ORIGINAL RESEARCH ARTICLE

Replication of Published Health Economic Obesity Models: Assessment of Facilitators, Hurdles and Reproduction Success

Björn Schwander[1,2] · Mark Nuijten[3] · Silvia Evers[1,4] · Mickaël Hiligsmann[1]

Abstract

Objectives This research aims to (1) replicate published health economic models, (2) compare reproduced results with original results, (3) identify facilitators and hurdles to model replicability and determine reproduction success, and (4) suggest model replication reporting standards to enhance model reproducibility, in the context of health economic obesity models.

Methods Four health economic obesity models simulating an adult UK population were identified, selected for replication, and evaluated using the Consolidated Health Economic Evaluation Reporting Standards (CHEERS) checklist. Reproduction results were compared to original results, focusing on cost-effectiveness outcomes, and the resulting reproduction success was assessed by published criteria. Replication facilitators and hurdles were identified and transferred into related reporting standards.

Results All four case studies were state-transition models simulating costs and quality-adjusted life-years (QALYs). Comparing original versus reproduction outcomes, the following deviation ranges were observed: costs − 3.9 to 16.1% (mean over all model simulations 3.78%), QALYs − 3.7 to 2.1% (mean − 0.11%), and average cost-utility ratios − 3.0 to 17.9% (mean 4.28%). Applying different published criteria, an overall reproduction success was observed for three of four models. Key replication facilitators were input data tables and model diagrams, while missing standard deviations and missing formulas for equations were considered as key hurdles.

Conclusions This study confirms the feasibility of rebuilding health economic obesity models, but minor to major assumptions were needed to fill reporting gaps. Model replications can help to assess the quality of health economic model documentation and can be used to validate current model reporting practices. Simple changes to actual CHEERS reporting criteria may solve identified replication hurdles.

✉ Björn Schwander
  bjoern.schwander@ahead-net.de

  Mark Nuijten
  mark@a2m.nl

  Silvia Evers
  s.evers@maastrichtuniversity.nl

  Mickaël Hiligsmann
  m.hiligsmann@maastrichtuniversity.nl

[1] Department of Health Services Research, CAPHRI-Care and Public Health Research Institute, Maastricht University, Maastricht, The Netherlands

[2] AHEAD GmbH-Agency for Health Economic Assessment and Dissemination, Waschhausgasse 17, 79540 Lörrach, Germany

[3] a2m-Ars Accessus Medica, Amsterdam, The Netherlands

[4] Trimbos Institute-Netherlands Institute of Mental Health and Addiction, Utrecht, The Netherlands

Key Points

Method replicability and result reproduction are common criteria of high-quality research to assure scientific rigor, but have received little attention in health economic evaluation so far.

Model replication and reproduction of the results of published health economic obesity decision models were conducted for the first time ever. Our study confirms the feasibility of replicating complex obesity models, although some challenges were identified.

Small changes to existing reporting criteria have the potential to increase the transparency of model reporting and may increase the reproduction success of health economic modeling results, which may subsequently increase the transparency and acceptance of health economic modeling studies.

1 Introduction

Method replicability and reproduction of results, which in other disciplines are common criteria of adequate research reporting to assure scientific rigor, are gaining importance in the field of health economic modeling, and have been the subject of recent studies [1, 2]. In the field of health economic modeling, the topics of research reporting, model transparency, and model quality have been commonly discussed and investigated in great detail; this is reflected in the availability and application of multiple quality and reporting standards for health economic assessments [3–5]. A recently published review investigated the definitions of replicability in other disciplines, and produced a set of definitions for the success of result reproduction in health economic modeling [6]. This approach goes beyond the usual topics of reporting standards, transparency, and quality. The issues of model replication and the reproduction of results have not yet been explored within health economic obesity decision models, which is especially relevant because obesity is a complex disease with several comorbidities. Consequently, complex modeling frameworks simulated over long-term horizons are required, and these carry the potential risk of errors by the modeler and/or misinterpretations by the reader. In order to investigate the reproducibility of results in this context, we have selected health economic obesity models for replication, on the basis of a previously published systematic review [7, 8] and on the basis of previously published structural quality criteria for health economic obesity models [9]. The field of obesity modeling is in general very diverse; this is driven by multiple preventive and therapeutic approaches and multiple complications and comorbidities, which have triggered the development of various unique modeling approaches. Of 87 systematically identified obesity model publications, 69 (79% of the total) were based on unique modeling approaches [7], whereas in type 2 diabetes, of 78 systematically identified published models, only 20 (26% of the total) were based on unique modeling approaches [10]. This observed difference might also be based on the fact that there are currently no attempts to align, compare, and validate obesity modeling approaches, such as the still ongoing Mount Hood challenge for type 2 diabetes [11–13]. Furthermore, it was found that most of these unique obesity models lack an external event validation [8], making the replication of obesity models specifically an interesting research exercise.

According to previously published research in the field of model replication, comprehensive replicability is generally perceived to be desirable in health economic models [1], but additional work is needed to understand how to improve model transparency and in turn increase the chances of successful result reproduction [2]. These existing publications state that further work is needed to better understand facilitators and hurdles, and to define standards that could ultimately increase the chances of replication. Accordingly, our research goes beyond currently published approaches and investigates model replication and result reproduction in complex obesity models, with a special focus on a systematic assessment of results reproduction success and on identifying solutions for improving current reporting standards to enhance model replicability.

Therefore the objectives of our research were (1) to replicate published health economic models in obesity, (2) to compare the reproduction results to the original results, (3) to determine facilitators, hurdles, and challenges of the replication process and to assess the reproduction success measured by different definitions suggested by McManus et al. 2019 [6], and finally (4) to suggest model replication reporting standards to enhance model reproducibility.

2 Data and Methods

2.1 Model Selection and Model Overview

Based on a previous systematic review identifying 87 health economic obesity models [7, 8], the models for replication were selected using an expert panel consensus [9]. The panel assessed the key structural modeling approaches applied in published obesity models, and provided an expert consensus to improve the methodology and consistency of the application of decision-analytic modeling in obesity research. In order to select high-quality obesity models, the related minimal structural requirements for health economic obesity models were applied, consisting of the following criteria: (1) simulation time horizon: long-term (lifetime or comparable) [$n = 55$ of 87]; (2) model type: state transition model (STM) or discrete event simulation (DES) [STM $n = 74$ or DES $n = 2$ of 87]; and (3) events simulated: at least coronary heart disease, type 2 diabetes and stroke ($n = 39$ of 87). To assure that the models were simulating a comparable setting and patient population, the United Kingdom (UK) country setting ($n = 15$ of 87) and the adult population were used ($n = 70$ of 87) as final model selection criteria, which resulted in four health economic obesity models [14–17]. Additional details of this step-wise model selection process are presented in the appendix (Online Resource 1, Table 6; see the electronic supplementary material).

