Validation of Microsimulation Models Used for Population Health Policy

Fernando Alarid-Escudero, Roman Gulati, Carolyn M. Rutter

1. Overview

This chapter discusses validation of simulation models used to inform health policy. We focus on microsimulation models that simulate the life histories of individual agents in discrete time, though many considerations discussed below apply equally to systems dynamics, agent-based, and discrete-event models. Here the term "validation" means determining whether a model is sufficiently credible, accurate, and reliable to be used for its intended applications.[1-3] In short, can we trust the model?

There is no definitive one-size-fits-all criterion to decide whether a model is validated, but confidence in a model's validity can be weaker or stronger depending on several factors. These factors include verifying whether model specifications were implemented correctly, evaluating the extent to which model-predicted results are consistent with empirical results, and examining whether model predictions are robust to alternative structural assumptions. Transparent reporting of model implementation and a greater number (and higher quality) of comparison targets increases confidence, while an opaque summary or inadequate inspection of sensitivity to a model's structural assumptions decreases confidence. In general, greater scrutiny is necessary when empirical results are not or cannot be observed and when applications involve extrapolation to new settings or over longer time horizons.

A common consideration underlying the assessment of factors that determine model validity is: *Does it make sense?* Is the model structure reasonable given our understanding of the disease or the effects of an intervention? Are the empirical data used to inform the model representative of the intended population or policy setting? Are model predictions consistent with clinical expectations, including plausible magnitudes of an intervention's harms and benefits? Ultimately, having confidence in a model's validity requires acceptance of data inputs and technical implementation, rigorous interrogation of structural features, demonstrated fidelity to empirically observed outcomes, and a general coherence of the model components to give scientifically grounded results.

This chapter examines common factors and considerations that can collectively be used to gauge the extent to which a model is validated for a given application. It reviews types of validation, discusses related concepts, takes a deeper dive into cancer model validation studies, and concludes with questions that consumers of models should ask (and modelers should answer) to inform judgment about a model's fitness for purpose. Final judgments about when model results can be trusted ultimately rely on the evolving understanding of the disease and intervention effects, available data relevant to the application, and access to reporting of model validation exercises.

2. Types of validation

2.1 Face Validity

Face validity refers to whether the model and model predictions "make sense." This type of validity should be considered at every stage of model development and use, including conceptualization of the model, internal design and implementation, selection of empirical data used to estimate model parameters via model calibration, calibration algorithm, evaluation of nearness of predictions to their targets, and conclusions about the degree of success of a validation for a given application. A model has face validity when it passes the proverbial "smell test" at each stage.

2.2 Code Validity

Code validity refers to evaluation and error-checking of model code. Code validity can be facilitated by following best practices of software design, including version control systems to manage and archive the iterative development of source code and unit testing to identify and correct unexpected effects of changes to code. While many modelers who perform validations are not software engineers by training, best practices of programming should be learned and applied. Publicly available and open source code, and code that has been made available during peer review in scientific journals, generally lends greater confidence to a model's code validity. This type of validation is often referred to as model verification.[4-6]

2.3 Internal Validity

Evaluating internal validity is part of the model calibration process. Model calibration is the process of estimating the parameters of a microsimulation model by identifying parameter values that result in model predictions that are sufficiently near observed clinical or epidemiological data (often called calibration targets) according to a formal distance measure. Internal validation can be thought of as the model's goodness-of-fit to calibration targets because it quantifies the performance of model predictions, using parameter values selected via calibration, relative to the calibration targets.

2.3.1 Calibration approaches

There are many approaches to model calibration. Some approaches focus on identification of a single best parameter vector (i.e., a single vector of values for unknown parameters). These methods often use hill-climbing approaches, such as the Nelder-Mead algorithm for deterministic models[7] and simulated annealing or genetic algorithms for stochastic models. A set of parameter vectors may be identified when the calibration problem has more than one solution.[8] A best set of parameter vectors can also be chosen based on a χ^2 goodness-of-fit test, where parameter vectors are selected if they are not statistically different from the calibration targets.[9] The χ^2 statistic is the squared difference between model-predicted targets and calibration targets, divided by either the model-predicted or calibration targets with degrees of freedom based on the number of independent observations associated with each target.[10] The χ^2 goodness-of-fit test can only be used when the degrees of freedom exceed the number of calibrated parameters.

