available at www.sciencedirect.com

SCIENCE DIRECT®

journal homepage: www.elsevier.com/locate/jval

The CHD challenge: Comparing four cost-effectiveness models

David Turner, MSc[a,], James Raftery, PhD, MA[a], Keith Cooper, PhD, MSc[a], Eleanor Fairbank, MSc[a], Stephen Palmer, MSc[b], Sue Ward, BA[c], Roberta Ara, MSc[c]*

[a] *Wessex Institute University of Southampton, Southampton, UK*
[b] *Centre for Health Economics, University of York, York, UK*
[c] *School of Health and Related Research (ScHARR), University of Sheffield, Sheffield, UK*

Keywords:
Coronary disease
Cost-utility analysis
Economic modeling
Model validation

ABSTRACT

Objectives: To compare four UK models evaluating the cost-effectiveness of interventions in coronary heart disease (CHD), exploring the relative importance of structure and inputs in accounting for differences, and the scope for consensus on structure and data.

Methods: We compared published cost-effectiveness results (incremental cost, quality-adjusted life year, and cost-effectiveness ratio) of three models conforming to the National Institute for Health and Clinical Excellence guidelines dealing with three interventions (statins, percutaneous coronary intervention, and clopidogrel) with a model developed in Southampton. Comparisons were made using three separate stages: 1) comparison of published results; 2) comparison of the results using the same data inputs wherever possible; and 3) an in-depth exploration of reasons for differences and the potential for consensus.

Results: Although published results differed by up to 73% (for statins), standardization of inputs (stage 2) narrowed these gaps. Greater understanding of the reasons for differences was achieved, but a consensus on preferred values for all data inputs was not reached.

Conclusions: We found that published guidance on methods was important to reduce variation in important model inputs. Although the comparison of models did not lead to consensus for all model inputs, it provided a better understanding of the reasons for these differences, and enhanced the transparency and credibility of all models. Similar comparisons would be aided by fuller publication of models, perhaps through detailed web appendices.

Introduction

Reimbursement agencies are continuously assessing new health-care technologies and need to ensure robustness, transparency, and consistency in their decisions. Health economic models can help inform these decisions and aid the efficient use of limited National Health Service (NHS) resources. However, if these models are to be used they must also be credible to those making policy decisions. This credibility can be hindered because these models are often complex and are the result of the different expertise of statisti-

Funding: The authors have no other financial relationships to disclose.

* *Corresponding author*. David Turner, Wessex Institute University of Southampton, Alpha House, Southampton Science Park, Chilworth, Southampton SO16 7NS, UK.
E-mail address: dturner@soton.ac.uk (David Turner).

1098-3015/$36.00 – see front matter
doi:10.1016/j.jval.2010.10.009

cians, clinicians, and health economists. There may also be a range of models that evaluate the same technology but use different methods and produce different results.

One way to increase the perceived credibility of a model is to test and demonstrate its validity. Four methods of validation have been suggested and these have been summarized by Philips et al. [1]. First, internal consistency implies that the practical model should behave as the theoretical model predicts, and that it is "debugged." Second, external consistency implies that the model should demonstrate face validity; that the outputs of a model are consistent with our knowledge of a disease or intervention. Third, between-model consistency implies that different independent models addressing the same question should give similar results. Fourth, predictive validity involves testing the results of a model against observable data or a prospective study to ascertain that the results are similar.

The fourth test of validity could be thought of as a "gold standard test" and would be the most credible evidence for the veracity of a model. If the results of a model matched real-world observations then we might have greater confidence. However, this is not generally possible because these models are often used to combine evidence from multiple sources, to extrapolate the results of a short-term clinical trial to the lifespan of patients or to generalize results to "real-world" settings. In these situations, data are unlikely to be available to formally assess the predictive validity of models and checking for between-model consistency may be the most feasible way of validating a model. If structure, inputs, and results are similar between models it implies there is general agreement on how to model a particular intervention or disease area; hence, it may mean that there is a clear preference for methods and data inputs. If there are disagreements between models in data inputs or structure then checking between-model consistency can highlight the important differences in terms of altering results. Effort can then be concentrated on those differences which may lead to either consensus over the use of existing data or more effort to generate better data. It is increasingly common to identify sets of alternative models that address a similar decision problem or disease area. For example, in a recent assessment of the cost-effectiveness of antivirals for the treatment of influenza, a total of 22 separate studies were identified (including 7 from a UK perspective) [2].The existence of alternative models provides an important opportunity to explore similarities and differences between models.

