Pharmacoeconomics 2011; 29 (1): 51-62
1170-7690/11/0001-0051/$49.95/0

Calibrating Models in Economic Evaluation

A Comparison of Alternative Measures of Goodness of Fit, Parameter Search Strategies and Convergence Criteria

Jonathan Karnon[1] and *Tazio Vanni*[2]

1 University of Adelaide, Adelaide, South Australia, Australia
2 London School of Hygiene and Tropical Medicine, London, UK

Abstract

Background: The importance of assessing the accuracy of health economic decision models is widely recognized. Many applied decision models (implicitly) assume that the process of identifying relevant values for a model's input parameters is sufficient to prove the model's accuracy. The selection of infeasible combinations of input parameter values is most likely in the context of probabilistic sensitivity analysis (PSA), where parameter values are drawn from independently specified probability distributions for each model parameter. Model calibration involves the identification of input parameter values that produce model output parameters that best predict observed data.

Methods: An empirical comparison of three key calibration issues is presented: the applied measure of goodness of fit (GOF); the search strategy for selecting sets of input parameter values; and the convergence criteria for determining acceptable GOF. The comparisons are presented in the context of probabilistic calibration, a widely applicable approach to calibration that can be easily integrated with PSA. The appendix provides a user's guide to probabilistic calibration, with the reader invited to download the Microsoft® Excel-based model reported in this article.

Results: The calibrated models consistently provided higher mean estimates of the models' output parameter, illustrating the potential gain in accuracy derived from calibrating decision models. Model uncertainty was also reduced. The chi-squared GOF measure differentiated between the accuracy of different parameter sets to a far greater degree than the likelihood GOF measure. The guided search strategy produced higher mean estimates of the models' output parameter, as well as a narrower range of predicted output values, which may reflect greater precision in the identification of candidate parameter sets or more limited coverage of the parameter space. The broader convergence threshold resulted in lower mean estimates of the models' output, and slightly wider ranges, which were closer to the outputs associated with the non-calibrated approach.

Conclusions: Probabilistic calibration provides a broadly applicable method that will improve the relevance of health economic decision models, and simultaneously reduce model uncertainty. The analyses reported in this paper inform the more efficient and accurate application of calibration methods for health economic decision models.

Background

Decision models are now an expected framework for cost-effectiveness analyses of healthcare technologies, as demonstrated by guidelines for submissions to reimbursement bodies in the UK and Australia.[1,2] Guidelines for the conduct of decision models have highlighted the importance of assessing the accuracy of model predictions,[3,4] but investigation of processes for undertaking such assessments is an area that has been largely overlooked in the cost-effectiveness literature. Model calibration involves the identification of input parameter values that produce model output parameters that best predict observed data. Criteria for an acceptable goodness of fit (GOF) of the model's outputs and observed data can be specified, so that we can be confident that the model has achieved an acceptable level of accuracy.

An accompanying article by Vanni et al.[5] in this issue splits the calibration process into seven stages, and presents a theoretical discussion on the options available to the analyst at each of these stages. Three key stages include the measure of GOF used to represent the relative accuracy of different sets of input parameter values, the search strategies that select sets of input parameter values to be tested in the calibration process, and the threshold for convergence (acceptable GOF). This article compares alternative approaches to these three stages, using a published cohort Markov model that compared adjuvant therapies for early breast cancer.[6]

Another key issue identified by Vanni et al.[5] concerned the integration of the calibration process with the final analysis of the economic model. Probabilistic calibration, a process developed by one of the authors (JK),[7-11] applies probability weights to sets of input parameter values that adequately predict a range of calibration outputs (or targets). This approach is a widely applicable approach to calibration that can be easily integrated with probabilistic sensitivity analysis (PSA). The comparisons of GOF measures, parameter search strategies and convergence thresholds are presented in the context of probabilistic calibration.

The following section summarizes the early breast cancer model structure, and the data sources used to populate the model. The calibration process is then described, stepping through the methods used in each of the seven stages, highlighting the alternative approaches tested in stages 3 (GOF measures), 4 (parameter search strategies) and 5 (convergence thresholds). The results show the differences of the alternative approaches with respect to the outputs of the calibrated model, followed by a discussion of the implications. The appendix provides a user's guide to probabilistic calibration, with the reader invited to download the Microsoft® Excel-based model reported in this article (see the Supplemental Digital Content 1, http://links.adisonline.com/PCZ/A94).