Bayesian calibration approaches focus on synthesizing prior information about model parameters, including expert opinion, and empirical evidence from calibration targets.[11-17] Bayesian approaches specify prior distributions for calibrated model parameters to characterize prior knowledge and distributions for observed calibration targets to characterize the uncertainty in these targets. Bayesian calibration yields a set of simulated draws from the posterior distribution of model parameters. The posterior distribution represents the joint distribution of the model parameters given prior information, calibration targets, and the assumed distribution of calibration targets. Bayesian approaches result in an estimated posterior distribution for model parameters. When the posterior distribution is approximately symmetric and unimodal, the estimated posterior mean vector, given by the average across simulated posterior draws, can provide a good single parameter vector. Bayesian model parameter estimation is more complicated when the posterior distribution is asymmetric or multimodal.

When calibration yields a single best parameter vector, internal validation can be based on the distance between the targets and corresponding predicted values. This may be summarized by the deviance statistic, which is defined as two times the negative log-likelihood,[19] or by the χ^2 statistic. Models that more closely fit the targets have smaller distance statistics. It is less clear how to internally validate models when calibration results in a set of parameter vectors. One approach is to summarize goodness of fit across the set based on the average and variation of goodness-of-fit statistics. Similar issues arise when Bayesian calibration approaches are used, but because the set of simulated draws represents a sample from the posterior distribution, it can be used to estimate posterior predicted distributions of calibration targets. While overall comparisons may focus on average predicted targets, measures of variation (e.g., standard deviations, interquartile ranges) and information about the distribution of predictions (e.g., using histograms) provide additional information about model behavior and fit. A common practice uses graphical comparisons of model predicted and observed calibration targets (e.g., see Rutter and Savarino, 2010).[20]

2.3.2 Nonidentifiability

Nonidentifiability is an undesirable property of some statistical models and can be a common problem for complex models. Identifiability is a statistical concept that refers to the unique mapping of model parameters to model-predicted targets. When a model is nonidentifiable, multiple unique parameter vectors result in the same model predictions, so a single best parameter vector cannot be identified.[21] When multiple unique parameter vectors result in *similar* model predictions, the model is weakly identifiable. Because microsimulation model predictions are simulated and include stochastic variation, it can be difficult to distinguish between nonidentifiable and weakly identifiable models. Models may be nonidentifiable for one set of calibration targets and identifiable or weakly identifiable for a second set of targets. In this case, the model is over-specified relative to the first set of calibration targets (i.e., there are too many calibrated parameters) and can be resolved using the second set.

Nonidentifiability can also result from functional relationships among parameters, and in this case no amount of data will resolve the identifiability problem. In this case, nonidentifiability

can also be resolved by simplifying the model by fixing or removing calibrated parameters. The presence of nonidentifiability poses a problem because different distinct parameter vectors can produce the same predictions but may imply different conclusions.[21] For example, a model that simulates colorectal cancer may calibrate well to data on both prevalence of adenomas (precursor lesions) and colorectal cancer incidence with well-identified parameters describing onset of preclinical disease. However, if the model fits equally well with both short and long sojourn times (also called preclinical screen-detectable periods), this indicates that sojourn time is not identifiable from these data. In this example, sojourn times may be identified through the addition of calibration targets, such as information about which cancers were detected by screening. However, if identifiability is not resolved, the parameter vector that produces longer sojourn times will predict less benefit for more frequent screening tests. Sensitivity analysis is important for parameters that are structurally important but are nonidentifiable.

2.4 External Validity

External validity refers to the comparison of model predictions with validation targets not used for calibration, without "tuning" the model's calibrated parameters. When carrying out external validation, other model inputs may be deterministically changed to match predicted targets, e.g., by using demographic or risk factor distributions that are consistent with the target population. External validation provides stronger evidence of how well the model would perform when applied to policy questions *because* the model has not been tuned to predict the validation targets. External validation addresses the potential for overfitting, which can be a problem with highly parameterized models. A complex model with many calibrated parameters can be calibrated to closely reproduce targets but may not perform well when externally validated or when used to evaluate policies or interventions that deviate from those used in the calibration setting.

