Calibrating Parameters for Microsimulation Disease Models: A Review and Comparison of Different Goodness-of-Fit Criteria

Alex van der Steen, MSc, Joost van Rosmalen, PhD, Sonja Kroep, PhD, Frank van Hees, MSc, Ewout W. Steyerberg, PhD, Harry J. de Koning, MD, PhD, Marjolein van Ballegooijen, MD, PhD, Iris Lansdorp-Vogelaar, PhD

Background. Calibration (estimation of model parameters) compares model outcomes with observed outcomes and explores possible model parameter values to identify the set of values that provides the best fit to the data. The goodness-of-fit (GOF) criterion quantifies the difference between model and observed outcomes. There is no consensus on the most appropriate GOF criterion, because a direct performance comparison of GOF criteria in model calibration is lacking. ***Methods.*** We systematically compared the performance of commonly used GOF criteria (sum of squared errors [SSE], Pearson chi-square, and a likelihood-based approach [Poisson and/or binomial deviance functions]) in the calibration of selected parameters of the MISCAN-Colon microsimulation model for colorectal cancer. The performance of each GOF criterion was assessed by comparing the 1) root mean squared prediction error (RMSPE) of the selected parameters, 2) computation time of the calibration procedure of various calibration scenarios, and 3) impact on estimated cost-effectiveness ratios. ***Results.*** The likelihood-based deviance resulted in the lowest RMSPE in 4 of 6 calibration scenarios and was close to best in the other 2. The SSE had a 25 times higher RMSPE in a scenario with considerable differences in the values of observed outcomes, whereas the Pearson chi-square had a 60 times higher RMSPE in a scenario with multiple studies measuring the same outcome. In all scenarios, the SSE required the most computation time. The likelihood-based approach estimated the cost-effectiveness ratio most accurately (up to −0.15% relative difference versus 0.44% [SSE] and 13% [Pearson chi-square]). ***Conclusions.*** The likelihood-based deviance criteria lead to accurate estimation of parameters under various circumstances. These criteria are recommended for calibration in microsimulation disease models in contrast with other commonly used criteria.
Key words: *model calibration; microsimulation modeling; cancer screening; goodness-of-fit criterion.* ***(Med Decis Making 2016;36:652–665)***

Mathematical disease models are increasingly being used for economic evaluations to inform policy makers on long-term health outcomes.[1,2] These disease models are often mathematical representations of the disease's natural history. Microsimulation disease models are one type of mathematical model[3] in which individual life histories are generated with individuals at risk for disease. At the end of the simulation, individual results are summarized to a higher level of aggregation (e.g., total number of cases in a population). The development and progression of disease in microsimulation disease models depends on many parameters, for which data may be unavailable or not directly observable (e.g., the duration of preclinical disease states). This makes parameter estimation difficult and uncertain.

The estimation of parameters is usually performed using a calibration process. Calibration explores possible values of model parameters and compares

Received 16 December 2014 from the Departments of Public Health (AvdS, SK, FvH, EWS, HJdK, MvB, IL-V) and Biostatistics (JvR), Erasmus MC, Rotterdam, The Netherlands. Revision accepted for publication 20 January 2016.

Supplementary material for this article is available on the *Medical Decision Making* Web site at http://mdm.sagepub.com/supplemental.

Address correspondence to Alex van der Steen, Department of Public Health, Erasmus MC, PO Box 2040, 3000 CA Rotterdam, The Netherlands; e-mail: alex.vandersteen@erasmusmc.nl.

DOI: 10.1177/0272989X16636851

model outcomes with observed outcomes to identify the set (or sets) of values that provides the best fit to the data.[4,5] Calibration is a challenging and time-consuming process because of various parameter interactions and the fact that model outcomes are obtained using simulation. Calibration is performed during the development phase of a model and is ideally repeated each time new data become available.[6] Although several new methods have been proposed to reduce the computation time,[7,8] calibration remains time-consuming.

Vanni and others[4] suggested a good-practice approach for model calibration consisting of 7 steps. The first 2 steps are the identification of parameters to vary in the calibration process and the selection of model outcomes for comparison with observed outcomes. The difference between model outcomes and observed data is evaluated using a goodness-of-fit (GOF) criterion (step 3). Next, a parameter search algorithm is chosen (step 4) and used in an iterative process to move from the initial set of parameters to sets of parameters for which model outcomes are closer to the observed data (to be evaluated by the GOF criterion). The acceptance of a set of parameters depends on the acceptance criteria of step 5 (e.g., search for 1 best parameter set). The end of a calibration is determined by the stopping rule (step 6), which typically is convergence of model outcomes to observed data or a specified maximum number of parameter searches. The last step involves the integration of the calibration results into the model. Although potential methods for each step were discussed, no guidance was given on preferred methods. Because of this lack of guidance, Stout and others[9] already observed that the process of model calibration is currently often an art, rather than a science.

The GOF criterion is one of the key determinants of calibration because it is the criterion that measures the difference between model outcomes and observed data. There is no consensus on the most appropriate criterion of GOF,[4,10] because a direct comparison of the performance of the most commonly applied GOF criteria in model calibration is lacking.[11] Such comparisons are important for the advancement of model calibration and for disease simulation modeling in general.[9]

In ordinary statistical models (i.e., without microsimulation), it is known that maximum likelihood estimators are asymptotically efficient.[12] This means that for large data sets, a criterion based on maximization of the likelihood yields the smallest mean square error among all possible estimators. However, when using a microsimulation model, the likelihood function can only be estimated through simulation. In a microsimulation model, one can thus only find the parameter estimates that maximize the simulated likelihood function. If the simulated likelihood function is a good approximation of the true likelihood function (i.e., if the simulated population is large enough) and the true likelihood function is sufficiently smooth (e.g., unimodal), then one would expect the parameter estimates that maximize the simulated likelihood function to be close to the parameter estimates that maximize the true likelihood function. However, it is not clear whether the aforementioned conditions are met in a practical microsimulation model and thus whether the asymptotic efficiency of maximum likelihood estimators also applies in microsimulation modeling.

In this study, we compare the efficiency of GOF criteria based on maximization of the likelihood with other commonly used GOF criteria to determine which of these criteria can best be used for calibration of microsimulation disease models.

METHODS

We compared 3 commonly used GOF criteria. First, we give an overview of the statistical properties of these GOF criteria. Then, we evaluate the performance of the GOF criteria in a practical calibration situation in which selected parameters of a microsimulation model for colorectal cancer were estimated.