Methods

The Early Breast Cancer Model

The model structure is similar to many published models developed for the evaluation of adjuvant therapies for early breast cancer.[12] As presented in figure 1, women commence the model in a disease-free state (following primary surgery). They may remain disease free or experience a new breast cancer-related event, categorized as a contralateral primary tumour (a new tumour in the opposite breast to the original tumour), a locoregional recurrence (reappearance of cancer in the same breast or the lymph nodes) or metastatic recurrence (when the cancer spreads to other parts of the body). Metastatic recurrence

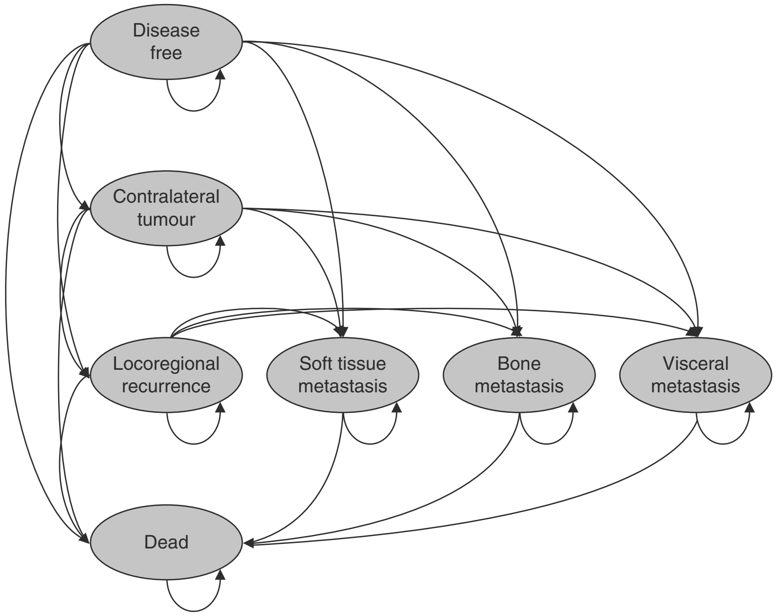

Fig. 1. Early breast cancer model structure (reproduced from Karnon et al.,[6] with kind permission from Springer Science+Business Media).

has the worst prognosis, and is further categorized by the site at which it is diagnosed – soft tissue, bone and visceral. Women can die from other causes from any state, but may only die due to breast cancer from the metastatic recurrence states.

Data sources to inform independent values for the clinical parameters have been reported elsewhere,[6] and include a large meta-analysis to inform probabilities of relapse,[13] and clinical trials to inform probabilities of different event types, progression post-non-metastatic relapse and mortality rates post-metastases.

Calibration Stages

Stage 1: Which Parameters Should be Varied in the Calibration Process?

As noted by Vanni et al.,[5] calibration can be used to identify values for unobservable parameters, or to identify the best fitting sets of parameter values, including observed and unobserved parameters, in the calibration process. The latter approach is relevant to probabilistic calibration as the aim of probabilistic calibration is to represent the correlations between the widest possible range of input parameters. The widest range is defined as all input parameters that contribute to the predicted values of the calibration targets. In this case, the calibration target is the age-dependent overall survival rate (see next section), which is affected by all of the model's clinical parameters, and so all of these parameters are varied in the calibration process.

Stage 2: Which Calibration Targets Should be Used?

The model does not include input parameters describing overall survival, which is estimated indirectly as a function of parameters describing relapse rates, progression and death post-relapse, and other-cause mortality. Thus, overall survival was defined as an appropriate calibration target.

A population-based longitudinal study of post-menopausal, hormone-positive, early breast cancer patients with a planned treatment schedule of 5 years' treatment with tamoxifen, conducted in the jurisdiction of interest, with no loss to follow-up, would have been the preferred data source for the calibration targets. However, no such study was identified. Data presented by the early breast cancer trialists group's meta-analysis were selected,[8] which were drawn from many multinational trials that had evaluated the effectiveness of 5 years' early adjuvant tamoxifen therapy. The limitations of clinical trials as calibration sources are recognized, including likely differences in the characteristics of patients who consent to enter trials and the broader patient population, as well as the potential for selective follow-up. However, no solid basis for adjusting reported survival rates could be devised and so the data were used as reported to inform the calibration targets.

Data from the meta-analysis described the number of deaths and number of life-years at risk of death within years 0–4, years 5–9 and beyond year 10 post-primary surgery for (postmenopausal) women aged ≥50 years who were allocated to receive 5 years' tamoxifen. The mean age in recent trials of adjuvant therapy for postmenopausal women with early breast cancer has been around 60 years,[14,15] and so these data are useful as a target for a modelled patient population of mean age 60 years.

Beyond year 15 post-primary surgery, survival data were presented only for all women allocated to 5 years' tamoxifen, regardless of age. The mean age in this group will be lower, and so may be less reflective of overall survival in the model population. Lower patient age reduces non-breast cancer mortality, although cancer in a younger cohort may be more aggressive and so breast cancer mortality may be increased. The increased uncertainty beyond 15 years is reflected to some extent by the increased width of the 95% confidence intervals (CIs) for the mortality estimates in this time period. Given the time span of the data, the use of overall survival as a calibration target is capped at 20 years post-primary surgery.

