A Review on Calibration Methods of Cancer Simulation Models

Authors: Yichi Zhang, BA, Oguzhan Alagoz, PhD

Author affiliations

Oguzhan Alagoz and Yichi Zhang are in the Department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, WI.

Orcid ID for Oguzhan Alagoz is 0000-0002-5133-1382

Word Count: Text: 2868
Abstract: 229
Tables: 6
Figures: 2

Funding source: This work was supported by the National Institutes of Health (NIH) under National Cancer Institute Grant R01CA251566. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report.

Conflict of Interest statement: Dr. Alagoz reports grants from NIH, during the conduct of the study; personal fees from Bristol Myers Squibb, personal fees from Exact Sciences, other from Innovo Analytics, outside the submitted work. Mr. Zhang has no conflict of interest to report.

Address for correspondence and reprint requests:

Oguzhan Alagoz, PhD
University of Wisconsin-Madison
1513 University Avenue
Madison, WI 53706
Phone: 608.890.0399
Email: alagoz@engr.wisc.edu

Abstract

Calibration, a critical step in the development of simulation models, involves adjusting unobservable parameters to ensure that the outcomes of the model closely align with observed target data. This process is particularly vital in cancer simulation models with a natural history component where direct data to inform natural history parameters are rarely available. This work reviews the literature of cancer simulation models with a natural history component and identifies the calibration approaches used in these models with respect to the following attributes: calibration target, goodness-of-fit (GOF) measure, parameter search algorithm, acceptance criteria, and stopping rules. After a comprehensive search of the PubMed database from 1981 to June 2023, 68 studies were included in the review. Nearly all (n=66) articles specified the calibration targets, and most articles (n=56) specified the parameter search algorithms they used, whereas goodness-of-fit metric (n=51) and acceptance criteria/stopping rule (n=45) were reported for fewer times. The most frequently used calibration targets were incidence, mortality, and prevalence, whose data sources primarily come from cancer registries and observational studies. The most used goodness-of-fit measure was weighted mean squared error. Random search has been the predominant method for parameter search, followed by grid search and Nelder-mead method. Machine learning-based algorithms, despite their fast advancement in the recent decade, has been underutilized in the cancer simulation models. More research is needed to compare different parameter search algorithms used for calibration.

Key points

- This work reviewed the literature of cancer simulation models with a natural history component and identified the calibration approaches used in these models with respect to the following attributes: calibration target, goodness-of-fit (GOF) measure, parameter search algorithm, acceptance criteria, and stopping rules.
- Random search has been the predominant method for parameter search, followed by grid search and Nelder-mead method.
- Machine learning-based algorithms, despite their fast advancement in the recent decade, has been underutilized in the cancer simulation models. Furthermore, more research is needed to compare different parameter search algorithms used for calibration.

1 Introduction

Computer simulation models have been increasingly used to address cancer control problems. For example, the National Cancer Institute (NCI)'s Cancer Intervention Modeling and Surveillance Modeling Network (CISNET) simulation models have been used to inform the US Preventive Services Task Force screening recommendations for breast cancer ([1-3]), colorectal cancer [4], and lung cancer [5].

A crucial component of cancer simulation models is natural history, which represents the trajectory of cancer in an individual over time in the absence of a medical intervention. While a few of the natural history model parameters such as prevalence of cancer subtype could be estimated from primary data, most components such as the average tumor growth rate and the proportion of the tumors that regress remain unobservable. Consequently, in the absence of direct available data to estimate such parameters, models can determine the values of these parameters such that the model results match observed outcomes.

The most commonly used method to estimate directly unobservable parameters is *calibration*, which refers to the process of adjusting unobservable parameter values to ensure that the model's outcomes align closely with observed target data such as observed incidence and mortality [6]. As contemporary cancer simulation models grow in complexity, resulting in a large number of natural history parameters that need to be estimated through calibration, and computational demands rise, modelers face the challenge of conducting efficient calibration. This requires an optimal compromise between parameter combinations that mirror clinical data and computational time and resource demands.

The simplest approach to implement calibration is to conduct a full-scale grid search for the entire parameter space,

which involves discretizing the continuous parameters and running the simulation model for all possible combinations of the unobservable input parameters. While this is an easy-to-implement method, it requires major computational time due to the complexity of the models and a large number of input parameter combinations that need to be evaluated. For example, one study on calibrating an established breast cancer simulation model notes that a single replication of the model takes approximately 10 minutes on a stand-alone computer [7]. Considering that the study needed to evaluate approximately 400,000 input parameter combinations, the calibration procedure could take over 70 days to complete, which may not be computationally feasible. Not surprisingly, due to the need for speeding up the calibration in cancer simulation models, there is a growing interest in developing efficient strategies to search the parameter space for the calibration. A rich body of literature suggested using metaheuristic and structured methods such as grid search, random search, Nelder-Mead algorithm, neural networks, and Bayesian optimization for calibration [7-12].