GOF Criteria

We compared the performance of the sum of squared errors (SSE),[4] the Pearson chi-square,[13] and a likelihood-based approach using deviance functions because these are the most frequently used GOF criteria in cancer screening models.[9] For the likelihood-based deviances, we evaluated the Poisson and binomial deviance[14] because most data sets used in disease modeling contain observations that are assumed to be either Poisson (e.g., number of interval cancers) or binomially distributed (e.g., number of screen-detected cancers).[15]

All GOF criteria compare model outcomes with observed data according to a mathematical formula (Table 1). We define *obs* as an observed outcome (in absolute number of cases; e.g., the total number of screen-detected cancers), *sim* as the model outcome (in absolute number of cases), *n* as the number of evaluated persons in the observed outcomes relevant to *obs*, and *m* as the number of evaluated persons in

Table 1 Overview of Goodness-of-Fit Criteria

Goodness-of-Fit Criterion	Formula
Sum of squared errors	$(obs - sim)^2$
Pearson chi-square[a]	$\frac{(obs - sim)^2}{sim}$
Poisson deviance[a,b]	$2\left[obs\left(ln\left(\frac{obs}{sim}\right)\right) - (obs - sim)\right]$
Binomial deviance[a,b]	$2\left[obs\left(ln\left(\frac{obs}{n}\right) - ln\left(\frac{sim}{m}\right)\right)\right] + 2\left[(n - obs)\left(ln\left(\frac{n - obs}{n}\right) - ln\left(\frac{m - sim}{m}\right)\right)\right]$

Note: The scaling of *sim* is according to the ratio of the number of evaluated persons at risk in the observed outcomes and in the model for the SSE and Pearson chi-square criteria. The binomial deviance has this scaling embedded. For the Poisson deviance, the scaling is according to the ratio of the person-years of evaluated persons at risk in the observed outcomes and in the model. *obs* is observed outcome (in absolute number of cases), *sim* is the model outcome (in absolute number of cases), *n* is the number of evaluated persons in observed data relevant to *obs*, and *m* is the number of evaluated persons in the model relevant to *sim*.

a. Formula is for values of $sim > 0$ (in the simulation model, in case $sim = 0$, we assumed $sim = 0.5$).

b. In case $obs = 0$, then $\lim_{x \to 0} xln(x) = 0$ is assumed.

the model relevant to *sim*. The formulas for the SSE and Pearson chi-square criteria are based on squared distances between model and observed outcomes. Using the Pearson chi-square criterion, this squared difference is scaled by the model outcome. The Poisson and the binomial deviances are based on the likelihood functions of the Poisson and the binomial distribution, and the formulas of these GOF criteria follow from the theory of generalized linear models.[14] For all included criteria, lower values indicate a better fit to the data. In model calibration, multiple model and observed outcomes are usually compared simultaneously (e.g., the total number of screen-detected cancers divided over various age groups).[16] In this case, individual GOF criteria (as in Table 1) should be combined to an overall GOF. In addition, the simulation model output has to be scaled to match the sample size of the observed outcomes for comparability except for the binomial deviance, which has this scaling embedded (Table 1).

Statistical Properties of GOF Criteria

In the case of only 1 model outcome and 1 observed outcome, the SSE, the Pearson chi-square and the Poisson deviance criteria will have a minimum value at a model outcome *sim* equal to *obs* and the binomial deviance if $sim/m = obs/n$ (Table 1); therefore, all criteria will be optimal. In practice, multiple studies are often available for the same outcome (e.g., multiple autopsy studies informing adenoma prevalence). We investigated the robustness of the GOF criteria in this situation by determining the expected value of a GOF [$E(GOF)$] in case of an infinite number of observations that exactly follow a known distribution with certain parameters. For each GOF, $E(GOF)$ conditional on a model outcome was calculated as shown in Equation 1:

$$E(GOF|sim, \theta) = \sum_{obs=0}^{\infty} p(obs|\theta)^* GOF(obs|sim), \quad (1)$$

where $p(obs|\theta)$ is the probability mass function of observed outcomes and θ consists of the model parameters.

$E(GOF|sim, \theta)$ was calculated, assuming that the observations were either Poisson or binomially distributed (and using the Poisson or binomial deviance accordingly). The Poisson distribution depends on the probability distribution parameter λ, which is the expected value of the observations (e.g., expected number of interval cancers).[12] We performed the analysis with λ equal to 1, 10, 100, and 1000, and with Equation 2:

$$p(obs|\lambda) = \frac{\lambda^{obs} e^{-\lambda}}{obs!} for\ obs = 0, 1, 2, ..; \lambda > 0 \quad (2)$$

The binomial distribution depends on 2 parameters, *n* (study size) and *p* (probability of success; e.g., screen detection).[12] The expected value of the binomial distribution is equal to n^*p. We conducted the analysis for 4 combinations of *n* (1000 and 10,000) and *p* (0.1% and 25%), while using Equation 3:

$$p(obs|n, p) = \left(\frac{n!}{obs!(n - obs)!}\right) p^{obs}(1 - p)^{n - obs}$$
$$for\ obs = 0..n;\ 0 \le p \le 1 \quad (3)$$

We calculated $E(GOF|sim,\theta)$ for model outcome *sim* in the range of −5% to 5% of the expected value of the observations (λ or n^*p) with increments of 1%. The resulting GOF values were then plotted. If unbiased, the SSE, Pearson chi-square, and Poisson deviance criteria would have their minimum value when the model outcome (*sim*) equals the expected value of the observed outcomes ([E(*obs*)], and for the binomial deviance when $sim/m = E(obs)/n$. In addition, we mathematically derived the value of the model outcome *sim* for which $E(GOF|sim,\theta)$ was minimal (Appendix 1).

Performance of GOF Criteria in Simulation Practice

We evaluated the performance of the GOF criteria in a practical setting by calibrating selected parameters of the MISCAN-Colon model, a model used for evaluation of colorectal cancer screening.[17]

MISCAN-Colon model

MISCAN-Colon is a well-established microsimulation model for evaluating the effect of screening and other interventions on colorectal cancer incidence and mortality. The model has been described in detail in previous studies and a standardized model profile can be found online.[17–20] Briefly, the model simulates life histories of a large population of individuals from birth to death. In some individuals, colorectal cancer develops according to the adenocarcinoma sequence. Screening either prevents cancer through the detection and removal of adenomas, or it detects cancers early, possibly improving prognosis.

The parameters of MISCAN-Colon are calibrated using an adapted version of the Nelder-Mead simplex parameter search algorithm.[21] The calibration starts with the construction of a simplex consisting of a number of parameter sets equal to the number of parameters plus 1, based on the initial set of random starting values by varying the parameters one by one. For each of these parameter sets, the GOF value is calculated. The search algorithm is sequential with the evaluation of 1 new parameter set at a time (iteration). The rationale behind this is that the new parameter set is better (in terms of GOF value) than and therefore replaces the worst of the existing parameter sets. A new simplex is formed and the calibration continues until the improvement of the GOF value is insignificant or when 500 iterations are completed.

Observed outcomes: hypothetical data set of observations for calibration

In regular calibrations, observed outcomes from studies are used for parameter estimation. However, the underlying parameter values will then be unknown and the deviation of the estimated parameters from the underlying parameters cannot be determined. Therefore, in this analysis, we first used the MISCAN-Colon model with underlying, known parameters to generate a hypothetical data set of observations for calibration. We simulated a cohort of people born in 1955 who were screened 5 times for colorectal cancer, every 2 years starting at age 55, using the fecal immunochemical test. The simulation was performed with a sufficiently large sample size (1 billion individuals) to ensure that the hypothetical data set was consistent with the parameters of interest and that there was virtually no simulation uncertainty in the hypothetical data. The hypothetical data set was scaled by a factor 0.0001 to obtain a data set of 100,000 individuals, which is in line with the size of randomized controlled trials for colorectal cancer.[22–24]

The hypothetical data set consisted of 3 types of data: 1) the screen-detected cancers at each screening, 2) the screen-detected large adenomas at each screening, and 3) the number of interval cancers in the first 2 years after each screening. An adenoma is the premalignant precursor of colorectal cancer and its removal prevents cancer. An interval cancer is a preclinical cancer diagnosed in a period of time after a negative screening and before the next scheduled colonoscopy.