It is noted that the independently specified age-specific survival probabilities are proxy targets for a continuous survival curve. The specification of convergence criteria that requires all age-specific survival probabilities to be hit simultaneously (see below) introduces some dependence between the survival probabilities, but the applied approach remains a second-best solution to the preferred use of a survival curve as an overall target.

Stage 3: What Measure of Goodness of Fit (GOF) Should be Used?

As discussed by Vanni et al.,[5] there is no consensus on the most appropriate measure of GOF. In this application, as multiple calibration targets were specified, it was deemed appropriate to use GOF measures that account for the precision of the empirical data informing different targets. As there was no *a priori* reason to choose between the chi-squared measure of GOF, and the likelihood approach, both measures are used and the resulting calibrated outputs of the model compared.

To create an overall measure of GOF, it is necessary to combine the individual GOF measures across the four age-specific survival probabilities. As the four probabilities describe similar outputs and because the timing of events is relevant, an equal weight was applied to the GOF measures for each age-specific survival probability.

Stage 4: What Parameter Search Strategy Should be Used?

The parameter search strategy refers to the process of identifying parameter sets whose corresponding model outputs produce an acceptable fit to observed output values.

The starting point for any search strategy is the space within which the search will occur. Some health economic decision models have defined an open space,[16] where all parameters are able to take any value. This approach may be suitable for calibration processes that aim to identify a limited number of best fitting parameter sets and candidate parameter sets can be manually inspected. However, in the context of probabilistic calibration, where all convergent parameter sets are maintained, it is not practical to inspect all parameter sets and so there is a risk that some convergent parameter sets could accurately predict calibration targets using infeasible values for one or more input parameters.

To avoid this problem, the parameter space is defined by probability distributions representing the uncertainty around each input parameter (as required for PSA). This limits the selection of parameter values to those within the defined probability distributions, and means that values closer to the expected value of each parameter are more likely to be selected.

Given the parameter space to be searched, there are various search strategies that can be applied to non-linear models, as are typically used in economic evaluation, but a broad distinction can be drawn between guided and random search strategies. A random search strategy selects parameter values at random from the defined parameter space. A guided strategy learns from previously selected parameter sets and moves towards better fitting parameter sets.[5]

The reported calibration process compares the random search method with a guided search strategy. The choice of guided search strategy was informed by a desire to use a Microsoft® Excel-based approach that could be generally applied by most cost-effectiveness analysts, but one that also produces outputs that can be integrated with a PSA. The in-built Excel optimization tool is Microsoft® Excel Solver, which can identify the best fitting set of parameter values, but lacks the facility to automatically record and store multiple parameter sets and their corresponding GOF measures.

Using Microsoft® Excel, an alternative approach involves the use of specialist add-ins, including extended versions of Microsoft® Excel Solver (Frontline systems[17]). The Risk Solver Platform offers a range of guided search strategies but – as most health economic decision models are non-linear – linear programming or Quadratic Solvers are generally not appropriate. The generalized reduced gradient (GRG) method was selected as a practicable approach. From each starting point in the parameter space, the GRG method moves along a gradient until it reaches a minimum point. Using the Risk Solver Platform, the GRG method can be set to start at multiple starting points that are systematically selected to cover the parameter space, which identifies the parameter sets associated with the multiple local minimum (optimum) points. Further start points are selected until no further improvements in fit are identified.

Stage 5: What Determines Acceptable GOF Parameter Sets (Convergence)?

Convergence criteria describe the minimum level of accuracy of model outputs compared with the observed values of the calibration targets. In the context of probabilistic calibration, convergence criteria determine whether each tested set of parameter values will be assigned a probability weight and used in the PSA, or whether the outputs are too divergent from the targets and the parameter set should be discarded from further analysis.

There is no consensus on the specification of convergence thresholds, but a pragmatic approach is to specify that the model output must lie within a specified CI of the observed data used as calibration target.

In this application, two alternative convergence criteria were defined and tested. First, parameter sets were excluded when one or more of the model output values lay outside the 95% CI of the relevant calibration target. An additional convergence criterion was then added, which accounted for the relevance of parameter sets that achieve high levels of accuracy for most targets, but fail with respect to a minority of targets. Here, additional parameter sets were included if three of the four targets were within the respective 90% CIs.

Stage 6: What Determines the Termination of the Calibration Process (Stopping Rule)?

The stopping rule for the parameter search determines the number of different parameter sets that are selected and tested as part of the calibration process, which can influence the level of accuracy of the fitted parameter values. In the context of probabilistic calibration, the stopping rule determines the number of different parameter sets that will be included in the PSA (i.e. the number of parameter sets that meet the convergence criteria).