Differences between models can be due to differences in parameter values, methods, and structures. These are similar to the sources of uncertainty in models [3]. Parameter differences could include state transition probabilities and the quality-adjusted life year (QALY) losses associated with those states. Data on the costs of being in particular states and the costs of transition between those states could also be included in this category as could the data and assumptions used to characterize uncertainty in a probabilistic model. Many potential sources of data exist, particularly for a disease which is both common and well researched. Different data sets may be particularly suited for answering specific questions.

Differences between models may also be attributed to methodological differences including: the methods used to derive utility values, the perspective of the analysis (either NHS or a societal perspective), and the discount rate used. Some of these will be reduced by closer adherence to guidelines [1] and the National Institute for Health and Clinical Excellence (NICE) reference case [4]. Differences may also arise due to structural issues, including the modeling approach used (e.g., Markov, decision tree, or discrete event simulation). They may also be because of differences in the questions addressed or the health states included in models. Models may cover a single technology or may model a disease or population.

Some differences between models may be legitimate because two models set up to answer different questions may have different structures. Also, some differences may occur because analysts have correctly followed different sets of guidelines applicable to separate jurisdictions, or the analyses have been conducted at different times. Other differences may arise because there is no obvious "best" approach; these may require a need for clarification and future research to obtain more reliable or appropriate sources of data. Identifying these differences would be a useful outcome of any checks of between-model consistency.

To examine the feasibility and usefulness of a check of between-model consistency, we compared the Southampton CHD treatment model with three previously published models. The research question of the Southampton treatment model was "What are the relative cost-effectiveness ratios of a wide range of commonly used treatments for coronary heart disease for a UK population?" This involved using data on the clinical effectiveness and cost of a number of coronary care interventions and meant that comparisons could be made with many other CHD models, providing they addressed any of the interventions covered by the Southampton model. The Southampton treatment model was developed as part of a study that modeled CHD [5,6].

Comparator models were selected from the literature, restricting comparisons to models specific to the UK and conforming to the NICE reference case. Furthermore, we restricted comparisons to models covering one of the interventions evaluated by the Southampton treatment model. This meant that each model focused on NHS practice and followed similar guidelines for economic modeling. This increased the comparability of the models as there were a number of characteristics that would be shared; two examples are the use of a cost per QALY approach and a health service perspective. It also meant that there were differences between models published at different times.

One of the comparator models was developed at the School of Health and Related Research at the University of Sheffield to look at the cost-effectiveness of statins (School of Health and Related Research [ScHARR]-statins model) [7]. Specifically, this model was constructed to answer the question: "at what level of CHD/cardiovascular disease (CVD) risk are statins cost effective in the United Kingdom?" The other two comparator models were developed at the University of York. The York percutaneous coronary intervention (York-PCI) model [8] was designed to "explore the cost effectiveness of thrombolysis compared to primary angioplasty in acute myocardial infarction (MI) patients." The last model (York-clopidogrel) [9] was designed to "explore the cost-effectiveness of clopidogrel plus

Table 1 – Summary descriptions of all models compared.

