International Journal of Technology Assessment in Health Care

cambridge.org/thc

Method

Cite this article: Radeva D, Hopkin G, Mossialos E, Borrill J, Osipenko L, Naci H (2020). Assessment of technical errors and validation processes in economic models submitted by the company for NICE technology appraisals. *International Journal of Technology Assessment in Health Care* **36**, 311–316. https://doi.org/10.1017/S0266462320000422

Received: 31 July 2019
Revised: 3 May 2020
Accepted: 1 June 2020
First published online: 3 July 2020

Key words:
Cost-effectiveness analysis; Errors; Modeling; Health technology assessment; Economic models; Validation

Author for correspondence:
Gareth Hopkin,
E-mail: ghopkin@ihe.ca

Assessment of technical errors and validation processes in economic models submitted by the company for NICE technology appraisals

Demi Radeva[1,2], Gareth Hopkin[1,3], Elias Mossialos[1], John Borrill[4], Leeza Osipenko[1] and Huseyin Naci[1]

[1]Department of Health Policy, London School of Economics and Political Science, London, UK; [2]United Health Group, Eden Prairie, Minnesota, USA; [3]Institute of Health Economics, Edmonton, Alberta, Canada and [4]Bristol-Myers Squibb, Uxbridge, London, UK

Background. Economic models play a central role in the decision-making process of the National Institute for Health and Care Excellence (NICE). Inadequate validation methods allow for errors to be included in economic models. These errors may alter the final recommendations and have a significant impact on outcomes for stakeholders.
Objective. To describe the patterns of technical errors found in NICE submissions and to provide an insight into the validation exercises carried out by the companies prior to submission.
Methods. All forty-one single technology appraisals (STAs) completed in 2017 by NICE were reviewed and all were on medicines. The frequency of errors and information on their type, magnitude, and impact was extracted from publicly available NICE documentation along with the details of model validation methods used.
Results. Two STAs (5 percent) had no reported errors, nineteen (46 percent) had between one and four errors, sixteen (39 percent) had between five and nine errors, and four (10 percent) had more than ten errors. The most common errors were transcription errors (29 percent), logic errors (29 percent), and computational errors (25 percent). All STAs went through at least one type of validation. Moreover, errors that were notable enough were reported in the final appraisal document (FAD) in eight (20 percent) of the STAs assessed but each of these eight STAs received positive recommendations.
Conclusions. Technical errors are common in the economic models submitted to NICE. Some errors were considered important enough to be reported in the FAD. Improvements are needed in the model development process to ensure technical errors are kept to a minimum.

Health technology assessment (HTA) agencies across Europe provide recommendations to support payer and prescriber decisions on the adoption, reimbursement, and use of therapeutic agents and devices (1). In England and Wales, the National Institute for Health and Care Excellence (NICE) is responsible for assessing new and existing medical technologies from both a health benefit and economic perspective. Ultimately, NICE makes recommendations that guide the National Health Service (NHS) coverage and reimbursement decisions across different disease areas. The NICE single technology appraisal (STA) process was introduced in early 2005 as a mechanism to provide a prompt appraisal of new healthcare technologies. This process aimed to better align the STA timelines with those adopted by the European Medicines Agency to allow people in England and Wales to have faster access to the most cost-effective treatments (1;2).

NICE relies on economic models, also known as decision analytic models, to inform its funding recommendations. These models use an explicit mathematical framework which represents clinical decision problems and incorporates evidence from a variety of sources to estimate costs and health outcomes(s) of the interventions under appraisal (2). For the STA process, the models are developed by the company or a consulting organization subcontracted by the company. Generally, models are built in Microsoft Excel, however the use of other software tools is allowed (e.g. R, WinBugs) (2). Independent evidence review groups (ERGs), based at academic centers and commissioned by NICE, are responsible for assessing and critiquing the companies' models (2). The robustness and credibility of the economic model and its results are dependent on a number of factors including whether the structure adequately reflects the underlying disease process, the best available evidence has been used to inform the model, and whether the model is computationally accurate (2).

