AdViSHE

Assessment of the Validation Status of Health-Economic decision models

AdViSHE is a questionnaire that modellers can complete to report on the efforts performed to improve the validation status of their health-economic (HE) decision model. It is not intended to replace validation by model users but rather to inform the direction of validation efforts and to provide a baseline for replication of the results. In addition to using it after a model is finished, the modellers can use AdViSHE to guide validation efforts during the modelling process.

The modellers are asked to comment on the validation efforts performed while building the underlying HE decision model and afterwards. Many of the questions simply refer to the model documentation. AdViSHE is divided into five parts, each covering an aspect of validation:

- Part A: Validation of the conceptual model (2 questions)
- Part B: Input data validation (2 questions)
- Part C: Validation of the computerized model (4 questions)
- Part D: Operational validation (4 questions)
- Part E: Other validation techniques (1 question)

No final validation score is calculated, as the assessment of the answers and the overall validation effort is left to the model users. It is assumed that the model has been built according to prevailing modelling and reporting guidelines. For instance, the model builders would presumably adhere to the ISPOR-SMDM[†] Modeling Good Research Practices (Caro et al., 2010) and/or CHEERS[†] Statement (Husereau et al., 2013). Some questions may not be applicable to a particular model. If this is the case, the model builder should take the opt-out option and provide a justification of why this item is not deemed applicable.

Part A: Validation of the conceptual model (2 questions)

Part A discusses techniques for validating the conceptual model. A conceptual model describes the underlying system (e.g., progression of disease) using a mathematical, logical, verbal, or graphical representation. Please indicate where the conceptual model and its underlying assumptions are described and justified.

The conceptual model and its underlying assumptions are described and justified in the section 'Model structure'.

[†]: ISPOR: International Society For Pharmacoeconomics and Outcomes Research, SMDM: Society for Medical Decision making, CHEERS: Consolidated Health Economic Evaluation Reporting Standards

A1/ Face validity testing (conceptual model): Have experts been asked to judge the appropriateness of the conceptual model?
If yes, please provide information on the following aspects:
- Who are these experts?
- What is your justification for considering them experts?
- To what extent do they agree that the conceptual model is appropriate?

If no, please indicate why not.

The conceptual model has been developed and discussed with 4 clinical experts: members of the subcommittee on inflammatory bowl diseases (IBD) of the Dutch association of gastroenterology (NVMDL). These experts are [redacted] Their approval of the conceptual model has been logged in meeting notes. The experts have supplied access to the most appropriate registries for time to active disease, quality of life and costs in The Netherlands.

Aspects to judge include: appropriateness to represent the underlying clinical process/disease (disease stages, physiological processes, etc.); and appropriateness for economic evaluation (comparators, perspective, costs covered, etc.).

A2/ Cross validity testing (conceptual model): Has this model been compared to other conceptual models found in the literature or clinical textbooks?
If yes, please indicate where this comparison is reported.
If no, please indicate why not.

A systematic literature search has been conducted (see appendix 1 of the manuscript for search terms). The search yielded 502 results, 153 full text inclusions, 9 treatment sequence models, 8 surgery models, 12 diagnostic models, 8 biomarker models, 57 pharmacotherapeutic models and 59 'other' models.
The treatment sequence models did not incorporate long term real world data, and often included a health state for 'response' which was not deemed appropriate for this cost-effectiveness model as there is no consistent reporting of the 'response' outcome across trials and clinical practice does not consider 'response' a treatment goal, resulting in treatment switches within a year if 'remission' is not achieved. Generally, health states 'remission', 'active disease', and in some instances 'remission due to surgery' are common across cost-effectiveness models for Crohn's disease.

Part B: Input data validation (2 questions)

Part B discusses techniques to validate the data serving as input in the model. These techniques are applicable to all types of models commonly used in HE modelling.