The details of these health economic models are presented in Table 1.

The first obesity model (case study 1) is based on extensive research, informed by a systematic review, a mixed treatment comparison, and a lifetime health economic Markov modeling approach, consisting of 13 health states. This research was funded by the UK National Institute for

Health Research Health Technology Assessment (HTA) program and is presented in a full-length HTA report, including an appendix with the health economic model, published in *Health Technology Assessment* [14]. The second obesity model (case study 2) is based on a systematic review focusing on interventions based on food purchasing patterns, and used a long-term health economic Markov modeling approach consisting of nine main health states. Although the model description in the original paper, published in *Nutrition & Diabetes*, is very brief, the publication is accompanied by an extensive appendix in which all relevant information on the modeling approach and on the underlying input data can be found [15].

The third obesity model (case study 3), funded by an industry research grant, is based on intervention-related clinical trials simulated over a lifetime horizon, using a health economic Markov modeling approach consisting of five main health states. All relevant information on the modeling approach and the input values was provided in the original paper, published in the *Journal of Medical Economics* [16]. Of the four case studies, this was the only one that presented information on internal and external validation of the model.

The fourth obesity model (case study 4), funded by an industry research grant, uses an intervention-related clinical trial, a company dataset, and a lifetime health economic

Markov modeling approach based on nine health states. All relevant information on the modeling approach and the input values was provided in the original paper, published in *Clinical Obesity* [17].

2.2 Replication of Health Economic Obesity Models

To prepare the replication of a specific model, a predefined data/information availability check was performed, and the results were recorded in table format for each selected model. This initial check was supplemented by the documentation of all identified issues, hurdles, and facilitators observed during the model replication (this process is described in Online Resource 1, and the results of this two-step procedure are presented in Online Resource 1, Tables 7–10; see the electronic supplementary material). The replication was performed in TreeAge Pro Healthcare (version 2020; TreeAge Software, Inc.) by one modeler. This specialized modeling software was used in order to minimize potential programming errors, as all relevant calculations are automated once the model structure and inputs have been defined by the modeler. A summary of identified replication facilitators and hurdles is provided in the result section (Table 2); details for each model are provided in the appendix (Online Resource 1; see the electronic supplementary material).

Table 1 Health economic obesity models selected for the replication process

Coding	Reference.	Model type	Time horizon/ cycle length	Events	Intervention; comparator	Country/per-spective	Health outcomes(s)	Software
Case study 1	Ara et al. (2012) [14]	Cohort STM	Lifetime/1 year	MI, T2D, stroke, death	Diet and exercise advice plus orlistat, sibutramine, rimonabant; placebo	UK/NHS	QALYs	Simul8
Case study 2	Au et al. (2013) [15]	Cohort STM	40 years*/6 months	MI, T2D, stroke, CC, OA death	SBT; SBT combined with meal plans and shopping lists (SBT+list); do nothing	UK/NHS	QALYs	Excel
Case study 3	Caro et al. (2007) [16]	Cohort STM	Lifetime/1 month	MI, stroke, AP, TIA, T2D, death	Diet and exercise advice plus rimonabant; placebo	UK/NHS	QALYs	Excel
Case study 4	Meads et al. (2014) [17]	Cohort STM	Lifetime/1 year	MI, T2D, stroke, death	Commercial weight loss program; usual care	UK/NHS	QALYs	Excel

AP angina pectoris, *CC* colorectal cancer, *MI* myocardial infarction, *NHS* National Health Service, *OA* osteoarthritis, *QALY* quality-adjusted life-year, *SBT* standard behavioral therapy, *STM* state transition model, *T2D* type 2 diabetes, *TIA* transient ischemic attack, *UK* United Kingdom

* Regarded as comparable to lifetime

2.3 Comparison of Reproduction Results to the Original Results

For each replicated model, model simulations were performed according to those presented in the original paper. The results were then compared to the original results, focusing on the health economic model outcomes, namely costs, clinical effects (especially quality-adjusted life-years [QALYs]), and cost-utility (as all models used QALYs as the effectiveness parameter). For each case study all published long-term comparisons were analyzed and the related costs, QALY, and cost-effectiveness (CE) results were presented as average CE ratios (for each alternative) and as incremental CE ratios (ICER). These health economic outcomes are presented together with the deviation of results between the replication and the original (absolute and as a percentage) in table format. In order to achieve a better rating of the deviation between original and reproduction results, the incremental costs and the incremental QALY results are visualized for all comparisons of the underlying case studies in the incremental CE coordinate plane.

2.4 Assessment of the Reproduction Success

A recently published review, presented in 2019 by McManus et al., investigated published definitions for replicability in health economics and other disciplines and produced a set of potential definitions for result reproduction success in health economic models, based on definitions from other scientific disciplines [6]. These definitions are (1) the same conclusions for the intervention's CE were reached; (2) costs and outcomes replicated for some treatment pathways/model scenarios and not others; (3) results for the costs and outcomes vary by only a specific percentage and are consistent with the original conclusions in comparison with the original; (4) the calculated ICER varies by only a specific percentage in comparison with the original; (5) CE figures could be reproduced to a reasonable degree of success (for example, the ICER plane or the CE acceptability curve); and (6) identical results are produced. The findings according to these success criteria are presented in table format for each case study. On the basis of these findings, the different replication success criteria are interpreted and combined in order to allow a final overall assessment of the success of the model reproduction of results. For each case study all published long-term comparisons were analyzed and the related results of the reproduction success assessment are indicated by "yes" (assessment criteria are fulfilled) and by "no" (assessment criteria are not fulfilled). For all criteria that are investigating a relative variation, expressed as a percentage, we investigated thresholds of 5%, 10%, and 20%, for the intervention, the comparator, and the incremental results, in order to see how this might influence the rating of the reproduction success.

2.5 Assessment of Model Replication Reporting Standards

The selected case studies were appraised for quality of reporting using the Consolidated Health Economic Evaluation Reporting Standards (CHEERS) checklist [18]. One reviewer assessed the reporting quality of the included studies. The twenty-four items of the CHEERS checklist were scored using "yes" (reported in full), "part" (partially reported), "no" (not reported), and "not applicable." According to a previously published approach [19], a score of 1 was assigned if the requirement of reporting for a specific item was fulfilled completely, 0.5 for partial fulfillment, and 0 otherwise, resulting in a maximum score of 24 for an article that reported all information completely.