Model (Publication year)	Southampton treatment (2008*)	ScHARR-statin (2007)	York-PCI (2007)	York-clopidogrel (2004)
Study population	CHD population	CHD cohort	First acute MI	Unstable angina or non-ST elevated MI
Intervention	A range of interventions were modeled	Statins compared to no treatment	Thrombolysis compared to primary angioplasty	Clopidogrel compared to standard therapy
Treatment duration	Varied depending upon intervention	Lifetime	6 months	1 year
Model time horizon	Lifetime	Lifetime	Lifetime	Lifetime
Treatment effect on mortality	Relative risks. Statin = 0.72; primary PCI = 0.68 changed to 0.469, thrombolysis = 0.69, clopidogrel = 0.8	Males secondary prevention. Relative risk = 0.7192	Long-term odds ratio = 0.7	Long-term relative risk = 0.71 (0.60 to 0.84)
Discount rate	3.50%	6% costs, 1.5% benefits	3.50%	6% costs, 1.5% benefits
Model type	Markov	Markov	Decision tree plus Markov	Decision tree plus Markov
No. of states	Modeled 8 states: well, unstable angina, angina, post-MI (first year), post-MI subsequent years, heart failure, CHD death, non-CHD death. Transition probabilities reported in web appendix at: 10.1016/j.jval.2010.10.009	Modeled 24 states: event-free, MI, stable angina, unstable angina, CHD death, transient ischemic attack, stroke, CVD death, other death, (post each non-fatal, split into history of CHD or history of CVD). Transition probabilities reported in Table 52, p90 of HTA report	Modeled 4 states: dead, non-fatal MI, non-fatal stroke, alive with ischemic heart disease. Transition probabilities reported in Table 2, p1240 of article	Modeled 4 states: well, MI (first cycle), post-MI, dead. Transition probabilities reported in Table 27, p41 of HTA report
Baseline cohort	Population by age and sex	Age 55, male cohort	Age 61, male cohort	Age 64–68, mixed cohort
Base case results		ICER = £13,900/QALY	ICER = £9241/QALY	ICER = £6078/QALY
Intervention costs (Per year)	Clopidogrel = £460; aspirin = £9; statin = £148; thrombolysis = £316; primary angioplasty = £3377	Statins = £317	Thrombolysis = £600; primary angioplasty = £4097	Clopidogrel = £464; aspirin = £3.47
Price year†	2005/6	2004	2004	2001/2
Utility‡	Post-MI (1st year) = 0.68; post-MI subsequent years = 0.72; unstable angina = 0.77; angina = 0.81; heart failure = 0.66	Utility by age: stable angina = 0.808; unstable angina = 0.770; MI = 0.760; transient ischemic event = 1, stroke = 0.629	MI first year = 0.683; MI subsequent years = 0.718; non-disabling stroke = 0.740; disabling stroke = 0.380; combined stroke = 0.612. Valuations from patients	Well = 0.8, MI (first cycle) = 0.8, post-MI = 0.8, dead = 0

CHD, coronary heart disease; CVD, cardiovascular disease; ICER, incremental cost-effectiveness ratio; MI, myocardial infarction; PCI, percutaneous coronary intervention; QALY, quality-adjusted life year; ScHARR, School of Health and Related Research.

* The Southampton treatment model has not previously been published. The year here refers to when analysis was completed.

† No adjustments have been made to present the model results using a common price year.

‡ Utility values derived from health state valuations from patients valued using societal norms except for values from Main et al. which were derived from assumption.

standard care compared to standard care alone in the treatment of non-ST-segment elevated acute coronary syndrome." The structure of both York models was derived from earlier work on a glycoprotein IIb/IIIa model [10]. A summary of the core structure of each model is provided (Table 1), and each model adhered closely to published guidelines for cost-effectiveness models [1]. In addition to Table 1, detailed model information for the Southampton treatment model is available as a web appendix found at: 10.1016/j.jval.2010.10.009. This article presents the findings of an assessment of model validity using Philips' third method of between-model consistency. This exercise had three aims. First, to compare the estimated cost-effectiveness of interventions for CHD obtained from different models. Second, to identify those differences in inputs and structural components which were most important to the outputs calculated, and to explore the relative importance of structure and inputs. Finally, to identify areas where there was agreement between the teams in terms of the optimal value for model struc-

Table 2 – Overview results of comparison (mean per patient).

	Model	Incremental cost (£)	Incremental QALYs	ICER (£/QALY)
Statins comparison	Southampton-statin, stage 1 result	1900	0.505	3760
	Southampton-statin, stage 2 results	3430	0.350	9800
	ScHARR-statin	7860	0.565	13,900
PCI comparison	Southampton-PCI, stage 1 results	800	0.065	12,250
	Southampton-PCI, stage 2 results	1020	0.088	11,590
	York-PCI	2680	0.290	9240
Clopidogrel comparison	Southampton-clopidogrel, stage 1 results	130	0.013	9930
	Southampton-clopidogrel, stage 2 results	480	0.058	8210
	York-clopidogrel	470	0.077	6080

PCI, percutaneous coronary intervention; QALY, quality-adjusted life year; ScHARR, School of Health and Related Research.

tures and inputs. This process would also identify areas where there was disagreement, and in these cases we aimed to explore the reason for these differences.