In the scaled hypothetical data set, there were 279 screen-detected cancers at the first screening, and approximately one-half of that at repeat screening (Table 2, rounded figures; unrounded figures were used in the calculation of the GOF). In the first screening, 3495 large adenomas were detected, declining to 2014 in screening number 5. In the first and second year after the first screening, the number of interval cancers totaled 22 and 32, respectively; the sum of the interval cancers in the first and second year after repeat screenings was 62 and 109, respectively.

Estimated parameters

We used the scaled hypothetical data set in several calibration scenarios to estimate the sensitivity of the screen test for detecting colorectal cancers (underlying value of 0.684), the sensitivity for detecting large adenomas (underlying value of 0.179), and the sojourn time of preclinical colorectal cancer (underlying value of 6.7 years). The sensitivity of the screen test for detecting cancers or adenomas is the probability of a positive test result in case of the presence of

Table 2 The Hypothetical Data Set Based on MISCAN-Colon Model Output With Underlying Parameters (Run Size: 1 Billion People; Sample Size: 100,000 People)

	First Screening	Repeat Screening (Screenings 2, 3, 4, and 5)			
Screen-detected cancers	279	161	135	129	126
Screen-detected large adenomas	3495	3036	2647	2311	2014
Interval cancers first year after screening	22	62[a]			
Interval cancers second year after screening	32	109[a]			

Note: Although rounded figures are presented, the scaled hypothetical data set consisted of unrounded figures.
a. For interval cancers, a summation of the repeat screenings in screenings 2, 3, 4, and 5 was used.

a cancer or adenoma. Consequently, the higher the sensitivity of the test, the higher the number of screen-detected cancers/adenomas and the lower the number of interval cancers. The sojourn time of colorectal cancer is the time period in which the cancer is possibly detectable by the screen test but symptoms are absent. The length of this time period influences the number of screen-detected cancers and the number of interval cancers.

Calibration scenarios

We conducted 6 alternative calibration scenarios. The scenarios differed with respect to the number and types of parameters estimated, relative weighting of different outcomes in the GOF calculation and the data set used for calibration (Table 3). In each calibration scenario, 1 or 2 of the 3 parameters (sensitivity for detecting cancers, sensitivity for detecting large adenomas, and the sojourn time) were estimated. We used the number of screen-detected cancers and the number of interval cancers when calibrating the sensitivity for detecting cancers and/or the sojourn time, and the number of screen-detected adenomas when calibrating the sensitivity for detecting large adenomas. Calibration scenarios 1 to 4 were performed using the scaled hypothetical data set (Tables 2 and 3). The likelihood-based deviance consisted of the binomial deviance for binomially distributed data (number of screen-detected cancers and large adenomas) and the Poisson deviance in case of Poisson data (number of interval cancers). The overall GOF was calculated as the unweighted sum of the individual GOF criteria, except in calibration scenario 4.

The 6 scenarios are as follows. In scenario 1 (2 weakly correlated parameter estimates), the sensitivity of the screen test for detecting cancers and the sensitivity of the screen test for detecting large adenomas were simultaneously estimated. These parameter estimates are expected to be at most weakly correlated. In scenario 2 (the 1-parameter estimate), only the sensitivity of the screen test for detecting cancers was estimated. For scenario 3 (2 highly correlated parameter estimates), the sensitivity of the screen test for detecting cancers and the sojourn time of cancer were simultaneously estimated. These parameter estimates are expected to be highly correlated. For scenario 4 (the weighted GOF calculation), the screen-detected cancers and the screen-detected large adenomas contributed equally (weight: 50%) to the overall GOF in this weighted calibration. For scenario 5 (considerable differences in the values of observed outcomes), we estimated the sensitivity of the screen test for detecting cancers and large adenomas, and we scaled the screen-detected and interval cancers by 0.00001 (instead of 0.0001 in the scaled hypothetical data set). In this way, we created a large difference in the values of the observed outcomes of the number of large adenomas (relatively large values) and the number of screen-detected and interval cancers (relatively small values). Finally, in scenario 6 (multiple studies for the same outcome), we created 10 different hypothetical data sets for the number of screen-detected adenomas. We sampled the 10 data sets with replacement from the hypothetical data set. Each of the 10 data sets consisted of 100 persons, which is in line with sizes in autopsy studies.[25–27] The average probability of screen-detected adenomas over the 10 data sets was made equal to the probability in the hypothetical data set to ensure that the underlying parameters remained the same. Subsequently, we estimated the sensitivity of the screen test for detecting large adenomas on these 10 distinct sets simultaneously. This scenario is similar to the situation in the statistical properties of the GOF criteria sections.

We repeated each calibration with 100 unique sets of random starting values for the parameters that were calibrated to address parameter and stochastic uncertainty. Together this resulted in 1800 calibrations ([3 GOF criteria $\times$ 6 calibration scenarios] $\times$ 100 starting values) performed with the MISCAN-Colon model. All simulations with MISCAN-Colon were

Table 3 Overview Calibration Scenarios

Scenario	Parameter 1	Parameter 2	Hypothetical Data and Sample Size Used
1	Sensitivity screen test for detecting cancers	Sensitivity screen test for detecting large adenomas	Number of screen-detected cancers and large adenomas and number of interval cancers (sample size: 100,000)
2	Sensitivity screen test for detecting cancers	n/a	Number of screen-detected cancers and number of interval cancers (sample size: 100,000)
3	Sensitivity screen test for detecting cancers	Sojourn time of cancer	Number of screen-detected cancers and number of interval cancers(sample size: 100,000)
4	Sensitivity screen test for detecting cancers	Sensitivity screen test for detecting large adenomas	Number of screen-detected cancers and large adenomas and number of interval cancers (sample size: 100,000)
5	Sensitivity screen test for detecting cancers	Sensitivity screen test for detecting large adenomas	Number of screen-detected cancers and large adenomas and number of interval cancers (sample size screen-detected large adenomas: 100,000; sample size screen-detected and interval cancers: 10,000)
6	Sensitivity screen test for detecting large adenomas	n/a	Number of screen-detected large adenomas (10 data sets with sample size 100)

n/a, not applicable.

performed with a sample size of 10 million people and identical initial random seeds. The ratio of the number of evaluated persons at risk in the model and in the observed outcomes was used for the binomial deviance and in the scaling of model to observed outcomes for the SSE and Pearson chi-square criteria. For the Poisson deviance, the scaling is performed according to the ratio of the person-years of evaluated persons at risk in the model and observed outcomes.

Outcomes

The performance of each GOF criterion in simulation practice was assessed using 1) root mean squared prediction error (RMSPE)[28,29] of the selected parameters, 2) computation time of the calibration procedure, and 3) impact on estimated cost-effectiveness ratios.