Table 2 Summary of key facilitators and key hurdles for model replication

Category	Detailed description of key replication facilitator/hurdle	Applies to specific case study (yes/no)			
		#1	#2	#3	#4
Key facilitators	Model structure and possible state transitions were presented in a state transition diagram	Yes	Yes	Yes	Yes
	Overview of input parameters was provided in table format	Yes	Yes	Yes	Yes
Key hurdles	PSAs were performed	Yes	No	Yes	Yes
	Relevant PSA values for PSA result reproduction were provided (type of distribution and either mean and standard deviation or distribution parameters were provided)	No	No	No	No
	Clinical event simulation results were provided (which are very helpful to guide potential assumptions to be made for rebuilding the model and which provide an additional means of testing the fit of the replication)	No	No	Yes	Yes
	Relevant details on the underlying life tables were provided (including year of data)	Yes	Yes	No	No
	Several self-created regression equations were introduced but without details on how to apply/solve the provided regressions correctly	Yes	No	No	No

PSA probabilistic sensitivity analysis

On the basis of the assessed quality of reporting, how successful the reproduction of results is, and the identified facilitators and hurdles, specific model replication reporting standards are suggested. The detailed health economic model reporting recommendations, provided in the CHEERS statement, are then used as the basis for evaluating whether and which changes of these existing reporting criteria would enhance the reproducibility of model results.

3 Results

3.1 Replication Process: Facilitators and Hurdles

It was possible to replicate all selected models, but in all cases there were hurdles, which needed to be overcome by specific assumptions, which potentially influenced the reproduction of results. A summary of the key facilitators and the key hurdles, identified during the publication review and during the model replication process, is presented in Table 2.

3.2 Comparison of Reproduction Results to the Original Results

The reproduced results, the original results, and the comparison of both results as absolute and as relative (presented as percentage) variation are presented in Table 3 for all four obesity case studies. In addition, the incremental cost-utility analysis (CUA) results are visualized as a CE coordinate plane (Fig. 1), presenting the ICER as cost per QALY gained, for both the original model and the replication.

In summary, the intervention and comparator cost of the replication showed quite good results when compared to the original values; in case studies 2, 3, and 4, the variation in costs between the reproduction costs and the original costs was always < 5%. This was also observed for the comparator in case study 1, but here the various intervention costs showed higher deviations (between 5.2% and 16.1%). Looking at the QALY result reproduction of the intervention and comparators, the variation observed was always < 5%. However, when looking at the incremental cost and QALY results, the relative deviation (in percent) increases substantially in all case studies. This comes about because the absolute incremental numbers are quite low, and hence only a small deviation in absolute numbers translates into a much higher relative deviation. The same issue is observed when looking at the key outcome of the case studies, namely the ICER.

In Fig. 1, it could be seen that the incremental costs were fairly comparable between the replication and the original (presented by the very similar height of the ICER point estimates for the replication and the original shown in the coordinate plane). This picture changes if looking at the incremental QALYs, where, especially in case study 1, a strong deviation is observed (presented by the horizontal distance between the ICER point estimates for the reproduction results and the original results). This distance is considerably smaller for case studies 2, 3, and 4, showing the best fit of reproduction results for case studies 3 and 4, in which the ICER point estimates almost overlap.

3.3 Assessment of the Success of Result Reproduction

The success ratings of reproduced results, according to the different criteria proposed in a recently published literature review [6], are presented in Table 4.

In summary, the same conclusion for CE (in all studies defined as an ICER per QALY < 20,000 GBP) was reached in each investigated case study comparison; this reflects the broadest definition of reproduction success (success criterion #1).

With regard to assessing a different degree of success in reproducing results, considering the different scenarios analyzed within one case study, for case study 1, the best reproduced results are observed for the comparison of orlistat versus placebo; a worse result fit was observed for all the other comparative scenarios (10/15 mg rimonabant and sibutramine vs. placebo), whereas no such issues were identified for case studies 2, 3, and 4 (success criterion #2).

A smaller variation of 5%, 10%, or 20% in intervention and comparator costs, utilities, and (intervention-specific) average CE ratios was observed in many cases. However, looking at incremental costs, utilities, and the ICERs as well, this situation was rarely observed. This is due to the smaller absolute numbers when looking at incremental results; even small absolute variations might lead to a strong relative variation. A good example of this issue is observed in case study 2 for the comparison of standard behavioral therapy combined with provision of detailed meal plans and corresponding shopping lists (SBT+list) versus standard behavioral therapy (SBT). Here the original incremental costs are GBP-10, and in the reproduction, the incremental costs are GBP-21, a result that is to be rated as quite comparable considering the 40-year simulation time horizon. However, when expressed as a percentage, the relative variation comparing the original to the replication for this example is 110% (success criteria #3 and #4).

Therefore, for the assessment of incremental costs, QALYs, and ICERs, the calculation of relative variations may be misleading. This issue could be overcome by another success criterion, such as visualizing the original and reproduction of the incremental costs and QALYs in the CE coordinate plane. Here, the distance between the mean ICER estimates can be used to rate whether the result could be reproduced within a reasonable degree. On the basis of

Table 3 Cost, utility, and CU results: original versus reproduced results by case study

