Salah Ghabri, Matt Stevenson, Jörgen Möller, J. Jaime Caro

Trusting the results of model-based economic analyses: is there a pragmatic validation solution?

Article (Accepted version) (Refereed)

Original citation:
Ghabri, Salah and Stevenson, Matt and Möller, Jörgen and Caro, J. Jaime (2018) *Trusting the results of model-based economic analyses: is there a pragmatic validation solution?* PharmacoEconomics. ISSN 1170-7690

DOI: 10.1007/s40273-018-0711-9

Trusting the Results of Model-Based Economic Analyses: Is There a Pragmatic Validation Solution?

Salah Ghabri[1] PhD, Matt Stevenson[2] BSc, PhD, Jörgen Möller MEng[3], J. Jaime Caro[3,4,5], MDCM

1. French National Authority for Health (HAS), Saint-Denis, France

2. School of Health and Related Research, University of Sheffield, Sheffield, UK

3. Evidera, London, UK

4. McGill University, Montreal Canada

5. London School of Economics, London, UK

Word count: 2,888

Running title: Trusting Model results

Corresponding author:

J. Jaime Caro

39 Bypass Rd

Lincoln MA

01773

j.caro@lse.ac.uk

Jaime.caro@mcgill.ca

Jaime.caro@evidera.com

+1 978 760 4627

Acknowledgements

The authors would like to thank Isaac Corro Ramos, Pepijn Vemer, George A.K van Voorn, Maiwenn J. Al, Talitha L. Feenstra and Chloé Herpin for their useful comments and suggestions.

S Ghabri drafted sections 2 and 3.1; M Stevenson drafted section 3.2; J Möller drafted section 3.3; JJ Caro reviewed these materials and integrated them into the paper and all authors participated in writing and editing sections 1, 2, and 4.

Compliance with ethical standards

No funding was received for the preparation of this manuscript. S. Ghabri and M. Stevenson report no conflicts of interest. J Möller and J Caro are employed by, Evidera, a company that provides consulting and other research services to pharmaceutical, device, government, and non-government organizations. The opinions expressed in this article are those of the authors and do not necessarily represent the views of their institutions.

Abstract

Models have become a nearly essential component of health technology assessment. This is because the efficacy and safety data available from clinical trials are insufficient to provide the required estimates of impact of new interventions over long periods of time and for other populations and subgroups. Despite more than five decades of use of these decision-analytic models, decision makers are still often presented with poorly validated models and thus trust in their results is impaired. Among the reasons for this vexing situation are the artificial nature of the models, impairing their validation against observable data, complexity in their formulation and implementation, lack of data against which to validate the model results, and the challenges of short timelines and insufficient resources. This paper addresses this crucial problem of achieving models that produce results that can be trusted and the resulting requirements for validation and transparency, areas where our field is currently deficient. Based on their differing perspectives and experiences, the authors characterize the situation, outline requirements for improvement and pragmatic solutions to the problem of inadequate validation.

Key points

Although models are frequently used to inform health technology assessments, they tend to be poorly verified and largely unvalidated. This makes it difficult to trust their results.

Validation of these models is challenging because there are few opportunities to check their results against applicable real-world observations.

Practical steps to improved model validity include transitioning from bespoke to open-source models that leverage standard modules and detailed, transparent documentation of all aspects including steps taken to verify and validate the model.

1. Background & objective

The decision-analytic model is an important tool for health technology assessment (HTA). These models have been used to extend the efficacy and safety information obtained from clinical trials to broader populations, time horizons and outcomes of interest (e.g. the quality-adjusted life year, QALY). Despite the extensive use of these decision-analytic models, their ability to predict the outcomes of interest in a reasonably accurate and unbiased way is usually unknown because it has not been assessed (i.e., validation) [1-2].. Moreover, these models are often constructed under severe time and other constraints and complexly formulated and, thus, it is very difficult to ensure that they perform as intended (i.e., verification) [3]. In part, this situation results from the lack of sufficient data to both populate the model and validate it [4], but with the increasing availability of large datasets this reason is diminishing. Nevertheless, the short timelines and unreasonable pressures to keep these models simple and transparent often impair their accuracy and usefulness [5].