Please indicate where the description and justification of the following aspects are given:

- search strategy;
- data sources, including descriptive statistics;
- reasons for inclusion of these data sources;
- reasons for exclusion of other available data sources;
- assumptions that have been made to assign values to parameters for which no data was available;
- distributions and parameters to represent uncertainty;
- data adjustments: mathematical transformations (e.g., logarithms, squares); treatment of outliers; treatment of missing data; data synthesis (indirect treatment comparison, network meta-analysis); calibration; etc.

Search strategy: see appendix of the manuscript
Data sources:
Efficacy:
Methods, Patient population (IBD south limburg registry)
Methods, Treatment efficacy and side-effects: meta-analysis and network meta-analysis
Methods, Background probabilities for loss of remission (IBD & ICC registries)
Appendix 3 (NMA results) and appendix 4 (survival analysis)
Costs:
Methods, Costs (COIN study, IBD-WORK study)
Quality of life:
Methods, Utilities (IB-Dream registry) + appendix 6

Mortality
Methods: mortality (Statistics Netherlands & HR for excess mortality

B1/ Face validity testing (input data): Have experts been asked to judge the appropriateness of the input data?
If yes, please provide information on the following aspects:
- Who are these experts?
- What is your justification for considering them experts?
- To what extent do they agree that appropriate data has been used?

If no, please indicate why not.

Input data has been selected and discussed with 4 clinical experts: members of the subcommittee on inflammatory bowl diseases (IBD) of the Dutch association of gastroenterology (NVMDL). These experts are [redacted]. The experts have suggested the registry data utilized in the model for long-term efficacy, costs and quality of life and have co-developed (and are co-authors on) the network meta-analysis that informs the efficacy estimates. The experts have judged the input data to be appropriate, but acknowledge that modelling requires a simplification of reality and that a cost-effectiveness model cannot predict outcomes for an individual patient.

Aspects to judge may include but are not limited to: potential for bias; generalizability to the target population; availability of alternative data sources; any adjustments made to the data.

B2/ Model fit testing: When input parameters are based on regression models, have statistical tests been performed?
If yes, please indicate where the description, the justification and the outcomes of these tests are reported.
If no, please indicate why not.

Survival models were selected based on a combination of expert opinion and model fit (AIC & BIC), see appendix. See results, background probabilties
Utility regression models were assessed on mean absolute error. See results, utilities and appendix 6.

Examples of regression models include but are not limited to: disease progression based on survival curves; risk profiles using regression analysis on a cohort; local cost estimates based on multi-level models; meta-regression; quality-of-life weights estimated using discrete choice analysis; mapping of disease-specific quality-of-life weights to utility values.
Examples of tests include but are not limited to: comparing model fit parameters (R^2, Akaike information criterion (AIC), Bayesian information criterion (BIC)); comparing alternative model specifications (covariates, distributional assumptions); comparing alternative distributions for survival curves (Weibull, lognormal, logit); testing the numerical stability of the outcomes (sufficient number of iterations); testing the convergence of the regression model; visually testing model fit and/or regression residuals.

Part C: Validation of the computerized model (4 questions)

Part C discusses various techniques for validating the model as it is implemented in a software program. If there are any differences between the conceptual model (Part A) and the final computerized model, please indicate where these differences are reported and justified.

C1/ External review: Has the computerized model been examined by modelling experts?
If yes, please provide information on the following aspects:
- Who are these experts?
- What is your justification for considering them experts?
- Can these experts be qualified as independent?
- Please indicate where the results of this review are reported, including a discussion of any unresolved issues.

If no, please indicate why not.

The TECH-VER protocol has been executed by an external review group, led by dr Talitha Feenstra of Groningen University. Minor coding errors were identified and adjusted in the model following this validation

Aspects to judge may include but are not limited to: absence of apparent bugs; logical code structure optimized for speed and accuracy; appropriate translation of the conceptual model.

C2/ Extreme value testing: Has the model been run for specific, extreme sets of parameter values in order to detect any coding errors?
If yes, please indicate where these tests and their outcomes are reported.
If no, please indicate why not.

Included in extensive TECH-VER testing of the model (see C1)

Examples include but are not limited to: zero and extremely high (background) mortality; extremely beneficial, extremely detrimental, or no treatment effect; zero or extremely high treatment or healthcare costs.