Examples of external validation include simulating previously conducted trials or simulating cohorts or populations with known receipt of interventions. When model calibration results in a single best parameter vector, the model can be used to simulate a trial or population using a very large sample to drive down stochastic variation. Simulating many trials of the observed sample size can be used to determine if the model reproduces the sampling variability in the validation study. In the context of Bayesian calibration, validation should reflect the way the model is used. If the model is implemented using either the estimated posterior mean parameter vector (calculated by averaging across all the simulated draws from the posterior distribution) then validation can proceed using this single parameter. However, if the model is implemented using a set of simulated draws from the posterior distribution, then validation should replicate this process, with validation targets simulated for each draw and the model prediction based on the estimated posterior mean predicted value, given by the average across these model predictions. Because microsimulation models are highly nonlinear, these two approaches may result in different predictions.

2.5 Predictive Validity

Predictive validity is a type of external validation in which the model predictions are generated and archived before the target results are revealed. Because the true values are not known,

predictive validity tests the model in a context similar to the settings which models are developed and used. For this reason, predictive validity provides the strongest evidence of model reliability.[22]

Predictive validations are uncommon. Once a high-quality study demonstrates that an intervention is effective, modeling may be considered to refine and personalize the intervention,[23] and in this case, the model development typically incorporates all relevant data including data from the original study. Consequently, in addition to the general challenges of external validation, predictive validations are further restricted by opportunity.

2.6 Comparative Validity

Comparative validity, or collaborative modeling, refers to comparison of model predictions across multiple models to targets that are either observed (such as disease incidence) or unobservable (such as sojourn time).[24,25] Fair comparison requires that the compared targets should be used in the same way across all models (i.e., as calibration targets or external validation targets). Comparative validity can provide insight into similarities and differences across models, although, unless models are compared to empirical validation targets, comparative validity cannot determine model veracity. Even models that perform differently from each other, or with varying success in terms of matching calibration targets can be valuable for some applications and may provide insights that are not otherwise available.[26]

3. Validation targets

As noted above, calibration targets are used for estimating the model parameters and for internal validation. Targets not used for calibration could be used for external validation. Targets are generally summary statistics that are selected to represent features of processes described by the model that the model should be able to reproduce. For example, in models describing the natural history of colorectal cancer, calibration targets can include statistics describing the prevalence of adenomas, the size of detected adenomas, and incidence and mortality rates from cancer registries before and after the introduction of screening. Targets can depend on characteristics of the entire population or of particular subgroups captured by the model, such as age, gender, and race/ethnicity. Expert opinion can be used to identify high-quality targets.

Several factors need to be considered when selecting targets. One issue is whether the targets may be subject to biases, such as population bias or transferability bias.[27] Both biases arise when the sample used to derive the targets does not represent the target population that would be subject to policy recommendations. Population bias may occur when the sample is not a representative subset of the target population; for example, due to differences in age, sex, or health status. In some cases, it is possible to address population bias when simulating targets by using the model to simulate a similar population, e.g., by matching the age and sex distribution of the sample. In other cases, this may be more difficult to address; for example, if individuals who participate in clinical trials that produce targets are systematically different from the target population and these differences cannot be easily reproduced by the model.

Transferability bias may occur when information is taken from a sample of a population that is different from the population subject to policy recommendations; for example, using data on cervical cancer incidence from a country where there is no screening and vaccination and using it as a calibration target for a model that will be used to inform policy in a country where there is screening in place. In summary, if there is a discrepancy between the sample informing the targets and the population of interest, the targets used for either calibration or validation may be biased. If there is information about how two populations differ, targets could be bias-corrected with quantitative bias analysis techniques.[28] In addition, care must be taken when using observational studies to inform microsimulation models. Observational studies may be subject to biases that result from confounding while microsimulation models specify causal pathways. Because it is generally difficult to simulate the mechanisms that cause bias in observational studies, the most direct approach to address this issue is to use statistical approaches to estimate causal effects from observational data.[29,30]

Because calibration involves simulation of targets for a potentially large number of parameter vectors, this condition requires the ability to efficiently simulate the data generating processes within the model. Because of this, studies with complex designs or that require simulation of bias corrections may be best suited for model validation, which requires simulation of targets for a single parameter vector or a limited set of parameter vectors.

