Title: The Barriers and Facilitators to Model Replication within Health Economics

Running title: Replicating decision models in health economics

Abstract (249 words)

Objective:

Model replication is important as it enables researchers to check research integrity, transparency and, potentially, to inform the model conceptualisation process when developing a new or updated model. The aim of this study was to evaluate the replicability of published decision analytic models, and to identify the barriers and facilitators to replication.

Methods:

Replication attempts of five published economic modelling studies were made. The replications were conducted using only publicly available information within the manuscripts and supplementary materials. The replicator attempted to reproduce the key results detailed in the paper, for example the total cost, total outcomes and if applicable, the incremental cost effectiveness ratio reported. Whilst a replication attempt was not explicitly defined as a success or failure, the replicated results were compared in terms of percentage difference to the original results.

Results:

In conducting the replication attempts, common barriers and facilitators emerged. For the majority of the case studies, the replicator needed to make additional assumptions when recreating the model. This was often exacerbated by conflicting information being presented in the text and the tables. Across the case studies, the variation between original and replicated results ranged from -4.54% to 108.00% for costs and -3.81% to 0.40% for outcomes.

Conclusion:

This study demonstrates that whilst models may appear to be comprehensively reported, it is often not enough to facilitate a precise replication. Further work is needed to understand how to improve model transparency and in turn to increase the chances of replication, thus ensuring future usability.

Highlights (2 to 3 brief summary statements)

- The subject of replicability is widely discussed in other scientific disciplines, but has not yet been thoroughly explored within health economic decision models.

- Facilitators included clear model diagrams and tables of input parameters. Whilst the main barrier to replication was conflicting information presented in the text and tables, which required additional assumptions.

- Greater variation was found in the replicated costs than in the outcomes, suggesting more detail surrounding how costs are assigned may be warranted.

Introduction:

Computer programs have been labelled as 'black boxes'[1] due to their underlying code frequently being withheld. In a similar vein, health economic decision models may be perceived as 'black boxes' given that often, the underlying equations, assumptions and input parameters, are sparsely reported. As well, industry submissions to health technology assessments may include proprietary data inputs with limited accessibility.[2] This limits the ability of others to interrogate and replicate models, potentially damaging research integrity, as reporting and methodological quality cannot easily be assessed. With this in mind, maintaining a high level of transparency in the development and reporting of decision models is essential. Increasingly, there are calls for greater transparency in the way modelling studies are reported[3-5] and a proxy for such transparency could be to assess whether or not the model is replicable.

There is growing consensus that an independent modeller should be able to reproduce the results of a model, on the basis of only the published information.[5,6] Publications by the International Society for Pharmacoeconomics and Outcomes Research & The Society for Medical Decision Making (ISPOR-SMDM) Task Force provide an explicit definition of a transparent model. Significantly, this definition cites the process of replication, suggesting that the ability to replicate a model may indicate that it is transparent:

> "Transparency serves two purposes: 1) to provide a non-quantitative description of the model … and 2) to provide technical information to readers who want to evaluate a model at higher levels of mathematical and programming detail, and possibly replicate it".[5]

In another publication by the Task Force, the importance of transparency is repeated, stating that "a model should not be a 'black box' for the end-user but be as transparent as possible, so that the logic behind its results can be grasped at an intuitive level".[6] Sampson et al. suggest that whilst calls for model transparency have been numerous, "there are few signs of improvement in practice".[7] This is echoed in a study focusing on models used within oncology, concluding that, "there is a need for elevated rigor and transparency of reporting"[3] although exactly how this might be achieved is not discussed.

Aside from increased transparency, replicable models have practical benefits, in terms of potentially reducing research waste as well as researcher time if they can be used when developing future models, as advocated by the REWARD alliance.[8] For example, if existing models were easily replicable, future modellers might be able to use such models as a springboard in developing others, leaving more time to devote to validation work. Furthermore, Chilcott et al.[9] discuss the concept of replication as a method to check the face validity (ensuring results and structure make intuitive sense[10]) of models in the development process, as well as citing the potential benefits of replicating a model using different software.

Currently, the existing literature looking at replicating the results of published modelling studies is small.[11-13] Most recently, the collaborative diabetes modelling group, Mt Hood, published the results of their 2016 conference meeting focusing on 'Research Transparency' which set modellers the challenge of replicating two diabetes simulation models.[13,14] Due to the difficulties incurred during these replications, Mt Hood published a checklist designed to facilitate the reporting of model inputs specific to diabetes. More generally, decision models are subject to quality and reporting checklists, such as CHEERS[15] or Philips.[10]

This study looks to replicate several decision models across a variety of disease areas, examining the reasons facilitating or preventing replication and also exploring whether current reporting standards adequately facilitate replication. In doing so, it is hoped that this paper will highlight ways in which future modelling studies can be reported, for greater understanding by readers and to facilitate replication.

