter et al, 2016 [16], and UKPDS 72 [17]) based only on publicly available information. The purpose was to determine what is required for replicability and ultimately, collectively among the modeling groups, develop reporting guidelines on what should be included in published articles to increase transparency (e.g., in terms of reported inputs, assumptions, settings)

2. Communicating outcomes: Each modeling group were given standard sets of different baseline patient characteristics (cohorts) and pre-defined settings and were asked to run these cohorts in their model, and generate outcomes in form of life-expectancy. The purpose of this challenge was to enable comparison of model outcomes based on the same inputs)

- Modeling groups presented their results, which were then compared and contrasted by all groups and public
- A consensus statement on how to report outcomes of modelling in diabetes is in preparation.

**D2: Have the model outcomes been compared to the outcomes of other models that address similar problems?**

Yes, ECHO-T2DM has been routinely compared and contrasted with other simulation models of diabetes. For example, ECHO-T2DM has been represented at the four most recent Mt. Hood Challenges (see response to A2). At Mt. Hood Challenge 5 (2010), 8 different models of diabetes participated in at least one of those Challenges.

- “The general consensus on the results presented at the Fifth Mount Hood Challenge was that, in general, the models performed reasonably well in terms of predicting the relative risk of interventions versus control treatments” [10] (p. 679)

We have also carried out a cross-validation vs. the NIH model and presented results in [18]. We have updated this and the results are included in the manuscript under submission.

NICE review of canagliflozin (ECHO-T2DM), empagliflozin (UKPDS-OM1), and dapagliflozin (Cardiff diabetes model) comparing to their AG analysis (UKPDS-OM1) in 2016 [2] (p. 37):
• “The committee acknowledged there were different advantages and disadvantages to the different modelling approaches, and it agreed with the AG that all models submitted were appropriate and of a reasonable quality”

• **D3:** Have the model outcomes been compared to the outcomes obtained when using alternative input data?
  
  Yes. In the manuscript under submission, validation results were generated for each of the four CVD risk equations supported in ECHO-T2DM (UKPDS 82 [19], UKPDS 68 [20], ADVANCE [21], and Swedish NDR [22]) and the results were compared and contrasted.

• **D4:** Have the model outcomes been compared to empirical data?
  
  Yes, formal external validation of ECHO-T2DM against empirical data is the explicit subject of the manuscript under submission. The list of studies that were used for the comparisons as part of dependent and independent validation is found in Supplementary Table 6.

  In addition, NICE Evidence Review Group (ERG) conducted by Southampton Health Technology Assessment Centre (SHTAC) independently concluded for the previous version of ECHO-T2DM that:

  o “The model has been very well validated against external data” (p. 14-15)
  o “Overall the ERG considers that the model is internally consistent and very well-validated” (p. 120) [6]

  Additionally, comparison of model results to empirical data is a common theme of the Mt. Hood Challenges. In particular, ECHO-T2DM participated in the Fifth Mt Hood conference, which found that:

  o “The general consensus on the results presented at the Fifth Mount Hood Challenge was that, in general, the models performed reasonably well in terms of predicting the relative risk of interventions versus control treatments” [10] (p. 679)

  As part of ECHO-T2DM participation at the Mt. Hood Challenge 6 (2012), where the tasks were to predict CVD and mortality patterns observed in naturalistic data (Kaiser Permanente Northwest [KPNW] in the US and the Swedish National Diabetes Registry [NDR] in Sweden) and an RCT (sub-groups in ADVANCE), the results of ECHO-T2DM simulations were publicly compared to these empirical findings.

  As part of ECHO-T2DM participation at the Mt. Hood Challenge 7 (2014), where the tasks were to (1) replicate the Look AHEAD study, and (2) replicate mortality after various CVD events from an observational study [12], the results of ECHO-T2DM simulations were publicly compared to these empirical findings.

**Part E: Other Validation Techniques**

• **E1:** Have any other validation techniques been performed?
  
  The formal validation of ECHO-T2DM described above is comprehensive, including examination of face validity, debugging and stress testing, cross-validation, and dependent and independent validation. Back-of-the-envelope calculations is regularly used to evaluate the face validity of model updates, and double programming is used to double-check the ECHO-T2DM implementation of risk equations.
References


2. NICE. Final appraisal determination - Canagliflozin, dapagliflozin and empagliflozin as monotherapies for treating type 2 diabetes National Institute for Health and Care Excellence; 2016.


4. SMC. Canagliflozin, 100mg and 300mg film-coated tablets (Invokana). Scotland: Scottish Medicines Consortium; 2014 09 May.

5. NCPE. Cost Effectiveness of canagliflozin (Invokana®) for adults with type 2 diabetes mellitus to improve glycaemic control as monotherapy or add-on therapy with other anti-hyperglycaemic agents including insulin, when these, together with diet and exercise, do not provide adequate glycaemic control. 2014; Available from: http://www.ncpe.ie/wp-content/uploads/2013/04/Summary-Canagliflozin2.pdf


15. Akpo E, Neslusan C, Joris K. An audit approach to the validation of health economics simulation models: The case of the Economic and Health Outcomes Model for Type 2 Diabetes Mellitus (ECHO-T2DM) - Manuscript in preparation2016.


Willis, M., C. Asseburg, J. He, and C. Neslusan. "Validation of Economic Simulation Models Using Clinical Trial Results: Experiences from Type 2 Diabetes Mellitus (T2DM)." Podium presentation at iHEA World Congress. Toronto, Canada. 10-13 July. 2011.


Willis, M., C. Asseburg, and J. He. 2013. 'Validation of economic and health outcomes simulation model of type 2 diabetes mellitus (ECHO-T2DM)', J Med Econ, 16: 1007-21.


Bacon, T., M. Willis, P. Johansen, C. Neslusan, S. Nuhoho, and M. Worbes-Cerezo. "The Cost-Effectiveness of Canagliflozin Versus Insulin Secretagogues (Sulphonylureas) or Insulin in Patients with Type 2 Diabetes Mellitus as Add-on to Metformin in Ireland." Poster


Neslusan, Cheryl, Anna Teschemaker, Pierre Johansen, Michael Willis, Atanacio Valencia-Mendoza, and Andrea Puig. 2015. 'Cost-Effectiveness of Canagliflozin versus Sitagliptin as add-on to Metformin in patients with type 2 diabetes mellitus in Mexico', Value in Health Regional Issues, 8: 8-19.


