PharmacoEconomics (2019) 37:1391–1408
https://doi.org/10.1007/s40273-019-00844-y

ORIGINAL RESEARCH ARTICLE

TECH-VER: A Verification Checklist to Reduce Errors in Models and Improve Their Credibility

Nasuh C. Büyükkaramikli[1] · Maureen P. M. H. Rutten-van Mölken[1] · Johan L. Severens[1] · Maiwenn Al[1]

Published online: 8 November 2019

Abstract

Background In health economic literature, checklists or best practice recommendations on model validation/credibility always declare verification of the programmed model as a fundamental step, such as 'is the model implemented correctly and does the implementation accurately represent the conceptual model?' However, to date, little operational guidance for the model verification process has been given. In this study, we aimed to create an operational checklist for model users or reviewers to verify the technical implementation of health economic decision analytical models and document their verification efforts.

Methods Literature on model validation, verification, programming errors and credibility was reviewed systematically from scientific databases. An initial beta version of the checklist was developed based on the checklists/tests identified from the literature and from authors' previous modeling/appraisal experience. Next, the first draft checklist was presented to a number of health economists on several occasions and was tested on different models (built in different software, developed by different stakeholders, including drug manufacturers, consultancies or academia), each time leading to an update of the checklist and culminating in the final version of the TECHnical VERification (TECH-VER) checklist, introduced in this paper.

Results The TECH-VER necessitates a model reviewer (preferably independent), an executable and transparent model, its input sources, and detailed documentation (e.g. technical report/scientific paper) in which the conceptual model, its implementation, programmed model inputs, and results are reported. The TECH-VER checklist consists of five domains: (1) input calculations; (2) event-state (patient flow) calculations; (3) result calculations; (4) uncertainty analysis calculations; and (5) other overall checks (e.g. validity or interface). The first four domains reflect the verification of the components of a typical health economic model. For these domains, as a prerequisite of verification tests, the reviewer should identify the relevant calculations in the electronic model and assess the provided justifications for the methods used in the identified calculations. For this purpose, we recommend completeness/consistency checks. Afterwards, the verification tests can be conducted for the calculations in each of these stages by checking the correctness of the implementation of these calculations. For this purpose, the following type of tests are recommended in consecutive order: (i) black-box tests, i.e. checking if model calculations are in line with a priori expectations; (ii) white-box testing, i.e. going through the program code details line by line, or cell by cell (recommended for some crucial calculations and if there are some unexpected results from the black-box tests); and (iii) model replication/parallel programming (recommended only in certain situations, and if the issues related to the identified unexpected results from black-box tests could not be resolved through white-box testing). To reduce the time burden of model verification, we suggest a hierarchical order in tests i–iii, where going to the next step is necessary when the previous step fails.

Conclusions The TECH-VER checklist is a comprehensive checklist for the technical verification of decision analytical models, aiming to help identify model implementation errors and their root causes while improving the transparency and efficiency of the verification efforts. In addition to external reviews, we consider that the TECH-VER can be used as an internal training and quality control tool for new health economists, while developing their initial models. It is the authors' aim that the TECH-VER checklist transforms itself to an open-source living document, with possible future versions, or 'bolt-on' extensions for specific applications with additional 'fit-for-purpose' tests, as well as 'tips and tricks' and some demonstrative error examples. For this reason, the TECH-VER checklist and the list of black-box tests created in this paper and a few model verification examples is uploaded to an open access, online platform (github and the website of the institute), where other users will also be able to upload their original verification efforts and tests.

https://publons.com/researcher/3106588/maureen-rutten-van-molken/.

Extended author information available on the last page of the article

Key Points for Decision Makers

Model verification is an integral part of the model validation process and aims to ensure that the model calculates what it intends to calculate. The TECHnical VERification (TECH-VER) checklist aims to provide a documenting framework to improve efficiency in the model verification efforts and to help identify modeling errors and their root causes in a systematic way.

The TECH-VER checklist presents an objective, transparent way to assist the end user of the model in forming a judgment on the model's verification status, with the end-goal to improve trust from various stakeholders (e.g. patients, payers) in the model outcomes in health technology assessment decision-making processes.

1 Introduction

Economic evaluations are undertaken in several jurisdictions around the world to inform decision making regarding the reimbursement of particular medications and/or other interventions. These economic evaluations are often based on decision analytic models that synthesize evidence from different sources, link intermediate outcomes to final outcomes, extrapolate short-term evidence to longer-term outcomes, and make the effects/trade-offs surrounding the treatment choices, as well as the key assumptions, explicit.

As the common aphorism generally attributed to George Box suggests, all models are wrong, but some are useful. All models are wrong since, by definition, all involve simplification and mathematical abstraction of the complex reality, which leads to deviation from the reality. On the other hand, these simplifications enable us to understand complex phenomena in the reality that have not been or cannot be observed empirically. Henceforth, some models are useful and we need them in health economics just like in other scientific fields.

In order to use health economic models for decision making, their results should be trustable. Therefore, in the work by Caro et al. [1] and Tappenden and Chillcott [2], the importance of model credibility was emphasized. In both studies, the model credibility was defined as the notion, which determines the extent to which the results of a model can be trusted and the level of confidence that can be placed in the model during the decision-making process. Tappenden and Chillcott [2] developed a taxonomy, which segregates threats to model credibility into three broad categories: (1) unequivocal errors; (2) violations; and (3) matters of judgment. In the same paper, the taxonomized credibility threats were mapped across the main elements of the model development process (which can be seen in Fig. 1), and a range of suggested processes and techniques were listed for avoiding and identifying these credibility threats.

Well-documented model validation and verification processes do not guarantee the correctness of the results of a model, but might improve the model credibility among its users and decision makers. There is no consensus on the definitions of model validation and verification in the literature, however the most commonly accepted definition from the software/operations research literature (e.g. Sargent [3]) describe validation as the act of evaluating whether a model is a proper and sufficient representation of the system it is intended to represent, in view of a specific application. On the other hand, verification focuses on whether a model is implemented correctly and assures that the model implementation (i.e. programming in the relevant software platform) accurately represents the conceptual model description [3]. Following from these definitions, it can be interpreted that the verification of a model should be considered as a prerequisite, as well as a constituent, of the validation process. Model validity is conditional on the status of its verification because a wrongly implemented model would automatically fail to represent the real world/system accurately, even though the conceptualization of the real world/system was right.

Despite the acknowledgment of the importance of verification in the health technology assessment (HTA) literature, there are currently no established guidelines on how to test and verify decision-analytic models for health economic evaluations. However, in a broader context, a large body of literature exists on software verification, mainly focusing on box-based testing approaches (e.g. Limaye [23]). Generally, software verification tests can be categorized into *white-box testing* (those performed by the software developer with the knowledge of the interworkings of the functions comprising the code base, such as scrutinizing the code line by line) and *black-box testing* (those performed without having to have knowledge of the interworkings of the code, such as running the code with a specific setting and given inputs and assessing if the output of the code is in line with the expectations or not).

The current study aims to create an operational checklist and documenting structure to support the technical verification process of health economic decision analytical models. For this purpose, a number of black-box and white-box tests for health economic models were collated and the TECHnical VERification (TECH-VER) checklist was developed, which encapsulates the necessary verification efforts for the major components of a health economic model.

The TECH-VER checklist aims to improve transparency and efficiency by streamlining the verification process and its documentation, and to help identify modeling errors and

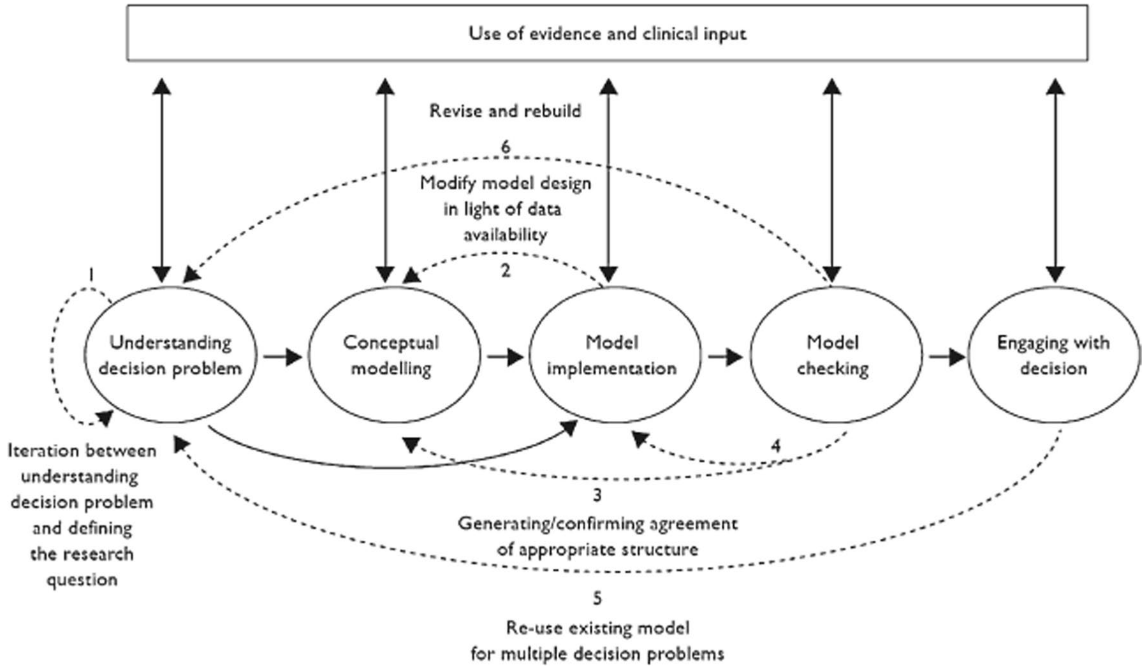

Use of evidence and clinical input
Revise and rebuild
6
Modify model design in light of data availability
2
1
Understanding decision problem
Conceptual modelling
Model implementation
Model checking
Engaging with decision
Iteration between understanding decision problem and defining the research question
4
3
Generating/confirming agreement of appropriate structure
5
Re-use existing model for multiple decision problems

Fig. 1 Model development process taken from Tappenden and Chillcott [2]

their root causes in a more systematic manner. Furthermore, new health economic modelers can use TECH-VER as a self-learning tool since they can integrate some of the verification tests while they are building the model.