Model	Scenario	Definitions	Costs			QALYs			CU ratio		
			INT	CP	INC	INT	CP	INC	INT	CP	INC
Case study 1	LT (per patient); orlistat vs. placebo	LT (original)	3097	2806	291	15.303	15.128	0.175	202	185	1663
		LT (reproduction)	3258	2937	321	15.133	15.107	0.026	215	194	12,346
		Difference	161	131	30	− 0.170	− 0.021	− 0.149	13	9	10,683
		Difference in %	5.2%	4.7%	10.3%	− 1.1%	− 0.1%	− 85.1%	6.4%	4.9%	642.4%
	LT (per patient); rimonabant vs. placebo	LT (original)	3478	2806	672	15.317	15.128	0.189	227	185	3556
		LT (reproduction)	3689	2937	752	15.157	15.107	0.050	243	194	15,040
		Difference	211	131	80	− 0.160	− 0.021	− 0.139	16	9	11,484
		Difference in %	6.1%	4.7%	11.9%	− 1.0%	− 0.1%	− 73.5%	7.0%	4.9%	322.9%
	LT (per patient); sibutramine 10 mg vs. placebo	LT (original)	3011	2806	205	15.376	15.128	0.248	196	185	827
		LT (reproduction)	3448	2937	511	15.161	15.107	0.054	227	194	9463
		Difference	437	131	306	− 0.215	− 0.021	− 0.194	31	9	8636
		Difference in %	14.5%	4.7%	149.3%	− 1.4%	− 0.1%	− 78.2%	15.8%	4.9%	1044.3%
	LT (per patient); sibutramine 15 mg vs. placebo	LT (original)	2967	2806	161	15.418	15.128	0.290	192	185	555
		LT (reproduction)	3,445	2,937	508	15.190	15.107	0.083	227	194	6120
		Difference	478	131	347	− 0.228	− 0.021	− 0.207	35	9	5565
		Difference in %	16.1%	4.7%	215.5%	− 1.5%	− 0.1%	− 71.4%	18.2%	4.9%	1002.7%
Case study 2	40Y (per patient); SBT+list vs. SBT	40Y (original)	9359	9369	− 10	27.180	27.070	0.110	344	346	− 91
		40Y (reproduction)	9255	9276	− 21	27.700	27.610	0.090	334	336	− 233
		Difference	− 104	− 93	− 11	0.540	0.540	− 0.020	− 10	− 10	− 142
		Difference in %	− 1.1%	− 1.0%	110.0%	1.9%	2.0%	− 18.2%	− 2.9%	− 2.9%	156.0%
	40Y (per patient); SBT+list vs. do nothing	40Y (original)	9,359	9,302	57	27.180	26.840	0.340	344	347	168
		40Y (reproduction)	9,255	9,261	− 6	27.700	27.390	0.310	334	338	− 19
		Difference	− 104	− 41	− 63	0.520	0.550	− 0.030	− 10	− 9	− 187
		Difference in %	− 1.1%	− 0.4%	− 110.5%	1.9%	2.0%	− 8.8%	− 2.9%	− 2.6%	− 111.3%
Case study 3	LT (per patient); rimonabant vs. diet and exercise	LT (original)	3054	2496	558	14.173	14.108	0.065	215	177	8581
		LT (reproduction)	2933	2464	470	13.658	13.583	0.075	215	181	6263
		Difference	− 120	− 32	− 88	− 0.515	− 0.525	0.010	0	4	− 2,318
		Difference in %	− 3.9%	− 1.3%	− 15.8%	− 3.6%	− 3.7%	15.4%	0.0%	2.3%	− 27.0%
Case study 4	LT (per patient); Slimming World vs. usual care	LT (original)	9065	9988	− 923	12.580	12.360	0.220	721	808	− 4195
		LT (reproduction)	9481	10,392	− 911	12.779	12.550	0.229	742	828	− 3978
		Difference	416	404	12	0.199	0.190	0.009	21	20	217
		Difference in %	4.6%	4.0%	− 1.3%	1.6%	1.5%	4.1%	2.9%	2.5%	− 5.2%

40Y 40-year time horizon (simulated by the original model), *CP* comparator, *CU* cost-utility, *INC* incremental, *INT* intervention, *LT* lifetime time horizon, *QALY* quality-adjusted life-year, *SBT* standard behavioral therapy, *SBT+list* standard behavioral therapy combined with provision of detailed meal plans and corresponding shopping lists
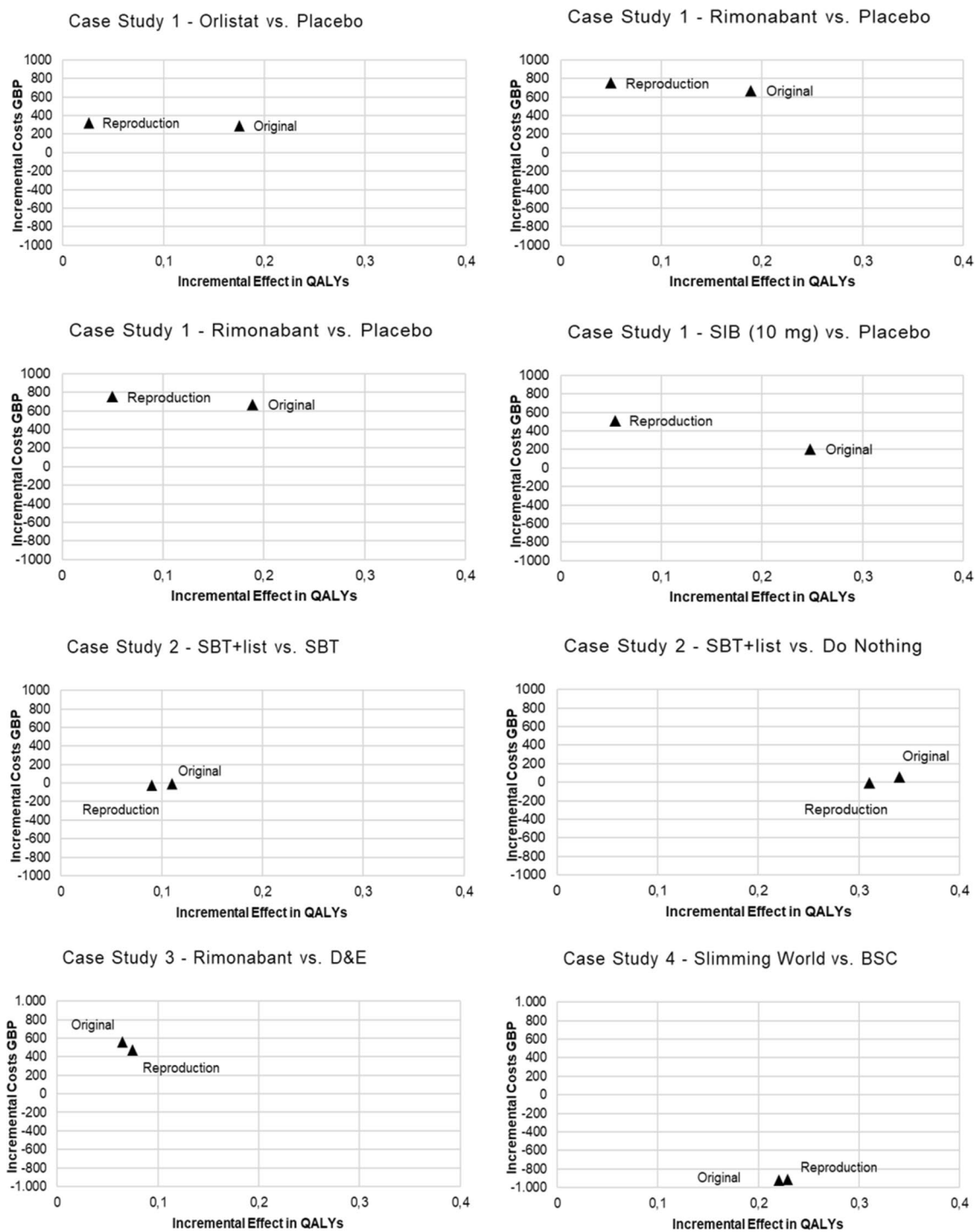
Fig. 1 Incremental cost-effectiveness results—original vs. reproduction by case study and comparison. *BSC* best supportive care/usual care, *D&E* diet and exercise, *QALY* quality-adjusted life-year, *SBT* standard behavioral therapy, *SBT+list* standard behavioral therapy combined with provision of detailed meal plans and corresponding shopping lists, *SIB* sibutramine

this approach, the rating of a successfully reproduced result was finally made for case studies 2, 3, and 4, but not for case study 1, where the variation of incremental QALYs was regarded as too strong. However, a partially successful reproduction could be seen in quite comparable incremental costs (success criterion #5).