Methods

Five modelling studies were selected for replication, with each of the replication attempts detailed as case studies. The rationale for selecting these original publications varied. For case studies 1-3, each of the models were identified through a systematic review of models within that disease area.[16] The models selected for replication were considered to be the most thoroughly reported (in terms of the most number of Philips Checklist criteria satisfied[10]) and therefore were considered most likely to be replicable. The final model replications, case studies 4-5, were pragmatic, arising from the need to develop a model for a similar decision problem. The original model detailed in case study 4 was chosen as it had characteristics and health states that made it suitable for adaptation to address a specific question. Whereas the model replicated within case study 5 was selected as it was the most up to date and contextually relevant.

The primary focus of the replication attempts was to recreate the key figures presented in the text, such as total cost, outcomes and if applicable, the incremental cost effectiveness ratio (ICER), for the base case analyses. The replication did not extend to sensitivity analysis, however for one of the pragmatic case studies, probabilistic sensitivity analysis was replicated, and the results of this are discussed. The degree to which the original results varied in comparison to the replicated results were calculated as a percentage difference.[12]

As with the manuscripts in case studies 1-3, the Philips Checklist was also completed for case studies 4-5, with the intention of exploring whether any barriers and facilitators identified in the replications related to items within the checklist. For each of the criteria, a subjective response of ‘yes’, ‘no’, ‘partial’ or ‘not applicable’ was given. If it were found that the barriers were not picked up within the checklist or indeed that they were deemed to be suitably reported, it might suggest that existing reporting criteria are insufficient.

The replication studies were conducted by separate authors, and were done so using only the information presented in the referenced publications. Where possible, the same software as in the original publication was used in the replication. Importantly, publications were not selected to intentionally single out individual authors, journals, institutions, clinical areas or modelling methodologies. It is also important to note that the inability to replicate a model does not necessarily infer errors within the model, merely a lack of information within the report or even, the inability of the person replicating the study. In the same respect, the ability to replicate may not necessarily indicate model quality, given that a model may be transparently reported, but with inappropriate assumptions given the clinical condition or population being modelled.

Results

The general characteristics of the models, the scenarios selected for replication and the characteristics of the replicator are detailed in Table 1. Responses to the Philips Checklist criteria are provided in Table 2. It was found that the majority of criteria were fulfilled. The items that were considered as partially or not completed amongst the studies were mostly concerning the quality of the model, as opposed to the reporting detail which would facilitate replication. The results of the replications compared to the original publication, along with any differences, are shown in Table 3.

The replications highlighted the following general facilitators to replication:

- Detailed diagram of the model structure.
- Example calculations provided (for both costs and outcomes).
- Clear tables of assumptions.

Conversely, the following factors were found to be barriers to replication:

- Conflicting or inadequate information to inform model parameters or model structure.
- Uncertainty about assumptions made about parameters or model structure.
- For models with longer time frames, any difference in assumptions due to lack of information was compounded, resulting in greater variation between original and replicated results.
- Inability to clarify a model with the original author.

These are illustrated within the context of the five case studies below.

Case Study 1

The first replication was of a state-transition model developed using Microsoft Excel, described in a Health Technology Assessment monograph.[17] The model evaluated calcineurin inhibitors for the treatment of eczema in both adults and children, across eight scenarios. These scenarios were divided into population: adults or children, location of eczema: facial or body, as well as severity: mild to moderate or moderate to severe. For this case study, two scenarios were chosen for replication, which encompassed the widest range of options: the first evaluated pimecrolimus for the treatment of mild to moderate facial eczema within adults, whilst the second evaluated tacrolimus for moderate to severe body eczema amongst children. The models were constructed using treatment states comprised of different disease severity mixes, allowing for the fact that different disease severities could receive the same treatment yet have different utilities.

The results of the adult scenario replication varied by -2.84% to 16.49% for costs, 0.00% to 0.31% for outcomes, and the same overall conclusion regarding cost effectiveness was found (topical corticosteroids dominated). Whilst replicating the childhood scenario however, it became evident that numerous additional assumptions were required. Due to this and the extended time horizon of 14 years, any differences between the original and replicated model per cycle, were amplified, resulting in the replicated model returning values that were far removed from the original results (costs varying by 108% of those reported originally). Therefore no attempts to replicate other treatment pathways within this scenario were made. Whilst the costs were far removed, the outcomes replicated were relatively close to the original values (varying by -2.29%), which may suggest that rather than outright replicator error, the variation may have resulted from misinterpretation of costing assumptions.