Model validity can be classified in many ways; however, key elements include face validity, external validity, and technical or internal validity (3). Face validity relates to the validity of the model concept and technique used. This must all be consistent with best practices established as related to the modeling for a particular disease and its treatment. External validity

https://doi.org/10.1017/S0266462320000422 Published online by Cambridge University Press

determines whether or not the model correctly reflects reality and technical or internal validity ensures that the model is doing what it is intended to do, including that the logic is properly implemented, with an absence of errors. The term "technical errors" comes from this final type of validity. Our study focuses on technical errors which are more readily quantifiable whilst face validity and external validity are more subjective. These errors can therefore be better operationalized. This approach is in line with previous discussion about how to treat validation of economic models in the absence of prescriptive guidelines for model development (4). While errors are an inevitable part of the model development process, there is a need for the submitting company to eliminate all errors from their final submission and for academic ERGs to identify and correct any remaining errors throughout the review process. To do so, both parties could ensure validation strategies are in place to identify and avoid errors but it is not clear to what extent this is currently the case.

Across several studies, it has been estimated that up to 94 percent of large spreadsheet models have at least one technical error (5). Whilst this study was not specific to health economic evaluation, other works suggested that similar error rates are present in models submitted to HTA agencies. More specifically, research examining the quality of models submitted to the Australia's Pharmaceutical Benefit Advisory Committee (PBAC) in 2000 reported 37 percent of models had major flaws in technical aspects of the model (6). The study was replicated at a later date in 2008 and the analysis found 83 percent of models reviewed by PBAC were "flawed in some respect" suggesting an increase over time (7). For England and Wales, Trueman and Livings (8) aimed to estimate the incidence of technical errors in economic evaluations submitted by companies and appraised by the ERGs. Over an unspecified time period, they report errors in 39 of 102 (38 percent) STAs with a total of forty-seven errors. However, their evaluation was limited to information recorded in the committee minutes and did not include an assessment of ERG reports and other publicly available documents.

To our knowledge, there are no recent systematic studies that evaluated the technical errors identified in economic models developed for NICE STAs. Our primary objectives were to quantify the frequency, type, and implications of technical model errors found by ERGs. We also considered it important to examine whether models have undergone some form of validation by manufacturers, and therefore, a second objective was to identify variation in types of validation methods present in the economic models used in the STAs.

Methods

All NICE appraisals completed during 2017 were identified on the NICE Web site. Out of these appraisals, we excluded a total of thirty-eight, for the following reasons: terminated appraisals ($n = 12$), multiple technology assessments ($n = 10$), any appraisals that were not originally published in 2017 ($n = 9$), cancer drugs fund rapid reconsiderations ($n = 4$), rapid reviews ($n = 1$), fast track ($n = 1$), or any technology which was withdrawn from assessment by the submitting company prior to a recommendation ($n = 1$). The remaining forty-one STAs were included in our study (full details of included STAs are available in Supplementary Material). All of the STAs were on medicines. For each of the forty-one assessed STAs, all available documentation, including appraisal consultation documents (e.g. public committee slides, committee papers, and notes), final appraisal document (FAD), and any other applicable documentation (e.g. company submission, ERG report, factual accuracy check), was retrieved. NICE provided missing documentation that could not be found on their Web site (e.g. TA457). Our analysis focused on 2017 which was the latest year of data available at initiation of the project. For this single year, we found and reviewed over 300 publicly available documents, including 41 ERG reports, 41 FADs, and multiple sets of public committee slides. This single-year focus allowed us to conduct a systematic, in-depth evaluation of available documents, all of the ERGs ($n = 10$), a variety of companies, and all of the appraisal committees ($n = 4$).

The following information was extracted from each assessed STA: the nonproprietary and brand name of the product being appraised, company name, indication, disease category, ERG name, model type, model validation activities reported by the company, the number, type, and magnitude of errors, and whether errors were reported in the FAD and the final NICE recommendation. The ERG reports were read in full for all forty-one STAs and data were extracted by one investigator (DR) while a key word search was performed on all other available documentation (e.g. committee notes from first, second, third meetings, FAD) to help identify additional technical errors discovered during the STA appraisal process. In order to identify which words should be used in this key word search, a sample of ten STAs, randomly selected using randomizer software in Excel, was used to assess the terms commonly used to describe errors. The final list of key words included error(s), wrong, incorrect, discrepancy, discrepancies, inconsistency, inconsistencies, omitted, omission, mistake(s), problem(s), and flaw(s).