C3/ Testing of traces: Have patients been tracked through the model to determine whether its logic is correct?
If yes, please indicate where these tests and their outcomes are reported.
If no, please indicate why not.

Yes, visually, see appendix 10. Time on treatment has been presented to the four participating gastroenterologists and was considered credible

In cohort models, this would involve listing the number of patients in each disease stage at one, several, or all time points (e.g., Markov traces). In individual patient simulation models, this would involve following several patients throughout their natural disease progression.

C4/ Unit testing: Have individual sub-modules of the computerized model been tested?
If yes, please provide information on the following aspects:
- Was a protocol that describes the tests, criteria, and acceptance norms defined beforehand?
- Please indicate where these tests and their outcomes are reported.

If no, please indicate why not.

The Darth modelling framework was used of which the non-user-written functions contain unit testing

Examples include but are not limited to: turning sub-modules of the program on and off; altering global parameters; testing messages (e.g., warning against illegal or illogical inputs), drop-down menus, named areas, switches, labelling, formulas and macros; removing redundant elements.

Part D: Operational validation (4 questions)

Part D discusses techniques used to validate the model outcomes.

D1/ Face validity testing (model outcomes): Have experts been asked to judge the appropriateness of the model outcomes?
If yes, please provide information on the following aspects:
- Who are these experts?
- What is your justification for considering them experts?
- To what extent did they conclude that the model outcomes are reasonable?

If no, please indicate why not.

All model results were presented to four expert gastroenterologists in a meeting for which meeting notes are available. Both input and output of the model was presented. The model was considered valid and the time on treatment lines plausible.

Outcomes may include but are not limited to: (quality-adjusted) life years; deaths; hospitalizations; total costs.

D2/ Cross validation testing (model outcomes): Have the model outcomes been compared to the outcomes of other models that address similar problems?
If yes, please provide information on the following aspects:
- Are these comparisons based on published outcomes only, or did you have access to the alternative model?
- Can the differences in outcomes between your model and other models be explained?
- Please indicate where this comparison is reported, including a discussion of the comparability with your model.

If no, please indicate why not.

Described in 'Discussion'. Treatment sequence models in the literature have somewhat smaller QALY gains between the best and worst option than our model, but they also have higher discount rates and a shorter time horizon, which explains these differences.

Other models may include models that describe the same disease, the same intervention, and/or the same population.

D3/ Validation against outcomes using alternative input data: Have the model outcomes been compared to the outcomes obtained when using alternative input data?
If yes, please indicate where these tests and their outcomes are reported.
If no, please indicate why not.

No, there are no other high quality background probability data available for The Netherlands

Alternative input data can be obtained by using different literature sources or datasets, but can also be constructed by splitting the original data set in two parts, and using one part to calculate the model outcomes and the other part to validate against.

D4/ Validation against empirical data: Have the model outcomes been compared to empirical data?
If yes, please provide information on the following aspects:
- Are these comparisons based on summary statistics, or patient-level datasets?
- Have you been able to explain any difference between the model outcomes and empirical data?
- Please indicate where this comparison is reported.

If no, please indicate why not.

D4.A/ Comparison against the data sources on which the model is based (dependent validation).

No other input registry data has been used in the model

D4.B/ Comparison against a data source that was not used to build the model (independent validation).

Yes, see section 'validation' for validation against claims data and a clinical study. There were very limited differences between the modelled time on line 2 and the observed time on line 2 in claims data.

Part E: Other validation techniques (1 question)

E1/ Other validation techniques: Have any other validation techniques been performed?
If yes, indicate where the application and outcomes are reported, or else provide a short summary here.

Double (4-eyes) programming + TECH-VER protocol by external validation group

Examples of other validation techniques: structured "walk-throughs" (guiding others through the conceptual model or computerized program step-by-step); naïve benchmarking ("back-of-the-envelope" calculations); heterogeneity tests; double programming (two model developers program components independently and/or the model is programmed in two different software packages to determine if the same results are obtained).