In order to apply the TECH-VER checklist completely, one needs a model reviewer (preferably different from the model developer to prevent possible bias), a fully accessible and conceptually valid electronic model (e.g. without any hidden/locked parts) accompanied with a transparent technical report/evidence dossier, which details the conceptual model design, inputs used with their references, description of all the calculations in the model, and the results. It should be noted that the steps in TECH-VER can also be applied partially to a work-in-progress model, throughout the development process. The authors consider that it would be most efficient if TECH-VER is used in conjunction with other validation efforts (e.g. conceptual or face validity) as these efforts are highly interdependent and the verification of a conceptually wrong model (e.g. a model that estimates total life expectancy of a cohort by rolling a die) would be a waste of time.

In the next section, the methods used while building the TECH-VER checklist are described.

2 Methods

Development of the checklist consisted of the following steps:

- literature review to identify studies on model credibility, validation, and verification;
- development of the initial list of verification steps from both the literature and authors' experiences;
- iterative revision after each application of the TECH-VER checklist to a different model;
- revision after feedback from discussions with other health economists.

2.1 Literature Review

A literature search was conducted on the EMBASE and MEDLINE databases using the interface in Embase.com. The search filters were targeting for identifying some specific keyword variations of 'checklist', 'guideline', 'validation', 'verification', 'error', 'questionnaire' and 'credibility', in the titles and abstracts of the published economic modeling studies for cost-effectiveness analysis/HTA. The search strategy is presented in Appendix Table 3. The database output, including all indexed fields per record (e.g. title, authors, abstract), was exported to Endnote version X7.4, where the hits were de-duplicated.

From the articles retrieved from the MEDLINE and EMBASE libraries, the relevant references were selected using a two-step selection procedure, based on the following.

1. Screening of the title and abstract: This step yielded the articles that were assessed in full-text. The major topics of the articles were assessed on relevancy for the

objectives by title and abstract. In this step, articles that seemed to contain relevant data for the objectives were selected for full-text screening, while articles that did not seem to contain relevant data were not selected for full-text assessment.

2. Screening of the full article: The articles selected during the first phase were assessed in full text. PDF files of the original articles were downloaded and stored. Articles were included if the reported information was relevant, based on the inclusion and exclusion criteria, and of sufficient quality and sample size.

The process of selection and inclusion and exclusion of articles was registered in an Endnote library by one of the researchers (NCB), and the inclusion of the articles was supervised by another researcher (MA). The exclusion criteria applied in the selection procedure are reported in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flowchart. The list of inclusion criteria applied during the selection process is explained below.

- Period of publication: From inception of the databases until May 2019.
- Country/language: No restriction.
- Study design/type: Checklists, questionnaires, tests or systematic reporting of the efforts on model errors/model verification technical validation for HTA or cost-effectiveness analysis.

2.2 Initial (Beta Version) Checklist

An initial checklist was formed, based on the findings from the existing studies identified from the literature review, as well as the authors' own experience with model checking and reviewing. The black-box tests extracted from the literature (i.e. Tappenden and Chilcott [2], Dasbach and Elbasha [6], Hoogendoorn et al. [24]) were elaborated by additional de novo black-/white-box and replication-based tests created by the authors, and all these tests were reordered according to the framework created by the authors, which compartmentalizes the model-based calculations to different modules.

2.3 Iterative Revision of the Beta Checklist After Each Application

After the initial checklist had been developed, it was applied in several health economic models by several model reviewers. The models varied in terms of their developers (pharmaceutical companies, academic institute, or consultant firms), their purpose (to support reimbursement submissions, give early strategic advice in market access, for academic publication) and underlying clinical evidence maturity (e.g. the clinical effectiveness claim was based on systematic synthesis of randomized controlled trials (RCTs), a single RCT, evidence based on a non-randomized study, or even just clinical expectations used in sampling power calculations). TECH-VER was also applied in some of the students' models. These models were developed for the graduation-level course on health economic modeling or for the Master's Degree theses, both at Erasmus School of Health Policy and Management (ESHPM) in Erasmus University Rotterdam.

After all applicable black-box-type tests of the most recent version of the checklist were performed, each model was checked cell by cell (if the model implementation was conducted in a spreadsheet environment), or line by line (if the implementation was conducted using a programming code such as R or Visual Basic Applications for Excel®), as part of the white-box testing procedure. If additional errors or problems in the model implementation were identified during these white-box testing efforts, new black-box test-type questions were added or the existing black-box test-type questions were revised, in order to increase the sensitivity of the black-box type tests of the checklist. This iterative process was performed each time the checklist was filled in for a new model.

2.4 Revision Based on Feedback from Other Health Economists

The checklist was discussed with other health economists in various instances.

1. A workshop organized with health economists at Erasmus University Rotterdam.
2. An elaborate discussion of the checklist with several health economists from other Dutch/Belgian academic centers in the Low Lands Health Economic Study Group (lolaHESG) in May 2016.
3. A workshop organized with health economists from industry and international academic centers at the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Annual Congress in November 2016.
4. Revision based on comments from the referee(s) and the editor(s) after submitting the manuscript for publication in *PharmacoEconomics*.

2.4.1 Workshop Organized with Health Economists at Erasmus University Rotterdam

The checklist was distributed to an audience of health economists 1 week before a workshop (2.5 h) on model verification, where the checklist was presented and discussed. This workshop was conducted at the Erasmus School of Health Policy and Management and the Institute for Medical Technology Assessment of the Erasmus University Rotterdam, with a group of health economists (n = 12) with varying levels of experience in health economic modeling (minimum experience of 2 years; all participants had developed their own economic models; a majority worked only with decision trees/state transition models in a spreadsheet environment; most had published their own models or were in the process of publishing; and some were very experienced modelers involved in the appraisal processes of medical technologies for different national HTA bodies such as National Institutes for Health and Care Excellence [NICE] from the UK or the Healthcare Institute [ZiN] from The Netherlands). Their feedback on the checklist and its usability was collected in a separate form. Elicited feedback and suggestions from the audience played a key role in the revision of the checklist.

2.4.2 Discussion of the TECHnical VERification (TECH-VER) at the Low Lands Health Economics Study Group Annual Meeting (May 2016)

The Low Lands Health Economics Study Group (lolaHESG) is inspired by the Health Economists' Study Group (HESG) in the UK, and supports and promotes the work of health economists working at Dutch and Belgian research institutes. During the annual meetings, participants meet in groups to discuss papers that had been precirculated. These 1-h discussions were led by a discussant who presented the paper.

The TECH-VER checklist was presented and discussed during the lolaHESG meeting in May 2016, among a group of health economics researchers ($n \approx 25$) from various Dutch/Belgian academic centers. The feedback from the discussant and the audience were incorporated into the revision of the checklist.

2.4.3 Model Verification Workshop at the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Annual European Congress (November 2016)

A workshop on model verification was organized together with a health economist from a consultancy that provides services to the pharmaceutical industry and a health economist from an academic institution in UK, which also participates in NICE Technology Appraisals as an independent academic Evidence Review Group (ERG). TECH-VER was presented to an audience of health economists ($n \sim 120$) from international academic centers, as well as professionals from the industry and from the government. The feedback from this audience, which was elicited by interactive questions and follow-up communication, were incorporated into the revision of the checklist. After the workshop, the full version of the checklist was shared via hard copy/email with those attendants who had indicated their interest, whose feedback was collected in person/via email.

2.4.4 Revision Based on the Comments from the Referee(s) and the Editor(s)

After submitting the previous version of this manuscript for publication in *PharmacoEconomics*, we received a list of suggestions to improve the functionality of the TECH-VER. At each revision round, the authors attempted to incorporate the feedback suggestions from the referee(s) and the editor(s), which led to the most recent updates in the TECH-VER. These suggestions were mostly related to the rearrangement and prioritization of the verification efforts in the TECH-VER to improve its potential usability in time-constrained projects.

3 Results

3.1 Literature Review Results

Based on the search, a total of 3451 unique records were identified from the MEDLINE and EMBASE databases (Fig. 2). Of those, 3383 records were excluded based on their title and/or abstract. Sixty-eight articles were screened in full-text and 15 articles were included after applying the inclusion and exclusion criteria [2, 4–16, 24]. The resulting PRISMA diagram can be seen in Fig. 2. The main reasons for exclusion were:

- studies were checklist/guidance/recommendations on topics other than model validation/verification, for instance the checklists on the reporting of economic evaluations, or checklists to evaluate the overall quality of economic evaluations ($n = 21$);
- studies reported some model validation efforts but no model verification effort ($n = 21$);
- studies on other topics on model validity theory, for instance statistical tests for external validation ($n = 10$);
- full-text was not accessible or the study was not related to HTA ($n = 2$).