For the random search strategy, the aim was to identify sufficient numbers of convergent parameters to be confident that the parameter space was adequately covered by the search strategy. An iterative stopping rule process tested the adequacy of increasing numbers of convergent input parameter sets. Initially, the search strategy was run until 1000 convergent input parameter sets were generated, followed by another search that generated a further 1000 convergent input parameter sets. If the lower CI values for all calibration targets lay below the mean values for the corresponding targets in the alternative set of convergent input parameters, and the upper CI values lay above the mean values in the alternative set, the stability in the calibration outputs was judged to demonstrate the adequacy of the coverage of the parameter space. If not, the two sets of 1000 were combined, an additional 2000 convergent input parameter sets were generated, and the process was repeated.

The duration of the GRG algorithm used in the Risk Solver Platform is determined by a series of calibration controls, for which the default settings were used in the calibration of the early breast cancer model. These settings included the definition of local optima if the absolute value of the relative change in the objective function is less than 0.0001 for the last five iterations. The selected multistart option generates multiple starting points, which are grouped into clusters, with separate searches commencing from a representative point within each cluster. After each local optima is identified, an updated estimate of the most probable total number of locally optima is defined. The multistart search ceases when the number of identified local optima is within one unit of the most probable total number of locally optimal solutions.

The program only extracts the parameter values associated with each local optima, rather than the much larger number of parameter sets that would have been tested along the route to each local optima.

Stage 7: How Should Model Calibration Results and the Main Model Analysis be Integrated?

A simple way to integrate the model calibration results and main model analysis is to use only the best fitting parameter set in the economic evaluation; however, in order to account for parameter uncertainty, we should estimate probability weights for each convergent input parameter set, which represent the probability that each set is the best fitting parameter set.

The probability weight assigned to each convergent parameter set is estimated as the reciprocal of the sum of the GOF measures for the calibration targets for each parameter set, divided by the sum of the reciprocals across all convergent parameter sets. The reciprocal is used because smaller absolute measures of GOF represent better fitting parameter sets (requiring higher probability weights).

The main analysis of the model then involves sampling large numbers of convergent parameter sets on the basis of the assigned probability weights, and recording the associated model outcomes. These outputs are analysed to inform mean estimates of each model output, as well as facilitating analysis of uncertainty via a PSA. The process for estimating the probability weights is described in the Appendix.

Analysis

The reported early breast cancer model is analysed to compare two alternative approaches within each of the following components of the probabilistic calibration process: measures of GOF (chi-squared vs likelihood), parameter search strategies (random vs guided) and convergence criteria (all targets within 95% CI vs all targets within 95% CI or three targets within 90% CI). The alternative convergence criteria are compared only in the context of a random search strategy, as the GRG-guided search strategy retains only the identified local optima within the defined parameter space, all of which met the tighter convergence threshold.

The alternative approaches are compared with respect to the main health outcome of the calibrated model – life-years gained, including both the mean and 95% CIs for this parameter.

Results

Using the random search strategy and the narrow convergence criteria, 4328 parameter sets were sampled in order to identify 1000 convergent parameter sets. Under the broad convergence criteria, 2472 parameter sets were sampled in order to identify 1000 convergent parameter sets. The random search strategy demonstrated adequacy of the coverage of the parameter space with 1000 convergent parameter sets.

Using the GRG-guided search strategy, all recorded parameter sets were convergent to the narrow convergence criteria; 549 parameter sets representing local optima within the parameter space were identified.

Figure 2a presents the distributions of the sum of the chi-squared GOF measures for convergent parameter sets for the guided and random search strategies. The guided search strategy produces a lower range of GOF measures, with a mean of 2.16 and no GOF values over 10. The random search strategy, combined with narrow convergence criteria has a mean of 6.3 and a tail going up to 14. The broad convergence criteria for the random search strategy results in a mean GOF measure of 8, with a longer tail going out to 24.

Figure 2b presents similar data using the log-likelihood measure of GOF. Values closer to zero are indicative of greater accuracy, and a similar picture emerges to the chi-squared GOF measure with respect to the three search strategy/convergence criteria combinations.