Table 4 Assessment of the success of reproduced results according to the criteria proposed by McManus et al. 2019 [6]

Reproduction success definitions	#	Case study 1				Case study 2		Case study 3	Case study 4
		Orlistat vs. placebo	Rimonabant vs. placebo	SIB 10 mg vs. placebo	SIB 15 mg vs. placebo	SBT+list vs. SBT	SBT+list vs. do nothing	Rimonabant vs. D&E	Slimming World vs. BSC
Same conclusions for intervention CE were reached	(1)	Yes**	Yes**	Yes**	Yes**	Yes**	Yes**	Yes**	Yes**
Costs and outcomes reproduced for some treatment pathways/model scenarios and not others	(2)	Best fit for simulating "orlistat vs. placebo"; worse result fit for the other intervention (rimonabant and sibutramine 10/15 mg)				Same direction of results for all pathways/model scenarios		Only one scenario	Only one scenario
Results for costs and utility vary by only 5%/10%/20% compared to the original and are consistent with the original conclusions	(3)	INT/CP/INC	INT/CP/INC	INT/CP/INC	INT/CP/INC	INT/CP/INC	INT/CP/INC	INT/CP/INC	INT/CP/INC
< 5% variation (costs)		No/yes/no	No/yes/no	No/yes/no	No/yes/no	Yes/yes/no	Yes/yes/no	Yes/yes/no	Yes/yes/yes
< 10% variation (costs)		Yes/yes/no	Yes/yes/no	No/yes/no	No/yes/no	Yes/yes/no	Yes/yes/no	Yes/yes/no	Yes/yes/yes
< 20% variation (costs)		Yes/yes/yes	Yes/yes/yes	Yes/yes/no	Yes/yes/no	Yes/yes/no	Yes/yes/no	Yes/yes/yes	Yes/yes/yes
< 5% variation (utility)		Yes/yes/no	Yes/yes/no	Yes/yes/no	Yes/yes/no	No/no/no	Yes/yes/no	Yes/yes/no	Yes/yes/yes
< 10% variation (utility)		Yes/yes/no	Yes/yes/no	Yes/yes/no	Yes/yes/no	Yes/yes/no	Yes/yes/yes	Yes/yes/no	Yes/yes/yes
< 20% variation (utility)		Yes/yes/no	Yes/yes/no	Yes/yes/no	Yes/yes/no	Yes/yes/yes	Yes/yes/yes	Yes/yes/no	Yes/yes/yes
Calculated CE-Ratio varies by only 5%/10%/20% vs. the original	(4)	INT/CP/INC	INT/CP/INC	INT/CP/INC	INT/CP/INC	INT/CP/INC	INT/CP/INC	INT/CP/INC	INT/CP/INC
< 5% variation (CE ratio)		No/yes/no	No/yes/no	No/yes/no	No/yes/no	Yes/yes/no	Yes/yes/no	Yes/yes/no	Yes/yes/no
< 10% variation (CE ratio)		Yes/yes/no	Yes/yes/no	No/yes/no	No/yes/no	Yes/yes/no	Yes/yes/no	Yes/yes/no	Yes/yes/yes

Table 4 (continued)

Reproduction success definitions	#	Case study 1				Case study 2		Case study 3	Case study 4
		Orlistat vs. placebo	Rimonabant vs. placebo	SIB 10 mg vs. placebo	SIB 15 mg vs. placebo	SBT+list vs. SBT	SBT+list vs. do nothing	Rimonabant vs. D&E	Slimming World vs. BSC
< 20% variation (CE ratio)		Yes/yes/no	Yes/yes/no	Yes/yes/no	Yes/yes/no	Yes/yes/no	Yes/yes/no	Yes/yes/no	Yes/yes/yes
CE figures could be reproduced to a reasonable degree of success*	(5)	No	No	No	No	Yes*	Yes*	Yes*	Yes*
Identical results are produced	(6)	No	No	No	No	No	No	No	No

BSC best supportive care/usual care, *CE* cost-effectiveness, *CP* comparator, *D&E* diet and exercise, *ICER* incremental cost-effectiveness ratio, *INC* incremental, *INT* intervention, *SBT* standard behavioral therapy, *SBT+list* standard behavioral therapy combined with provision of detailed meal plans and corresponding shopping lists, *SIB* sibutramine

*CE coordinate plane was reproduced

**ICER < GBP20,000

The strictest criterion, namely the production of identical reproduction results, was observed in no case (success criterion #6).

In order to rate the success of the final results of the reproduction, a combination of different broader and more specific criteria seems to be the most adequate approach. As a successful replication of a health economic model needs to result in the same CE conclusions, success criterion #1 needs to be considered. Furthermore, the assessment of the relative deviation of costs and utilities as well as of the average CE ratios (success criteria #3 and #4) should be considered (here, the acceptable deviation could be set to 5%); whereas incremental results should not be assessed in a relative manner, due to the issue of small numbers described above. The application of this success factor assures that the reproduced results for the single interventions are within an acceptable error range. Finally, the ICER results should be visualized in the CE plane in order to determine if the deviation presented is to be regarded as acceptable or not (success criterion #5), assuring that the ICER results are fairly comparable.

The proposed combination of success criteria were all clearly fulfilled for case studies 2, 3, and 4. In contrast, case study 1 shows strong variations (< 10%) in relative cost, utility, and CE ratios, and also fails to present fairly acceptable ICER results (as visualized in Fig. 1); accordingly, case study 1 needs to be rated as a failure in reproducing results.

3.4 Assessment of Model Replication Reporting Standards

The results of assessing the reporting quality according to the CHEERS checklist are presented in Table 5 for each case study of obesity model replication. With regard to the CHEERS total scoring outcomes, there was no relevant difference in reporting quality observed between the case studies (the CHEERS score ranges between 18.0 and 20.0; the maximum possible CHEERS score is 24).

The description of study (input) parameters (CHEERS item #18) is one of the most sensitive topics for a model replication; here, case studies 1, 3, and 4 were rated as reporting the relevant data in part, whereas case study 2 was rated as reporting the relevant data in full.

Very specific information is required in order to enhance a successful model replication. Considering the identified key hurdles and applying the CHEERS guidance [18] on the quality of reporting related to these issues, it is determined whether the current consensus on reporting is adequate for successful model replications.