In this study, a technical error was defined as an error which results from the actions of a modeler during the development process that is objectively incorrect. We did not consider differences in opinion about the validity of assumptions underlying the model as errors even if the ERG altered these assumptions due to the subjective nature of these decisions. While ERGs would often identify errors in the models, it is rare that they would label them with a specific type (e.g. logic, transcription). We classified the type of each error based on all available information about the nature of the error, its causes, and its impact on the model. The errors were characterized as follows: computational, logic, data handling, transcription, interpretation, other, or unknown. A description of each error type and examples from the data extraction are presented in Table 1. This classification was in line with previous studies on errors in HTA modelling (6;8). We also categorized technical errors as either minor or major in accordance with the ERGs definition of each type of error. If the specific terminology—"major" or "minor"—was not used in the documentation but the ERG had made a clear statement about the impact of the error on the incremental cost-effectiveness ratio (ICER), this information was used to determine the magnitude of the error. In cases where ERG comments were not clear, the severity of errors was recorded as "not reported." The presence, type, and magnitude of errors were determined by one investigator (DR) through a review of all publicly available information. If the nature, type, or magnitude of the errors were unclear, then other team members were consulted and a consensus was reached on how the error should be coded.

Information on validation efforts by companies was extracted in a similar fashion. The presence and nature of validation steps undertaken were independently determined by one investigator (DR). While there is no standard guidance for model validation, for the purposes of this study, a taxonomy of validation types

https://doi.org/10.1017/S0266462320000422 Published online by Cambridge University Press

Table 1. Description and examples of error types

Error types	Description
Computation[a]	Programming errors which are objectively incorrect (e.g. failure to anchor cells correctly before application to column of data)
Logic[a]	Errors in the context of health economic evaluation (e.g. illogical crossing of time to event curves such that a higher proportion of people are disease free than were initially alive)
Data handling[a]	Failure to handle the data appropriately (e.g. interquartile ranges being used to calculate standard errors)
Transcription[a]	Typographical errors occur either when transferring data between sources and model, or model and submission document
Interpretation	Failure to interpret the data appropriately (e.g. the direction of an odds ratio from a meta-analysis was incorrectly interpreted during translation into a treatment effect within a model)
Unknown[a]	Error causes could not be determined
Other	Errors that do not fall in the above classification (e.g. incorrect text citation provided, preventing the ERG from verifying the evidence supporting model assumptions)

[a]Informed by Chilcott et al. and Trueman and Livings (7;8).

was constructed based on previous research. Validation practices were classified as one of the following: the use of auditing software, face validity, model behavior, internal consistency, external consistency, cell-by-cell checks, internal peer review, external peer review, internal double programming, external double programming, cross-check inputs, cross-check outputs, clinical advisory panel, economics advisory panel, and technical validation. Descriptions of each validation type from the data extraction are given in Table 2. To stay consistent, the validation types were named and defined in line with research by Kim and Thompson, Chilcott et al., and Trueman and Livings (3;7;8). As with errors, the classification of validation efforts was independently determined by one investigator and if the classification of an error was unclear, then another member of the research team was consulted.

Results

In terms of frequency of errors, the total number of errors across the forty-one STAs was 198. Only two (5 percent) of the reviewed STAs had no reported errors. Out of the forty-one STAs, nineteen (46 percent) STAs had between one and four errors and sixteen (39 percent) had between five and nine errors, four (10 percent) had more than ten errors. The most common type of errors included transcription ($n = 58$; 29 percent) and logic ($n = 58$; 29 percent) followed by computation ($n = 50$; 24 percent) and data handling ($n = 22$; 21 percent). Errors were reported in the FAD for eight (20 percent) STAs with one error per assessment.

In terms of the severity of errors, the ERGs listed only forty-three (22 percent) as minor errors and nine (5 percent) as major. The remaining 73 percent of errors were not classified as minor or major by ERGs and not enough information about the errors was provided for the authors to make this judgement. These gaps in information stem from the lack of requirement for ERGs to formally list or classify all errors they find according to their impact. For example, in TA475, it is reported by the ERG that “the company model suggests it applies a trial period of 16 weeks but due to a coding error it applies the secukinumab trial period duration of 12 weeks,” and in TA489, it is reported by the ERG that “the cost of a GP visit … uses the cost of a dermatologist visit instead of a GP visit.” Neither of these examples, like the majority of descriptions of errors, included a classification of the significance of the error and thus, could not be assessed by the research team.