In this opinion piece, we address the crucial topic of trusting the results of economic analyses based on decision analytic models. Based on our different perspectives and experiences, we characterize the situation leveraging a pilot study on the cost-effectiveness analyses (CEA) submitted to the French National Authority for Health ("Haute Autorité en Santé", or HAS), as well as the experience of modeling experts, one providing services to the National Institute for Health and Care Excellence (NICE) through one of its independent Evidence Review Group (ERG), and another constructing models at a consulting company. We conclude by discussing the prospects for possible options to the problem of inadequate validation.

2. What are the approaches to model validation?

The ISPOR-SMDM Modeling Good Research Practices Task Force [1] proposes five important questions regarding model verification and validation:

1. Is the model (its assumptions and structure) consistent with current knowledge on the question and the history and management of the pathology under study (face validity)?
2. Has the model been verified to ensure it was technically implemented without errors (internal validity)?
3. Do the outputs of the model correctly reflect the results provided by external (ideally independent) sources (external validity)?
4. Does the model predict accurately what will be observed in ongoing studies (predictive validity)?

5. Does the model accord with other models that use other approaches to the problem (cross validity, noted to be potentially misleading as other models may also be invalid)?

Current HTA guidelines reference these questions but without providing a practical framework for addressing them [2]. For example, the Canadian [6] and the NICE [7] guidelines recommend describing the validation undertaken, while the Belgian guidelines [8] insist on the importance of checking the model against other models for the same intervention and HAS guidelines [9] emphasize that a model should produce results that are "suitable" for decision-making. Australian and Dutch guidelines [10-11] do explicitly ask applicants to add structured reporting of model validation tests to their dossier using, for example, the Validation-Assessment Tool of Health-Economic Models for Decision Makers and Model Users (AdViSHE) [12].

The following experiences with model validation show that these requirements are usually not fulfilled.

3. Experiences with model validation

3.1. The French pilot study

In France, the process of economic evaluation of drugs and medical devices was implemented in October 2013 [13]. A pilot study was carried out in June 2017 to identify the main issues regarding validation of decision-analytic models [14]. Two health economists independently reviewed each of the 77 manufacturers' CEAs received at HAS between 2013 and 2017. Based on prior publications [1, 15-16], validation in each one was classified as: neither undertaken nor discussed; undertaken, but not discussed; not undertaken but discussed; undertaken and discussed. Of the 77, 45% were in oncology, 17% in infectious disease; 13% in cardiology and 25% in other therapeutic areas. Generally, there was a lack of a validation plan or statistical rationale including, for example, a transparent documentation of any attempts to adjust model parameters to fit particular known outcomes (i.e., calibration). Less than 50% reported a validation process and only 35% performed any external validation. These results were similar to those obtained in a comprehensive literature review [15].

This pilot study highlighted that finding a balance between building trustworthy models and the resource constraints and national regulatory deadlines is a challenge, especially for HTA agencies where appraisals are done within the organization. The paucity of model validation among those 77 submissions suggests that this is a difficult step, perhaps because it takes additional time to complete and requires finding external experts without conflicts of interest. By contrast with verification, external validation remains a major concern as it is not possible to externally validate all outcomes of

a model because independent external data cannot be easily found, especially in the case of new health technologies (e.g. immuno-oncology therapies) [17-18]. Another issue raised by this study was whether HTA agencies should provide specific guidelines on model validation. This will be addressed in the forthcoming update of the HAS guidelines on economic evaluation.

3.2. Experience of an ERG and NICE Committee member with models submitted by industry

Model verification and validation by an independent group is an essential step in the appraisal process undertaken by the National Institute for Health and Care Excellence (NICE) [19]. There are two major reasons for this: first, models submitted to support reimbursement of health technologies can be complex and it is not surprising that mistakes are made in their construction; second, where two assumptions are equally plausible, it is common that the company whose product is under evaluation selects the assumption that is more favorable to their product. In this second circumstance, it can be argued that the company would be negligent to their shareholders if they did not take the option to position the intervention in a more positive light. As far back as 2005, [20] a review of multiple technology appraisals (MTAs) indicated that there was systematic bias in the incremental cost effectiveness ratio (ICERs) estimated by the submitting company compared with those estimated by the independent academic group, with the ICERs estimated by the company being systematically more favorable (p value <0.001).