Among the 15 included articles, the majority of the studies ($n = 10$) were model-based cost-effectiveness analysis studies, which reported some limited amount of verification efforts [7–16]. The reporting of the verification efforts in none of these studies was systematic. The majority of the studies just reported a very brief summary description of the verification efforts, such as "the model has been thoroughly tested and de-bugged, including artificial simulations designed to reveal errors in both logic and programming" (Willis et al. [14]) or "The technical functioning was tested by means of an extensive sensitivity analysis. Extreme values of the input variables were used, and the model's actual outputs were compared with expected outcomes."

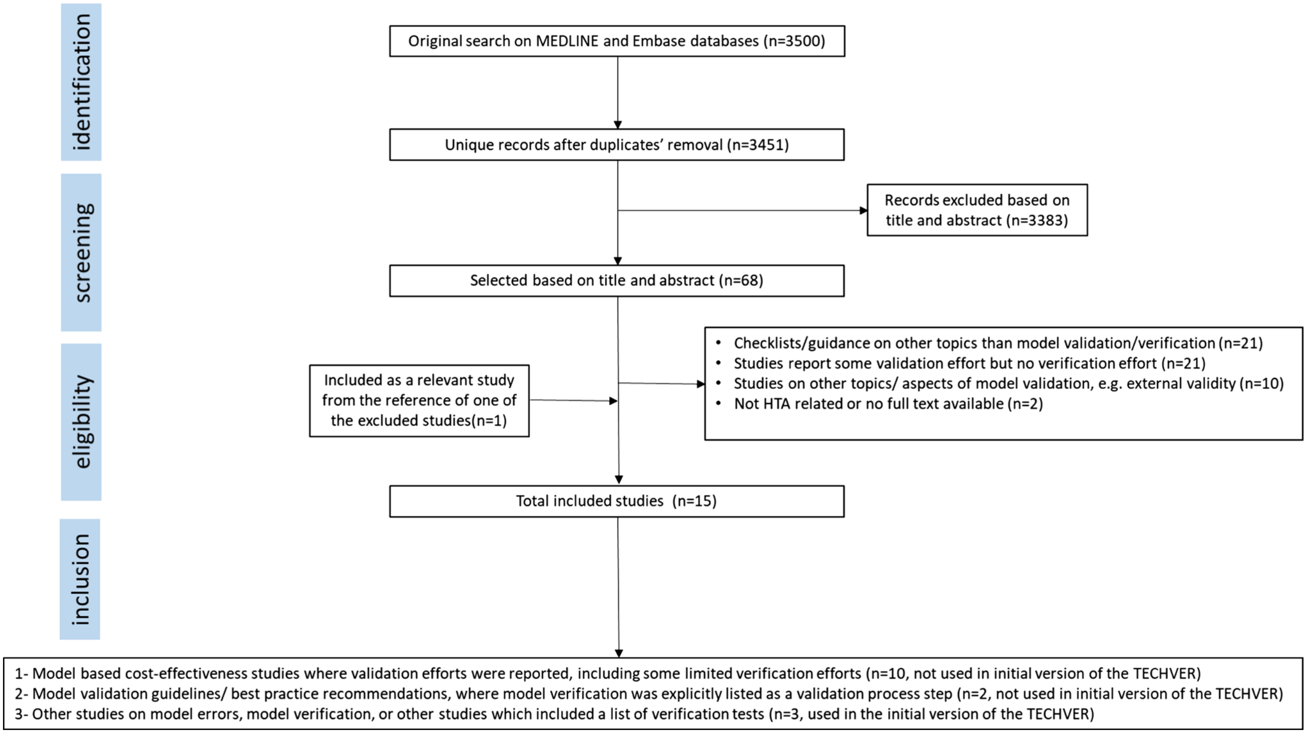

identification
screening
eligibility
inclusion
Original search on MEDLINE and Embase databases (n=3500)
Unique records after duplicates' removal (n=3451)
Records excluded based on title and abstract (n=3383)
Selected based on title and abstract (n=68)
• Checklists/guidance on other topics than model validation/verification (n=21)
• Studies report some validation effort but no verification effort (n=21)
• Studies on other topics/ aspects of model validation, e.g. external validity (n=10)
• Not HTA related or no full text available (n=2)
Included as a relevant study from the reference of one of the excluded studies(n=1)
Total included studies (n=15)
1- Model based cost-effectiveness studies where validation efforts were reported, including some limited verification efforts (n=10, not used in initial version of the TECHVER)
2- Model validation guidelines/ best practice recommendations, where model verification was explicitly listed as a validation process step (n=2, not used in initial version of the TECHVER)
3- Other studies on model errors, model verification, or other studies which included a list of verification tests (n=3, used in the initial version of the TECHVER)

Fig. 2 PRISMA flowchart for the literature review on model verification. *PRISMA* Preferred Reporting Items for Systematic Reviews and Meta-Analyses, *HTA* health technology assessment

(Hammerschmidt et al. [8]). A few studies reported only one type of verification effort, for instance the study by Cox et al. [7] attempted to replicate the base case of another published model in the literature, and compared the literature- and replication-based results with each other. Hence, these 10 studies were not used while forming the initial version of TECH-VER.

Two of the included articles were guidelines/best practice recommendations on the model validation process [4, 5]. In these studies, model verification was explicitly mentioned as one of the components of the model validation process, however no operational procedure/black-box test list was provided for model verification. Therefore, these two studies were not used while forming the initial version of TECH-VER.

Additionally, we included three other studies in our literature review [2, 6, 24]. These papers explicitly reported various types of black-box testing [2, 6, 24]. The first study, by Tappenden and Chilcott [2], provided a taxonomy of model errors and credibility threats, and linked them to model development processes. Together with this error taxonomy, Tappenden and Chilcott also provided a list of black-box tests and recommendations to help avoid and identify model errors/credibility threats.

The second study, by Hoogendoorn et al. [25], presented a dynamic population-level model for assessing the cost effectiveness of smoking cessation programs in patients with chronic obstructive pulmonary disease. Although the paper itself was excluded for not reporting any validation/verification efforts, another scientific report by the same authors [24] was cited. This cited report documented their validation and verification efforts, including a list of black-box tests, and was therefore included [24].

The third study, by Dasbach and Elbasha [6], pinpointed the necessity and the lack of a verification guidance, and gave a brief overview of verification methods used in software engineering and other fields. These researchers also uploaded a list of black-box tests, for model verification purposes, to an open-access software coding platform (GitHub).

The black-box tests found from these studies identified from the literature review were considered in the formation of the initial beta version of the TECH-VER, which was extended by additional black-box tests and restructured based on the feedback suggestions from other health economists, as outlined in Sects. 2.3 and 2.4.

3.2 TECH-VER Checklist

The TECH-VER checklist (Table 1) consists of the following five verification stages.

1. Model input (pre-analysis) calculations.

2. Event/state calculations.
3. Result calculations.
4. Uncertainty analysis calculations.
5. Overall validation/other supplementary checks.

The first four stages represent the calculations of the 'compartmentalized modules' of a typical decision analytical model.

In the verification stages of the *compartmentalized modules (stages 1–4)*, as a prerequisite, the reviewer is first asked to identify the relevant calculations in the electronic model and to assess the appropriateness of the provided justifications for the methods/approaches/assumptions used in each of the identified calculations. For this prerequisite step, we recommend the reviewer checks the completeness and consistency of the relevant calculations.

Completeness/consistency test: This part involves locating the calculations to be tested in the model and assessing their completeness (e.g. no hidden sheets/ranges/modules, no links to external programs, no hardcoded values, and no password protection) and assessing the consistency of the calculations (model implementation versus description in the report versus description/values in the reference source).

After this prerequisite step is completed, the correctness of the implementation of the model calculations needs to be checked. For this correctness check, the following tests should be applied in a consecutive and hierarchical order.

- *Black-box testing:* This involves checking if the related model calculations show results in line with *a priori* expectations, not only for plausible parameter inputs but also for extreme value parameters or probabilistic inputs.
- *White-box testing:* This involves checking the detailed model calculations that are being inspected, such as by going through the related code carefully, line by line, or by scrutinizing the formulae in all related ranges in a spreadsheet, cell by cell.
- *Replication-based tests*: These involve replication efforts of the calculations being inspected. The reviewer will try to replicate/re-perform the calculations using the same or different software (or even by pen and paper, if possible).

The tests listed above are sorted in an increasing order in terms of the expected time investment required for each test type, if they are conducted on the whole model. For model reviews that need to be conducted under a time constraint, we would like to limit the white-box and replication-based testing for specific, essential calculations only and under certain circumstances. Therefore, the tests can be conducted based on the suggested hierarchical order, visualized in Fig. 3.

For each verification stage of the compartmentalized modules, after the completeness/consistency checks are completed, a wide-ranging list of model-specific black-box tests needs to be prepared by the model reviewer. We present a list of black-box tests that can be useful, while model reviewers are creating their own lists, in Table 2. These types of tests aim to detect unexpected model behaviors, which might be caused by one (or more) error(s) in the electronic model.