The ultimate aim of model calibration is to obtain unbiased estimates of the underlying parameter values of the decision model, so that these parameters can be used to make valid predictions of for example the cost-effectiveness of an intervention. Every GOF criterion leads to the best fit of the observed outcomes according to its own definition; therefore, comparing the values of different GOF criteria directly is not possible. For that reason, we compared the performance of different GOF criteria by using a criterion function for the parameter estimates. We used the RMSPE to evaluate the differences between the true and the estimated parameters, based on the assumption that the parameter estimates are asymptotically normally distributed. The RMSPE is independent of the applied GOF criterion and is defined as the square root of the mean squared error (MSE). The MSE equals the sum of 2 parts: 1) the variance of the estimated parameter and 2) the squared bias of the estimated parameter.[30]

To calculate the RMSPE, we evaluated the Euclidian distance (shortest distance between 2 points[31]) between estimated and underlying parameter value(s) for each model calibration. The square root of the mean of the Euclidian distances of the 100 calibrations (RMSPE) with unique starting values was used as a measure for the overall deviation between the estimated and underlying parameters. To ensure that each parameter had a similar influence on the RMSPE in calibrations with 2 parameters, the individual estimated and underlying parameter differences were converted to percentage deviations before calculating the Euclidian distance.

For each calibration, the number of iterations needed to identify the set that provided the best fit to the data was stored. The average number of iterations needed for 100 calibrations was used as an estimate for the required computation time for each GOF criterion in all calibration scenarios.

To assess the policy impact of differences in estimated parameters between the GOF criteria, we performed a cost-effectiveness analysis for all 100 calibration outcomes per GOF criterion and per scenario. Based on these parameters, the MISCAN-Colon model was used to predict the impact of 5 rounds of

biennial fecal immunochemical testing starting at age 55 for a population of 10 million people. For comparison, the model was also run with the underlying parameter values. Subsequently, we calculated the difference between the ''true'' (i.e., based on underlying parameter values) undiscounted life-years gained, total costs and cost-effectiveness of screening, and those estimated based on the different calibration outcomes. The cost-effectiveness ratio was compared by using the average difference of the 100 runs per GOF criterion and per scenario. The assumptions for the costs used in this analysis can be found in Appendix 2.

Alternative parameter search algorithm: grid search

To assess the robustness of our results to the parameter search algorithm, we repeated all calibrations using a grid search algorithm (instead of the Nelder-Mead algorithm). The model was run for a grid of possible parameter values for the 6 calibration scenarios, using a sample size of 100 million people for each point on the grid. The GOF values resulting from these runs were kept and plotted. In this way, for each scenario and criterion the parameter value(s) leading to the minimum GOF value from these runs can be visualized.

RESULTS

Statistical Properties of GOF Criteria

For Poisson distributed data, the minimum GOF value for all GOF criteria, except the Pearson chi-square, occurred at a model outcome equal to the expected value of the observed outcomes (Figure 1A). The bias using the Pearson chi-square criterion was relatively large for small parameter values (λ equal to 1 and 10) and was negligible for large parameter values of λ (λ equal to 100 and 1000). Similarly, for binomially distributed data the Pearson chi-square was the only criterion that led to biased estimated parameters, with more bias for small n^*p (Figure 1B). We derived that the minimum GOF value for the Pearson chi-square criterion occurred with a model outcome equal to $\sqrt{\lambda^2 + \lambda}$ in case of Poisson distributed data and $\sqrt{n(n-1)p^2 + np}$ in case of binomially distributed data (versus the expected value of λ and n^*p, respectively) (Appendix 1).

Performance of GOF Criteria in Simulation Practice

The RMSPE using the likelihood-based deviance was lowest in 4 of 6 calibration scenarios and was close to best in the other 2 (Figure 2 and Table 4). The mean estimated parameters for the sensitivity of the screen test for detecting cancers and for large adenomas, and for the sojourn time using the likelihood-based deviance was close to the underlying values in all calibration scenarios.

The use of the SSE criterion in scenario 5 (considerable differences in the values of observed outcomes) resulted in a significantly higher RMSPE (Table 4). The RMSPE of the SSE was equal to 0.2373, whereas the RMSPE for the other criteria did not exceed 0.01. In the case of estimating 1 parameter (scenarios 2 and 6) and with a weighted calibration (scenario 4), the SSE criterion accurately estimated the parameters.

In a situation with multiple studies for the same outcome (calibration scenario 6), the RMSPE of the Pearson chi-square criterion was over 60 times higher compared with the other criteria (0.0469 versus 0.0008 and 0.0008). This was attributable to a biased estimation of the sensitivity for large adenomas (average value of 0.2259 versus the underlying value of 0.179).

In all scenarios, the SSE criterion required the most computation time. The average number of iterations was about 3% to 9% higher for the SSE compared with the other 2 criteria, except for calibration scenario 2, where it was similar.

With the underlying model parameters, the MISCAN-Colon model estimated that 1.3 million life-years could be saved at a cost of $6.15 billion. In most calibration scenarios, the average difference between life-years gained and costs as predicted with the estimated parameters and these underlying values was less than 0.5% (Table 5). For example, in calibration scenario 1 for 2 weakly correlated parameter estimates, the difference between predicted and underlying cost-effectiveness of screening was approximately 0.01% for all 3 GOF criteria. In the scenario with considerable differences in the values of the observed outcomes (calibration scenario 5), the parameter estimates based on the SSE criterion underestimated life-years saved and costs of screening somewhat more, but the difference was still modest at 1%. However, in the scenario with multiple studies for the same outcome (calibration scenario 6), the parameter estimates based on the Pearson chi-square criterion resulted in a 5% overestimation of life-years gained with screening and