The NICE appraisal process has now largely moved to single technology appraisals (STAs). In STAs, in contrast to MTAs, the ERGs do not construct their own model but critique the evidence submitted by the company and make amendments, where possible, to the company model. Due to time constraints, however, it is common that the ERG only highlights structural limitations but does not amend the model. Based on the experience of one of the authors (MS), who is both an appraisal committee member and director of an ERG, the change to STAs has not altered the typical rank-ordering of the company-submitted ICER and the ICER deemed most plausible by the ERG. One recent example of a model shown to be particularly favorable was the STA of azacytidine for acute myeloid leukemia [21], where the company's ICER of approximately £21,000/QALY gained increased to approximately £63,000 when undisputed errors in the model construction were corrected; and incorporating further amendments resulted in an ERG-preferred ICER of about £273,000. Another example is regorafenib for previously treated advanced hepatocellular carcinoma [22], where the company's initial submission resulted in a more favorable ICER because it did not pool data from two surveys as believed appropriate by the NICE cancer drugs fund committee and failed to mention one of these two surveys in their submission. Yet another recent example, is cenegermin for treating

neurotrophic keratitis where the NICE Appraisal Committee described the submitted model as "structurally flawed" and stated that the results "cannot be considered reliable" [23]. Numerous appraisals have been undertaken where the companies' assumptions relating to the most appropriate survival curves have been judged to be too optimistic by NICE appraisal committees. It is clear from this experience that an independent assessment of a company's model, particularly in the absence of credible validation, is essential. The exception is for interventions that are part of a class and which have similar costs and effectiveness as recommended comparators. In this instance, it should be the case that similar results are produced for each intervention regardless of the model structure, assumptions and parameter values. An example of this is golimumab for treating non-radiographic axial spondyloarthritis [24].

3.3. View of a modeler working at a consulting company

Building a model is always a challenge because it requires understanding enough of the complex process to be modeled to ensure that the simulation behaves in such a way that it corresponds reasonably well to reality. For HTA, this requires sufficient grasp of the intricacies of a disease and the effects of the interventions that may be implemented, some of which may be as yet poorly known. There is also the additional challenge that the context to be modeled is deliberately artificial. In the other fields from which HTA has borrowed modeling techniques (e.g., decision trees [25]; state-transition models [26]; dynamic transmission models [27], and system dynamics [28]) there is usually a real world against which to validate the model. This is obvious with weather forecasting, modeling of physical contexts like an airport luggage system, the integrity of a new airplane design, the trajectory of a missile, and so on. In HTA, however, the models simulate an idealized view of the world, where one or more of the following apply:

- All patients start at time 0
- There are no constrained resources, and thus no queues
- All treatment options are instantly available
- Variability is severely reduced
- Physicians and patients behave optimally
- Development of new interventions is frozen in time
- Time horizons are well beyond the observable immediate future

Without the real world as a validity check, and in the face of the often-conflicting demands of sponsors, good practice guidelines and tight timelines, HTA modeling must depend on following best-practices for construction of the model, expressly verifying all of it, and validating as much of it as possible. Best practices for construction and verification include:

Construction

- Not hiding anything
- Minimizing the use of custom macros or bespoke software subroutines
- Using pre-verified components (e.g., a standard module to select a death time from a specified distribution) as much as possible
- Permitting users to change the inputs
- Ensuring clear, detailed documentation (data decisions, sources, model structural choices, all assumptions and their consequences)

Verification

- Evaluating all equations separately to ensure that they are implemented correctly and yield expected results. If they feed values into each other the testing must cover this.
- Careful checking of the implementation to ensure that the model behaves as intended and doesn't have 'bugs' such as Markov leaks (when people disappear or reappear illogically) or event sequences happen in an impossible order.