After the black-box tests are conducted, white-box tests are recommended only for a priori selected essential calculations (such as calculation of the cycle-based technology acquisition costs, transition probabilities, or how these probabilities informed the transitions in certain cycles, etc.). Additionally, the white-box tests are recommended to locate the root cause of an unexpected model behavior, signaled by the black-box tests. Replication tests are recommended only in certain circumstances, such as when white-box tests fail to detect the root cause of an unexpected model behavior, or when white-box tests seem to be more challenging due to the non-transparent, complex programming in the reviewed model.

Note that some of the software-specific built-in functions/tools (e.g. 'Error checking' and 'Evaluate formula' options in Microsoft Excel®, or debug toolbox in Visual Basic for Applications for Excel®) can be used during the white-box testing efforts. However, these tools and others do not negate the necessity of detailed cell-by-cell or line-by-line inspection by the reviewer during the white-box tests. Similarly, the Inquire® tool embedded to Excel can prove to be useful in the completeness/consistency checks in spreadsheet environments as this tool can identify hidden sheets/ranges and hardcoded formulae in spreadsheet models.

In the next subsections, we briefly describe the scope and importance of each verification stage, give a couple of typical error examples that could potentially be detected by that verification stage, and provide further guidance on how to report the description and results of the verification efforts.

3.2.1 Verification Stage 1: Model Input/Pre-Analysis Calculations

This verification stage focuses on the pre-analysis calculations that yield direct-to-use, cycle-based, or event-based model inputs from the reference source inputs. Often, these pre-analysis calculations might be performed outside the electronic model, using another statistical software (e.g. R, STATA). If these calculations are inaccessible to the reviewer, they should be reported as missing during the completeness and consistency checks.

Some real-life examples

- *Errors in the pre-analysis calculations for deriving transition probabilities:* In the NICE company submission for secukinumab for ankylosing spondylitis [17], an error was detected in the network meta-analysis (NMA) that was used in generating the treatment-specific response probabilities for the electronic model. The error was detected after the ERG realized that the company's NMA could not be replicated.
- *Errors in the pre-analysis calculations for deriving transition probabilities:* In one of the student group's models for the health economics graduate-level course project, errors were detected after a strong dissimilarity between the overall survival (OS) extrapolation from the students' model and the corresponding Kaplan–Meier curve from the trial was observed. Furthermore, the Weibull OS extrapolation from the students' model did not demonstrate the characteristics of the distribution (i.e. survival function was not non-increasing and the associated hazard rate did not demonstrate a monotone behavior in time). Later, it was found that this dissimilarity and the strange behavior of the extrapolation was caused by problems with the X-Y digitization process while generating pseudo patient-level data, and by the wrong translation of the log-transformed regression outputs to use in the corresponding distribution function for the survival extrapolation.
- *Errors in the pre-analysis calculations for deriving cost inputs:* In the NICE appraisal of pomalidomide with dexamethasone for relapsed/refractory multiple myeloma [18], the ERG identified an error in the weekly resource use cost estimation of the company, when the calculated resource use estimate of the company differed significantly from the estimate obtained from the back of the envelope calculations from the ERG. Later, it was discovered that the error was due to the fact that instead of dividing the annual resource use estimate by 52, the company wrongly converted the annual estimate to the weekly probability of using a resource.
- *Errors in the pre-analysis calculations for deriving cost and utility inputs:* In one of the student group's models for the health economics graduate-level course project, errors were detected in the treatment acquisition cost calculations. These costs should be dependent on weight; the errors were detected when the cycle-based costs for drug acquisition remained unchanged for different weight assumptions, and the 'no wastage' scenario generated the same result as the 'with wastage' scenario. It was later found that a constant drug acquisition cost was always assumed. In the same model, another error in calculating the adverse event utility decrement was identified. The overall utility value corresponding to the cycle during which this utility decrement was applied was lower than zero. It was later realized that the students did not use the same time units while scaling the average duration for that adverse event (in days) to the cycle length duration (1 week), which overestimated the adverse event-associated utility decrements.

The pre-analysis calculations can be categorized as follows:

- transition probabilities (e.g. NMA, other statistical models, survival analysis techniques, etc.).
- costs (e.g. calculating cycle/event-based estimates).
- utilities/other health outcomes (e.g. calculating cycle/event-based estimates).
- other calculations.

While checking the correctness of the type of calculations above, the model reviewer should follow the outlined steps in Table 1 (prerequisite checks, black-box, white-box and replication-based tests in hierarchical order). Crucial calculations for white-box testing in this stage can be calculation of the treatment (e.g. drug/device) acquisition cost per cycle, or how the transition probabilities are generated and how the treatment effectiveness is applied.

3.2.2 Verification Stage 2: Event/State Calculations

This verification stage focuses on the event/state-based calculations in the electronic model. These calculations might involve (but are not limited to) the following calculations:

- unfolding of decision and chance nodes in a decision tree.
- calculation of the distribution of cohorts among different health states at a given cycle in a state transition model (e.g. Markov trace).
- determining when/which type of event will occur next, at a given time in a discrete event simulation model.
- assignment of costs/QALYs/other health outcomes to the relevant states or events in the electronic model.

The verification checks of these calculations are essential as the errors that can be detected in this verification stage might have a substantial impact on the decision. Crucial calculations for white-box testing in this stage can be the way in which transition probabilities are used in calculating the number of patients in each state (or experiencing an event) for a couple of cycles, and how costs and utilities are assigned to these states and events.

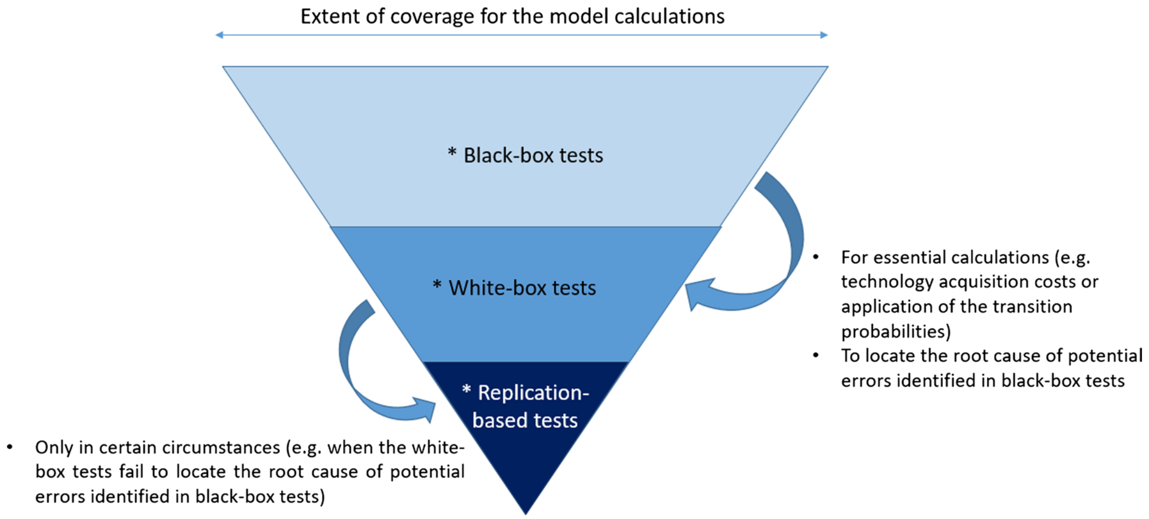
Table 1 TECH-VER checklist

<table>
<tr><td><b>1- Model input (pre-analysis) calculations: this verification stage checks the pre-analysis calculations that yield direct model inputs (e.g. transition probabilities, cycle-based or event-based costs and utilities) from reference source inputs</b></td><td rowspan="4"><b>1-4: Verification stages of the compartmentalized modules (Follow these steps for all the identified calculations in stages 1-4)</b></td><td rowspan="4"><b>Pre-requisite for conducting verification tests:</b>
Locating the calculations in the model and the explanation of the calculations in the report. Assess the appropriateness of the methods. Check the <i>completeness and consistency</i> of the calculations in the model
• (Report the sheets/ranges/coding lines where the corresponding calculations are carried out in the electronic model, report any provided justification for the methods/ assumptions used. Assess if these are appropriate with respect to the published methodological guidelines. Document the consistency checks that are conducted)
<b>Verification tests after the pre-requisite steps are complete:</b>
Check if the implementation of these calculations is correct using <i>black-box type, white-box type</i> and <i>replication-based</i> tests, in a consecutive order, following the hierarchical order in Figure 3 under a time constraint.
• (Report all the necessary details of any test conducted, so that it can be reproduced by another reviewer, for each of the identified calculations in the electronic model.)</td></tr>
<tr><td><b>2- Event/state calculations: this verification stage checks the event/state calculations that determine the patient flow/disease progression stage as well as the assignment of costs/QALYs or other relevant health/economic outcomes at a given cycle/time</b></td></tr>
<tr><td><b>3- Result calculations: this verification stage checks the result calculations that yield the undiscounted/ discounted total and incremental results (e.g. costs, QALYs, other relevant health or economic outcomes and ICER)</b></td></tr>
<tr><td><b>4- Uncertainty analysis: this verification stage checks the uncertainty analysis calculations (e.g. one-way, multi-way, probabilistic sensitivity, value of information and scenario analyses)</b></td></tr>
<tr><td colspan="3"><b>5- Overall tests (validation or other supplementary tests): these tests include validation efforts from other sources and tests that are applied to the whole model and efforts that do not specifically belong to one of the compartmentalized modules</b>
• Compare the model outcomes with clinical inputs used in the model, findings from the literature, clinical expert knowledge and other model outcomes (Outline the conducted comparisons between the electronic model and the other sources and report if there is any inconsistency)
• Check the other aspects of model implementation that does not fall under the scope of the other stages, such as the interface, programming and data storage efficiency, etc. (Report all the necessary details of any test conducted, so that it can be reproduced by another reviewer)</td></tr>
</table>