Methods

Comparisons between models were made in three stages. In all cases changes were made to the Southampton treatment model only. Stage 1 involved a comparison of the main published (base-case) results. The only change applied at this stage was to alter the Southampton treatment model to take a cohort approach (a cohort of 1000 men, 55–64 years old) to enable a comparison with an equivalent cohort approach used in the three alternative models. Stage 2 involved further changes to the Southampton treatment model. These comprised structural changes, such as removing particular health states (e.g., heart failure in the statin comparison), as well as changing relevant data inputs to match the approaches used in the alternative models. The Southampton treatment model was thus standardized to reflect the other models in terms of discount rate, utilities, costs, effectiveness, mortality rates, prevalence, and assumptions. The effects of these changes on costs, life years, QALYs, and incremental cost-effectiveness ratios (ICER) were recorded. After all feasible changes were made, the final outputs were compared.

Stage 3 aimed to reach understanding of the reasons for the differences between models and for the choices made with respect to model structure and data sources using a series of one-way sensitivity analyses to evaluate the effect of each change on the model results. These sensitivity analyses involved changing the parameter values for a range of data inputs in the Southampton model to the values used in the comparator model. These changes were carried out sequentially, with each parameter returned to its original value before the value assigned to the next parameter was altered. This demonstrated the effect of each change and its relative importance in reducing the differences between the Southampton treatment model and the comparator model for estimates of cost, QALYs, and cost per QALY. Stage 3 also involved collaboration with experts from all three comparator models. After the desk-based comparisons we exchanged documents and held three teleconferences between all three centers to explore reasons for differences in results. Some additional two-way teleconferences were also held. An initial draft paper was prepared by the Southampton group, and then shared and iteratively developed with York and Sheffield colleagues.

Results

The results of the analyses are shown in Table 2. This table provides the results for each comparator model. Also provided are the stage 1 results, which give the unadjusted Southampton treatment model results for each of the three comparisons. For statins, the incremental cost-effectiveness ratio predicted by the Southampton treatment model was very different (73% lower) than the value predicted by the ScHARR-statin model. The difference was primarily in the cost, with the incremental cost in the Southampton model one-quarter of that in the ScHARR model. This was partly but not exclusively due to the lower yearly cost for statins in the Southampton model (£148 compared to £316). For the PCI comparison the Southampton-treatment model produced a higher ICER, being 33% greater than the value for the York-PCI model. However, this difference hides the extent of the difference in estimated incremental costs and QALYs, which were 70% and 78% lower, respectively. For the clopidogrel comparison the stage 1 results gave a higher ICER, being 63% higher, but the estimated costs and QALYs were much lower, being 72% and 83% lower, respectively. The absolute values of the incremental costs and QALYs were comparatively small in both clopidogrel models.

Table 2 also shows the results for the stage 2 analysis and gives the effects of including parameter values from the comparator models in the Southampton model results. For the statin comparison, after stage 2, the ICERs were much closer; the Southampton treatment model was 29% lower. This convergence was caused by an increase in the estimated incremental costs in the Southampton model. The difference in the incremental QALYs widened at this stage. For the PCI comparison, the stage 2 changes had only a minor impact on the ICER because the value changed to 25% greater than the York model. However, the values for both incremental costs and incremental QALYs were still much lower for the Southampton treatment model. For the clopidogrel comparison incremental costs, QALYs and the ICER were all closely matched after stage 2; the ICER for the Southampton treatment model was 35% higher. This difference was caused by differences in

Table 3 – One-way sensitivity analysis showing the effect of the changes made to each variable.

	Changes made to Southampton treatment model*	Cost	QALY	C/QALY
Statins comparison	Cohort of 1000 men, 55–64 years old	1900	0.505	3760
	Adjustments to effectiveness of intervention, statin cost, utility values discount rates	3130	0.711	4400
	Remove heart failure	2050	0.424	4830
	CHD mortality in the post-MI first year state	1900	0.501	3790
	CHD mortality in the MI subsequent years health state	1920	0.453	4240
	CHD mortality in unstable angina	1900	0.523	3630
	CHD mortality angina	1930	0.436	4430
	Prevalence of angina and MI†	1960	0.455	4310
	Utility rates adjusted for age‡	1900	0.407	4670
PCI comparison	Cohort of 1000 men, 55–64 years old, discount rates 3.5%	800	0.065	12,250
	All included costs§	740	0.065	11,390
	Utility values used in health states	800	0.065	12,280
	Prevalence of angina, MI, and heart failure	1010	0.082	12,360
Clopidogrel comparison	Cohort of 1000 men and women, 65–74 years old, 50% of MI = NSTEMI	130	0.013	9930
	Discount rates 6% = costs, 1.5% = benefits	110	0.016	6690
	Utility values used in health states	130	0.014	9280
	Prevalence of angina, MI, and heart failure‖	140	0.014	10,320

CHD, coronary heart disease; MI, myocardial infarction; NSTEMI, non-ST elevation myocardial infarction; PCI, percutaneous coronary intervention; QALY, quality-adjusted life year.