Regardless of how well best practices for construction and verification are followed, the need for validation remains paramount.

Validation requires, at a minimum:

- Testing the model against the datasets used to develop it. Although this is entirely dependent validation, it supports the careful verification.
- Actively searching for independent data against which to validate model outputs.

4. A pragmatic solution?

In the preceding section, there is already the glimmer of a pragmatic solution. Following good practices will reduce the degree to which models are a "black box" and make it more difficult to present only those model implementations that favor a particular point of view. Increasing and formalizing efforts at verification [3] will decrease the flagrant errors and bolster confidence that the model is at least producing results consistent with its intended implementation. But, are these initiatives enough to bring about trust in those results? In the absence of a reality against which to validate, how do we get to a collective level of confidence that allows us to rely on the modeling results to inform the difficult decisions at issue in HTA?

The first step is for all involved to acknowledge that these models are not aiming to produce predictions for the real world. At best, the objective is to compare interventions in well-defined but artificial scenarios to get an idea of their potential cost-effectiveness. Once that is accepted, steps can be taken to ensure that the comparisons are not biased by choices made during construction or by failure to detect errors during verification. This requires that the modeling community stop reinventing the wheel—bespoke, product-specific models, including their parameterization, must become a thing of the past [3, 16].

A far more efficient and confidence-building approach is to engage in producing standard modules that can be extensively verified individually. These should be well-documented and freely-available to any modeler to use in creating a model. Then, when assembling them to construct a new model it is only the assembly that needs to be tested—a much lower hurdle than verifying the entire model. This also implies that if all stakeholders, including the HTA agencies have approved the modules, some of the problems described in the previous sections would be minimized.

De novo construction of models, even with pre-verified modules, should eventually become the exception, however [16]. For as many indications as possible, we need to progress to well-established, detailed models that can handle most contexts, interventions, patient populations and other important variations. These standard models should leverage as many of the validated modules as feasible; and portions that require something new should be taken as a prompt for the addition of new modules to the common library. Needless to say, the resulting models should also be available to anyone who wants to use them. This is in line with recent calls for “open-source” models [29]. Having many eyes poring over the conceptualization, the structure, its implementation, and all other aspects of the model over broad periods of time has a much better chance of uncovering problems than the rushed verification typical of bespoke models. After Mozilla’s transition of its Firefox browser to open source, for example, more than 7,000 bugs were reported in the six months between releases of an updated version [30]. The salutary effect of knowing that one’s work will be closely scrutinized is also undeniable.

This transition to open-source models should also substantially improve their validity, at least with respect to the objective of assessing potential cost-effectiveness. By the very nature of open-source, the models will also be more generalizable, work for multiple interventions and allow for all relevant events. They should, as well, become more transparent and well-documented. With multiple modelers contributing, the risk of a model having to be shelved due to a key developer disappearing is also minimized. While open-source may increase trust in the results of model-based economic analyses, the academic community, HTA organizations and private companies face major challenges

in terms of funding construction of models, protecting commercial or academic “in confidence” data, as well as technical barriers to implementation of open source models (e.g., agreement on the programming language).

Thus, until we move fully to open-source, we propose the following pragmatic steps to the vexing problem of trusting decision-analytic models intended for HTA:

1. Construct the model using, to the extent possible, pre-verified standard modules.
    - Those aspects that cannot be modeled using standard modules should be used as the basis for developing new modules, which then undergo full verification.
2. Build the model in incremental steps with documented checks of model behavior between each step.
3. Predict what is expected from the next model step and review the results of the next round based on that prediction.
4. Test the model against the datasets used to develop and document the tests.
5. Seek independent data against which to validate model outputs and document the searches, results and validation exercises.
6. Compare results against those of other models, if they exist, and try to explain any differences.
7. Make models intended for supporting health-care decisions available to anyone who wishes to review them (possibly under a non-disclosure agreement). Specifically, provide:
    - All code, equations, data sources
    - All analyses
        - Base case
        - Uncertainty
        - Scenarios
    - Detailed technical report
    - Verification reports
    - Validation exercises.