All forty-one STAs underwent some type of model validation. Only five STAs had one ($n = 1$) type of validation, eighteen TAs had more than one and less than four, and eighteen TAs had between five and eight types of validation types conducted. The most common validation methods used in the forty-one STAs included external (12 percent, $n = 18$) and internal consistency (9 percent, $n = 13$), cross-checked inputs (9 percent, $n = 13$) and outputs (11 percent, $n = 16$), model behavior (9 percent, $n = 14$), and a clinical advisory panel (10 percent, $n = 15$). The least used validation methods were internal (1 percent, $n = 1$) and external double programming (1 percent, $n = 1$). Only eleven of the forty-one STAs explicitly stated that they had used a checklist as a validation method. Checklist types included Tappenden and Chilcott ($n = 11$), Philips et al. ($n = 10$), Drummond and Jefferson ($n = 9$), and a few were listed as general or unspecified. Ten different ERGs were responsible for assessing the company submissions. No clear relationship between the use of validation methods and occurrence of technical errors could be established due to all models using some form of validation and the wide variety and combination of approaches used.

Only three of the forty-one STAs were not recommended for reimbursement. The main reasoning behind negative recommendations by NICE included uncertainty in the modeling assumptions, concerns with the use of surrogate outcomes, and conclusions that technologies did not provide value for money according to established thresholds. All STAs with errors reported in the FAD ($n = 8$), regardless if they were classified as “major” or “minor” by the ERGs, received a positive recommendation. None of the STAs that received a negative recommendation included reasoning explicitly linked to errors but errors may have contributed to an uncertainty regarding modeling and value.

With regards to other characteristics, a total of twenty-six companies were responsible for the forty-one STA submissions with Bristol Myers-Squibb ($n = 4$), Amgen ($n = 3$), Eli Lilly ($n = 3$), Roche ($n = 3$), and Janssen ($n = 3$) with the most submissions per company. There were a variety of disease areas represented in the STAs reviewed, including cancer ($n = 25$), blood and immune system ($n = 6$), digestive system ($n = 3$), respiratory

Table 2. Description of validation types

Validation types	Description
Model auditing tools[a]	Auditing software such as Operis Analysis Kit can simplify some aspects of quality control, allowing the analyst to focus on more challenging aspects
Face validity[a]	Testing model behavior meets expectations or simple "back-of-envelope" calculations and inclusion of key features of the disease and intervention[b]
Model behavior[a]	Check whether parameter changes in model have appropriate effect on outcomes (e.g. stress testing, extreme value testing, scenario analysis)
Internal consistency[a]	Also known as internal validation, it compares model outputs with data used in the model-building process[b]
External consistency[a]	Also known as prospective validation or external validation, this type of validation uses extended follow-up data from the studies that informed the model or utilizes data sources not used in the model-building process, respectively[b]
Cell-by-cell checks[a]	Usually performed by the original analyst
Internal peer review[a]	Testing performed by an analyst not external to the model development process
External peer review[a]	Testing performed by analyst external to model development
Internal double programming [a]	Re-programming of model by original analyst
External double programming[a]	Testing performed by analyst external to model development
Cross-check inputs[a]	Cross-check of model inputs and data sources
Cross-check outputs[a]	Cross-check model outputs and corresponding documents and reports
Clinical advisory panel	Clinical experts consulted on methods and inputs used (usually in the form of workshops and ad-hoc consultations)
Economics advisory panel	Economics experts consulted on methods and inputs used (usually in the form of workshops and ad-hoc consultations)
Technical validation	Internal quality control and validation performed by external consultancy

[a]Informed by Chilcott et al. and Trueman and Livings (7;8).
[b]Informed by research from Kim and Thompson (3).

($n = 3$), central nervous system ($n = 1$), eye ($n = 1$), endocrine system ($n = 1$), and infectious diseases ($n = 1$). The most commonly used models were Markov ($n = 14$), Partitioned ($n = 13$), a combination of two or more models ($n = 5$), Semi-Markov ($n = 3$), Discrete Event Simulation (DES) ($n = 3$), and others ($n = 3$).

From a disease area perspective, cancer appraisals had the highest total number appraisals with an average of 4.7 errors per appraisal. Other disease areas had limited counts but had the following average number of errors per appraisal: digestive system (mean 7.6), infectious (mean 7), respiratory (mean 6.7), blood and immune system (mean 4.6), central nervous system (mean 1), endocrine (mean 1), and eye (mean 1). In terms of errors reported in the FAD, we found one error per appraisal for each disease area apart from eye which had none.

In terms of model types, Markov models had 4.2 errors per appraisal and partitioned models had 3.2 per appraisal. Of the model types which were used less frequently, DES had 12 errors per appraisal which was the highest across the model types, other models had 7.3 per appraisal, and semi-Markov models had 7 errors per appraisal. Where more than one type of model was used within a submission, the average number of errors was 6.4.