* Changes made in each comparison adjust the values used in the Southampton treatment model to those used in the relevant comparator model.

† We calculated percentages using the proportions given. To calculate the percentage of angina patients, we used the proportion of stable angina plus the proportion of unstable angina patients, divided by the total number of patients (not including the transient ischemic attack [TIA] or stroke patients).

‡ These age-related utilities were multiplied by the health state utilities to give a combined utility value.

§ In the base case comparison only the cost of the intervention itself was changed.

‖ Angina, long-term MI, short-term MI, and heart failure; no unstable angina.

incremental QALYs because the two estimates of incremental costs were almost identical after stage 2.

As part of stage 3, a series of one-way sensitivity analysis showed the effect on the baseline incremental costs, QALYs, and ICER of each of the changes made (Table 3). For the statin model a comparison of a number of changes was made (effectiveness of the intervention, statin cost, utility values, and discount rates). Taken together these had large effects on both the incremental cost and incremental QALYs (increases of 65% and 41%, respectively). However, the percentage effect on the ICER was lower (increase by 17%). The change that had the single largest effect on the ICER was removing the heart failure state. Age-adjusted utility values and changing the rates of CHD-related mortality in health states, particularly for stable angina, also had an important impact. For the PCI comparison, the most important changes made to the ICER were to the costs of the intervention. Changing the prevalence had large effects on both the incremental costs and incremental QALYs; however, because these were of the same sign and magnitude (increase of 26%), they had a negligible effect on the ICER. For the clopidogrel comparison, changes in the discount rate produced the largest effect on the results.

Discussion

We found that the predicted ICERs for all comparisons were comparatively close, particularly after input values had been adjusted in the Southampton treatment model. The conclusion for all models compared was that the intervention (in the evaluated patient group) would have been cost effective using NICE threshold values [4]. These results were consistent with the existing literature. Mauskopf and colleagues [11] published a review of the cost-effectiveness of clopidogrel. Part of this review covered eight studies based on the Clopidogrel in Unstable Angina to Prevent Recurrent Events trial [12], including the York-clopidogrel model. All these studies were considered to be within the cost-effectiveness thresholds for their country of analysis. Ward et al [7] carried out a literature review and identified three UK studies that included statins for secondary prevention. Reported ICERS ranged from £5291 to £42,483 per life year gained. However, these estimates were made using prices obtained while all statins were within patent and, hence, would overestimate current ICERs. Wailoo and colleagues [13] adapted the York-PCI model to use cost and treatment delay estimates derived from the National Infarct Angioplasty Project (NIAP). They found a similar estimate of the ICER for angioplasty of £4520 per QALY.

The Southampton treatment model used a disease-based approach to model a wide range of CHD interventions. Each of the three comparator models was a single technology model that evaluated a single intervention or group of closely linked interventions. A difficulty faced by the Southampton model was that the same structure had to be used for a variety of interventions. For this reason the model structures used for intervention-specific models are likely to be more appropriate

for that intervention. In addition, single intervention models were more flexible and could answer specific questions. For example, the York-PCI model presented scenario analysis modeling the effect of altering the additional time delay until angioplasty as well as differences in the initial length of stay. These types of scenario analysis would have been difficult to replicate in the Southampton treatment model. The technology models are also likely to be more parsimonious, which has been identified as a desirable attribute of models [14].

In contrast, the Southampton model could estimate the comparative cost-effectiveness of different interventions and the economic impact of policies which would cover many aspects of CHD care. This may have appeal to decision makers as it would allow the evaluation of a wider range of different technologies on a common framework. The Southampton treatment model incorporated data on the UK prevalence of CHD disease states and annual incidence of new CHD cases; therefore, it also could address issues of the effect of strategies on the present and future burdens of CHD. The choice between models would depend on the question being addressed.