In addition, a standard platform for model implementation should be developed, perhaps under the auspices of a broadly representative consortium. These practical and feasible steps will make major strides in bolstering everyone’s trust in the results of model-based economic evaluations.

Although not the subject of this paper, it should be noted that successful validation of economic models depends on access to data. Thus, data holders can contribute to trust in modeled results by permitting modelers to use their data in validation exercises. Moreover, data on a product's

effectiveness, especially longer term, are often lacking. This can only be remedied with extended data collection, and perhaps by leveraging large data sets.

5. Conclusion

In this paper, we have summarized the experiences with model validation and verification from 3 differing viewpoints. It is clear that there is much progress to be made before models can be routinely trusted. We have suggested pragmatic steps which we believe would improve the validity of models. We are aware that their implementation not only will take time, but will also require a broad consensus among all public and private organizations using economic models for supporting health-care decisions.

References

1. Eddy DM, Hollingworth W, Caro JJ, Tsevat J, McDonald KM, Wong JB, ISPOR-SMDM Modeling Good Research Practices Task Force. Model transparency and validation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7. Med Decis Making 2012; 32:733-43.
2. Karnon J. Model Validation: Has it's Time Come? Pharmacoeconomics 2016; 34:829–31
3 Dasbach EJ, Elbasha EH. Verification of Decision-Analytic Models for Health Economic Evaluations: An Overview. PharmacoEconomics 2017; 35:673-83.
4. Ghabri S, Hamers F, Josselin J M, Exploring Uncertainty in Economic Evaluations of Drugs and Medical Devices: Lessons from the First Review of Manufacturers' Submissions to the French National Authority for Health. PharmacoEconomics 2016;34:617-24
5. Caro J, Jörgen Möller, Decision-Analytic Models: Current Methodological Challenges. PharmacoEconomics 2014; 32: 943
6. Canadian Agency for Drugs and Technologies in Health. Guidelines for the economic evaluation of health technologies: Canada. 4th ed. Ottawa: Canadian Agency for Drugs and Technologiesin Health; 2017. https://www.cadth.ca/dv/guidelineseconomic- evaluation-health-technologies-canada-4th-edition. Accessed 15 Oct 2017.
7. National Institute for Health and Care Excellence (NICE). Single technology appraisal: user guide for company evidence submission template. 2015. https://www.nice.org.uk/process/pmg24/chapter/cost-effectiveness. Accessed 15 Oct 2017.
8 Belgian Health care Knowledge Centre (KCE). Belgian Guidelines for economic evaluations and budget impact analysis.2015. https://kce.fgov.be/sites/default/files/page_documents/KCE_183_economic_evaluations_second_edition_Report.pdf. Accessed 6 Aug 2018.
9 Haute Autorité de Santé (HAS). Choices in methods for economic evaluation. 2012. https:// https://www.has-sante.fr/portail/upload/docs/application/pdf/2012-10/choices_in_methods_for_economic_evaluation.pdf Accessed 15 Oct 2017.
10 Department of Health, Commonwealth of Australia. Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee (PBAC), version5.0. September 2016. Available at https://pbac.pbs.gov.au/content/information/files/pbac-guidelines-version-5.pdf Accessed 6 Aug 2018.
11 Zorginstitut Nederland. Guideline for the Conduct of Economic Evaluations in Healthcare. 2016. Available at https://english.zorginstituutnederland.nl/publications/reports/2016/06/16/guideline-for-economic-evaluations-in-healthcare Accessed 6 Aug 2018.
12 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;34:349-61.
13 Journal Officiel de la République Francaise. Décret n°2012-1116 du 2 octobre 2012 relatif aux missions médico-économiques de la Haute Autorité de Santé. JORF n°0231 du 4 octobre 2012; page 15522 texte n°8. Available at https://www.legifrance.gouv.fr/;jsessionid=EEAEBD3C6A9DB9AFF25516B14F46DBDC.tplgfr32s_3 Accessed 6 Aug 2018
14 Ghabri S, Herpin C. Economic model validation: A pilot study on manufacturers submissions. Presented at ISPOR 20th Annual European congress, 2017.Available at https://www.ispor.org/docs/default-source/presentations/1328.pdf?sfvrsn=bba5258b_1 Accessed 6 Aug 2018.
15 Afzali HH, Gray J, Karnon J. Model performance evaluation (validation and calibration) in model-based studies of therapeutic interventions for cardiovascular diseases: a review and suggested reporting framework. Appl Health Econ Health Policy 2013;11:85-93
16 Afzali HH, Karnon J, Merlin T. Improving the accuracy and comparability of model-based economic evaluations of health technologies for reimbursement decisions: a methodological framework for the development of reference models. Med Decis Making 2013;33:325-32
17 Ciani O, Buyse M, Drummond M, Rasi G, Saad ED, Taylor RS. Time to review the role of surrogate end points in health policy: state of the art and the way forward. Value Health 2017;20:487–95.
18 Huang M, Latimer N, Zhang Y et al, Estimating the long-term outcomes associated with immuno-onclogy therapies: challenges and approaches for overall survival extrapolations. Value & Outcomes Spotlight 2018. :28-30.