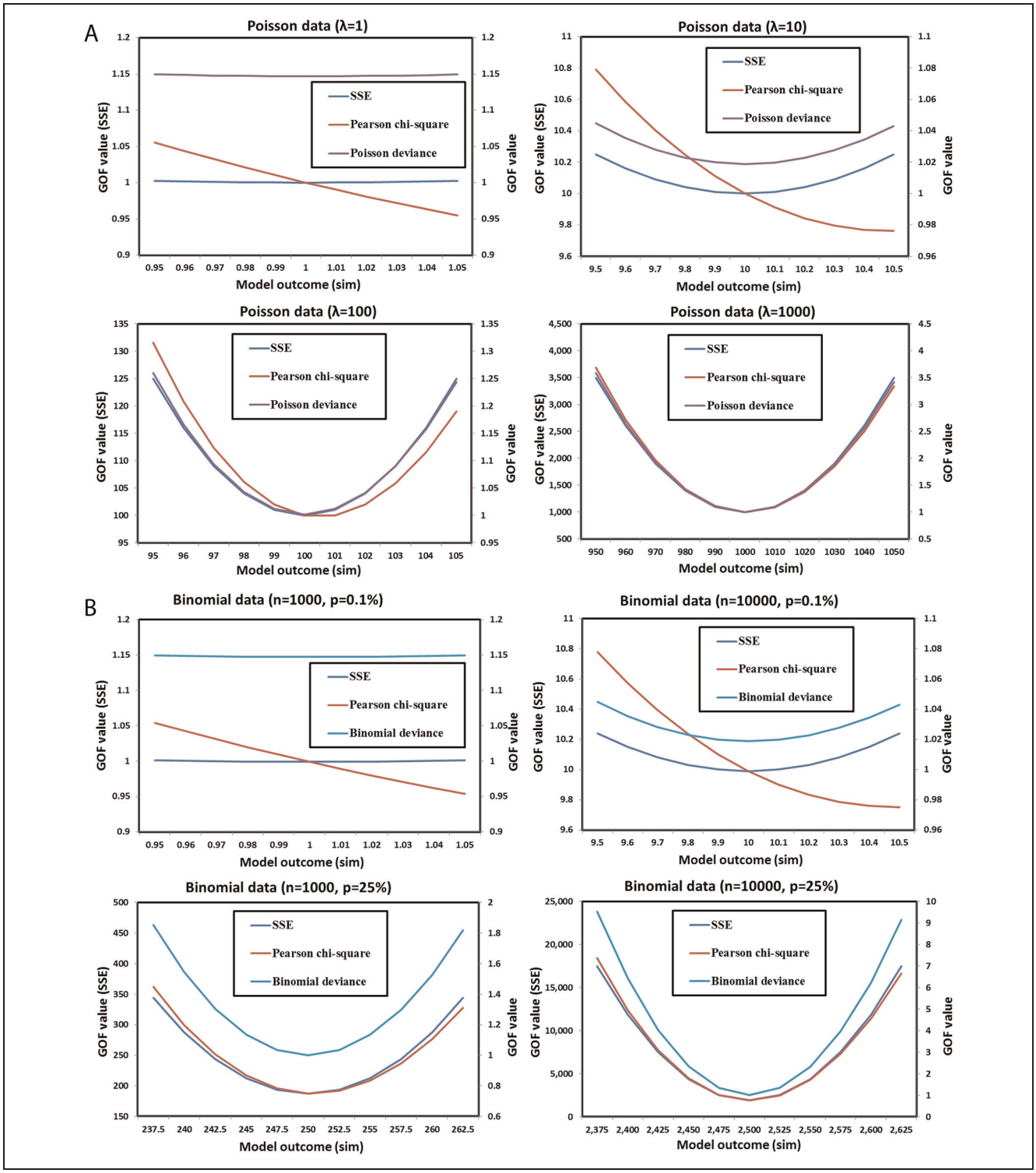

Figure 1 Overview of the statistical properties. (A) Poisson data, input parameter λ = *1, 10, 100, and 1000 (SSE, primary axis; Pearson chi-square and Poisson deviance, secondary axis). (B) Binomial data, input parameters* n = *1000 and 10,000, and* p = *0.1% and 25% (SSE, primary axis; Pearson chi-square and binomial deviance, secondary axis). GOF, goodness of fit; SSE, sum of squared errors.*

Table 4 Results of Calibration Scenarios for MISCAN-Colon Analysis

Scenario	Sensitivity[a]		RMSPE	Iterations[b]
	Cancers	Large Adenomas		
Calibration scenario 1: 2 weakly correlated parameter estimates				
Underlying value	0.684	0.179		
SSE	0.6841 (0.6710–0.7046)	0.1790 (0.1776–0.1800)	0.0110	38.90 (29–54)
Pearson chi-square	0.6841 (0.6756–0.6956)	0.1790 (0.1780–0.1803)	0.0070	36.17 (30–44)
Likelihood-based deviance[c]	0.6832 (0.6754–0.6955)	0.1790 (0.1777–0.1806)	0.0067	36.64 (31–46)
Calibration scenario 2: 1 parameter				
Underlying value	0.684			
SSE	0.6838 (0.6735–0.6930)		0.0039	16.50 (16–18)
Pearson chi-square	0.6842 (0.6767–0.6922)		0.0031	16.44 (16–18)
Likelihood-based deviance[ll]	0.6833 (0.6772–0.6943)		0.0035	16.46 (16–18)
Calibration scenario 3: 2 highly correlated parameter estimates[d]				
Underlying value	0.684	6.7		
SSE	0.6853 (0.6604–0.7102)	6.66 (6.38–6.95)	0.0246	46.60 (31–85)
Pearson chi-square	0.6830 (0.6665–0.7061)	6.70 (6.42–6.96)	0.0206	45.35 (32–88)
Likelihood-based deviance	0.6851 (0.6658–0.7043)	6.69 (6.49–6.93)	0.0183	45.28 (32–81)
Calibration scenario 4: weighted calibration				
Underlying value	0.684	0.179		
SSE	0.6850 (0.6729–0.6957)	0.1790 (0.1778–0.1803)	0.0068	37.91 (30–44)
Pearson chi-square	0.6840 (0.6765–0.6965)	0.1788 (0.1754–0.1814)	0.0085	35.89 (30–46)
Likelihood-based deviance	0.6841 (0.6750–0.6926)	0.1786 (0.1757–0.1812)	0.0081	36.16 (31–46)
Calibration scenario 5: considerable differences in the values of observed outcomes				
Underlying value	0.684	0.179		
SSE	0.6482 (0.2029–0.9468)	0.1790 (0.1779–0.1805)	0.2373	39.32 (24–55)
Pearson chi-square	0.6841 (0.6670–0.6978)	0.1790 (0.1780–0.1802)	0.0095	37.30 (29–45)
Likelihood-based deviance	0.6839 (0.6689–0.6971)	0.1790 (0.1779–0.1803)	0.0092	37.34 (30–45)
Calibration scenario 6: multiple studies for the same outcome				
Underlying value	0.179			
SSE	0.1790 (0.1773–0.1807)		0.0008	17.72 (16–24)
Pearson chi-square	0.2259 (0.2237–0.2280)		0.0469	16.30 (16–18)
Likelihood-based deviance	0.1790 (0.1772–0.1810)		0.0008	16.38 (16–18)

Note: More detailed information on the calibration scenarios can be found in the calibration scenarios section. RMSPE, root mean squared prediction error; SSE, sum of squared errors.

a. Sensitivity values are presented as averages, with minimum and maximum values of the 100 calibrations given in parentheses.

b. Iterations are presented as the average number of iterations needed, with minimum and maximum values of the 100 calibrations given in parentheses.

c. Likelihood-based deviance: 1) sensitivity of cancers: binomial deviance (screen-detected cancers) and Poisson deviance (interval cancers); 2) sensitivity of large adenomas: binomial deviance (screen-detected large adenomas); and 3) sojourn time: the Poisson deviance (interval cancers).

d. Values for sojourn time are given in lieu of sensitivity for large adenomas. For the SSE for calibration 3 with outlier: 0.6883 (0.6604–0.9841) and 6.66 (4.02–6.95); RMSPE: 0.0642.

a 9% underestimation of total costs. Therefore, the estimated cost-effectiveness of screening was 13% more favorable than the underlying cost-effectiveness.