The main barrier within this replication was the way in which the multiple scenarios, based on modifications of a general model, were presented. Although the transition probabilities were given, they were for all of the eight different scenarios together, with no clear labelling as to which transition probabilities related to which scenario. In addition, some of the transition probabilities were instead presented as the likelihood of patients being offered different treatments, having previously failed a treatment. This was further complicated by conflicting information within the text and the transitions presented in the table. For example, when recreating the adult scenario, it was stated that following a failed treatment of low-potency steroids, the probability that a patient would receive pimecrolimus was 0.85, mid-potency steroids, 0.1, and high-potency steroids, 0.05. However, this conflicted with information in the text, stating high-potency steroids were not a treatment option within this scenario. Consequently, this left a 0.05 probability to be allocated to a treatment with no description about how this should be done. An author of the original publication was contacted to provide clarification, however they were unable to help, citing the time that had passed since the publication, and current workload.

Facilitators:

- Detailed diagrams of the model structures, clearly depicting possible transitions.

Barriers:

- Areas where the text and tables conflicted.
- Grouping of transition probabilities, instead of presenting the values for each of the scenarios individually.
- Extended modelling time horizon meant any differences between the replicated and original model were amplified over time (in the childhood scenario).
- Unable to obtain clarification from the authors.

Case Study 2

This case study replicated a decision tree, modelling the use of proton pump inhibitors for the maintenance therapy of erosive reflux oesophagitis over a one-year time horizon,[18] using Microsoft Excel (instead of TreeAge, as used in the original study). The manuscript included a figure clearly showing the tree structure as well as a table of the probabilities used. These enabled the replicated model to closely compare to the original, outcomes were matched exactly, and costs ranged only by -4.43% to 1.20%.

Despite the simplicity of the decision tree, there were still some barriers to successful replication. These included discrepancies between the text and branch structure presented in the model diagram (which was later assumed to be purely descriptive) along with a lack of clarity surrounding how to cost the treatments used as maintenance therapy.

Facilitators:

- Simplistic model structure.

Barriers:

- Conflicting information between the text and model diagram.
- Lack of clarity regarding how costs were attributed during maintenance therapy.

Case Study 3

Affleck et al.[19] described a 15 state, state-transition model, built to evaluate treatment approaches for scalp psoriasis using four weekly cycles over a one year time horizon. Treatment pathways 1 and 5 were replicated.

The model was described comprehensively with tables of the transition probabilities, utilities, and descriptions of the health states being provided. In addition, a detailed diagram of both the model and the different treatment pathways being evaluated, was given. This enabled the replicated outcomes to vary from the original publication by only 0.37% and 0.40% across the two pathways, whereas the costs varied by -4.54% and 0.07%

Only minor barriers to replication were found, involving the way some of the costs, assigned to each of the health states, were described. Particularly, it was stated that a “weighted average of treatment modalities” was costed, although the weightings were not given.

Facilitators:

- A comprehensive table was provided which detailed the different health states of the model at baseline, any assumptions as well as the possible transitions from the state.

Barriers:

- Ambiguity surrounding the “weighted average” used when calculating the cost of treatments.

Case Study 4

In this case study, a state-transition model developed by Chambers et al.[20] evaluating the use of aspirin for stroke survivors was replicated. Additionally, some values were taken from a later paper by the same authors.[21] This model, run in the base case over 5 years and in other iterations, over a 2 and 25 year time horizon. Attempts were made to replicate the results from both the 5 and 25-year analyses, using Microsoft Excel.

In the base case, costs were replicated to within -0.91% and -0.67%, and outcomes were within 1.79% and 1.81%, in comparison to the original. Increased variation in costs was seen when the time horizon was extended to 25 years, with variation of 3.93% and 4.13%.

There was uncertainty relating to some parameters. This was due to the table giving a range for each of the parameters, instead of listing individual values for each of the time points. This simplified reporting, but made it unclear as to what value was used in particular cycles. Additionally, some values were reported with limited numbers of decimal places. In the model replication, total long-term costs over 25 years were overestimated by approximately 4%. Although total estimates of life years were very similar there were small discrepancies as disabled life years were slightly overestimated and disability free life years were slightly underestimated. Long term care costs were the largest cost and estimates per cycle were much higher for disabled stroke survivors so this would account for the additional estimate of cost. Therefore, very small discrepancies in the number of individuals in disabled states had the potential for larger discrepancies in expected costs.

Facilitators:

- Tables detailing how the main cost parameters were derived, along with a complete table of costs entered in the model, greatly facilitated the model replication.

Barriers:

- Ranges were given for the parameters, instead of individual values.