19 NICE. Guide to the processes of technology appraisal. Available at https://www.nice.org.uk/Media/Default/About/what-we-do/NICE-guidance/NICE-technology-appraisals/technology-appraisal-processes-guide-apr-2018.pdf Accessed 6 Aug 2018
20 Miners AH, Garau M, Fidan D, Fischer AJ. Comparing estimates of cost effectiveness submitted to the National Institute for Clinical Excellence (NICE) by different organisations: retrospective study. BMJ 2005; 330:65-8.
21 Tikhonova I, Hoyle MW, Snowsill TM, Cooper C, Varley-Campbell JL, Rudin CE, Mujica Mota RE. Azacitidine for Treating Acute Myeloid Leukaemia with More Than 30 % Bone Marrow Blasts: An Evidence Review Group Perspective of a National Institute for Health and Care Excellence Single Technology Appraisal. Pharmacoeconomics 2017; 35:363-73.
22 NICE. Regorafenib for previously treated advanced hepatocellular carcinoma. Available at https://www.nice.org.uk/guidance/ta514/documents/final-appraisal-determination-document Accessed 6 Aug 2018.
23 NICE. Cenegermin for treating neurotrophic keratitis. Available at https://www.nice.org.uk/guidance/gid-ta10131/documents/appraisal-consultation-document Accessed 6 Aug 2018
24 NICE. Golimumab for treating non-radiographic axial spondyloarthritis. Available at https://www.nice.org.uk/guidance/ta497 Accessed 6 Aug 2018
25 Ransohoff DF, Feinstein AR. Editorial: Is Decision Modeling Useful in Clinical Medicine. Yale Journal of Biology and Medicine 1976; 41:761-7.
26 Beck JR, Pauker SG. The Markov process in medical prognosis. Med Decis Making 1983;3:419-58.
27 Pitman R, Fisman D, Zaric GS, Postma M, Kretzschmar M, Edmunds J, Brisson M, ISPOR-SMDM Modeling Good Research Practices Task Force. Dynamic transmission modeling: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--5. Value Health 2012;15:828-34.
28 Brailsford SC, Hilton NA. A comparison of discrete event simulation and system dynamics for modelling health care systems. In, Riley, J. (ed.) 2001. Planning for the Future: Health Service Quality and Emergency Accessibility. Operational Research Applied to Health Services (ORAHS) Glasgow Caledonian University.
29 Dunlop W, Mason N, Kenworthy J, Akehurst R. Benefits, challenges and potential strategies of open source health economic models. Pharmacoeconomics. 2017;35:125–8.
30 Wang J, Carroll JM. Behind Linus's Law: A preliminary analysis of open source software peer review practices in Mozilla and Python. Proceedings of the 2011 International Conference on Collaboration Technologies and Systems 2011;117-24.