Discussion

In this study, we found that the number of technical errors in economic models submitted by companies to NICE in STAs is high, and for 2017, all but one of the companies' models contained one or more errors. Of the forty-one STAs that were reviewed, there were a total of 198 errors identified with an average of 4.8 errors per submission. These findings suggest that errors are more common in STAs submitted to NICE than may have been suggested by earlier work (8) and are higher than has been reported for other international agencies (6;7). This high number of errors were present despite the widespread use of validation and suggests that current methods of validation are not adequately eliminating errors prior to submission. When the type of errors is examined, the majority are from transcription, data handling, and computation categories and these are errors that could be identified and corrected by in-depth validation methods. Our initial intent was to examine whether there was a relationship between methods of validation and the number of type of errors present in economic models but the range in type and combination of validation methods meant this was not possible.

The large number of errors seen in company submissions presents issues for appraisal processes. For ERGs, identifying and fixing errors within models is time consuming and resource intensive, and if a large number of errors are present, this may reduce a committee's confidence in the submission provided by a company or lead to extended timelines due to the need for additional consultations. This is an important set of findings, and stakeholders across HTA in England and more widely should consider how errors can be identified and eliminated prior to review within the HTA processes.

Alongside findings on the number and type of errors, there appears to be some evidence that some model types, like DES, may have a higher number of errors. Our ability to make conclusions on this is limited by the small sample size of the study but it is worth considering whether these types of computationally complex models are more prone to errors in design and coding (9).

https://doi.org/10.1017/S0266462320000422 Published online by Cambridge University Press

Also, it appears that disease areas with more complex natural histories, for example, for digestive or respiratory systems compared to cancer, may have a higher number of errors. Again, this should be interpreted with reference to the small number of cases within the disease area. For both model type and disease area, validation in these highlighted areas may need particular focus.

A key strength of this study is that for the year of the analysis, all publicly available documents were systematically reviewed to identify errors in models submitted to NICE as part of the technology appraisal process. The benefit from this in-depth and thorough review of all available documentation may be the reason that a higher number of errors were identified in this study and this was a key rationale for using this approach. Prior studies have relied on a more superficial review of a higher number of appraisals and this may have led them to underreport errors (8). The use of this approach does, however, introduce some limitations that should be considered. With a single year of data, we are not able to assess whether this level of errors is consistent over time or whether there are important trends across years in changing numbers and types of errors. In addition, the association of the number and type of errors with particular characteristics of the STAs could not be assessed due to small cell counts and a lack of inferential power once characteristics of the forty-one STAs were tabulated. For example, it would be valuable to know if companies have varying numbers or types of errors, whether there are varying levels of identification across ERGs or committees, or if other characteristics (e.g. proximity of base case ICER to thresholds) had systematic differences, but this was not possible.

Another limitation relates to our use of published documentation, as we had no access to submitted models or confidential information. This is problematic for several reasons. First, the frequency and type of errors shown in this paper reflect errors that were identified during review in the appraisal process. It is possible that some errors were not identified during review and the true number of errors within submission may be higher than is reported here. This may particularly be the case where errors are not prominent or do not impact the functionality and face validity of results. ERGs also grouped some similar errors in their descriptions rather than fully outlining each individual error and this supports the idea that the total number of errors is underreported here. Second, our approach meant we were reliant on the ERG description of an error for classification of the importance of an error and there is a high level of variability in reporting. As our results show, this means an assessment of impact could not be made for the majority of errors and this is problematic both for research and the process of review in the real world. Taken together, this limitation means our study may underreport the number of errors and may underestimate the frequency of major errors.

There were also some limitations related to other parts of our methodology. A single reviewer extracted information and reported errors from available documentation and the majority of errors were coded by this reviewer. Duplication of review may have been a preferable approach but significant problems related to this approach were addressed by consulting with a second reviewer if the appropriateness of including an error or its classification was unclear to the first reviewer.

Despite these limitations, our findings can provide several recommendations regarding validation and errors in submissions to NICE and can provide guidance on future research. Given the increasing complexity and computational requirements of models developed for NICE and the increased capability demands for the STA program, the need for companies to reduce the number of errors in submissions is evident. To this end, there have been a number of validation techniques that have been developed in recent years. Compared to earlier approaches, these newer techniques better capture the increasing complexity of economic modeling software used by companies in their STA submissions and they provide structured ways to validate models. In our study, one of the STAs noted the use of a validation technique that was developed in the 1980s, which no longer seems appropriate. Whilst older validation techniques can provide a strong base to assess some aspects of validity, they lack reference to modeling software which is used today and do not cover checking of program coding which is increasingly required. Thus, companies should transition to using structured approaches which capture the complexity of modern economic modeling and which are the most appropriate validation techniques for their given model. In addition, the scale of errors suggests that independent, confidential model review processes may be needed to ensure internal validity of models prior to HTA submission. If NICE or other HTA agencies were to encourage this or make this a requirement of submission, they could help standardize validation approaches that are used and increase transparency and speed of validation during ERG review.