Despite a drastic increase of interest in calibration from the cancer modeling community in the last decade, no study conducted a systematic review of the methods used for calibration in cancer simulation models, which is the focus of the present review. To our knowledge, only one previous study conducted a systematic review of the calibration methods used in the cancer simulation models and included studies published until 2006 [6]. That study did not focus on the optimization and heuristic methods used for the parameter search algorithm, instead, it primarily analyzed the studies based on modeled tumor types, metrics used for measuring goodness-of-fit of the models, and validation strategies. Compared to that study, we included more recent studies, which is crucial since there has been a major increase in the number of studies utilizing calibration methods and the diversity of the calibration approaches. In addition, our emphasis in this work is to classify the studies with respect to the calibration methods used, which provides insight into the preferred calibration methods by modelers in cancer

research.

2 Methods

2.1 Search Strategy

We conducted a systematic review on calibration methods employed in cancer simulation models with a natural history component. A comprehensive search of the PubMed database from 1981 to June 2023 was performed for articles published in English. Our search criteria only considered papers that contain "simulation" and "cancer" in title/abstract and "calibration" in the main text. Note that PubMed automatically includes synonymous or related keywords. For instance, in the case of 'cancer', PubMed also identifies words like 'neoplasm' or 'cancers.' In situations where multiple papers were published based on the same study, only one was included in our review. Additional studies were found through manual searches of the studies that cited tutorial papers describing simulation calibration and the previous systematic review article.

Our inclusion criteria consisted of articles that developed a cancer simulation model with a natural history component and calibrated the natural history component to match specific calibration targets. We excluded preprint/nonrefereed articles as well as articles focusing on models of disease other than cancer, with the exception of Human Papillomavirus due to its profound association with cervical cancer. We also excluded models devoid of a natural history component or those that do not calibrate their parameters. Tutorial articles that describe specific calibration methods were also excluded.

2.2 Classification of the Studies and Reporting of the Results

Our search focused on the following attributes of the calibration articles: calibration target, goodness-of-fit

measure, parameter search algorithm, acceptance criterion and stopping rule. We briefly describe them here.

Calibration targets are observed empirical data, such as cancer incidence and mortality rates, that can be directly estimated. As the name implies, these targets serve as benchmarks that a model aims to replicate during the calibration process. Typical calibration targets include incidence (the rate of new cancer cases within a specific period), mortality (the death rate due to cancer), survival (the proportion of patients living for a certain time after diagnosis), and stage distribution (the breakdown of cancer cases by cancer stage at diagnosis). These critical data are sourced from cancer registries, which collect comprehensive cancer patient information; observational studies, which monitor subjects in natural conditions without intervention; and randomized controlled trials, where subjects are randomly assigned to experimental or control groups to test the efficacy of treatments.

Goodness-of-fit (GOF) of a simulation model reflects how well the results of the running the model using an input parameter combination align with calibration targets. A straightforward GOF measure involves visual comparison where model outputs are manually contrasted with calibration targets to evaluate their fit, which is not ideal due to the subjectivity, hence, most models employed quantitative measures. Predominantly used GOF measures are mean squared error (MSE), weighted mean squared error, likelihood, and confidence interval envelope. Choosing a GOF measure that is compatible with simulation model is crucial for the model's success. We provide a formal description for each GOF metric used in the articles included in our study in **Appendix A**.

Parameter search algorithms refer to the methods to identify parameter combinations that are sufficiently close to the calibration targets. The degree of closeness between parameter combinations and calibration targets is measured by the GOF measure, serving as an error function in the parameter search algorithms. This problem can be conceptualized as finding the parameter combination that minimizes the GOF measure across the parameter

space, a topic that has been comprehensively studied. Additionally, the parameter space is typically expansive and non-convex due to the nonlinear nature of the simulation model or constraints on its parameters, rendering it challenging to find a global optimum using an algorithm with a reasonable runtime. Despite this, a selection of alternative heuristic algorithms can be utilized to find parameter combinations that reasonably approximate the calibration targets within an acceptable runtime. We describe each parameter search algorithm used in the articles included in our study in **Appendix B**.