Alternative parameter search algorithm: grid search

The grid search algorithm led to similar results as the Nelder-Mead parameter search algorithm. For calibration scenarios 1 to 4, the lowest GOF values for all criteria were found near the underlying values (Appendix 3). For calibration scenario 5, the SSE criterion had a minimum GOF value at 0.659 for the

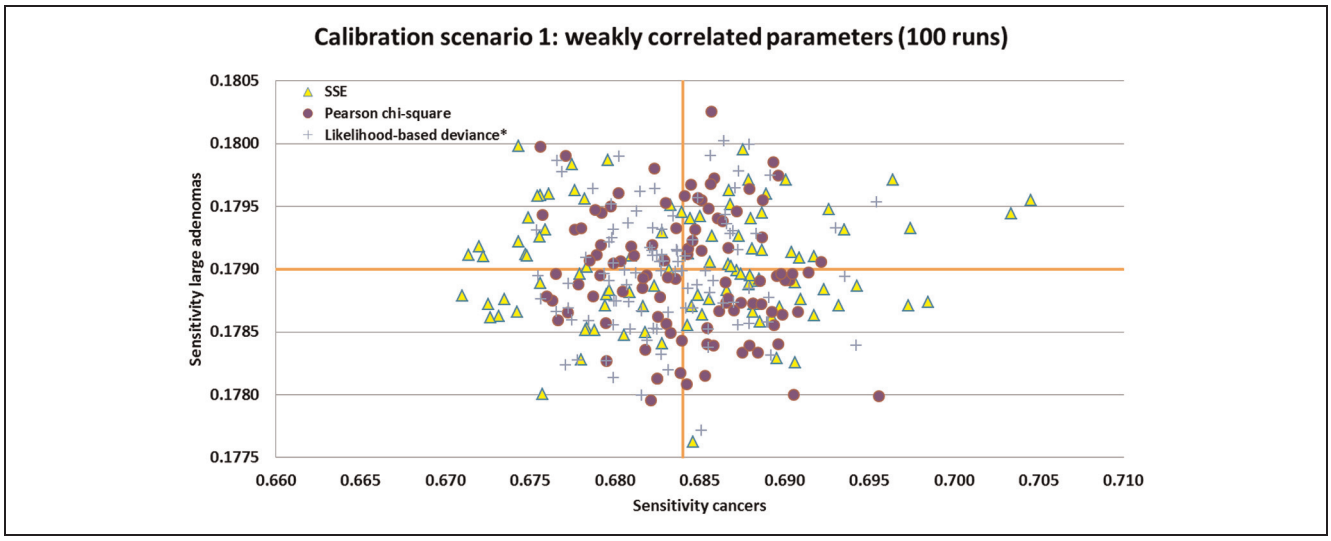

*Figure 2 Results of MISCAN-Colon analysis for calibration scenario 1 (2 weakly correlated parameter estimates). Orange lines indicate underlying values. *Likelihood-based deviance: 1) sensitivity cancers: binomial deviance (screen-detected cancers) and Poisson deviance (interval cancers); and 2) sensitivity large adenomas: binomial deviance (screen-detected large adenomas). SSE, sum of squared errors.*

sensitivity of the screen test for detecting cancers (underlying value of 0.684). The minimum GOF value when using the Pearson chi-square criterion in calibration scenario 6 occurred at a value of 0.226 for the sensitivity of the screen test for detecting large adenomas (underlying value of 0.179).

DISCUSSION

This study demonstrates that the choice of a GOF criterion in model calibration is important: there are situations in which an erroneous choice of criterion leads to improper estimation of model parameters and therefore incorrect information for policy makers. Among the most applied GOF criteria, from both a theoretical and a practical point of view, a likelihood-based approach (deviance functions such as Poisson deviance and binomial deviance) is the best in microsimulation disease models. The performance of this likelihood-based deviance was generally good in all explored calibration scenarios.

The Pearson chi-square criterion can lead to a bias in the estimated parameters. For large observed outcomes, this bias is negligible; however, for small observed outcomes, this bias can be relatively large (see Appendix 1 for derivation of the bias). It is already known that the Pearson chi-square statistic does not exactly follow the chi-square distribution and is therefore inaccurate for small numbers of observations.[12] In MISCAN-Colon calibration scenario 6, we reproduced this situation. The results from calibration scenario 6 (both when using Nelder-Mead and grid search as parameter search algorithm) clearly showed a relatively high RMSPE for the Pearson chi-square criterion due to a biased parameter estimation. Moreover, the use of this biased parameter estimation resulted in a 13% mean difference in the estimated cost-effectiveness ratio compared with the cost-effectiveness ratio based on the underlying parameters. If there is only 1 observed outcome (instead of 10 as in calibration scenario 6) or when the mean is taken over multiple observed outcomes, the Pearson chi-square criterion leads to unbiased parameter estimates.

Overall, the SSE was the worst performing criterion in simulation practice. Especially in the situation of considerable differences in the values of observed outcomes (calibration scenario 5), its RMSPE was high compared with the other criteria. The estimation of the sensitivity for detecting cancers differed substantially from the underlying value using both parameter search algorithms. Although the SSE is a very straightforward criterion, it favors larger observations when constructing an overall GOF because it is based on squared distances between model and observed outcomes for comparison. As a consequence, parameters that are very important but for which observations are rare may be estimated incorrectly. When multiple observed outcomes are included in the calibration, the relative weight is determined by the criterion. Assigning

Table 5 Results of MISCAN-Colon Cost-Effectiveness Analysis

	LY Gained ($\times$1000)[a]	Difference in LY Gained (%)[a]	Total Costs (in Billion $)[b]	Difference in Total Costs (%)[b]	CE Ratio ($)[c]	Difference in CE Ratio
Calibration scenario 1: 2 weakly correlated parameter estimates						
Underlying value	1278		6.15		4812	
SSE	1278 (1273–1285)	0.01	6.15 (6.13–6.17)	−0.01	4811	−0.01
Pearson chi-square	1278 (1275–1281)	0.00	6.15 (6.13–6.17)	0.01	4812	0.00
Likelihood-based deviance[d]	1278 (1274–1280)	−0.01	6.15 (6.14–6.16)	0.01	4812	0.01
Calibration scenario 2: one parameter						
Underlying value	1278		6.15		4812	
SSE	1278 (1275–1281)	0.00	6.15 (6.14–6.16)	0.00	4812	0.00
Pearson chi-square	1278 (1275–1280)	0.01	6.15 (6.14–6.16)	0.00	4812	−0.01
Likelihood-based deviance	1278 (1275–1281)	−0.01	6.15 (6.14–6.16)	−0.01	4812	0.01
Calibration scenario 3: 2 highly correlated parameter estimates						
Underlying value	1278		6.15		4812	
SSE	1279 (1259–1300)	0.09	6.13 (5.70–6.48)	−0.35	4791	−0.44
Pearson chi-square	1277 (1260–1298)	−0.04	6.15 (5.75–6.51)	0.11	4819	0.15
Likelihood-based deviance	1278 (1262–1293)	0.05	6.14 (5.84–6.47)	−0.10	4805	−0.15
Calibration scenario 4: weighted calibration						
Underlying value	1278		6.15		4812	
SSE	1278 (1274–1282)	0.03	6.15 (6.13–6.17)	0.01	4811	−0.01
Pearson chi-square	1277 (1272–1284)	−0.02	6.15 (6.13–6.21)	0.05	4815	0.07
Likelihood-based deviance	1277 (1271–1281)	−0.04	6.15 (6.13–6.20)	0.11	4819	0.14
Calibration scenario 5: considerable differences in the values of observed outcomes						
Underlying value	1278		6.15		4812	
SSE	1262 (1085–1344)	−1.25	6.08 (5.34–6.35)	−1.15	4817	0.11
Pearson chi-square	1278 (1273–1282)	0.00	6.15 (6.13–6.17)	0.00	4812	0.00
Likelihood-based deviance	1278 (1273–1282)	−0.01	6.15 (6.13–6.17)	0.01	4813	0.01
Calibration scenario 6: multiple studies for the same outcome						
Underlying value	1278		6.15		4812	
SSE	1278 (1275–1280)	−0.01	6.15 (6.13–6.18)	0.01	4813	0.02
Pearson chi-square	1336 (1334–1339)	4.59	5.59 (5.58–5.62)	−9.01	4186	−13.01
Likelihood-based deviance	1278 (1275–1280)	−0.01	6.15 (6.13–6.18)	0.02	4813	0.02