Additionally, our findings suggest that guidance on the description and assessment of errors as minor or major may be useful for NICE and HTA agencies with similar systems. It was not possible to assess the magnitude of errors with current reporting for the majority of errors and identifying whether errors were minor or major could provide useful context to appraisal committees. It could also be useful to outline criteria for when an error should be reported in a FAD and provide more context for their inclusion. Few errors in FAD were highlighted as major, and in some cases, errors identified as minor were included in the FAD and it was not clear why their inclusion was deemed necessary. Further information on this would provide clarity on the importance of the error and the reasons for its inclusion in a FAD.

Finally, we recommend that further research extends this line of work and addresses outstanding questions that will help provide a more informed understanding of errors within NICE processes and HTA in other settings. Conducting reviews of additional years of NICE STAs would be able to capture whether there are trends in errors and validation over time and would also build a larger sample which could be used to test associations between errors and the characteristics of STAs. HTA agencies in other jurisdictions may also be interested in replicating this work for submissions of economic models to their own processes to confirm whether trends are present across settings.

Conclusions

Our findings demonstrate that despite widespread use of validation exercises, almost all economic models in STAs had errors, and in several STAs, these errors were significant enough to be reported in the FAD. Economic models have become an integral part of the modern decision-making process in healthcare policy (10). The frequency, magnitude, and severity of the errors found in such models submitted to NICE underscore the need for more rigorous systematic validation efforts. Consideration is needed of what role NICE should play in this move to standardized procedures for model validations and how to monitor the impact of changes.

Funding. No funding was received for this research.

https://doi.org/10.1017/S0266462320000422 Published online by Cambridge University Press

Conflict of interest. Leeza Osipenko and John Borrill were employed by the National Institute for Health and Care Excellence (NICE) at the time of project initiation, data collection and analysis, and initial drafting of the paper.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0266462320000422.

References

1. **Angelis A, Lange A, Kanavos P**. Using health technology assessment to assess the value of new medicines: Results of a systematic review and expert consultation across eight European countries. *Eur J Health Econ*. 2017;**19**:123–52.
2. **National Institute of Health and Care Excellence**. PMG19: Guide to the processes of technology appraisal. London, 2018.
3. **Kim LG, Thompson SG**. Uncertainty and validation of health economic decision models. *Health Econ*. 2010;**19**:43–55.
4. **McCabe C, Dixon S**. Testing the validity of cost-effectiveness models. *Pharmacoeconomics*. 2000;**17**:501–13.
5. **Panko RR.** What we don't know about spreadsheet errors. In: *The European Spreadsheet Risks Interest Group 16th Annual Conference. London: European Spreadsheet Risks Interest Group*, 2015.
6. **Hill S, Mitchell A, Henry D**. Problems with the interpretation of pharmacoeconomic analyses: A review of submissions to the Australian Pharmaceuticals Benefit Scheme. *J Am Med Assoc*. 2000;**283**:2116–21.
7. **Chilcott J, Tappenden P, Rawdin A, Johnson M, Kaltenthaler E, Paisley S, et al.** Avoiding and identifying errors in health technology assessment models: Qualitative study and methodological review. *Health Technol Assess*. 2010;**14**:1–107.
8. **Trueman D, Livings C**. Technical errors in cost-effectiveness models: Evidence from the single technology appraisal programme in England and Wales. *Value Health*. 2013;**16**:A592.
9. **Van Gestel A, Severens JL, Webers CAB, Beckers HJM, Jansonius NM, Schouten J**. Modelling complex treatment strategies: Construction and validation of a discrete event simulation model for glaucoma. *Value Health*. 2010;**13**:358–67.
10. **Philips Z, Ginnelly L, Sculpher M, Claxton K, Golder S, Riemsma R, et al.** Review of guidelines for good practice in decision-analytic modelling in health technology assessment. *Health Technol Assess* 2004;**8**:36.

https://doi.org/10.1017/S0266462320000422 Published online by Cambridge University Press