4. Extended examples

4.1 Internal validation example

In a recent study, a simulation model of type-specific high-risk human papillomavirus (HPV)-induced cervical carcinogenesis was used to compare the cost-effectiveness of several cervical cancer screening strategies recommended in the United States (U.S.) after incorporating women's preferences throughout the screening process. The model of the natural history of cervical cancer was calibrated using a Bayesian approach to type-specific HPV prevalence, cervical cancer incidence, and proportion of cancers attributable to high-risk types, all stratified by age.

The model was internally validated by simulating all calibration targets at each parameter vector drawn from the posterior distribution of the calibrated parameters and by comparing both the posterior model-predicted means and uncertainty intervals to the calibration targets. The model-predicted targets matched the calibration targets closely, especially HPV prevalence and cervical cancer incidence, and provided a good insight into the reasonable fit to the proportion of different HPV types by age.[31]

4.2 External validation example 1

In 2016, the U.S. Preventive Services Task Force recommended biennial mammography screening for breast cancer for women ages 50-74 and against regular screening for women ages 40-49.[32] The recommendation was based in part on projections from six cancer models

that found screening women ages 40-49 would modestly reduce mortality but substantially increase false positive tests.[33] The recommendation was controversial, with fears about rationing healthcare stoked in national media and an unprecedented intervention by the U.S. Congress to mandate that health insurance covers annual screening beginning at age 40.[34] While the scientific validity of the model projections was not the primary concern, the model-predicted results for younger women were externally validated only recently.[35]

In their recent validation study, five of the six cancer models predicted breast cancer incidence and mortality in the United Kingdom (U.K.) Age trial, which screened women ages 40-49 annually.[35] The models replicated the participant demographics and trial protocol but did not alter their model parameters, which were estimated via a synthesis of other data sources. As shown in Figure 1(a), all five models predicted 17-year mortality rate ratios comparing the intervention and control groups, and all predictions were within the 95% confidence limits of the observed results. Based on these graphical and numerical assessments, the authors concluded that the models successfully recapitulated the benefit of screening in younger women, and hence the models could be trusted to inform screening guidelines for these ages.

Although this validation focused on comparing each model and the model average to external targets, the consistency across models also illustrates the models' comparative validity concerning the effectiveness of mammography screening in younger women.

4.3 External validation example 2

Three models for colorectal cancer have been used to inform screening policy for the U.S. Preventive Services Task Force and for Centers for Medicare & Medicare Services. All three models include the development of adenomas, progression from adenoma to preclinical and clinical states, and risk of cancer-specific death. However, differences in data sources and low-level implementation details led to variable predictions about distributions of relative time spent in adenoma and preclinical states—differences with potentially clinically important implications concerning the benefit of more frequent screening in the population.

In a recent study, external validity of the three models was evaluated by assessing the performance of predicted effects of a one-time flexible sigmoidoscopy screening on 10-year colorectal incidence and mortality as studied in the U.K. Flexible Sigmoidoscopy Screening (UKFSS) trial.[22] In addition to these primary endpoints, secondary endpoints, including cancer stage at diagnosis, location within the colon, and the breakdown of screen versus interval detections, were examined. All three models predicted effects of the intervention that were within the trial's 95% confidence interval for effects on mortality, and, as shown in Figure 1(b), two of three models were within the 95% confidence interval for effects on incidence. The comparisons with the primary and secondary endpoints provided valuable insight into plausible distributions of the relative time spent in the adenoma and preclinical states. By helping to identify these phases of disease natural history, this validation not only supported previous predictions by the three models concerning colorectal screening recommendations but also directly indicated areas for improving the models' structural designs.