TECH-VER TECHnical VERification, *QALYs* quality-adjusted life-years, *ICER* incremental cost-effectiveness ratio

Some real-life examples

- *Error in the cohort trace:* In the electronic model attached to the NICE company submission of idelalisib for refractory follicular lymphoma [19], the ERG identified an error in the Markov trace, which led to a negative number of transitions between certain states of the model at certain times. This error could have been detected by the black-box tests suggested for verification stage 2 (event/state calculations), for instance by checking if all the transition probabilities are greater than or equal to zero.
- *Error in time to event generation:* In the electronic model used in the same appraisal, for each patient that has recently progressed, his/her time to death after progression is estimated from a parametric extrapolation curve fitted to post-progression survival data from the long-term follow-up from another trial. However, at monthly intervals after


Extent of coverage for the model calculations
* Black-box tests
* White-box tests
* Replication-based tests

- For essential calculations (e.g. technology acquisition costs or application of the transition probabilities)
- To locate the root cause of potential errors identified in black-box tests

- Only in certain circumstances (e.g. when the white-box tests fail to locate the root cause of potential errors identified in black-box tests)

Fig. 3 Representation of the recommended hierarchical order of the type of the verification tests under a time constraint for the first four stages of TECH-VER

Table 2 List of black-box tests

Test description (please also document how the test is conducted)	Expected result of the test
Pre-analysis calculations	
Does the technology (drug/device, etc.) acquisition cost increase with higher prices?	Yes
Does the drug acquisition cost increase for higher weight or body surface area?	Yes
Does the probability of an event, derived from an OR/RR/HR and baseline probability, increase with higher OR/RR/HR?	Yes
In a partitioned survival model, does the progression-free survival curve or the time on treatment curve cross the overall survival curve?	No
If survival parametric distributions are used in the extrapolations or time-to-event calculations, can the formulae used for the Weibull (generalized gamma) distribution generate the values obtained from the exponential (Weibull or Gamma) distribution(s) after replacing/transforming some of the parameters?	Yes
Is the HR calculated from Cox proportional hazards model applied on top of the parametric distribution extrapolation found from the survival regression?	No, it is better if the treatment effect that is applied to the extrapolation comes from the same survival regression in which the extrapolation parameters are estimated
For the treatment effect inputs, if the model uses outputs from WINBUGS, are the OR, HR, and RR values all within plausible ranges? (Should all be non-negative and the average of these WINBUGS outputs should give the mean treatment effect)	Yes
Event-state calculations	
Calculate the sum of the number of patients at each health state	Should add up to the cohort size
Check if all probabilities and number of patients in a state are greater than or equal to 0	Yes
Check if all probabilities are smaller than or equal to 1	Yes
Compare the number of dead (or any absorbing state) patients in a period with the number of dead (or any absorbing state) patients in the previous periods?	Should be larger
In case of lifetime horizon, check if all patients are dead at the end of the time horizon	Yes
Discrete event simulation specific: Sample one of the 'time to event' types used in the simulation from the specified distribution. Plot the samples and compare the mean and the variance from the sample	Sample mean and variance, and the simulation outputs, should reflect the distribution it is sampled from
Set all utilities to 1	The QALYs accumulated at a given time would be the same as the life-years accumulated at that time
Set all utilities to 0	No utilities will be accumulated in the model
Decrease all state utilities simultaneously (but keep event-based utility decrements constant)	Lower utilities will be accumulated each time
Set all costs to 0	No costs will be accumulated in the model at any time
Put mortality rates to 0	Patients never die
Put mortality rate at extremely high	Patients die in the first few cycles
Set the effectiveness-, utility-, and safety-related model inputs for all treatment options equal	Same life-years and QALYs should be accumulated for all treatment at any time
In addition to the inputs above, set cost-related model inputs for all treatment options equal	Same costs, life-years, and QALYs should be accumulated for all treatment at any time
Change around the effectiveness-, utility- and safety-related model inputs between two treatment options	Accumulated life-years and QALYs in the model at any time should also be reversed
Check if the number of alive patients estimated at any cycle is in line with general population life-table statistics	At any given age, the percentage alive should be lower or equal in comparison with the general population estimate
Check if the QALY estimate at any cycle is in line with general population utility estimates	At any given age, the utility assigned in the model should be lower or equal in comparison with the general population utility estimate

Table 2 (continued)

Test description (please also document how the test is conducted)	Expected result of the test
Set the inflation rate for the previous year higher	The costs (which are based on a reference from previous years) assigned at each time will be higher
Calculate the sum of all ingoing and outgoing transition probabilities of a state in a given cycle	Difference of ingoing and outgoing probabilities at a cycle in a state times the cohort size will yield the change in the number of patients at that state in that cycle
Calculate the number of patients entering and leaving a tunnel state throughout the time horizon	Numbers entering = numbers leaving
Check if the time conversions for probabilities were conducted correctly.	Yes
Decision tree specific: Calculate the sum of the expected probabilities of the terminal nodes	Should sum up to 1
Patient-level model specific: Check if common random numbers are maintained for sampling for the treatment arms	Yes
Patient-level model specific: Check if correlation in patient characteristics is taken into account when determining starting population	Yes
Increase the treatment acquisition cost	Costs accumulated at a given time will increase during the period when the treatment is administered
Population model specific: Set the mortality and incidence rates to 0	Prevalence should be constant in time
Result calculations	
Check the incremental life-years and QALYs gained results. Are they in line with the comparative clinical effectiveness evidence of the treatments involved?	If a treatment is more effective, it generally results in positive incremental LYs and QALYs in comparison with the less-effective treatments
Check the incremental cost results. Are they in line with the treatment costs?	If a treatment is more expensive, and if it does not have much effect on other costs, it generally results in positive incremental costs
Total life years greater than the total QALYs	Yes
Undiscounted results greater than the discounted results	Yes
Divide undiscounted total QALYs by undiscounted life years	This value should be within the outer ranges (maximum and minimum) of all the utility value inputs
Subgroup analysis results: How do the outcomes change if the characteristics of the baseline change?	Better outcomes for better baseline health conditions, and worse outcomes for worse health conditions, are expected
Could you generate all the results in the report from the model (including the uncertainty analysis results)?	Yes
Do the total life-years, QALYs, and costs decrease if a shorter time horizon is selected?	Yes
Is the reporting and contextualization of the incremental results correct?	The use of terms such as 'dominant'/'dominated'/'extendedly dominated'/'cost effective'. etc.. should be in line with the results In the incremental analysis table involving multiple treatments, ICERs should be calculated against the next non-dominated treatment
Are the reported ICERs in the fully incremental analysis non-decreasing?	Yes
If disentangled results are presented, do they sum up to the total results (e.g. different cost types sum up to the total costs estimate)?	Yes
Check if half-cycle correction is implemented correctly (total life-years with half-cycle correction should be lower than without)	The half-cycle correction implementation should be error-free. Also check if it should be applied for all costs, for instance if a treatment is administered at the start of a cycle, half-cycle correction might be unnecessary
Check the discounted value of costs/QALYs after 2 years	Discounted value = undiscounted/$(1+r)^2$
Set discount rates to 0	The discounted and undiscounted results should be the same
Set mortality rate to 0	The undiscounted total life-years per patient should be equal to the length of the time horizon
Put the consequence of adverse event/discontinuation to 0 (0 costs and 0 mortality/utility decrements)	The results would be the same as the results when the AE rate is set to 0
Divide total undiscounted treatment acquisition costs by the average duration on treatment	This should be similar to treatment-related unit acquisition costs

Table 2 (continued)

Test description (please also document how the test is conducted)	Expected result of the test
Set discount rates to a higher value	Total discounted results should decrease
Set discount rates of costs/effects to an extremely high value	Total discounted results should be more or less the same as the discounted results accrued in the first cycles
Put adverse event/discontinuation rates to 0 and then to an extremely high level	Less costs and higher QALYS/LYs when adverse event rates are 0, higher costs and lower QALYS/LYs when AE rates are extreme
Double the difference in efficacy and safety between the new intervention and comparator, and report the incremental results	Approximately twice the incremental effect results of the base case. If this is not the case, report and explain the underlying reason/mechanism
Do the same for a scenario in which the difference in efficacy and safety is halved	Approximately halve of the incremental effect results of the base case. If this is not the case, report and explain the underlying reason/mechanism
Uncertainty analysis calculations	
Are all necessary parameters subject to uncertainty included in the OWSA?	Yes
Check if the OWSA includes any parameters associated with joint uncertainty (e.g. parts of a utility regression equation, survival curves with multiple parameters)	No
Are the upper and lower bounds used in the one-way sensitivity analysis using confidence intervals based on the statistical distribution assumed for that parameter?	Yes
Are the resulting ICER, incremental costs/QALYs with upper and lower bound of a parameter plausible and in line with a priori expectations?	Yes
Check that all parameters used in the sensitivity analysis have appropriate associated distributions – upper and lower bounds should surround the deterministic value (i.e. upper bound ≥ mean ≥ lower bound)	Yes
Standard error and not standard deviation used in sampling	Yes
Lognormal/gamma distribution for HRs and costs/resource use	Yes
Beta for utilities and proportions/probabilities	Yes
Dirichlet for multinomial	Yes
Multivariate normal for correlated inputs (e.g. survival curve or regression parameters)	Yes
Normal for other variables as long as samples do not violate the requirement to remain positive when appropriate	Yes
Check PSA output mean costs, QALYs, and ICER compared with the deterministic results. Is there a large discrepancy?	No (in general)
If you take new PSA runs from the Microsoft Excel model do you get similar results?	Yes
Is(are) the CEAC line(s) in line with the CE scatter plots and the efficient frontier?	Yes
Does the PSA cloud demonstrate an unexpected behavior or have an unusual shape?	No
Is the sum of all CEAC lines equal to 1 for all WTP values?	Yes
Do the explored scenario analyses provide a balanced view on the structural uncertainty (i.e. not always looking at more optimistic scenarios)?	Yes
Are the scenario analysis results plausible and in line with a priori expectations?	Yes
Check the correlation between two PSA results (i.e. costs/QALYs under the SoC and costs/QALYs under the comparator)	Should be very low (very high) if different (same) random streams are used for different arms
If a certain seed is used for random number generation (or previously generated random numbers are used), check if they are scattered evenly between 0 and 1 when they are plotted	Yes
Compare the mean of the parameter samples generated by the model against the point estimate for that parameter; use graphical methods to examine distributions, functions	The sample means and the point estimates will overlap, the graphs will be similar to the corresponding distribution functions (e.g. normal, gamma, etc.)