Note: More detailed information on the calibration scenarios can be found in the calibration scenarios section. CE, cost-effectiveness; LY, life-years; SSE, sum of squared errors.

a. LY gained (undiscounted) are presented as averages, with minimum and maximum values of the 100 calibrations given in parentheses.

b. Total costs (undiscounted) are presented as averages, with minimum and maximum values of the 100 calibrations given in parentheses.

c. CE ratio is the total costs divided by LY gained.

d. Likelihood-based deviance: 1) sensitivity of cancers: binomial deviance (screen-detected cancers) and Poisson deviance (interval cancers); 2) sensitivity of large adenomas: binomial deviance (screen- detected large adenomas); 3) sojourn time: the Poisson deviance (interval cancers).

suitably chosen weights to the SSE is a solution to improve its performance (calibration scenario 4). From a theoretic point of view, the SSE criterion should be weighted by the inverse of the variance of the outcome, in which case the resulting estimator should have the asymptotic consistency of quasi-likelihood estimation.[32] Another issue is that the SSE criterion is based on the normal distribution, whereas in calibration practice it is also applied to data that are not normally distributed.

Ideally, parameter estimation is performed with an unbiased estimator and a low variance. The RMSPE will then be relatively small. To reduce the RMSPE

in the current study, increasing the number of simulations (e.g., 200 simulations per GOF instead of the 100 used in this study) will likely not have much beneficial effect. However, a longer computation time or using a larger number of life histories (e.g., 100 million instead of the 10 million used in this study) may lower the RMSPE because the resulting estimated parameter value will be closer to the underlying value. The gain of a more precise optimization is, however, limited by the bias for the Pearson chi-square criterion in the setting of calibration scenario 6. Furthermore, in the case of estimating 2 parameters, with an unweighted SSE criterion, the parameter associated with the smallest values of observed outcomes will likely still be biased.

Our findings clearly show that performing multiple model calibrations is a good practice because a single model calibration may yield poor results. In this study, we used 100 calibration runs with different starting values to err on the side of caution. This number of runs takes a long time and thus may be deemed infeasible in model calibration in day to day practice. Performing a smaller number of runs (say 5 or 10) probably still leads to better results than only a single calibration run and should be practically feasible, and performing multiple runs is therefore recommended. Note that the performance of the model calibration may also be improved by doing fewer calibration runs with a larger sample size.

Our analysis to evaluate the performance of the GOF criteria in an application of disease modeling included the use of the MISCAN-Colon model but we consider the results to be generalizable to other models. First and foremost, we investigated many distinct scenarios and the likelihood-based deviance performed well in all those scenarios. Second, in line with the statistical properties, unbiased parameter estimates were obtained in calibrating the MISCAN-Colon model. Third, this calibration was performed with (an adapted version of) the Nelder-Mead parameter search algorithm.[21] This algorithm uses the values of the objective function only (no derivatives are required) and is invariant with respect to transformations of the objective function (e.g., optimizing the square root of the GOF criterion would yield the same parameter estimates). In addition, the results from the calibrations performed with grid search were similar to those with Nelder-Mead. Therefore, we believe that the optimization algorithm had limited influence on the GOF criteria in our analysis.[21,33]

This study concerned the situation in which one searches for one best parameter set. However, our results are generalizable to other situations such as selecting several acceptable parameter sets or constructing confidence intervals.[10,34–36] One option to identify several acceptable parameter sets is to use a Pareto frontier approach.[37] This method does not require that GOF values for different outcomes are combined into a single value and will therefore not be sensitive to the biases of the SSE and Pearson chi-square criteria documented in this article. For the calculation of confidence intervals in microsimulation, one can use Bayesian statistics[38] or a profile likelihood approach.[34,39] To ensure the correct coverage of these confidence intervals, the GOF criterion should be based on the likelihood function; there is generally no theoretical support for the calculation of confidence intervals using other GOF criteria. The ability to calculate confidence intervals is thus an additional advantage of using likelihood-based GOF criteria. Profile likelihood currently seems to be the most attractive approach for our type of microsimulation model, because this method is relatively easy to implement.

Three limitations are noteworthy. First, our selection of GOF criteria for comparison was not exhaustive. We included the most commonly used GOF criteria in cancer simulation modeling accounting for almost all of the quantitative criteria used in the identified articles.[9] Second, our analysis was performed with a maximum of 2 parameters to be estimated. However, the results were similar in the calibration scenarios with 2 weakly correlated parameters, with 1 parameter, and with 2 highly correlated parameters. Therefore, we believe that the estimation of more than 2 parameters will not change the conclusions. Third, in the simulation analysis with the MISCAN-Colon model, only fairly large numbers of cases for both observations and simulations were used. In case of rare diseases or a high degree of detail, for possible gain of information (e.g., concerning multiple age groups or stratification of cancers into histologic type, stage, age, and sex), observed and simulated cases per group are considerably small. In case of zero occurrences in the simulation model, the Pearson chi-square, Poisson deviance, and binomial deviance criteria are undefined; in case of zero observed outcomes for the Poisson deviance and binomial deviance, an additional assumption [$\lim_{x\to 0} x\ln(x) = 0$] is necessary. Moreover, a large simulated population is needed in case of small numbers to obtain an accurate GOF criterion estimation. Therefore, an area of future research should be to develop different types of criteria that are able to deal with zero occurrences; in addition, those criteria should address data and

simulation uncertainty that can be applied simultaneously. As another area of future research, we consider the situation in which there is correlation between data sources. In this article, we assumed conditional test results to be independent (given disease status) and therefore that the outcomes were independent. This assumption enabled us to combine the GOF criteria for different outcomes using a simple summation. However, if the different outcomes are not independent, a more complex combination of GOF criteria for different outcomes would be needed.

We showed that even with large model and observed outcomes the choice of a GOF criterion is important. Best-practice methods include the use of the SSE and Pearson chi-square criteria.[9,40] Our study clearly shows that in some calibrations, these best-practice methods can lead to biased parameter estimates and consequently biased estimates of the health-economic impact of disease interventions. In conclusion, we found that a criterion based on maximum likelihood (likelihood deviance consisting of Poisson and/or binomial deviance) was the best among the most often applied GOF criteria in microsimulation disease models. This criterion leads to accurate estimation of parameters under various circumstances.

REFERENCES

1. Mandelblatt J, Schechter C, Levy D, Zauber A, Chang Y, Etzioni R. Building better models: if we build them, will policy makers use them? Toward integrating modeling into health care decisions. Med Decis Making. 2012;32:656–9.

2. Wilschut JA, Hol L, Dekker E, et al. Cost-effectiveness analysis of a quantitative immunochemical test for colorectal cancer screening. Gastroenterology. 2011;141:1648.e1–55.e1.