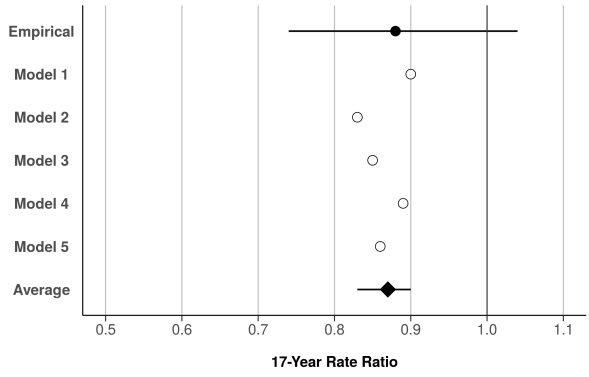
In contrast with the breast cancer validation, this study reported within-model uncertainty for the validation predictions. It focused on validation of the individual models rather than their average, though the results supported the value of model averaging. It is worth noting, however, that while "ensemble validation" may appear more favorable, it may obfuscate conclusions when there are outlier models, when comparison emphasizes point estimates but uncertainty is substantial, and when weighted averages of the ensemble do not adequately reflect variable uncertainty in the models' data sources, calibration procedures, or model parameters and structures. Little has been published about ensemble validation and comparative validity when external data are available. Based on our experience, we recommend reporting within-model uncertainty and undertaking individual model validations whenever possible.

4.4 Predictive validity example

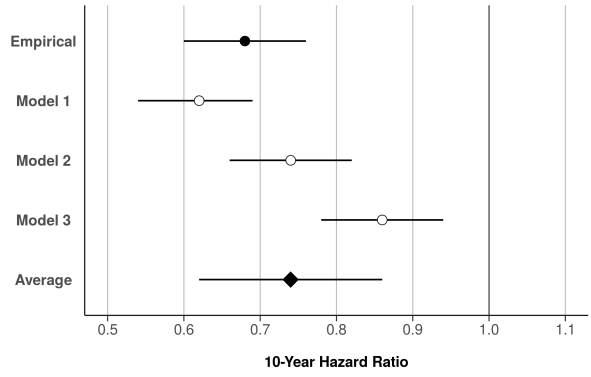
Before the publication of randomized trials of prostate-specific antigen screening for prostate cancer, a modeling study explored whether annual screening might substantially reduce prostate cancer mortality relative to biennial screening.[36] Under a working hypothesis for how early detection confers benefit, the authors concluded that the long natural history of prostate cancer implies that biennial screening is likely to achieve much of the benefit of annual screening. A decade later, results from the Prostate, Lung, Colorectal, and Ovarian cancer screening trial reported no significant difference in prostate cancer mortality among men randomized to the intervention group (and received an average of 5.0 tests over six years) compared to men randomized to usual care (and received an average of 2.7 tests over six years).[37,38] Despite coarse approximations to population demographics and intervention details, the assumed mechanism of screening benefit was concordant with the mortality results, and the general conclusion predicted by the modeling study appears to be borne out by the trial.

Figure 1. External and comparative validations of cancer interventions in microsimulation models of two randomized clinical trials.

(a) Effect of mammography screening on breast cancer mortality in the UK Age trial



(b) Effect of flexible sigmoidoscopy screening on colorectal cancer incidence in the UKFSS trial



5. Key takeaways

The overall validity of model predictions for a particular application can be assessed by examining several types of validity. For some applications, rigorous assessment of model validity may not be necessary; for example, an initial exploration of the budget impact of a range of public health interventions may require only back-of-the-envelope calculations for ballpark estimates This chapter focused on applications that require more detailed modeling and rigorous model assessment; for example, clinical recommendations of an intervention with a modest net benefit that affects millions of individuals requires high confidence in the accuracy of predictions.

Trust in results of a microsimulation model begins with face validity. Face validity of the model design, implementation, and correspondence to outside estimates is essential to the final assessment of model validity. If the model does not comport with the known reality of the disease or the effects of interventions, or the results don't make sense, the model cannot be trusted for that application.

Trust in a model is further built through intensive scrutiny of model performance in settings similar to the intended application. Careful implementation—supported by best practices in programming, rigorously tested code, sensitivity analyses, and documented internal checks—is necessary but not sufficient to assess model validity for a given application. When possible, external validations that are performed using high-quality data are invaluable to demonstrating that the model adheres to empirical evidence. When external validation is not possible, such as when the outcome of interest (e.g., overdiagnosis) is not observable, comparative validation can provide insight into robustness of predicted outcomes to a range of model assumptions.

While there is no universal prescription for determining when a model has been validated for an application, it is possible to list common questions that consumers of models should ask—and that modelers should answer. These questions are given in Table 1 with sample answers using the external validation examples described in this chapter. At a minimum, determining answers to these questions can help consumers to identify the quality and scope of the validation. In practice, answering these questions may also help consumers and modelers alike to identify gaps in the validation that were not addressed or discrepancies between the validation setting and the target application setting.