The lack of variation in the log-likelihood GOF measures, combined with the magnitude of the GOF values, means that there was very little variation in the derived probability weights. For

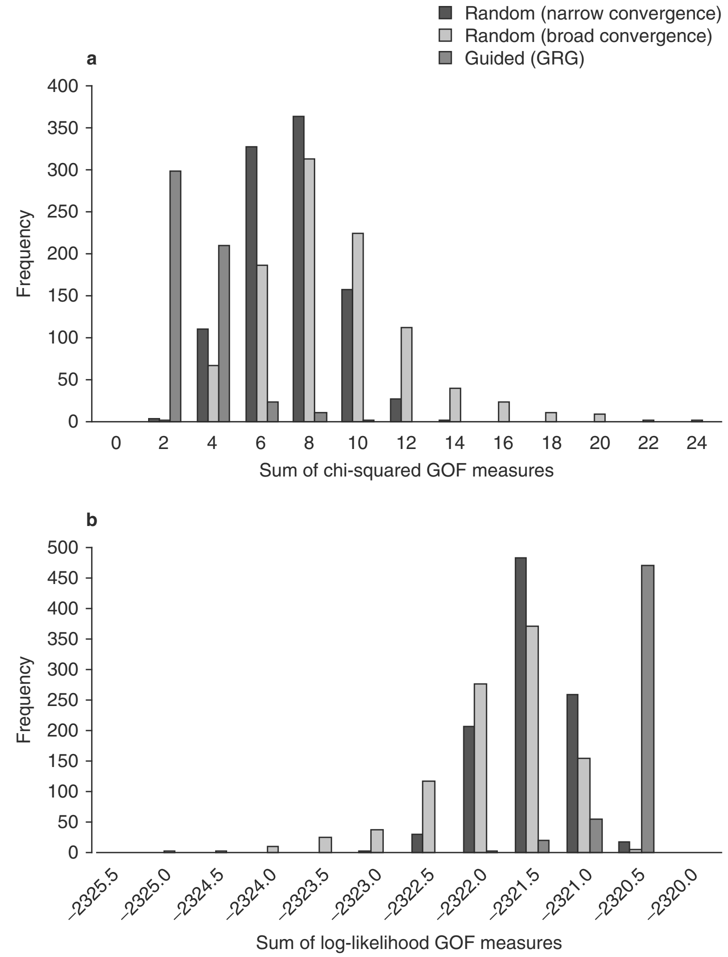

Fig. 2. Comparison of the distributions of the (**a**) chi-squared goodness-of-fit (GOF) measure using a random search strategy and a generalized reduced gradient strategy (GRG); and (**b**) log-likelihood GOF measure using a random search strategy and a GRG strategy.

Table I. Comparison of model outputs (life-years gained) for calibrated and non-calibrated models

Outputs	Mean (95% CI)
Non-calibrated	12.701 (12.439, 12.969)
Calibration approach (GOF measure used, search strategy, convergence criteria[a])	
chi-squared GOF, random search, narrow convergence	12.876 (12.733, 13.056)
likelihood GOF, random search, narrow convergence	12.849 (12.720, 13.027)
chi-squared GOF, random search, broad convergence	12.834 (12.685, 13.014)
likelihood GOF, random search, broad convergence	12.803 (12.659, 12.976)
chi-squared GOF, guided (GRG) search, narrow convergence	13.096 (12.913, 13.161)
likelihood GOF, guided (GRG) search, narrow convergence	13.091 (12.877, 13.167)

[a] All recorded parameter sets met the narrow convergence criteria for the guided search strategy.

GOF = goodness of fit; **GRG** = generalized reduced gradient.

the 1000 convergent parameter sets identified using the random search strategy, the minimum and maximum probability weights were 0.00099 and 0.001, respectively. The minimum and maximum probability weights derived from the chi-squared GOF measure were 0.0003 and 0.0038, respectively.

Table I and figure 3 present estimates of the output parameter ‘discounted life-years gained’ using the chi-squared GOF measures, as well as a non-calibrated PSA (comprising the first 1000 randomly sampled sets of input parameter sets). Table I shows that the non-calibrated model provides the lowest mean estimate of life-years gained, almost 0.4 life-years less than the maximum estimate (the guided search strategy with the chi-squared GOF measure). The narrow convergence criteria results in a higher mean estimate than the broad criteria, and the chi-squared GOF measure gives higher estimates than the likelihood GOF measure.

Figure 3 illustrates the uncertainty around the mean estimates using the chi-squared GOF analyses, showing that the calibrated PSAs produce narrower ranges of the output parameter than the non-calibrated analyses. The narrow convergence criteria provides slightly greater certainty than the broad convergence criteria, but the guided search strategy provides the most certainty with respect to the mean estimates of life-years gained.