With regard to the identified lack of reporting of standard deviations or distribution parameters, in order to enable the reproduction of probabilistic sensitivity analysis (PSA) results (in three case studies), the CHEERS statement asks

Table 5 CHEERS checklist results for all included obesity models/case studies

	Item #	Case study 1 [14]		Case study 2 [15]		Case study 3 [16]		Case study 4 [17]	
		Rating	Score	Rating	Score	Rating	Score	Rating	Score
Title and abstract									
Title	1	Yes	1	Yes	1	Yes	1	Yes	1
Abstract	2	Yes	1	Yes	1	Yes	1	Yes	1
Introduction									
Background and objectives	3	Yes	1	Yes	1	Yes	1	Yes	1
Methods									
Target population and subgroups	4	Yes	1	Yes	1	Yes	1	Yes	1
Setting and location	5	Yes	1	Yes	1	Yes	1	Yes	1
Study perspective	6	Yes	1	Yes	1	Yes	1	Yes	1
Comparators	7	Part	0.5	Yes	1	Part	0.5	Part	0.5
Time horizon	8	Part	0.5	Part	0.5	Part	0.5	Part	0.5
Discount rate	9	Part	0.5	Yes	1	Part	0.5	Part	0.5
Choice of health outcomes	10	Part	0.5	Part	0.5	Part	0.5	Part	0.5
Measurement of effectiveness	11a	NA		NA		NA		No	0
	11b	Yes	1	Yes	1	Part	0.5	NA	
Measurement and valuation of preference-based outcomes	12	Yes	1	Yes	1	Yes	1	Yes	1
Estimating resources and costs	13a	NA		NA		NA			
	13b	Part	0.5	Yes	1	Part	0.5	Part	0.5
Currency, price date and conversion	14	Yes	1	Yes	1	Yes	1	No	0
Choice of model	15	Part	0.5	Part	0.5	Part	0.5	Part	0.5
Assumptions	16	Yes	1	Yes	1	Yes	1	Yes	1
Analytical methods	17	Yes	1	Yes	1	Yes	1	Yes	1
Results									
Study parameters	18	Part	0.5	Yes	1	Part	0.5	Part	0.5
Incremental costs and outcomes	19	Yes	1	Yes	1	Yes	1	Yes	1
Characterizing uncertainty	20a	NA		NA		NA		NA	
	20b	Part	0.5	Part	0.5	Yes	1	Yes	1
Characterizing heterogeneity	21	NA	0	NA	0	NA	0	Yes	1
Discussion									
Study findings, limitations, generalizability and current knowledge	22	Yes	1	Yes	1	Yes	1	Yes	1
Other									
Source of funding	23	Yes	1	No	0	Yes	1	Yes	1
Conflicts of interest	24	Yes	1	Yes	1	No	0	Yes	1
Total CHEERS score			19.0		20.0		18.0		18.5

CHEERS Consolidated Health Economic Evaluation Reporting Standards, *NA not applicable*

modelers to "report the values, ranges, references, and if used, probability distributions for all parameters" [18]. However, it is not made clear in the related CHEERS example table that, in addition to the distribution type, the standard deviation is required to inform the PSA. This lack of clarity in the related CHEERS example table might have led to the observed situation, namely that all case studies that have applied a PSA (case studies 1, 3, and 4) have not provided all the required information. This resulted in the rating of "partial" compliance with the CHEERS criteria with regard to

the quality of reporting study (input) parameters (CHEERS item #18).

Two further identified key hurdles for model replication are also related to the reporting of input parameters, namely the lack of reporting of details on life tables (case studies 3 and 4) as well as the introduction of several self-created regression analyses without providing details on how to apply/solve the provided regressions correctly (case study 1). All those aspects are also related to the CHEERS criteria related to the quality of reporting study (input) parameters

(CHEERS item #18). These were already rated as being in "partial" compliance due to the PSA issue stated above.

With regard to the identified lack of reporting of clinical event results (in two case studies), the CHEERS statement offers no guidance or related requirements. Hence, the missing information on clinical event results (case studies 1 and 2) had no impact on the CHEERS rating on the quality of reporting results (CHEERS item #19). As generally health economic models are driven by clinical events and related mortality, we believe that this issue should be addressed in future adaptations of the reporting standard. In this context, it would be most helpful to present event and mortality results (for all simulated alternatives) over the whole time horizon of the model, for each model cycle; this would be most helpful for checking how adequately a model adaptation predicts the underlying clinical events. This information helps to identify whether a potential result deviation (between the original and reproduction) is driven by clinical events or by the related cost and utility valuation approach of these health states.

4 Discussion

This study confirms the feasibility of rebuilding four identified health economic obesity models. However, success in reproducing results was observed in only three out of the four studies, and some challenges were observed. The replication of health economic models is an important topic, especially as there is no broad application of open-source models, although these were proposed by several authors in order to enhance model transparency and result credibility [20–22]. Such open-source models would have the advantage of joint development, joint validation, and ongoing improvement by the scientific community, but to date only a view open-source models are available, mainly due to lack of funding and other challenges (e.g., organization, software, and intellectual property restrictions) of such initiatives [22, 23]. The replication of a health economic decision analytic model is a complex exercise, and one should keep in mind that the more information and results of a model that are provided, the more information is available to investigate whether a result reproduction was successful or not. From the perspective of a modeler performing a replication of quite complex long-term obesity models, it is extremely helpful if the authors publish the simulated clinical event output frequencies, as these make it possible to check whether the event simulation and hence the clinical heart of the replicated model is working correctly (or not). If the clinical event frequencies are comparable, the replication of the structure of the model and transition probabilities can be considered correct. If the ICER is then different, the reason can be due to the inappropriate replication of the costs

or utilities, or inappropriate reporting of costs or utilities by authors. This knowledge helped to determine the source of potentially observed mismatches between original and reproduction results, as it enabled the researchers to better locate the potential issue. It is no coincidence that this information was not provided for case study 1, for which we failed to perform a successful result reproduction.

However, it needs to be taken into account that the replication of a model itself is an error-prone exercise. Hence, a failed reproduction could be based on errors made during programming, and might not necessarily come from a lack of documentation or inadequate reporting by the original authors. In order to minimize this potential source of errors, we have used specialized modeling software (TreeAge Pro Healthcare) to rebuild the selected health economic obesity models. Consequently, potential errors might be due mainly to input data typos, as building the model structure (and related calculations) is widely automated. However, using TreeAge instead of the software used in the original study could also be an issue preventing a 1:1 reproduction of modeling results, due to the automatic application of some TreeAge features (e.g., automatic half cycle correction) as stated in detail in the appendix tables (Online Resource 1; see the electronic supplementary material). Furthermore, it needs to be considered that the success of a model replication is also influenced by the skills of the programmer; hence one limitation is that the replication was performed by only one modeler. However, this modeler has over 20 years of experience as a professional health economist and all critical issues were reviewed and discussed within the team, which included experienced health economic modelers. On the other hand, programming errors in the original publication could not be ruled out completely, as especially complex Excel models require complex testing and validation to assure the correctness of all calculations, and this might also impact the presented reproduction results.