Comparisons required changes to the structure of the Southampton treatment model. These were greatest for the comparisons with the York models, which included a two-stage process where a decision tree fed into a long run Markov model. The York-PCI model included stroke and revascularization, whereas the Southampton model included heart failure but not stroke. These differences could not be addressed without significant structural alterations to the Southampton model. For the York-clopidogrel model comparison further structural changes were needed. MI can be distinguished (on the basis of electrocardiograms) into ST elevation MI and non-ST elevation MI (NSTEMI). Guidelines recommend different management for these conditions [15], with clopidogrel recommended for NSTEMI patients [16]. The York-clopidogrel model included individuals with unstable angina and those with NSTEMI. However, the Southampton model did not distinguish between types of MI. Compatibility was achieved by assuming that 50% of MI patients in the Southampton model were NSTEMI, based on expert opinion.

In addition to structural differences, the present study was useful in identifying differences in the data used in the various models. The Southampton and York-PCI models used a 3.5% discount rate for both costs and benefits; the ScHARR-statin and York-clopidogrel models used 6% for costs and 1.5% for benefits. All teams used the rates recommended by NICE at the time of analysis and would use the most up-to-date values in any subsequent modeling (currently 3.5% for both costs and benefits). There was a large difference in the cost of drugs in the two statin models with the value used for the annual cost of statins in the Southampton model approximately half of that in the ScHARR statin model. Again, this reflects the timing of analysis with the cost being calculated in the Southampton model after certain statins had come "off patent." Costs were also an important difference for the PCI comparison; again this reflects timing of analysis as the York-PCI model used an earlier cost year (2003–2004 compared to 2005–2006). All teams were clear that they would use the most up-to-date and appropriate costs available at the time of analysis.

There were differences in the parameter values used in the models where it was harder to reach agreement between teams on the optimum values to be used. All models used utility values derived from a variety of sources. The choice of utility values could not be said to be entirely satisfactory in any of the models. These represent an opportunistic sample of values and were derived from different studies; using different age groups, proportions of men and women, and disease severities. The relative values attached to different disease states are likely to have distortions and inaccuracies. The comparison of models shown here has highlighted that utility values can lead to differences in model results. We are not aware of any work that would provide a credible and consistent set of utility values in CHD. This represents an area where better data are needed to improve models, for example, from a large-scale survey of people with different CVD conditions using appropriate methodology and recording relevant individual characteristics such as age, gender, disease, and severity. It may also be obtainable from a systematic review or a UK consensus of experts (both modelers and clinicians).

Another important difference in input values between models where no consensus was reached was in the choice of mortality rates. For CHD-related mortality rates, the Southampton treatment model used Scottish data [17,18] and information from the Echocardiographic Heart of England Screening Study [19]. The Southampton model adjusted data to allow for differences in standard mortality rates between England and Wales (combined data) and Scotland. Adjustments were also necessary because these data were collected after the introduction of relevant interventions, such as statins. The ScHARR-statin and both York models used the Nottingham Heart Attack Register [20]. These led to differences in the values used, particularly those for stable angina and 1-year post-MI. This is also an area which would benefit from achieving a consensus on the most appropriate data source for UK coronary models. The choice of the optimum data set to use would be a complex decision and should include the views of both modelers and clinicians as to which set of values best represents the current experience of UK individuals with CHD.

We are aware of three similar exercises covering diabetes, colorectal cancer screening, and rheumatoid arthritis modeling. The Mount Hood Challenge [21,22] compared international diabetes models by populating parameters using a common data set. This exercise required prior definitions of disease states and assumptions, and a new data set to model, and concluded that it was feasible to cross-validate and explain differences in dissimilar diabetes simulation models using standardized patients. In this example, wide differences in model results were observed and the authors concluded that this demonstrated the need for cross-validation [21]. They also concluded that performing systematic comparisons and validation exercises enabled the identification of key differences among the models, as well as their possible causes and directions for improvement in the future [22]. A similar exercise was carried out in colorectal cancer where participants used their models to address pre-specified screening scenarios [23]. The authors concluded that comparisons can identify critical sources of

variation. They also stated that the next steps to such an exercise would be modelers and clinicians continuing to work together to resolve differences identified. Four models of rheumatoid arthritis were also compared [24]. This was largely done from the published articles by comparing model descriptions and published sensitivity analysis. The authors had electronic access to one model; this was compared to one of the other models by inputting data values from a second model. However, this was not done for the other two models as these had different structures and the authors felt comparisons were not feasible. The authors found that output differences depended on structure, assumptions, and utilities, and concluded by emphasizing the need for transparency in reporting and for a continuing debate on model quality in order to reach consensus on difficult methodological issues.