Case Study 5

The final case study focused on an evaluation conducted by Ganesalingam et al.[22] comparing mechanical thrombectomy to standard care alone: Intravenous tissue-type plasminogen activator (IV-Tpa), in cases of acute stroke. Analyses were carried out using a combined short-term decision tree and state-transition, cohort model. The time horizon was 20 years with discounting of costs and outcomes at 3.5%. A hypothetical cohort of 1000 patients was used in a simulation (replicated using Microsoft Excel, original software not stated).

The replication resulted in costs that varied by 2.00% and 1.39% in comparison to the original, whereas the outcomes varied by 0.13% and -0.87%. This case study was the only one where the interventions were not dominant or dominated, and so the ICER was also replicated. The original ICER was £11,651 per QALY, in comparison to £12,051 when using the replicated values, a total of 3.43% variation.

The model was thoroughly reported in the publication, with a diagram being provided and all of the parameters required to recreate the main analyses being clearly listed in a table. The cost per cycle for two of the model states was also given, which further facilitated the replication.

Despite the parameters being comprehensively reported, several barriers to replication were still encountered which required additional assumptions. These included uncertainty about the allocation of treatment costs following recurrent stroke, as well as how discounting was applied. It was unclear whether the first cycle was considered as time zero (given that 3 months was meant to have elapsed within the decision tree) and whether the cycles within the first year were or were not discounted.

Moreover, when trying to recreate some of the probabilistic sensitivity analyses conducted, it was apparent that not all of the distribution parameters were included in the publication, and although some of these were available in online supplementary materials, additional assumptions about the distributions used were required. Furthermore, the shape parameters reported for the beta distributions to generate utilities were implausible, as they generated values that were far lower than the point estimates.

Whilst carrying out this replication, an attempt was made to contact two of the authors to ask if they would be willing to share the original model code, however no response was received.

Facilitators:

- Model parameters clearly listed.
- Example of costs per cycle were given for two of the model states.

Barriers:

- Ambiguity about the assumptions made with treatment costs following recurrent stroke.
- Lack of clarity surrounding the 3 month decision tree and how this affected subsequent cycle discounting.
- Not all of the distributions used in the probabilistic sensitivity analyses were listed. As well, implausible shape parameters were given for the beta distributions.
- Unable to obtain clarification from the authors.

Discussion

The above case studies have highlighted several common barriers and facilitators to model replication. Facilitators to model replication included the provision of clear model diagrams which detailed all potential transitions as well as clearly reporting transition probabilities alongside the treatment pathways. Moreover, documenting the cost per cycle for model states (as was seen in case study 5), and providing example calculations for transitions, greatly facilitated model replication.

Common barriers to replication included conflicting information being presented in the text in comparison to what was presented in model diagrams or tables. Moreover, whilst all of the papers provided some sort of table describing the model input parameters, it was often found that these were grouped for multiple treatment pathways or time horizons, instead of being specific. In doing this, it was difficult to appreciate which parameter referred to which model iteration or scenario, and thus was a common barrier to replicating the analysis. The case studies replicating models with longer time horizons also proved harder to replicate given that any, even minor, discrepancies in costs, outcomes or transition probabilities between the original and replicated model became amplified over time. Furthermore, given that in two of the case studies an author was unsuccessfully contacted for clarification, this study should also act as a reminder to modellers to archive and thoroughly annotate their work, so as to facilitate any future enquiries.

Another finding of the above replications, was the tendency for greater variation in the replicated costs than outcomes, for example, costs ranged from -4.54% to 108.00% whereas outcomes varied by -3.81% to 0.40%. This might suggest that additional emphasis is needed when reporting the unit costs and assumptions about resource use, to assist with greater accuracy in model replication.

Moreover, in all of the case studies, it initially appeared that the original publications comprehensively reported the technical aspects of the decision models chosen for replication, as shown by the fact that the majority of the Philips Checklist criteria received a favourable response. Given that several common barriers were still encountered however suggests that the Philips Checklist may not be an entirely accurate means by which to ensure studies are replicable. This might be because the Philips Checklist focuses more on model quality, in that appropriate justifications are given, as opposed to reporting clarity. These findings are similar to those reported by the Mt Hood conference focusing on transparency,[14] who referred to the commonly used Philips[10] and CHEERS[15] checklists as being potentially "overly general to satisfy the needs in complicated multifactorial disease areas" to facilitate their replication. It remains to be seen if the checklist they went on to develop, facilitates future replication attempts, or if indeed other strategies need to be considered.[13]

It is the opinion of the authors that if future model replications are to be facilitated, then a formalised process of how models are presented, instead of further checklists, is required. For example, whilst providing a table of input parameters may appear to be enough to satisfy a checklist item, it may not be apparent, until a replication is conducted, if any parameters have been omitted.