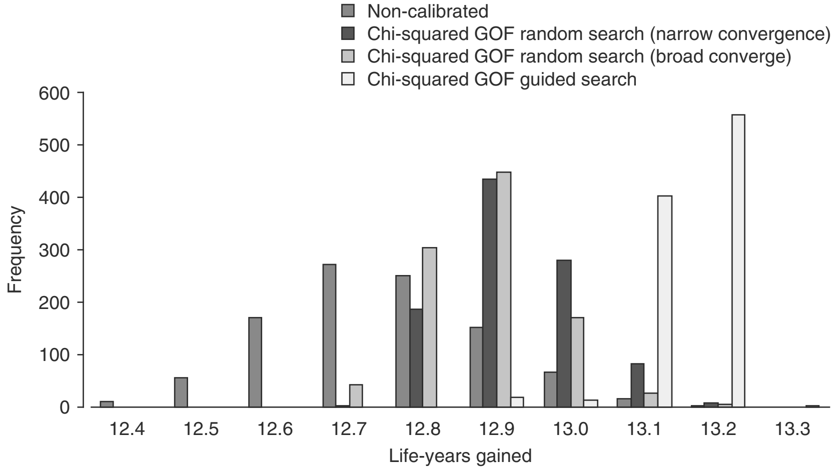

Fig. 3. Comparison of the distributions of discounted life-years produced by probabilistic sensitivity analyses using calibrated and non-calibrated input parameters. **GOF** = goodness of fit.

Discussion

This article has presented an empirical application and comparative analysis of model calibration methods emphasizing the use of probabilistic calibration. In a PSA, non-calibrated models sample individual input parameters independently of the value of other input parameters and so are more likely to include infeasible combinations of input parameter values. Probabilistic calibration defines the probability that different combinations of input parameter values are the true combinations. In the presented analyses, the calibrated models consistently provided higher mean estimates of the models' output parameter (life-years gained) than the non-calibrated model, which represents improvements in accuracy as the calibrated analyses exclude non-convergent parameter sets that predict output values outside an acceptable range, and assign greater weight to better predicting parameter sets. In the process, model uncertainty is also reduced, as demonstrated by the 95% CIs presented in table I.

Recently published models provide examples of applied calibration processes for decision models. Kim et al.[18] randomly sampled 550 000 parameter sets from uniform distributions over the identified range for each parameter, and then used a subset of 50 good fitting parameter sets to illustrate the range of the model's outputs. Jit et al.[16] used a combination of simulated annealing and a GRG algorithm, using a range of seed parameter sets, to be sure of identifying the global minimum (i.e. the optimum combination of input parameters). Neither of these approaches facilitated a PSA of the model, and hence the estimation of credible intervals around mean estimates of net benefit, cost-effectiveness acceptability curves and/or value of information analyses.

In probabilistic calibration, the measure of GOF is not used to identify the optimal set of input parameters, but to inform probabilities that each convergent parameter set provides the most accurate estimate of the defined calibration targets (such as life-years gained in the presented example). Using the chi-squared and likelihood GOF measures, the former differentiated between the accuracy of different parameter sets to a far greater degree than the likelihood GOF measure.

Table I shows that the life-years gained estimates using the likelihood GOF measures are consistently lower than the chi-squared output values. The lower values are closer to the non-calibrated outputs because there is less differentiation between convergent parameter sets (the magnitude of the estimated likelihood GOF measures is large compared with the differences in the GOF measures for different parameter sets). This lack of differentiation has been observed previously,[8] but it is possible that other models, involving more targets and/or more uncertainty in the data, could display greater differentiation using likelihood GOF measures – this is an empirical question.

There is no consensus in choosing between the chi-squared and likelihood GOF measures. However, we note that each approach will identify a single best fitting set of input parameters and that the model outputs of interest (e.g. incremental costs and benefits) associated with the best fitting set can be compared with the mean outputs of the probabilistic calibration (i.e. probabilistic analysis sampling parameter sets with respect to their defined probability weights). All else equal, we should prefer the GOF measure that produces mean outputs that are closest to the outputs associated with the best fitting set. In the breast cancer model presented, this decision rule would favour the chi-squared GOF measure, although the likelihood GOF measure may be preferred in other models.

The parameter search strategy was also shown to impact model outputs. In the breast cancer model, the guided (GRG) search strategy produced higher mean estimates of the models' output parameter, as well as a narrower range of predicted output values, than informed by a random search strategy. The guided search strategy recorded only local optima within the searched parameter space, and the reduced uncertainty around the model outputs may reflect greater precision in the identification of candidate parameter sets. Alternatively, the reduced uncertainty may reflect more limited coverage of the parameter space by the guided search strategy (i.e. a range of alternative local optima may have been missed). The multistart application of the GRG search strategy is intended to minimize the probability that the global optimum (or extrema) is missed, but there may be a greater chance that other local optima are missed. Further empirical comparisons of random and alternative guided search strategies will inform this issue but, at this early stage of the development process of probabilistic calibration, the use of a thorough random search strategy (with adequacy of coverage checks) is acceptable.

A third issue concerned the choice of the convergence threshold, where a narrow and a broad threshold were compared in the context of the random search strategy. The broader threshold resulted in lower mean estimates of the models' output, and slightly wider ranges, which were closer to the outputs associated with the non-calibrated approach. We recommend that, if sufficient numbers of convergent parameter sets can be identified and adequacy of coverage checks are passed, then tighter convergence criteria should be specified.