These studies show that a number of different approaches have been used to compare models. Comparing published results and sensitivity analyses would be the simplest approach, and would be feasible from public domain information without requiring access to any of the models used. This method is likely to be limited in its ability to explain the reasons for differences in results. Using a common data set or setting a series of scenarios for each model to address would require the cooperation of all modeling teams involved. However, it would give information on how each model is altered with changes in inputs and values, and what answers different models give to a common question. The current exercise represents a compromise – it did not require sharing models, instead it relied on changing the data input to one model only. This was facilitated by an exchange of information between modeling teams.

We found the current exercise feasible but demanding. Comparison was aided by the fact that two of three comparator models were published as Health Technology Assessment (HTA) monographs [7,9] and these are considerably more detailed than most journal articles. Word count limitations often restrict the level of detailed information possible, requiring in-depth discussion with the model owner to accurately repeat the scenario. In fact, even with these detailed reports, cooperation between teams was required in terms of answering specific questions and also supplying further information not available from the published data. We believe that modeling articles should include detailed web appendices to aid replication and checks of between-model consistency. However, it may also be desirable to go beyond this and have accepted standards for the reporting of models, similar to those in existence for clinical trials. This will ensure more consistency in the ways that models are reported and, hence, may make comparing models easier.

In the current exercise, comparison was aided as each model involved had been constructed to conform to NICE guidelines for technology appraisals [4]. However, differences between models were generated because recommended discount rates had changed over time. Although changes in guidelines will be necessary as methodology evolves and circumstances change, it should be recognized that this may make comparing models more difficult and recommended changes should be very carefully justified. Adhering to guidelines meant that there were a number of similarities between models, for example, in the use of an NHS cost perspective and a cost per QALY approach. This illustrates the value of guidelines in promoting consistency in key methods and parameters. However, adherence to guidelines would not help when there is genuine uncertainty over the best data source to use. For example, in the current study, it was not clear as to the best source of data for both utilities and mortality rates.

Checking between-model consistency requires a potentially large investment in researcher time, and it is important to consider situations where the exercise would be useful to modelers and decision makers. This is more likely to be the case where the condition poses a significant burden and where considerable uncertainties exist. CHD is a good example of this because it imposes a large health [25] and economic burden [26], and is a complex condition. Model comparison will also be indicated if there are large differences in model results, particularly if results from different models cross decision-maker's thresholds. Here there will be considerable uncertainty as to the implications of results to decision making. Checks of between-model consistency will also be useful as a development tool for modelers, as these can illustrate the model characteristics that are similar or different to those of existing models. Model comparisons were used for this reason in the current exercise as a way of both checking and developing the Southampton model. However, we feel it is important to remember that checks of between-model consistencies can show differences between models, but they will have limited use in demonstrating which one is the "best" model. Often there will be no clear indication that one data source or method is better than another. A check of between-model consistency cannot be used as a substitute for external validation.

Conclusion

The exercise indicated that it was feasible, but not straightforward, to compare models where there are structural differences between models. A check of between-model consistency was found to be a useful tool for model development and aided the development of the Southampton treatment model. We also found model comparison to be useful in identifying model inputs where there were weaknesses in the data available and, hence, could be useful in prioritizing future research needs. Organizations responsible for guidelines should be aware that changes to these guidelines may make it more difficult to compare models and this should be a consideration in the decision as to whether to make changes to preferred methods. Fuller publication of models, perhaps through detailed web appendices, could facilitate paper-based comparisons. A common standard for model reports similar to that required for trials would also facilitate the comparison of models. A variety of methods to compare models have been used in the literature. It is not currently clear as to the relative strength and weaknesses of different methods and more work exploring this issue would be useful. This exercise provides a useful guide to future CHD modelers and policy makers on areas requiring further exploration.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at 10.1016/j.jval.2010.10.009.

REFERENCES

[1] 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;8:1–158.

[2] Burch J, Paulden M, Conti, et al. Antiviral drugs for the treatment of influenza: a Systematic Review and Economic Evaluation. Health Technol Assess 2009;13:1–265.

[3] Briggs AH. Handling uncertainty in cost-effectiveness models. Pharmacoeconomics 2000;17:479–500.

[4] National Institute for Health and Clinical Excellence (NICE). NICE, guide to the methods of technology appraisal. Available from: URL: http://www.nice.org.uk [Accessed November 1, 2010].

[5] Cooper K, Davies R, Roderick P, et al. The development of a simulation model of the treatment of coronary heart disease. Health Care Management Science 2002;5:259–67.