Table 2 (continued)

Test description (please also document how the test is conducted)	Expected result of the test
Check if sensitivity analyses include any parameters associated with methodological/structural uncertainty (e.g. annual discount rates, time horizon)	No
Value of information analysis if applicable: Was this implemented correctly?	Yes
Which types of analysis? Were aggregated parameters used? Which parameters are grouped together? Does it match the write-up's suggestions?	Yes
Is EVPI larger than all individual EVPPIs?	Yes
Is EVPPI for a (group of) parameters larger than the EVSI of that (group) of parameter(s)?	Yes
Are the results from EVPPI in line with OWSA or other parameter importance analysis (e.g. ANCOVA)?	Yes
Did the electronic model pass the black-box tests of the previous verification stages in all PSA iterations and in all scenario analysis settings? (Additional macro can be embedded to the PSA code, which stops the PSA when an error such as negative transition probability is detected)	Yes
Check if all sampled input parameters in the PSA are correctly linked to the corresponding event/state calculations	Yes

OWSA one-way sensitivity analysis, *ICER* incremental cost-effectiveness ratio, *PSA* probabilistic sensitivity analysis, *WTP* willingness to pay, *CE* cost effectiveness, *CEAC* cost-effectiveness acceptability curve, *LY* life-years, *QALYs* quality-adjusted life-years, *OR* odds ratio, *RR* relative risk, *HR* hazard ratio, *SoC* standard of care, *EVPI* expected value of perfect information, *EVPPI* expected value of partial perfect information, *EVSI* expected value of sample information, *ANCOVA* analysis of covariance

progression, time to death after progression is again re-sampled, using the same distribution, as if the patient has just progressed. This led to an overestimation of the post-progression survival, when the extrapolation distribution of choice was different from exponential. This error could have been detected by the black-box tests suggested for verification stage 2 (event/state calculations), for instance by checking if the simulated time to event reflects the distribution the time to event is sampled from.

- *Errors in assignment of costs and utilities:* In one of the student group models for the health economics graduate-level course project, errors were detected in the assignment of cost and utilities to the relevant events/states. For instance, in one of the group's models, the expensive drug costs were assigned as long as the patients were alive even though the drug could be administered 10 cycles at maximum. In another example, it was detected that the treatment-specific utility estimates for interventions A and B were assigned in the opposite way it should have been. These errors were detected using white-box tests, after they were identified by replication-based tests, while comparing the utilities and costs accrued per cycle with those from parallel models (other black-box-type tests could have also identified these errors).
- *Errors in the assignment of costs:* In the NICE submission of ramucirumab for gastric cancer [20], hospitalization rates derived from the primary trial were assigned at each cycle. These hospitalizations also included those due to adverse events; however, the adverse event costs were assigned separately at each cycle, which also included costs due to hospitalization. This led to a double counting of adverse events associated with hospitalization costs. This error was detected during white-box testing and assessing the appropriateness of the calculations for verification stage 2, conducted for the event/state calculations.

3.2.3 Verification Stage 3: Result Calculations

This verification stage focuses on the result calculations in the electronic model, which can be categorized as follows:

- summation of the accumulated costs, QALYs, and life-years or other outcomes over time to obtain total costs, QALYs, and life-years or other total outcomes.
- the calculation and interpretation of the incremental results and ICER(s).
- applying half-cycle correction/discount rates.
- disaggregation of total costs and total QALYs.
- other calculations.

The errors that can be detected in this verification stage might have a direct impact on the incremental results and cost-effectiveness conclusions. The crucial calculations that need to be checked with white-box testing can be the way total discounted and half-cycle corrected costs and QALYs are summed up and the way ICERs are derived, and how the costs and QALYs are disaggregated.

Some real-life examples

- *Error in half-cycle correction:* In one of the previous models the first author of this study had developed, an error was detected in the implementation of the half-cycle correction. Instead of summing over *n-1* half-cycle-corrected intervals from *n* cells, where each cell represented the QALYs accrued at the corresponding cycle, the author summed over *n* half-cycle-corrected intervals from *n+1* cells. Unfortunately, the value at the $(n+1)^{th}$ cell was not the QALYs that would be accrued at the $(n+1)^{th}$ cycle, but was the sum of the total QALYs accrued in the first *n* cycles. This error could have been detected by some of the black-box-type tests from TECH-VER, such as checking if half-cycle corrected total QALYs were lower than total QALYs that were not half-cycle corrected.
- *Error in discount rate:* In one of the student group models for the health economics graduate-level course project, errors were detected in the discounting. Namely, in month *t*, an accrued cost of *c* was discounted as $\left(\frac{c}{(1+0.035/12)}\right)^{t}$, whereas it should have been discounted as $\left(\frac{c}{(1+0.035)}\right)^{t/12}$. This error could have been detected by checking if the discounted value from the model was $\frac{c}{(1+0.035)}$ when $t = 12$.
- *Error in ICER calculations:* In one of the student group models for the health economics graduate-level course project, cost calculations were, in general, correct, but when presenting the disaggregated costs, the students assigned the administration cost of chemotherapy under both the 'chemotherapy-related costs' and the 'resource use-related costs', which led to double counting. This error could have been detected if the sum of the disaggregated costs was compared with the total costs.
- *Error in ICER calculations:* In the thesis of one of the Masters students, an error in the fully incremental analysis calculations was detected. Instead of calculating the ICER with respect to the previous cost-effective technology, the student calculated the ICER with respect to the cheaper technology. This error could have been detected by the black-box-type tests in TECH-VER (e.g. checking if the reported ICERs in the fully incremental analysis were always non-decreasing).

3.2.4 Verification Stage 4: Uncertainty Analysis Calculations

In this part, the calculations related to uncertainty analysis in the model are tested. These analyses attempt to quantify different aspects of parametric, structural, and decision uncertainty. The calculations in these modules can be categorized as below:

- one-way/multiway sensitivity analysis calculations.
- probabilistic sensitivity analysis (PSA) calculations.
- scenario analysis calculations.
- value of information (VOI) analysis calculations.
- other types of calculations.

Some real-life examples

- *Error in PSA and one-way sensitivity (OWSA) calculations:* In the NICE company submission for glecaprevir/pibrentasvir for hepatitis C [21], 100% sustained viral response (primary effectiveness measure) was achieved for the genotype 5 subgroup, based on a small sample of patients. The company assumed the standard error for that parameter as zero since no events were observed for no response. This led to a substantial underestimation of the parametric uncertainty. The ERG detected this error in the 'completeness and consistency' checks for the prerequisite steps of verification stage 4, conducted for PSA calculations. After implementing a standard continuity correction for zero cells to obtain standard error estimates, the impact on the PSA and OWSA results were substantial (i.e. the probability that glecaprevir/pibrentasvir becomes cost effective for that specific genotype reduces by 66%, and, in the OWSA, the viral response rate became by far the most influential parameter).
- *Omitting the correlation between the regression coefficients:* In one of the student group models for the health economics graduate-level course project, the students omitted the correlation between the regression coefficients for the Weibull OS extrapolation. This led to overestimation of the parametric uncertainty to some extent. The error was detected based on assessing the PSA cloud from the students' model, which seemed to be substantially more scattered than the PSA cloud of the other group's model.
- *Errors in sampled progression-free survival (PFS) and OS curves:* In one of the student group models for the health economics graduate-level course project, the regression coefficients used for PFS and OS extrapolation were sampled independently. In some of the PSA iterations, the students had a negative number of cohorts in the progressed disease state, which was calculated from (*OS(t)-PFS(t)*). The error was detected by the lecturers because there was a substantial unexplainable gap between the deterministic base-case cost outcomes and mean cost outcomes from the PSA iterations; however, the error could have also been detected if an automated check for negative numbers in the cohort trace was integrated to the PSA for loop test.

The errors that can be detected in this verification stage might lead to biased estimation of the uncertainty in the cost-effectiveness outcomes from the model. For instance, if not all relevant parameters are sampled in the PSA, the parametric uncertainty will be underestimated and the model outcomes will appear to be more certain than they actually are. Similarly, if the confidence interval of an input parameter was wrongly calculated for an OWSA, the relevant importance for that parameter would be misleading. If a wrong distribution is used while sampling a parameter in a PSA and VOI analysis, the decision uncertainty results and subsequent future research recommendations may be unjustified.