Suggestions of ways to present models to facilitate replication might include a table with the total costs associated with each state. In providing these values, any implicit assumptions that the modeller had failed to document in the manuscript could be deciphered. Whilst these recommendations may be feasible when describing state-transition models, transparent reporting is likely to be far more challenging with complex models such as discrete event simulations. Given that models with a longer time horizon were found harder to replicate, it might be of use to provide a summary of results for a short term value (for example a year) so that a replicator could check against these before running the model over many more cycles (and thus inflating any inherent discrepancies).

A more thorough presentation of model structure and other elements could also be encouraged by changing workflow practices to give more consideration to replicability. Replication of the model programming by two separate individuals within the study team, both working from the same analysis plan, would encourage clear and unambiguous descriptions of model structure. This type of redundancy in workflow is already commonly practised by statisticians analysing clinical trial data as well as in software development (Chapter 9.7.9, Quality Control[23]).

Whilst this study has focused on how authors could present their modelling studies to facilitate easier replication, other factors such as journal data sharing policies, word limits and the use of supplementary materials should also be acknowledged as contributing factors. Additionally, if model registries[7,24] or the publishing of open source models[25] were more commonplace, replication may be more easily facilitated, given that a replicator could access and inspect model code (albeit that this would still require detailed annotation to be understood by a third party).

Strengths and Limitations

Though the spectrum of clinical conditions from which the modelling studies were selected with the intention of being broad, it is acknowledged that in the majority of case studies, a deterministic, state-transition model was replicated, and therefore the findings might not be generalisable to other, more complex, model types. There was also a limited range of modelling software used, and therefore the potential benefits or difficulties of using other software, such as R, were not explored. It should be noted that the results of the replications are highly dependent on the competencies of the replicator and therefore may not necessarily be a true reflection on the quality or replicability of the model within the original publication. Attempts were made to mitigate against this however, by using different modellers for each of the case studies, of whom had varying levels of experience. Importantly, it should be reiterated that none of the case studies were chosen to single out any author, institution or journal.

Conclusion

It is hoped that this study will act as a catalyst to review the ways in which models are currently presented. From the case studies above it is evident that even if a model appears to be largely well reported, it may still be difficult to precisely replicate the results presented, if even a single assumption or parameter value is omitted. Therefore, a review of how modelling studies are conducted and presented is required.

Conflicts of Interest

The authors report no conflicts of interest.

Acknowledgements

The authors would like to thank the reviewers of the manuscript for their valuable comments. They would also like to thank the HESG conference delegates (Gran Canaria, 2016) for their discussion of one of the replications in this paper.

Thanks also go to the authors whose models were replicated in this publication.

Funding Source

The study was funded by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care (CLARHC) East of England (project HE7: Dec 2015). Funding for EM and TS was also received via TS's Career Development Fellowship (CDF award-2014-07-006) supported by the National Institute for Health Research. The views expressed in this paper are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health and Social Care.

References:

1. Morin A, Urban J, Adams P, et al. Shining light into black boxes. *Science.* 2012;336(6078):159-160.
2. Drummond MF, Schwartz JS, Jönsson B, et al. Key principles for the improved conduct of health technology assessments for resource allocation decisions. *Int J Technol Assess Health Care.* 2008;24(3):244-258.
3. Beca J, Husereau D, Chan KK, Hawkins N, Hoch JS. Oncology Modeling for Fun and Profit! Key Steps for Busy Analysts in Health Technology Assessment. *Pharmacoeconomics.* 2017:1-9.
4. Padula WV, McQueen RB, Pronovost PJ. Can Economic Model Transparency Improve Provider Interpretation of Cost-effectiveness Analysis? Evaluating Tradeoffs Presented by the Second Panel on Cost-effectiveness in Health and Medicine. *Med Care.* 2017;55(11):909-911.
5. Eddy DM, Hollingworth W, Caro JJ, Tsevat J, McDonald KM, Wong JB. Model transparency and validation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7. *Value Health.* 2012;15(6):843-850.
6. Weinstein MC, O'brien B, Hornberger J, et al. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices—Modeling Studies. *Value Health.* 2003;6(1):9-17.
7. Sampson CJ, Wrightson T. Model Registration: A Call to Action. *PharmacoEconomics - Open.* 2017;1(2):73-77.
8. The Reward Alliance. The REWARD statement. http://researchwaste.net/reward-statement/. Published 2016. Accessed 19th April, 2016.
9. Chilcott J, Tappenden P, Rawdin A, et al. Avoiding and identifying errors in health technology assessment models: qualitative study and methodological. *Health Technol Assess.* 2010;14(25).
10. Philips Z, Ginnelly L, Sculpher M, et al. Review of guidelines for good practice in decision-analytic modelling in health technology assessment. *Health Technol Assess.* 2004;8.

11. Bermejo I, Tappenden P, Youn J-H. Replicating Health Economic Models: Firm Foundations or a House of Cards? *Pharmacoeconomics.* 2017:1-9.