[6] Cooper K, Davies R, Raftery J, Roderick P. Use of a coronary heart disease simulation model to evaluate the costs and effectiveness of drugs for the prevention of heart disease. J Oper Res Soc 2008;59:1173–81.

[7] Ward S, Lloyd Jones M, Pandor A, et al. A systematic review and economic evaluation of statins for the prevention of coronary events. Health Technol Assess 2007;11:1–160.

[8] Vergel YB, Palmer S, Asseburg C, et al. Is primary angioplasty cost effective in the UK? Results of a comprehensive decision analysis. Heart 2007;93:1238–43.

[9] Main C, Palmer S, Griffin S, et al. Clopidogrel used in combination with aspirin compared with aspirin alone in the treatment of non-ST-segment-elevation acute coronary syndromes: a systematic review and economic evaluation. Health Technol Assess 2004;8:1–141.

[10] Palmer S, Sculpher M, Philips Z, et al. Management of non-ST-elevation acute coronary syndromes: How cost-effective are glycoprotein IIb/IIIA antagonists in the UK National Health Service. Int J Cardiol 2005;100:229–40.

[11] Mauskopf JA, Boye KS, Schmitt C, et al. Adherence to guidelines for sensitivity analysis: cost-effectiveness analysis of dual oral antiplatetlet therapy. J Med Econ 2009; 12:141–53.

[12] Yusuf S, Zhao F, Mehta SR, et al. Effects of clopidogrel in addition to aspirin in patients with acute coronary syndromes without ST-segment elevation. N Engl J Med 2001;345:494–502.

[13] Wailoo A, Goodacre S, Sampson F, et al. Primary angioplasty versus thrombolysis for acute ST-elevation myocardial infarction: an economic analysis of the National Infarct Angioplasty project. Heart 2010;96:668–72.

[14] Karnon J, Brennan A, Akehurst R. A critique and impact analysis of decision modelling assumptions. Med Decis Making 2007;27:491–9.

[15] British Cardiac Society Guidelines and Medical Practice Committee and Royal College of Physicians Clinical Effectiveness and Evaluation Unit. Guideline for the management of patients with acute coronary syndromes without persistent ECG ST segment elevation. Heart 2001;85: 133–42.

[16] The Task Force on the Management of Acute Coronary Syndromes of the European Society of Cardiology. Management of acute coronary syndromes in patients presenting without persistent ST-segment elevation. Eur Heart J 2002;23:1809–40.

[17] Capewell S, Murphy NF, MacIntyre K, et al. Short-term and long-term outcomes in 133,429 emergency patients admitted with angina or myocardial infarction in Scotland, 1990-2000: population-based cohort study. Heart 2006;92: 1563–70.

[18] Murphy NF, Stewart S, Hart CL, et al. A population study of the long-term consequences of Rose angina: 20-year follow-up of the Renfrew-Paisley study. Heart 2006;92:1739–46.

[19] Hobbs FD, Roalfe AK, Davis RC, et al. Prognosis of all-cause heart failure and borderline left ventricular systolic dysfunction: 5 year mortality follow-up of the Echocardiographic Heart of England Screening Study (ECHOES). Eur Heart J 2007;28:1128–34.

[20] Gray D, Hampton JR. Twenty years' experience of myocardial infarction: the value of a heart attack register. Br J Clin Pract 1993;47:292–5.

[21] Brown JB, Palmer AJ, Bisgaard P, et al. The Mt. Hood challenge: cross-testing two diabetes simulation models. Diabetes Res Clin Pract 2000;50(Suppl. 3):S57–S64.

[22] The Mount Hood 4 Modeling Group. Computer modeling of diabetes and its complications: a report on the Fourth Mount Hood Challenge Meeting. Diabetes Care 2007;30:1638–46.

[23] Pignone M: Russell L, Wagner J, eds. Economic models of colorectal cancer screening in average-risk adults. Workshop summary. Washington DC: The National Academies Press; 2005.

[24] Drummond MF, Barbieri M, Wong JB. Analytic choices in economic models of treatments for rheumatoid arthritis: what makes a difference? Med Decis Making 2005;25:520–33.

[25] Petersen S, Peto V, Scarborough P, Rayner M. Coronary Heart Disease Statistics. London: British Heart Foundation, 2005.

[26] Luengo-Fernandez R, Leal J, Gray A, et al. Cost of cardiovascular diseases in the United Kingdom. Heart 2006; 92:1384–9.