Similar to the previous stages, after the prerequisite completeness and consistency checks have been completed, the model reviewer is advised to follow the hierarchical order outlined in Fig. 3, while conducting black-box, white-box and replication-based verification tests, in a hierarchical order.

3.2.5 Verification Stage 5: Overall Validation/Other Supplementary Tests

This verification stage involves checks that did not fall within the remit of the other verification stages, such as verification of the model interface or checking the model performance/programming inefficiencies.

These checks are important because a wrong interface switch button from an electronic model will lead to erratic results in the future. It should be ensured that all settings, switches, and other buttons used in the user interface of the electronic model operate correctly as per their intended use. Similarly, extremely slow models indicate inefficient programming or unnecessary data storage, which might lead to numerical inconsistencies or even program/computer crashes. For this purpose, the programming of the model should be assessed, the necessity of all stored data in the model should be reconsidered, and extra-long formulas and nested loops should be flagged for double checking.

In addition, some of the validation checks that can be considered beyond the scope of verification (e.g. internal validation or cross-validation) can be conducted at this stage. These validation efforts can also be helpful in identifying model errors.

The reviewer can prepare a number of validation-based tests in this verification stage. For example, if the outcomes of a clinical study have been used in the pre-analysis calculations, these can be compared with the model outcomes (e.g. comparing the median survival in the model versus median survival in the RCT used in generating the inputs of the model). Similarly, if the reviewer has another model in the same indication, and the two models have common comparators, the outcomes from the two models belonging to the common comparators can be contrasted.

Furthermore, if other cost-effectiveness results are available from the literature, the reviewer can try to check if the model under review can simulate the results from the literature when the baseline characteristics and the inputs used in the model are changed accordingly.

Some real-life examples

- *Comparing the model outcomes with another model outcome:* In the NICE submission for ribociclib in breast cancer [22], the company submission model generated noticeably different outcomes in comparison to the outcomes from the palbociclib NICE submission model, in the best supportive care treatment arms from both models. Even though this difference was later attributed to different model structure assumptions and not a modelling error, this difference in outcomes triggered the ERG to take an even more detailed look at the company's model implementation, which led to the identification of additional errors.
- *Model interface problems:* In one of the country adaptations of a model for an oncology product, the button on the main page of the model called "Back to the base-case assumptions" updated the input parameters that were different from the inputs used in the base-case.

If the model provides significantly different results than the results from the literature, this might indicate an error in the model implementation, although obviously it is also possible that the published model contains errors, or the difference in costs can be due to differences in concepts. Therefore, it should be noted that these tests should be conducted in coordination with the efforts associated with the other validation tests (such as AdVISHE) [5] to avoid overlap of validation/verification efforts.

4 Discussion and Conclusion

In this current study, we introduced the TECH-VER checklist, which is, to the best of the authors' knowledge, the first attempt to create an open source, *operational* checklist to support the technical verification process of health economic decision analytical models. However, it is recognized that many institutions creating health economic models may already use their own in-house verification checklists.

The TECH-VER checklist consists of five verification stages, each of which focuses on the verification of different aspects of a decision analytical model. For each compartmentalized module calculation-related verification stage, we suggested a list of different types of verification tests (completeness/consistency as a prerequisite check, black-box, white-box and replication-based tests), each should be designed and conducted by the reviewer/validator in a necessity-based order, taking time constraints into account. A hierarchical order is suggested by the authors, which prioritizes the black-box-type tests, and more time-consuming tests are reserved for a priori selected essential calculations or when black-box tests signal unexpected model behavior. A comprehensive list of black-box-type tests are also presented in this paper, which can be used by readers while they are designing their own black-box test suites specific for their models.

The TECH-VER checklist aims to provide a documenting framework for verification efforts, which would improve the transparency and efficiency of verification efforts and help identify modeling errors and their root causes in a more systematic manner. Furthermore, unexperienced, new health economic modelers can use TECH-VER as a self-learning tool since they can integrate some of the verification tests while they are building their model. A simplified version of this checklist is being used in the peer review process of a Masters-level course on health economic modeling at our institute. As these students are new to modeling, having a tool to guide them in the process of peer reviewing models has proven to be very valuable. In addition, the TECH-VER checklist (with different levels of detail requirements) can be integrated into multi-team validation efforts in the model development process, and can also be incorporated into HTA submissions for reimbursement purposes. Ultimately, our checklist might also be requested for manuscripts describing modeling studies submitted to scientific journals for publication.

The detailed documentation of the verification tests will improve the replicability of the verification efforts, meaning that in the next review rounds more time can be invested in thinking of additional tests/checks, such as extending the original list of black-box-type tests or extending the set of essential calculations to be scrutinized/replicated. The authors are aware that verification tests are model-specific and, without a doubt, other useful black-box tests are available that are not mentioned in this study. Therefore, initiatives such as sharing the verification tests in open-source platforms such as github, as suggested by Dasbach and Elbasha [6], are very important. For this reason, TECH-VER has also been published on an open-access online platform,[1] together with some applications from the authors. On this platform, the validators can upload their own tests, put comments, and share their own experiences on verification/model errors and other practical examples.

We acknowledge that consultations with the health economic modeling experts outside of our institute (Erasmus University Rotterdam) were conducted during scientific meetings and did not follow a formal Delphi method approach. Even though a Delphi approach would be more systematic, the level of detail involved in defining the checks and tests would make the realization of the Delphi approach extremely difficult. We also believe that different inputs from experts from other institutions, both from the Netherlands as well as from the other countries and stakeholders other than academia, were incorporated into the final version of the TECH-VER, and also believe that after publication of the TECH-VER, different stakeholders will be able to include their comments/feedback on the online platform where TECH-VER has been uploaded.

We would like to emphasize that the TECH-VER checklist is aimed to serve as a documentation framework. The reviewer/validator is flexible on the type and amount of tests to be conducted, however it is recommended to document the verification efforts, which are conducted and which are omitted, together with the validators' reasons for omission. We acknowledge that some of the verification tests are model-dependent and cannot be applicable for all models. The black-box list test we provided is comprehensive, but the reviewer/validator is flexible to choose the tests in a different priority order, based on what they consider to be the most useful in their specific setting. For instance, a reviewer/validator, after feeling confident about the expected value of partial perfect information (EVPPI) calculations, can first conduct an EVPPI analysis to identify the most influential parameters on decision uncertainty, and focus on verification of the calculations on these parameters. Similarly, a model reviewer/validator might choose to conduct replication-based testing for calculations, after the black-box tests resulted in unexpected model behaviors, and scrutinizing the calculations as part of white-box testing was not possible due to inaccessible codes in the economic model. The TECH-VER framework accomodates such flexibilities, as long as the reviewer/validator documents the conducted as well as the omitted verification efforts in detail, along with the reasons for the omission.

Some of the practical examples presented in this paper originated from students' models/projects from Erasmus University Rotterdam. The errors identified from the students' models can be reflective of the type of errors one might make as a junior modeler. Therefore, applying some of the tests provided in TECH-VER during the model development phase can decrease the amount of errors in the final model. Other practical examples from different stakeholders (e.g. submitted NICE appraisal models from industry/consultancy companies, or authors' own models) can be representative of the types of errors one might make as a more senior modeler, or with different motivation of biases. We believe that in the online platform where the TECH-VER has been published, different stakeholders will be able to share more practical examples and verification strategies.

The authors consider that it would be most efficient if TECH-VER is used in conjunction with other validation

[1] https://github.com/nasuhcagdas/TECHVER and http://www.imta.nl/techver.

efforts (e.g. conceptual or face validity) as these efforts are highly interdependent and verification of a conceptually wrong model (e.g. a model that estimates total life expectancy of a cohort by rolling a die) would be a waste of time. On the other hand, model validity is conditional on its verification because a wrongly implemented model would automatically fail to accurately represent the real world, even though its conceptualization was right.

Completion of the checklist steps and resolving the detected errors should not be interpreted as a guarantee that the model is free of errors. Error detection is dependent on multiple aspects, including the reviewer's/validator's experience, as well as the reviewer's attention level at the time of verification. Therefore, the TECH-VER checklist does not intend to score the verification status of a model as that might give a false sense of security and cause overconfidence in the model, or might lead to using such a score out of context, such as arguing that a model with a higher score is better than the other model.

Similar to the validation status of a model, final judgment of the verification status of the model will always be a subjective judgment by the end-user of the model. However, the TECH-VER checklist aims to present an objective, transparent way of documenting and sharing information between different model users and model reviewers. In addition, it aims to assist the end-user's final judgment on the model's verification status by using a transparent documenting framework. This framework includes replicable tests, encapsulating all relevant parts of a typical decision analytical model. Therefore, verification checklists such as the TECH-VER checklist should be an integral part of the lifecycle of health economic decision analytical models.

Acknowledgments The authors would like to thank all the other health economists, whose feedback led to the revision of the initial version of TECHVER, resulting in this final version, as well as the anonymous reviewers and the editorial team for this special issue.

Author Contributions NCB wrote the initial drafts of the manuscript and conducted the verification tests on the beta version. All authors contributed to the conception and planning of the work, and critically revised and approved the final submitted version of the manuscript.