Conclusions

The application of model calibration, particularly probabilistic calibration, as applied to health economic decision models is not widespread, and further investigation is required. In particular, a process of validation against more theoretically grounded approaches would be valuable (such as the Bayesian updating approach[19]). However, we believe that model calibration should not be restricted to applications that use less accessible modelling techniques, and probabilistic calibration provides a broadly applicable method that will improve the relevance of health economic decision models, and simultaneously reduce model uncertainty.

Acknowledgements

No sources of funding were used to conduct this study or prepare this manuscript. The authors have no conflicts of interest that are directly relevant to the content of this article.

Appendix

The Early Breast Cancer Model

The model is developed in Microsoft® Excel, contains seven worksheets and is available as Supplemental Digital Content. The first 'Model structure' sheet contains a diagram illustrating the structure of the model – the health states and the possible routes between the health states, as described in the main article. The remaining sheets are laid out in the order in which they are accessed during the calibration process. The following sections describe the content and actions required within each sheet.

Input Parameters

The 'Input parameters' sheet contains the data informing the probability distributions for all of the model's input parameters, and specifies the probability distributions from which sampled values are drawn.

The first nine rows determine the recurrent event experienced (contralateral; locoregional; or soft tissue, bone or visceral metastases) if a recurrence occurs. The sampled values of the Beta distributions defined for each recurrent event in column E are adjusted in columns F, G and H so that the probabilities for the relevant recurrent events for patients in the disease-free, contralateral and locoregional states sum to 1.

Column F of rows 12–16 contains Beta distributions describing the probability of a recurrent event in different time periods, which are combined with the probabilities for the relevant recurrent events to estimate separate probabilities of the different recurrent events (in columns H to L). Recurrent events from the contralateral and locoregional states are handled similarly in rows 19 and 20.

Rows 21–23 describe the annual probability of death for patients with the different forms of metastatic recurrence. The rows below describe other cause mortality rates, subtracting age-specific breast cancer mortality from general population female mortality rates.

Calibration Model

This sheet contains the version of the model that is used to calibrate the input parameter values. It is set up as a standard cohort Markov model, with a separate column for each health state, and a separate row for each Markov cycle, with the transitions between states linked to the input parameters sheet. The model outputs correspond to the defined calibration targets, with overall survival at 5, 10, 15 and 20 years being recorded in column K.

Calibration Outputs

The 'Calibration outputs' sheet calculates the GOF measures for each set of input parameters (row 15), and then stores the results (row 17 and below). First, the user must specify how many convergent parameter sets are required (cell M4). The calibration process is then started by hitting the 'Run Calibration' button (cells L6 to O11).

The mean, standard error and 95% CIs for the calibration targets are reported in columns D–G of rows 4–8. The model's predictions of the calibration targets from each iteration of the calibration model are captured in cells B15–E15. Cells F15–I15 contain the formulae for estimating the chi-squared GOF measures. If the convergence criteria are met (i.e. all predicted values lie within the 95% CI of the respective calibration targets), the sum of the four GOF measures is reported in cell J15. Cell K15 reports the reciprocal of the sum of the GOF measures.

Cells L15 and M15 are populated via a Visual Basic for Applications (VBA) macro that first estimates the probability that each convergent parameter set is the optimal parameter set as the reciprocal of the sum of the GOF measures for each set, divided by the sum of the reciprocals of the GOF measures (cell K12). Column M then estimates the cumulative probabilities, which are required to sample the convergent sets as part of the PSA.

The same process for the log-likelihood GOF is contained in columns O–V.

Calibrated Inputs

Each sampled set of the input parameter values in the calibration process are captured in columns B–O of row 3 (descriptions of the corresponding parameters are held in rows 1 and 2). As they are evaluated, each parameter set is stored consecutively from row 7 downwards (controlled by the VBA macro operated from the 'Calibration outputs' sheet).

When the macro is completed, the user must copy the full set of cumulative probabilities from the 'Calibration outputs' sheet into cell A7 of the 'Calibrated inputs' sheet. If the chi-squared GOF measure is being used, cumulative probabilities in column M should be copied, if the log-likelihood GOF measure is used, cumulative probabilities in column V should be copied.

When the 'Run final model' VBA button (in sheet 'Final model') is hit, Cell A1 randomly samples a value between 0 and 1, which is copied into cell A2 and which selects the corresponding input parameter set to be used in the PSA of the final model (i.e. the parameter set with the cumulative probability most directly above the sampled value). The selected set of input parameter values are copied into cell B5.