For assessing the success of the reproduction results, we have applied different criteria as defined and proposed in a recently published review on this topic [6]; to our knowledge these criteria were applied for the first time in this study to systematically assess the reproduction success. The six criteria applied range from very broad to very specific; accordingly it is easier or harder to fulfill them. The strictest criterion, namely that identical results be produced, was not achieved by any of the case studies. This is not unexpected considering that all obesity models were simulating a long-term time horizon, and hence a small deviation (even a rounding issue) will get more and more pronounced over time. Another reason may be the high complexity of obesity models, triggered by including all the relevant complications of obesity. The greater the complexity, the greater the chance of misinterpreting the data, assumptions, and model structure description in the original paper, combined

with a higher probability of errors by either the replicator or the original programmer. This strictest definition does not seem to be very helpful as identical results have not yet been achieved with regard to the publications of other model replications [1, 2]. Moreover, the other proposed criteria were not rated as sufficient to adequately define reproduction success, as all were rated as too soft to act as stand-alone reproduction success criteria. Therefore we have used a combination of various criteria in order to investigate and to determine the success of reproduction. Although this proposed combination does not assure identical results, it assures that the CE conclusion is identical, that the deviation in single components is acceptable (< 5%), and that the incremental CE results are fairly comparable. As this study was, to our knowledge, the first application of these replication success criteria, and hence also of this criteria combination, further research and scientific dialog is required to investigate and define how to best rate the success of a health economic model replication; we believe that the applied criteria developed by McManus et al. [6] and our research will help to inform this scientific dialog.

The identified key model replication facilitators were input data tables and model diagrams showing the model structure and possible state transitions. Key replication hurdles were missing standard deviations for performing probabilistic analysis, missing clinical event results, missing details on applied life tables, and missing formulas for equations based on own calculations. Whereas the key facilitators were quite in line with those identified by other research teams [1, 2], our identified key barriers seem to be more specific than those identified in previous research. This might be related primarily to the fact that we have selected long-term obesity models, whereas other research teams [1, 2] included a broader range of health economic models, including short- and long-term time horizons and different disease areas. This focus on only one disease area and on long-term models is also a limitation of our research; the transferability of our findings to other kinds of health economic models needs to be investigated by future research.

Looking specifically at the reproduction of the cost and utility results of single strategies, a previous study [2] found that there was a tendency for greater variation in the reproduced costs than in outcomes, which is also seen in our research; costs ranged from − 3.9 to 16.1% (mean over all model simulations 3.78%), whereas QALYs varied by − 3.7 to 2.1% (mean − 0.11%). However, looking at the comparison of reproduced results and original results in terms of incremental cost and QALYs (please refer to Fig. 1), which was done for the first time in our study, the observed variations in incremental QALYs were more pronounced than the variations in incremental costs; this highlights the importance of reporting and visualizing incremental replication results.

As one key facilitator McManus et al. [2] suggested that cost and outcome results should be presented over time in an additional table to enhance model replication. We agree that this information would be very helpful for replication, especially to see from which point in time deviations between reproduced results and original results are observed. However, on the basis of this information, it would not be clear where the replication error might be located, which is why we are suggesting that the clinical events be presented over time. If it is possible to reproduce the results of the clinical events, the structure and related transition probabilities are replicated correctly. If a deviation in costs or outcomes is then observed, this is related to costs or to the parameter values for costs and utilities, and the methodology of including these parameters. Hence, in the best case, all model outcomes, including the underlying event rates, would be presented to facilitate model replication.

We applied the CHEERS checklist [18] as it looks particularly at the quality of reporting, a core criterion for successful model replication, and as it was found to be the most commonly used checklist since 2017 in a recently published systematic review [3]. Other frequently applied checklists (such as the Phillips checklist [24] or the CHEC project [25]) assess the quality of conducting the health economic study, which was not our key focus. We investigated whether the CHEERS score might be predictive for the success of model replication, but this was not the case, with scores ranging from 18 to 20. The non-successful replication case study 1 rated a score of 19 (the maximum possible CHEERS score is 24). A comparable finding was observed by another research team that investigated the Phillips checklist [24] in the context of model replication; they found that the Phillips checklist was not reliable for ensuring that studies are replicable [2]. However, we believe that simple changes in the CHEERS reporting criteria might be adequate to solve the key hurdles for model replicability that we observed in our presented research. These proposed changes are (1) the probability distribution and all the necessary parameters to define its shape are to be presented for all input parameters, assuring the reproduction of PSA results; (2) when a model simulates (clinical) events, the event simulation results should be presented, to guide potential necessary assumptions and to better locate potential replication errors; and (3) for all included regressions/risk equations (whether published or unpublished) applied in the model, the calculation formula should be presented, preferably with an application example, to assure the correct replication of formula-based transition probabilities, costs, and outcomes. Although the current CHEERS statement covers parts of these aspects—namely it asks for "probability distributions for all parameters" to be included and for "outcomes of interest" to be reported— we believe that these aspects need to be made clearer to adequately guide reporting on the model.

To our knowledge, there are currently no other publications that suggest specific changes to CHEERS or other health economic reporting guidelines to enhance health economic model replication. However, we have identified a recently published paper that suggests a nine-item osteoporosis-specific addition to the CHEERS checklist, in order to address disease-specific issues adequately [26]. The further development of health economic reporting standards is an ongoing process, and there is a specific International Society for Pharmacoeconomics and Outcomes Research (ISPOR) task force currently working on an update of the CHEERS criteria.

5 Conclusion

Small changes to existing reporting criteria, as presented above, may increase both the transparency of health economic model reporting and the success of reproducing its consequent results. Proofing the replicability of our health economic simulation "experiments" might increase the scientific rigor and acceptance of our field. In conclusion, model replications can help to assess the quality of health economic model documentation, can be used to validate and refine current model reporting practices, and might subsequently increase the transparency and acceptance of health economic modeling studies.

Declarations

Conflict of interest The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the article.

Ethics approval Not applicable.

Consent to participate Not applicable.

Data availability statement The replicated models generated during the presented study are available from the corresponding author on reasonable request.

Code availability Not applicable.

Authors' contributions All authors contributed to the study conception and design. BS attempted to replicate all four models presented in the article and performed the CHEERS assessment. MN reviewed the programmed models. All authors were involved in writing the manuscript and have approved the final version submitted.

Consent for publication All authors have reviewed the final manuscript and gave their consent for publication.