Table 1. Key questions that should be asked consumers of models and answered by modelers to assess validity of a microsimulation model for a particular application.

Question	Sample answers using example 1	Sample answers using example 2
What is the relevant application?	Effect of mammography screening of women ages 40-49 on long-term breast cancer mortality	Effect of flexible sigmoidoscopy on 10-year colorectal cancer incidence and mortality

What type of validation was done?	External and comparative	External and comparative
How reliable is the validation target?	High-quality randomized trial for appropriate ages and follow-up with well-documented protocol and adherence	High-quality randomized trial with well-documented protocol and adherence
How was the validation evaluated?	Numerical and graphical comparison of point estimates from each model (and their average) to empirical point estimate; of the point estimate from each model to the empirical interval estimate	Numerical and graphical comparison of point estimates from each model to empirical interval estimate; of rates in intervention and control arms; of disease stage, colon location, and mode of detection
How close was the evaluation?	Each model point estimate was within the 95% confidence interval of the empirical point estimate after 17 years	Each model point estimate was within the empirical 95% confidence interval for mortality after 10 years; 2 of 3 model point estimates were within the empirical 95% confidence interval for incidence after 10 years
What is the scope of the validation?	Validation does not necessarily extend to other patient age groups, screening modalities, clinical care settings, and follow-up horizons.	Validation does not necessarily extend to other patient populations, screening modalities, clinical care settings, and follow-up horizons.

References

1. Kopec JA, Fines P, Manuel DG, et al. Validation of population-based disease simulation models: a review of concepts and methods. *BMC Public Health.* 2010;10:710.
2. Eddy DM, Hollingworth W, Caro JJ, et al. Model Transparency and Validation: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research.* 2012;15(6):843-850.
3. Vemer P, van Voom GA, Ramos IC, Krabbe PF, Al MJ, Feenstra TL. Improving model validation in health technology assessment: comments on guidelines of the ISPOR-SMDM modeling good research practices task force. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research.* 2013;16(6):1106-1107.
4. Oberkampf WL, Trucano TG, Hirsch C. Verification, validation, and predictive capability in computational engineering and physics. *Applied Mechanics Reviews.* 2004;57(5):345-384.
5. Oberkampf WL, Barone MF. Measures of agreement between computation and experiment: validation metrics. *Journal of Computational Physics.* 2006;217(1):5-36.
6. Frisch M. Calibration, Validation, and Confirmation. In: *Computer Simulation Validation.* Springer; 2019:981-1004.
7. Nelder J, Mead R. A simplex method for function minimization. *Computer Journal.* 1965;7:308-313.
8. Kong CY, McMahon PM, Gazelle GS. Calibration of disease simulation model using an engineering approach. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research.* 2009;12(4):521-529.
9. Kim JJ, Kuntz KM, Stout NK, et al. Multiparameter calibration of a natural history model of cervical cancer. *Am J Epidemiol.* 2007;166(2):137-150.
10. DeGroot MH, Schervish MJ. *Probability and statistics.* Pearson Education; 2012.
11. Rutter CM, Miglioretti DL, Savarino JE. Bayesian Calibration of Microsimulation Models. *Journal of the American Statistical Association.* 2009;104(488):1338-1350.
12. Menzies NA, Soeteman DI, Pandya A, Kim JJ. Bayesian methods for calibrating health policy models: a tutorial. *Pharmacoeconomics.* 2017;35(6):613-624.
13. Whyte S, Walsh C, Chilcott J. Bayesian calibration of a natural history model with application to a population model for colorectal cancer. *Medical decision making.* 2011;31(4):625-641.
14. Hawkins-Daarud A, Prudhomme S, van der Zee KG, Oden JT. Bayesian calibration, validation, and uncertainty quantification of diffuse interface models of tumor growth. *Journal of mathematical biology.* 2013;67(6-7):1457-1485.
15. Jackson CH, Jit M, Sharples LD, De Angelis D. Calibration of complex models through Bayesian evidence synthesis: a demonstration and tutorial. *Medical decision making.* 2015;35(2):148-161.
16. Welton NJ, Ades A. Estimation of Markov chain transition probabilities and rates from fully and partially observed data: uncertainty propagation, evidence synthesis, and model calibration. *Medical Decision Making.* 2005;25(6):633-645.