12. Smolen LJ, Klein TM, Kelton K. Replication Of A Published Markov Chronic Migraine Cost-Effectiveness Analysis Model For Purposes Of Early Phase Adaptation And Expansion. *Value Health.* 2015;18(3):A19.

13. Palmer AJ, Si L, Tew M, et al. Computer Modeling of Diabetes and Its Transparency: A Report on the Eighth Mount Hood Challenge. *Value Health.* 2018.

14. Mt Hood Diabetes Challenge. The Mount Hood 2016 Challenge, Switzerland. https://docs.wixstatic.com/ugd/4e5824_36cb1fd0aca94f1980d8a2228cf7e6e8.pdf. Published 2016. Accessed 14th November, 2017.

15. Husereau D, Drummond M, Petrou S, et al. Consolidated health economic evaluation reporting standards (CHEERS) statement. *Cost Effectiveness and Resource Allocation.* 2013;11(1):6.

16. McManus E, Sach T, Levell N. The Use of Decision–Analytic Models in Atopic Eczema: A Systematic Review and Critical Appraisal. *Pharmacoeconomics.* 2017:1-16.

17. Garside R, Stein K, Castelnuovo E, et al. The effectiveness and cost-effectiveness of pimecrolimus and tacrolimus for atopic eczema: a systematic review and economic evaluation. *Health Technol Assess.* 2005;9(29):1-230.

18. Dean BB, Siddique RM, Yamashita BD, Bhattacharjya AS, Ofman JJ. Cost-effectiveness of proton-pump inhibitors for maintenance therapy of erosive reflux esophagitis. *Am J Health Syst Pharm.* 2001;58(14):1338-1346.

19. Affleck AG, Bottomley JM, Auland M, Jackson P, Ryttov J. Cost effectiveness of the two-compound formulation calcipotriol and betamethasone dipropionate gel in the treatment of scalp psoriasis in Scotland. *Curr Med Res Opin.* 2011;27(1):269-284.

20. Chambers M, Hutton J, Gladman J. Cost-effectiveness analysis of antiplatelet therapy in the prevention of recurrent stroke in the UK. *Pharmacoeconomics.* 1999;16(5):577-593.

21. Chambers MG, Koch P, Hutton J. Development of a decision-analytic model of stroke care in the United States and Europe. *Value Health.* 2002;5(2):82-97.

22. Ganesalingam J, Pizzo E, Morris S, Sunderland T, Ames D, Lobotesis K. Cost-utility analysis of mechanical thrombectomy using stent retrievers in acute ischemic stroke. *Stroke.* 2015;46(9):2591-2598.

23. Medicines and Healthcare products Regulatory Agency. *Good Clinical Practice Guide.* 9th ed. London: The Stationery Office; 2018.

24. Arnold RJ, Ekins S. Ahead of our time: collaboration in modeling then and now. *PharmacoEconomics.* 2017;35(9):975-976.

25. Dunlop WC, Mason N, Kenworthy J, Akehurst RL. Benefits, Challenges and Potential Strategies of Open Source Health Economic Models. *Pharmacoeconomics.* 2017;35(1):125-128.

Case Study	**Replicator Category of Experience***	**First author (year)**	**Model Type**	**Disease**	**Population**	**Intervention / Comparator**	**Perspective**	**Software**	**Time Horizon (Cycle Length)**	**Health Outcome**
1	2	Garside (2005)	State-transition, Cohort	Atopic Eczema	Adults with mild to moderate facial eczema	Pimecrolimus / TCS	NHS	Microsoft Excel	1 year (4 weeks)	QALY
			State-transition, Cohort	Atopic Eczema	Children with moderate to severe body eczema	Tacrolimus / TCS	NHS	Microsoft Excel	14 years (4 weeks)	QALY
2	1	Dean (2001)	Decision Tree	Erosive Reflux Oesophagitis	“Ambulatory care patients”	Rabeprazole, Omeprazole, Lansoprazole	Third-party Payer	Data TreeAge	1 year (Not applicable)	Percentage of symptomatic recurrences prevented

3	1	Affleck (2011)	State-transition, Cohort	Psoriasis	Adults with moderately severe scalp psoriasis	TCF gel / First line therapy BMV	NHS in Scotland	Microsoft Excel	1 year (4 weeks)	QALY
4	3	Chambers (1999)	State transition, Cohort	Stroke	Stroke survivors	Antiplatelet therapy for prevention of recurrent stroke	Broad health and social care	Data TreeAge	Base case: 5 years, extended to 25 years (3 months)	Number of strokes, life years
5	3	Ganesalingam (2015)	Decision tree State-transition, Cohort	Stroke (acute)	Adults suffering acute stroke	Mechanical thrombectomy / IV-tPA	NHS	Software not stated	20 years (3 months)	QALY

Abbreviations: TCS; Topical corticosteroid. TCF; Two-compound formulation. BMV; betamethasone valerate. IV-tPA; Intravenous tissue-type plasminogen activator.