Data Availability Data sharing was not applicable to this article as no datasets were generated or analyzed during the current study. The TECH-VER checklist and the list of black-box tests created in this paper and a few model verification examples has been uploaded to an open access, online platform (https://github.com/nasuhcagdas/TECHVER), where other users will also be able to upload their original verification efforts and tests as well as on the institute's website (https://www.imta.nl/techver).

Compliance with Ethical Standards

Funding No funding was received for this study.

Conflict of interest Nasuh C. Büyükkaramikli, Maureen P.M.H. Rutten-van Mölken, Johan L. Severens, and Maiwenn Al have no conflicts of interest that are relevant to the content of this review.

Appendix

See Table 3.

Table 3 Literature review strategy using the Embase.com interface

#6	#5 AND ([embase]/lim OR ([medline]/lim NOT ([embase classic]/lim AND [medline]/lim))) AND ('Article'/it OR 'Article in Press'/it)	3500
#5	#3 AND #4	8152
#4	checklist:ab,ti OR valida*:ab,ti OR validi*:ab,ti OR verify*:ab,ti OR verifi*:ab,ti OR error:ab,ti OR guideline*:ab,ti OR credib*:ab,ti OR questionnaire*:ab,ti	2,272,728
#3	#1 AND #2	42,232
#2	'economic model' OR 'simulation'/de OR ('model'/de AND ('economics'/exp OR 'economic aspect'/exp)) OR 'decision tree'/de OR (((model OR modeling OR modelling OR simulation* OR microsimulation*) NEAR/6 (econom* OR pharmacoeconom* OR cost OR costs)):ab,ti) OR ((decision NEAR/3 (analy* OR tree OR trees)):ab,ti) OR 'discrete event*':ab,ti OR 'state transition':ab,ti OR markov:ab,ti OR (((individual* OR 'patient level*') NEAR/3 (sampl* OR simulation*)):ab,ti) OR ((dynamic NEAR/3 transmission*):ab,ti) OR probabilistic*:ab,ti OR 'partition* survival*':ab,ti OR 'he model*':ab,ti OR ((economic NEAR/1 submission*):ab,ti)	295,055
#1	'biomedical technology assessment'/exp OR 'economic evaluation'/exp OR 'quality adjusted life year'/exp OR 'program cost effectiveness'/de OR ((technology NEAR/3 assessment*):ab,ti) OR ((economic* NEAR/3 (evaluat* OR value)):ab,ti) OR (((cost OR costs) NEAR/3 (benefit* OR effectiv* OR efficien* OR efficac* OR minim* OR utilit* OR consequen*)):ab,ti) OR ((qualit* NEAR/3 adjust* NEAR/3 ('life year*' OR lifeyear*)):ab,ti) OR qaly*:ab,ti OR 'health econ*':ab,ti OR pharmacoeconom*:ab,ti	431,838

References

1. Caro JJ, Eddy DM, Kan H, Kaltz C, Patel B, Eldessouki R, et al. Questionnaire to assess relevance and credibility of modeling studies for informing health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. Value Health. 2014;17(2):174–82.
2. Tappenden P, Chilcott JB. Avoiding and identifying errors and other threats to the credibility of health economic models. Pharmacoeconomics. 2014;32(10):967–79.
3. Sargent RG. Verification and validation of simulation models. In: Henderson SG, Biller B, Hsieh M-H, Tew JD, Barton RR (eds) Proceedings of the 2007 Winter simulation conference. Piscataway: Institute of Electrical and Electronic Engineers Inc.; 2007. p. 124–37.
4. Eddy DM, Hollingworth W, Caro JJ, Tsevat J, McDonald KM, Wong JB, et al. Model transparency and validation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7. Value Health. 2012;15(6):843–50.
5. Vemer P, Corro-Ramos I, van Voorn GA, Al MJ, Feenstra TL. AdViSHE: a validation-assessment tool of health-economic models for decision makers and model users. Pharmacoeconomics. 2016;34(4):349–61.
6. Dasbach EJ, Elbasha EH. Verification of decision-analytic models for health economic evaluations: an overview.". PharmacoEconomics. 2017;35(7):673–83.
7. Cox ER, Motheral B, Mager D. Verification of a decision analytic model assumption using real-world practice data: implications for the cost effectiveness of cyclo-oxygenase 2 inhibitors (COX-2s). Am J Manag Care. 2003;9(12):785–96.
8. Hammerschmidt T, Goertz A, Wagenpfeil S, Neiss A, Wutzler P, Banz K. Validation of health economic models: the example of EVITA. Value Health. 2003;6(5):551–9.
9. Joranger P, Nesbakken A, Hoff G, Sorbye H, Oshaug A, Aas E. Modeling and validating the cost and clinical pathway of colorectal cancer. Med Decis Mak. 2015;35(2):255–65.
10. McEwan P, Foos V, Palmer JL, Lamotte M, Lloyd A, Grant D. Validation of the IMS CORE diabetes model. Value Health. 2014;17(6):714–24.
11. Möller J, Davis S, Stevenson M, Caro JJ. Validation of a DICE simulation against a discrete event simulation implemented entirely in code. PharmacoEconomics. 2017;35(10):1103–9.
12. Pichon-Riviere A, Augustovski F, Bardach A, Colantonio L, Latinclen Tobacco Research Group. Development and validation of a microsimulation economic model to evaluate the disease burden associated with smoking and the cost-effectiveness of tobacco control interventions in Latin America. Value Health. 2011;14(5):S51–9.
13. Van Gestel A, Severens JL, Webers CA, Beckers HJ, Jansonius NM, Schouten JS. Modeling complex treatment strategies: construction and validation of a discrete event simulation model for glaucoma. Value Health. 2010;13(4):358–67.
14. Willis M, Asseburg C, He J. Validation of economic and health outcomes simulation model of type 2 diabetes mellitus (ECHO-T2DM). J Med Econ. 2013;16(8):1007–21.
15. Willis M, Johansen P, Nilsson A, Asseburg C. Validation of the economic and health outcomes model of type 2 diabetes mellitus (ECHO-T2DM). Pharmacoeconomics. 2017;35(3):375–96.
16. Ye W, Brandle M, Brown MB, Herman WH. The Michigan model for coronary heart disease in type 2 diabetes: development and validation. Diabetes Technol Ther. 2015;17(10):701–11.
17. Wolff R, Büyükkaramikli N, Al M, Ryder S, Birnie R, Armstrong N, et al. Secukinumab for ankylosing spondylitis after inadequate response to non-steroidal antiinflammatory drugs or TNF-alpha inhibitors: a Single Technology Assessment. York: Kleijnen Systematic Reviews Ltd; 2016.
18. Büyükkaramikli NC, de Groot S, Fayter D, Wolff R, Armstrong N, Stirk L, et al. Pomalidomide with dexamethasone for treating relapsed and refractory multiple myeloma previously treated with lenalidomide and bortezomib: an evidence review group perspective of an NICE single technology appraisal. PharmacoEconomics. 2018;36(2):145–59.
19. Riemsma R, Büyükkaramikli N, Corro Ramos I, Swift SL, Ryder S, Armstrong N, et al. Idelalisib for treating refractory follicular lymphoma: a single technology assessment. York: Kleijnen Systematic Reviews Ltd; 2018.
20. Büyükkaramikli NC, Blommestein HM, Riemsma R, Armstrong N, Clay FJ, Ross J, et al. Ramucirumab for treating advanced gastric cancer or gastro-oesophageal junction adenocarcinoma previously treated with chemotherapy: an evidence review group perspective of a NICE single technology appraisal. PharmacoEconomics. 2017;35(12):1211–21.
21. Riemsma R, Corro Ramos I, Büyükkaramikli N, Fayter D, Armstrong N, Ryder S, et al. Glecaprevir-pibrentasvir for treating chronic hepatitis C; a single technology assessment. York: Kleijnen Systematic Reviews Ltd; 2017.
22. Büyükkaramikli NC, de Groot S, Riemsma R, Fayter D, Armstrong N, Portegijs P, et al. Ribociclib with an aromatase inhibitor for previously untreated, HR-positive, HER2-negative, locally advanced or metastatic breast cancer: an Evidence Review Group perspective of a NICE Single Technology Appraisal. PharmacoEconomics. 2019;37(2):141–53.
23. Limaye MG. Software testing. New York: Tata McGraw-Hill Education; 2009.
24. Hoogendoorn M, Rutten-van Mölken MPMH, Hoogenveen RT, et al. Working paper: comparing the cost-effectiveness of a wide range of COPD interventions using a stochastic, dynamic, population model for COPD. 2010. Available at: http://www.bmg.eur.nl/fileadmin/ASSETS/bmg/Onderzoek/Onderzoeksrapporten___Working_Papers/OR2010.01.pdf. Accessed 15 Aug 2019.
25. Hoogendoorn M, Feenstra TL, Hoogenveen RT, Rutten-van Mölken MP. Long-term effectiveness and cost-effectiveness of smoking cessation interventions in patients with COPD. Thorax. 2010;65(8):711–8.

Affiliations

Nasuh C. Büyükkaramikli[1] · Maureen P. M. H. Rutten-van Mölken[1] · Johan L. Severens[1] · Maiwenn Al[1]

✉ Nasuh C. Büyükkaramikli
buyukkaramikli@imta.eur.nl

[1] Institute for Medical Technology Assessment (iMTA), Erasmus School of Health Policy and Management (ESHPM), Erasmus University Rotterdam, Rotterdam, The Netherlands