3. Brennan A, Chick SE, Davies R. A taxonomy of model structures for economic evaluation of health technologies. Health Econ. 2006; 15:1295–310.

4. Vanni T, Karnon J, Madan J, et al. Calibrating models in economic evaluation: a seven-step approach. Pharmacoeconomics. 2011;29:35–49.

5. Weinstein MC. Recent developments in decision-analytic modelling for economic evaluation. Pharmacoeconomics. 2006;24:1043–53.

6. Goldhaber-Fiebert JD, Stout NK, Goldie SJ. Empirically evaluating decision-analytic models. Value Health. 2010;13:667–74.

7. Iskra I, Droste R. Application of non-linear automatic optimization techniques for calibration of HSPF. Water Environ Res. 2007; 79:647–59.

8. Liu Y, Ye WJ. Time-consuming numerical model calibration using genetic algorithm (GA), 1-nearest neighbor (1NN) classifier and principal component analysis (PCA). Conf Proc IEEE Eng Med Biol Soc. 2005;2:1208–11.

9. Stout NK, Knudsen AB, Kong CY, McMahon PM, Gazelle GS. Calibration methods used in cancer simulation models and suggested reporting guidelines. Pharmacoeconomics. 2009;27:533–45.

10. Karnon J, Vanni T. Calibrating models in economic evaluation: a comparison of alternative measures of goodness of fit, parameter search strategies and convergence criteria. Pharmacoeconomics. 2011;29:51–62.

11. Kong CY, McMahon PM, Gazelle GS. Calibration of disease simulation model using an engineering approach. Value Health. 2009;12:521–9.

12. Wackerly DD, Mendenhall W, Scheaffer RL. Mathematical Statistics With Applications, 5th ed. Pacific Grove (CA): Duxbury Press; 1996.

13. Waller LA, Smith D, Childs JE, et al. Monte Carlo assessments of goodness-of-fit for ecological simulation models. Ecol Model. 2003;164:49–63.

14. McCullagh P, Nelder M. Generalized Linear Models, 2nd ed. London: Chapman and Hall; 1989.

15. Flanders WD, Kleinbaum DG. Basic models for disease occurrence in epidemiology. Int J Epidemiol. 1995;24:1–7.

16. Freitas AA. A critical review of multi-objective optimization in data mining: a position paper. ACM SIGKDD Explorations Newsl. 2004;6:77–86.

17. Loeve F, Boer R, van Oortmarssen GJ, van Ballegooijen M, Habbema JD. The MISCAN-COLON simulation model for the evaluation of colorectal cancer screening. Comput Biomed Res. 1999; 32:13–33.

18. Lansdorp-Vogelaar I, Kuntz KM, Knudsen AB, Wilschut JA, Zauber AG, van Ballegooijen M. Stool DNA testing to screen for colorectal cancer in the Medicare population: a cost-effectiveness analysis. Ann Intern Med. 2010;153:368–77.

19. Knudsen AB, Lansdorp-Vogelaar I, Rutter CM, et al. Cost-effectiveness of computed tomographic colonography screening for colorectal cancer in the medicare population. J Natl Cancer Inst. 2010;102:1238–52.

20. van Hees F, Habbema JD, Meester RG, Lansdorp-Vogelaar I, van Ballegooijen M, Zauber AG. Should colorectal cancer screening be considered in elderly persons without previous screening? A cost-effectiveness analysis. Ann Intern Med. 2014;160:750–9.

21. Neddermeijer HG, van Oortmarssen G, Piersma N, Dekker R, Habbema JD. Adaptive extensions of the Nelder and Mead Simplex Method for optimization of stochastic simulation models. Erasmus University Rotterdam, Erasmus School of Economics (ESE), Econometric Institute; 2000.

22. Mandel JS, Bond JH, Church TR, et al. Reducing mortality from colorectal cancer by screening for fecal occult blood. Minnesota Colon Cancer Control Study. N Engl J Med. 1993;328:1365–71.

23. Scholefield JH, Moss SM, Mangham CM, Whynes DK, Hardcastle JD. Nottingham trial of faecal occult blood testing for colorectal cancer: a 20-year follow-up. Gut. 2012;61:1036–40.

24. Kronborg O, Fenger C, Olsen J, Bech K, Sondergaard O. Repeated screening for colorectal cancer with fecal occult blood test. A prospective randomized study at Funen, Denmark. Scand J Gastroenterol. 1989;24:599–606.

25. Vatn MH, Stalsberg H. The prevalence of polyps of the large intestine in Oslo: an autopsy study. Cancer. 1982;49:819–25.

26. Clark JC, Collan Y, Eide TJ, et al. Prevalence of polyps in an autopsy series from areas with varying incidence of large-bowel cancer. Int J Cancer. 1985;36:179–86.

27. Arminski TC, McLean DW. Incidence and distribution of adenomatous polyps of the colon and rectum based on 1,000 autopsy examinations. Dis Colon Rectum. 1964;7:249–61.

28. Picard RR, Cook RD. Cross-validation of regression-models. J Am Stat Assoc. 1984;79:575–83.

29. Mayer DG, Butler DG. Statistical validation. Ecol Model. 1993; 68:21–32.

30. Heij C, de Boer P, Franses P, Kloek T, van Dijk H. Econometric Methods With Applications in Business and Economics. Oxford (UK): Oxford University Press; 2004.

31. Lee SH, Lim JS, Kim JK, Yang J, Lee Y. Classification of normal and epileptic seizure EEG signals using wavelet transform, phase-space reconstruction, and Euclidean distance. Comput Methods Programs Biomed. 2014;116:10–25.

32. Wedderburn RWM. Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. Biometrika. 1974;61:439–47.

33. Barton RR, Ivey JS. Modifications of the Nelder-Mead simplex-method for stochastic simulation response optimization. In: 1991 Winter Simulation Conference Proceedings. IEEE Press, Piscataway, NJ. 1991;945–953.

34. Draisma G, Boer R, Otto SJ, et al. Lead times and overdetection due to prostate-specific antigen screening: estimates from the European Randomized Study of Screening for Prostate Cancer. J Natl Cancer Inst. 2003;95:868–78.

35. Salomon JA, Weinstein MC, Hammitt JK, Goldie SJ. Cost-effectiveness of treatment for chronic hepatitis C infection in an evolving patient population. JAMA. 2003;290:228–37.

36. Goldhaber-Fiebert JD, Stout NK, Salomon JA, Kuntz KM, Goldie SJ. Cost-effectiveness of cervical cancer screening with human papillomavirus DNA testing and HPV-16,18 vaccination. J Natl Cancer Inst. 2008;100:308–20.

37. Enns EA, Cipriano LE, Simons CT, Kong CY. Identifying best-fitting inputs in health-economic model calibration: a Pareto frontier approach. Med Decis Making. 2015;35:170–82.

38. Rutter CM, Miglioretti DL, Savarino JE. Bayesian calibration of microsimulation models. J Am Stat Assoc. 2009;104:1338–50.

39. Raue A, Kreutz C, Maiwald T, et al. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics. 2009;25: 1923–9.

40. Briggs AH, Weinstein MC, Fenwick EA, et al. Model parameter estimation and uncertainty analysis: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group-6. Med Decis Making. 2012;32:722–32.