*Replicator experience defined according to the following categories:

1. Recent training in decision modelling as part of an MSc in Health Economics. The replication was conducted as part of an MSc dissertation with supervision from two health economists of category 2/3 experience.

2. Early career researcher with a background in Mathematics and an MSc in Health Economics. Over 4 years’ experience.

3. Experienced health economist with significant experience of decision modelling (8 to 20 years’ experience).

All replicators were new to the clinical area of the model they replicated.

Table 1: Description of the models included within each of the case studies.

	Case Study				
Checklist Item	1	2	3	4	5
Is there a clear statement of the decision problem?	y	y	y	y	y
Is the objective of the evaluation and model specified and consistent with the stated decision problem?	y	y	y	y	y
Is the primary decision maker specified?	y	y	y	y	y
Is the perspective of the model stated clearly?	y	y	y	y	y
Are the model inputs consistent with the stated perspective?	y	y	y	y	y
***Has the scope of the model been stated and justified?**	**y**	**y**	**y**	**y**	**y**
Are the outcomes of the model consistent with the perspective, scope and overall objective of the model?	y	y	y	y	y
***Has the evidence regarding the model structure been described?**	**y**	**n**	**y**	**y**	**n**
Is the structure of the model consistent with a coherent theory of the health condition under evaluation?	y	y	y	y	y
Have any competing theories regarding model structure been considered?	y	n	n	n	n
***Are the sources of data used to develop the structure of the model specified?**	**y**	**n**	**y**	**n**	**n**
Are the causal relationships described by the model structure justified appropriately?	y	y	y	y	y
***Are the structural assumptions transparent and justified?**	y	y	y	y	y
Are the structural assumptions reasonable given the overall objective, perspective and scope of the model?	y	y	y	y	y
***Is there a clear definition of the options under evaluation?**	**y**	**y**	**y**	**y**	**y**
Have all feasible and practical options been evaluated?	y	y	y	y	y
Is there justification for the exclusion of feasible options?	y	n/a	n/a	n/a	n/a
Is the chosen model type appropriate given the decision problem and specified causal relationships within the model?	y	n	y	y	y
Is the time horizon of the model sufficient to reflect all important differences between options?	y	n	y	y	y
***Is the time horizon of the model, and the duration of treatment and treatment effect described and justified?**	**y**	**y**	**y**	**y**	**y**
Has a lifetime horizon been used? If not, has a shorter time horizon been justified?	n	n	n	y	y
Do the disease states (state transition model) or the pathways (decision tree model) reflect the underlying biological process of the disease in question and the impact of interventions?	y	y	y	y	y
***Is the cycle length defined and justified in terms of the natural history of disease?**	**p**	**n/a**	**y**	**p**	**p**
Are the data identification methods transparent and appropriate given the objectives of the model?	y	y	y	y	y
Where choices have been made between data sources, are these justified appropriately?	y	n/a	y	n/a	n/a
Has particular attention been paid to identifying data for the important parameters in the model?	y	y	y	y	y
Has the process of selecting key parameters been justified and systematic methods used to identify the most appropriate data?	y	y	y	y	y
Has the quality of the data been assessed appropriately?	y	y	y	y	n