Final Inputs

The 'Final inputs' sheet contains the model's input parameters, which are linked to the selected parameter values that are held in cells B5–O5 of the 'Calibrated inputs' sheet.

Final Model

The final sheet contains another copy of the breast cancer model, which is populated using parameter values sampled from the 'Calibrated inputs' sheet, and stored in the 'Final inputs' sheet. This model calculates the total number of life-years gained over a time horizon of 40 years (to a maximum age of 100 years), discounted at 3.5% per annum (controlled in cell K1).

The PSA of this model is started by hitting the 'Run final model' button. Parameter sets in the 'Calibrated inputs' sheet are sampled, and the corresponding model outputs stored in column N of the 'Final model' sheet. The VBA macro also estimates the mean and 95% CI for the life-years gained outputs in cells P10–P12.

References

1. Department of Health and Ageing. PBAC guidelines: guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee (PBAC) [version 4.3]. Woden (ACT): PBAC, 2008 [online]. Available from URL: http://www.pbs.gov.au/html/industry/static/how_to_list_on_the_pbs/elements_of_the_listing_process/pbac_guidelines [Accessed 2010 Sep 29]
2. National Institute for Health and Clinical Excellence. Guide to the methods of technology appraisal (reference N0515). London: NICE, 2004 [online]. Available from URL: http://

www.nice.org.uk/niceMedia/pdf/TAP_Methods.pdf [Accessed 2010 Sep 29]

3. 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 (36): 1-172
4. Weinstein MC. Recent developments in decision-analytic modeling for economic evaluation. Pharmacoeconomics 2006; 24 (11): 1043-53
5. Vanni T, Karnon J, Madan J, et al. Calibrating models in economic evaluation: a seven-step approach. Pharmacoeconomics 2011; 29 (1): 35-49
6. Karnon J, Delea TE, Barghout V. Cost utility analysis of early adjuvant letrozole or anastrozole versus tamoxifen in postmenopausal women with early invasive breast cancer: the UK perspective. Eur J Health Econ 2008; 9: 171-83
7. Karnon J, McIntosh A, Dean J, et al. A prospective hazard and improvement analytic approach to predicting the effectiveness of medication error interventions. Saf Sci 2007; 45: 523-39
8. Carlton J, Karnon J, Czoski-Murray C, et al. The clinical effectiveness and cost-effectiveness of screening programmes for amblyopia and strabismus in children up to the age of 4–5 years: a systematic review and economic evaluation. Health Technol Assess 2008; 12 (25): iii, xi-194
9. Karnon J, Jones R, Czoski-Murray C, et al. Cost-utility analysis of screening high risk groups for anal cancer. J Public Health 2008 Sep; 30: 293-304
10. Karnon J, Campbell F, Czoski-Murray C. Model-based cost-effectiveness analysis of interventions aimed at preventing medication error at hospital admission (medicines reconciliation). J Eval Clin Pract 2009; 15 (2): 299-306
11. Karnon J, Czoski Murray C, Smith KJ, et al. A hybrid cohort individual sampling natural history model of age-related macular degeneration: assessing the cost-effectiveness of screening using probabilistic calibration. Med Decis Making 2009; 29: 304-16
12. Karnon J. Cost considerations and cost effectiveness of aromatase inhibitors in breast cancer. Pharmacoeconomics 2006; 24 (3): 215-32
13. Early Breast Cancer Trialists' Collaborative Group. Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. Lancet 2005; 365: 1687-717
14. The Breast International Group (BIG) 1-98 Collaborative Group. A comparison of letrozole and tamoxifen in postmenopausal women with early breast cancer. N Engl J Med 2005; 353: 2747-57
15. Howell A, Cuzick J, Baum M, et al. Results of the ATAC (Arimidex, Tamoxifen, Alone or in Combination) trial after completion of 5 years' adjuvant treatment for breast cancer. Lancet 2005; 365 (9453): 60-2
16. Jit M, Gay N, Soldan K, et al. Estimating progression rates for human papillomavirus infection from epidemiological data. Med Decis Making 2010; 30: 84-98
17. Frontline Systems, Inc. [online]. Available from URL: http://www.solver.com/ [Accessed 2010 Sep 20]
18. Kim JJ, Kuntz KM, Stout NK, et al. Multiparameter calibration of a natural history model of cervical cancer. Am J Epidemiol 2007; 166: 137-50
19. Ades AE, Welton NJ, Caldwell D, et al. Multiparameter evidence synthesis in epidemiology and medical decision-making. J Health Serv Res Policy 2008; 13 Suppl. 3: 12-22

Correspondence: Dr *Jonathan Karnon*, Professor, University of Adelaide, Department of Public Health, Level 8, 10 Pulteney Street, Mail Drop 207, Adelaide, SA 5005, Australia.
E-mail: jonathan.karnon@adelaide.edu.au