References

1. Bermejo I, Tappenden P, Youn JH. Replicating health economic models: firm foundations or a house of cards? PharmacoEconomics. 2017;35(11):1113–21. https://doi.org/10.1007/s40273-017-0553-x.
2. McManus E, Turner D, Gray E, Khawar H, Okoli T, Sach T. Barriers and facilitators to model replication within health economics. Value Health. 2019;22(9):1018–25. https://doi.org/10.1016/j.jval.2019.04.1928.
3. Watts RD, Li IW. Use of checklists in reviews of health economic evaluations, 2010 to 2018. Value Health. 2019;22(3):377–82. https://doi.org/10.1016/j.jval.2018.10.006.
4. Walker DG, Wilson RF, Sharma R, Bridges J, Niessen L, Bass EB et al. AHRQ Methods for Effective Health Care. Best practices for conducting economic evaluations in health care: a systematic review of quality assessment tools. Rockville (MD): Agency for Healthcare Research and Quality (US); 2012.
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):843–50. https://doi.org/10.1016/j.jval.2012.04.012.
6. McManus E, Turner D, Sach T. Can you repeat that? Exploring the definition of a successful model replication in health economics. PharmacoEconomics. 2019;37(11):1371–81. https://doi.org/10.1007/s40273-019-00836-y.
7. Schwander B, Hiligsmann M, Nuijten M, Evers S. Systematic review and overview of health economic evaluation models in obesity prevention and therapy. Expert Rev Pharmacoecon Outcomes Res. 2016;16(5):561–70. https://doi.org/10.1080/14737167.2016.1230497.
8. Schwander B, Nuijten M, Hiligsmann M, Evers S. Event simulation and external validation applied in published health economic models for obesity: a systematic review. Expert Rev Pharmacoecon Outcomes Res. 2018;18(5):529–41. https://doi.org/10.1080/14737167.2018.1501680.
9. Schwander B, Nuijten M, Hiligsmann M, Queally M, Leidl R, Joore M, *et al*. Identification and expert panel rating of key

structural approaches applied in health economic obesity models. Health Policy Technol. 2020;9(3):314–22. https://doi.org/10.1016/j.hlpt.2020.03.005.

10. Yi Y, Philips Z, Bergman G, Burslem K. Economic models in type 2 diabetes. Curr Med Res Opin. 2010;26(9):2105–18. https://doi.org/10.1185/03007995.2010.494451.

11. Brown JB, Palmer AJ, Bisgaard P, Chan W, Pedula K, Russell A. The Mt. Hood challenge: cross-testing two diabetes simulation models. Diabetes Res Clin Pract. 2000;50(Suppl 3):S57-64. https://doi.org/10.1016/s0168-8227(00)00217-5.

12. Palmer AJ, Clarke P, Gray A, Leal J, Lloyd A, Grant D, et al. Computer modeling of diabetes and its complications: a report on the Fifth Mount Hood challenge meeting. Value Health. 2013;16(4):670–85. https://doi.org/10.1016/j.jval.2013.01.002.

13. Si L, Willis MS, Asseburg C, Nilsson A, Tew M, Clarke PM, et al. Evaluating the ability of economic models of diabetes to simulate new cardiovascular outcomes trials: a report on the ninth mount hood diabetes challenge. Value Health. 2020;23(9):1163–70. https://doi.org/10.1016/j.jval.2020.04.1832.

14. Ara R, Blake L, Gray L, Hernandez M, Crowther M, Dunkley A et al. What is the clinical effectiveness and cost-effectiveness of using drugs in treating obese patients in primary care? A systematic review. Health Technol Assess. 2012;16(5):iii–xiv, 1–195. https://doi.org/10.3310/hta16050.

15. Au N, Marsden G, Mortimer D, Lorgelly PK. The cost-effectiveness of shopping to a predetermined grocery list to reduce overweight and obesity. Nutr Diabetes. 2013;3:e77. https://doi.org/10.1038/nutd.2013.18.

16. Caro J, Stillman O, Danel A, Getsios D, McEwan P. Cost effectiveness of rimonabant use in patients at increased cardiometabolic risk: estimates from a Markov model. J Med Econ. 2007;10(3):239–54. https://doi.org/10.3111/13696990701438629.

17. Meads DM, Hulme CT, Hall P, Hill AJ. The cost-effectiveness of primary care referral to a UK commercial weight loss programme. Clin Obes. 2014;4(6):324–32. https://doi.org/10.1111/cob.12077.

18. Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS)-explanation and elaboration: a report of the ISPOR Health Economic Evaluation Publication Guidelines Good Reporting Practices Task Force. Value Health. 2013;16(2):231–50. https://doi.org/10.1016/j.jval.2013.02.002.

19. Hiligsmann M, Evers SM, Ben Sedrine W, Kanis JA, Ramaekers B, Reginster JY, et al. A systematic review of cost-effectiveness analyses of drugs for postmenopausal osteoporosis. PharmacoEconomics. 2015;33(3):205–24. https://doi.org/10.1007/s40273-014-0231-1.

20. Dunlop WCN, Mason N, Kenworthy J, Akehurst RL. Benefits, challenges and potential strategies of open source health economic models. PharmacoEconomics. 2017;35(1):125–8. https://doi.org/10.1007/s40273-016-0479-8.

21. Schramm W, Sailer F, Pobiruchin M, Weiss C. PROSIT open source disease models for diabetes mellitus. Stud Health Technol Inform. 2016;226:115–8.

22. Jansen JP, Incerti D, Linthicum MT. Developing open-source models for the US health system: practical experiences and challenges to date with the open-source value project. PharmacoEconomics. 2019;37(11):1313–20. https://doi.org/10.1007/s40273-019-00827-z.

23. Emerson J, Bacon R, Kent A, Neumann PJ, Cohen JT. Publication of decision model source code: attitudes of health economics authors. PharmacoEconomics. 2019;37(11):1409–10. https://doi.org/10.1007/s40273-019-00796-3.

24. Philips Z, Bojke L, Sculpher M, Claxton K, Golder S. Good practice guidelines for decision-analytic modelling in health technology assessment: a review and consolidation of quality assessment. PharmacoEconomics. 2006;24(4):355–71. https://doi.org/10.2165/00019053-200624040-00006.

25. Evers S, Goossens M, de Vet H, van Tulder M, Ament A. Criteria list for assessment of methodological quality of economic evaluations: consensus on Health Economic Criteria. Int J Technol Assess Health Care. 2005;21(2):240–5.

26. Hiligsmann M, Reginster JY, Tosteson ANA, Bukata SV, Saag KG, Gold DT, et al. Recommendations for the conduct of economic evaluations in osteoporosis: outcomes of an experts' consensus meeting organized by the European Society for Clinical and Economic Aspects of Osteoporosis, Osteoarthritis and Musculoskeletal Diseases (ESCEO) and the US branch of the International Osteoporosis Foundation. Osteoporos Int. 2019;30(1):45–57. https://doi.org/10.1007/s00198-018-4744-x.