Where expert opinion has been used, are the methods described and justified?	p	p	y	y	n/a
Are the pre-model data analysis methodology based on justifiable statistical and epidemiological techniques?	y	y	y	y	y
***Is the choice of baseline data described and justified?**	**y**	**y**	**y**	**y**	**y**
***Are transition probabilities calculated appropriately?**	**y**	**y**	**y**	**y**	**y**
***Has a half cycle correction been applied to both cost and outcome?**	**n**	**n/a**	**n**	**n**	**n**
If relative treatment effects have been derived from trial data, have they been synthesised using appropriate techniques?	y	y	y	y	y
***Have the methods and assumptions used to extrapolate short-term results to final outcomes been documented and justified? Have alternative assumptions been explored through sensitivity analysis?**	**y**	**y**	**y**	**y**	**y**
***Have assumptions regarding the continuing effect of treatment once treatment is complete been documented and justified? Have alternative assumptions been explored through sensitivity analysis?**	**y**	**y**	**n**	**y**	**y**
Are the utilities incorporated into the model appropriate?	y	n/a	y	n/a	y
***Is the source for the utility weights referenced?**	**y**	**n/a**	**y**	**n/a**	**y**
Are the methods of derivation for the utility weights justified?	y	n/a	y	n/a	y
***Have all data incorporated into the model been described and referenced in sufficient detail?**	**y**	**y**	**y**	**y**	**y**
Has the use of mutually inconsistent data been justified (i.e. are assumptions and choices appropriate)?	n/a	n/a	n/a	n/a	n/a
***Is the process of data incorporation transparent?**	**y**	**y**	**y**	**y**	**y**
***If data have been incorporated as distributions, has the choice of distribution for each parameter been described and justified?**	**n/a**	**n/a**	**n/a**	**n/a**	**p**
Have the four principal types of uncertainty been addressed?	p	p	p	p	p
If not, has the omission of particular forms of uncertainty been justified?	n	n	n	n	n
Have methodological uncertainties been addressed by running alternative versions of the model with different methodological assumptions?	n	n	n	y	n
Is there evidence that structural uncertainties have been addressed via sensitivity analysis?	n	y	n	n	y
Has heterogeneity been dealt with by running the model separately for different sub-groups?	y	n	n	n	n
Are the methods of assessment of parameter uncertainty appropriate?	y	y	y	y	y
Has probabilistic sensitivity analysis been done, if not has this been justified?	y	n	n	n	y
***If data are incorporated as point estimates, are the ranges used for sensitivity analysis stated and justified?**	**y**	**y**	**p**	**y**	**p**
Is there evidence that the mathematical logic of the model has been tested thoroughly before use?	n	n	n	n	n
Are the conclusions valid given the data presented?	y	y	y	y	y
Are any counterintuitive results from the model explained and justified?	n/a	n/a	n/a	n/a	n/a
If the model has been calibrated against independent data, have any differences been explained and justified?	y	n/a	n/a	n/a	n/a
Have the results of the model been compared with those of previous models and any differences in results explained?	y	n	y	y	y

Abbreviations: Y: Yes; N: No; P: partial; N/A: Not applicable.

Where it was unclear if a criteria was satisfied the reviewer erred on the side of caution and responded ‘No’.

*These were identified as criteria that might have the greatest influence on the replicability of a modelling study, given that they directly related to reporting of items needed for replication.

Table 2: Responses to the Philips Checklist Criteria[10] for all of the modelling studies replicated

		RESULTS					
		Cost per patient			Health Outcome per patient		
	Scenario Replicated	**Original**	**Replication**	**Difference (%)**	**Original**	**Replication**	**Difference (%)**
Case Study 1	Base case (adults, no pimecrolimus)	39.39	38.27	-1.12 (-2.84%)	0.968	0.968	0.000 (0.00%)
	Adults, pimecrolimus as second-line treatment	70.58	79.69	9.11 (12.91%)	0.961	0.964	0.003 (0.31%)
	Adults, pimecrolimus as first-line treatment	135.44	157.,78	22.34 (16.49%)	0.967	0.967	0.000 (0.00%)
	Base case (children, no tacrolimus)	956.47	1,989.44	1,032.97 (108.00%)	1.085	1.060	-0.248 (-2.29%)
Case Study 2	Rabeprazole	1414	1431	17 (1.20%)	86	86	0 (0.00%)
	Lansoprazole	1671	1597	-74 (-4.43%)	68	68	0 (0.00%)
	Omeprazole	1599	1581	-18 (-1.13%)	81	81	0 (0.00%)
Case Study 3	Base case (TCF gel as first line therapy)	241.86	230.89	-10.97 (-4.54%)	0.7818	0.7847	0.003 (0.37%)
	First line therapy as BMV	255.12	255.29	0.17 (0.07%)	0.7801	0.7832	0.003 (0.40%)

Case Study 4	No treatment (5 year time horizon)	15,093	14,955	-138 (-0.91%)	3.911	3.981	0.070 (1.79%)
	Aspirin (5 year time horizon)	14,817	14,717	-100 (-0.67%)	3.918	3.989	0.071 (1.81%)
	No treatment (25 year time horizon)	24,881	25,858	977 (3.93%)	7.607	7.585	-0.022 (-0.29%)
	Aspirin (25 year time horizon)	24,491	25,503	1,012 (4.13%)	7.664	7.643	-0.021 (-0.27%)
Case Study 5	Base case (IV-tPA)	52,495	53,545	1,050 (2.00%)	3.790	3.795	0.005 (0.13%)
	Thrombectomy	64,757	65,656	899 (1.39%)	4.842	4.800	-0.042 (-0.87%)

Abbreviations: TCS; Topical corticosteroid. TCF; Two-compound formulation. BMV; betamethasone valerate. IV-tPA; Intravenous tissue-type plasminogen activator.

% difference calculated using the following formula: ((Replication – Original) / Original) x 100%

Table 3: Per patient results of the replication attempts in comparison to the original results reported in the publication.