17. Rutter C, Ozik J, DeYoreo M, Collier N. Microsimulation model calibration using incremental mixture approximate Bayesian computation. *Annals of Applied Statistics.* 2018;in press.
18. Bernardo JM, Smith AF. *Bayesian theory.* Vol 405: John Wiley & Sons; 2009.
19. Czado C, Gneiting T, Held L. Predictive model assessment for count data. *Biometrics.* 2009;65(4):1254-1261.
20. Rutter CM, Savarino JE. An evidence-based microsimulation model for colorectal cancer: validation and application. *Cancer Epidemiology and Prevention Biomarkers.* 2010;19(8):1992-2002.
21. Alarid-Escudero F, MacLehose RF, Peralta Y, Kuntz KM, Enns EA. Nonidentifiability in Model Calibration and Implications for Medical Decision Making. *Med Decis Making.* 2018;38(7):810-821.
22. Rutter CM, Knudsen AB, Marsh TL, et al. Validation of Models Used to Inform Colorectal Cancer Screening Guidelines: Accuracy and Implications. *Med Decis Making.* 2016;36(5):604-614.
23. Mant D. Can randomised trials inform clinical decisions about individual patients? *Lancet.* 1999;353(9154):743-746.
24. van Ballegooijen M, Rutter CM, Knudsen AB, et al. Clarifying differences in natural history between models of screening: the case of colorectal cancer. *Medical Decision Making.* 2011;31(4):540-549.
25. Knudsen AB, Zauber AG, Rutter CM, et al. Estimation of benefits, burden, and harms of colorectal cancer screening strategies: modeling study for the US Preventive Services Task Force. *Jama.* 2016;315(23):2595-2609.
26. Kleindorfer GB, O'Neill L, Ganeshan R. Validation in simulation: various positions in the philosophy of science. *Management Science.* 1998;44(8):1087-1099.
27. Turner RM, Spiegelhalter DJ, Smith GC, Thompson SG. Bias modelling in evidence synthesis. *J R Stat Soc Ser A Stat Soc.* 2009;172(1):21-47.
28. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. *International journal of epidemiology.* 2014;43(6):1969-1985.
29. Murray EJ, Robins JM, Seage III GR, et al. Using observational data to calibrate simulation models. *Medical Decision Making.* 2018;38(2):212-224.
30. Murray EJ, Robins JM, Seage GR, Freedberg KA, Hernán MA. A comparison of agent-based models and the parametric g-formula for causal inference. *American journal of epidemiology.* 2017;186(2):131-142.
31. Sawaya GF, Sanstead E, Alarid-Escudero F, et al. Estimated quality of life and economic outcomes associated with 12 cervical cancer screening strategies: a cost-effectiveness analysis. *JAMA internal medicine.* 2019.
32. Siu AL, Force USPST. Screening for Breast Cancer: U.S. Preventive Services Task Force Recommendation Statement. *Ann Intern Med.* 2016;164(4):279-296.
33. Berry DA, Cronin KA, Plevritis SK, et al. Effect of screening and adjuvant therapy on mortality from breast cancer. *The New England journal of medicine.* 2005;353(17):1784-1792.

34. Healy M. Breast cancer screening recommendations clarify science but muddy political waters. *Los Angeles Times*2016.
35. van den Broek JJ, van Ravesteyn NT, Mandelblatt JS, et al. Comparing CISNET Breast Cancer Models Using the Maximum Clinical Incidence Reduction Methodology. *Med Decis Making.* 2018;38(1_suppl):112S-125S.
36. Etzioni R, Cha R, Cowen ME. Serial prostate specific antigen screening for prostate cancer: A computer model evaluates competing strategies. *J Urol.* 1999;162:741-748.
37. Andriole GL, Crawford ED, Grubb RLr, et al. Mortality results from a randomized prostate-cancer screening trial. *N Engl J Med.* 2009;360:1310-1319.
38. Pinsky PF, Black A, Kramer BS, Miller A, Prorok PC, Berg C. Assessing contamination and compliance in the prostate component of the Prostate, Lung, Colorectal, and Ovarian (PLCO) cancer screening trial. *Clin Trials.* 2010;7(4):303-311.