Acceptance criteria refer to the standards set by the modelers to determine if predictions made by the model using a particular input parameter combination align sufficiently well with the calibration targets. These criteria are typically measured using the GOF metrics. On the other hand, *stopping rules* refer to the thresholds or conditions that, once met, lead modelers to terminate the calibration process. Common stopping rules include identifying an adequate number of parameters that fulfill the acceptance criteria or reaching a pre-specified number of iterations/time during calibration.

3 Results

3.1 Search results

Our search resulted in 253 unique articles. Of these, 56 met the inclusion criteria, while the remainder were excluded for the reasons mentioned previously. An examination of the references within these articles led us to identify 12 additional articles that satisfied our criteria, bringing the total to 68 articles.

Prior to 2006, very few articles met our inclusion criteria. A notable uptick in articles that fit our criteria began after 2006, peaking at 8 articles per year in 2019 (**Figure 1**). Breast, colorectal, and cervical were the predominant

cancer types represented with these simulation models, with 13 (19.1%), 13 (19.1%), and 11 (16.2%) articles respectively, as shown in **Table 3**. This contrasts with the less common cancer types, such as skin or thyroid cancers, which were each modeled in only a single article.

In our literature review, 66 (97.1%) of the 68 articles mentioned their calibration targets. Among these, one article did not specify the data source for targets, and three articles did not specify the target type. The most commonly used calibration data source is the cancer registry, which was used by 43 (63.2%) articles (**Table 1)**, notably the United States National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) program. This is followed by observational studies, which include a variety of subcategories, including cohort studies, retrospective studies, and national surveys [13-15]. Thirteen (19.1%) articles utilized data from randomized controlled trials. The most frequently used target types are incidence, mortality, and prevalence, respectively (**Table 2**). Only 19 (28.0%) of the 68 articles used a single type of calibration target type.

Of the 68 articles, the weighted MSE was the most employed GOF metric (**Table 4**). Specifically, 7 (10.3%) articles utilized chi-squared error as their weight, and 10 (14.7%) articles incorporated other types of weights. Following the weighted MSE, the MSE without weights was the second most popular metric, being used in 15 (22.1%) articles. Likelihood was the third most frequent GOF metric, with 10 (14.7%) articles using it. 17 (25%) of the articles do not specify their choice of GOF metrics.

Most articles we found in the literature search describes their parameter search algorithms. Of the 68 included articles, random search emerged as the most favored parameter search algorithm, being referenced in 16 (23.5%) articles. Note that the specific usage of the random search method is occasionally inferred rather than explicitly

stated. For example, Hammer et al. (2019) [16] describes their parameter search algorithm as "the simulation was run 10,000 times, with each run using a number randomly drawn from the range of possible growth rates." Such descriptions hint at the possible use of random search, but also raise questions about the exact methodology applied since other more complicated algorithms such as simulated annealing also randomly draw the number from plausible range. It is possible that some models might be leveraging more nuanced algorithms but might not be detailing them adequately in their documentation.

Following random search in popularity are the grid search and Nelder-mead methods, each being adopted in 9 (13.2%) and 8 (11.8%) articles respectively. Interestingly, despite the rising interest in machine learning (ML) for analyzing large datasets, as is typical with simulation models, only five articles utilized ML-based strategies. Furthermore, 12 (17.6%) articles did not specify their calibration algorithm, often leaving other calibration details ambiguous as well.

In our search, 21 (30.8%) articles reported to have terminated the calibration process after evaluating a designated number of combinations. These articles then selected the top-performing combinations without explicitly stating their acceptance criteria. A total of 14 (20.6%) articles clearly specified both acceptance criteria and stopping rule. Six and three articles only specify stopping rule and acceptance criteria, respectively. A total of 23 (33.8%) articles did not provide any information on acceptance criteria or stopping rule.

4 Discussion

Calibration of natural history parameters in cancer simulation models is typically the most time-consuming component of building a model, making a careful selection of efficient parameter search algorithms necessary. In

addition, the choice of suitable GOF metrics, acceptance criteria, and stopping rule with the algorithm is also crucial for calibration. Our review aims to summarize the strategies adopted by modelers in selecting these essential components, providing insights into current practices and potential improvements in simulation model calibration.

MSE, in both weighted and unweighted forms, is the most commonly used GOF measure, and it is considered as the default method in many implementations. However, prior to finalizing the choice of a GOF measure for a model, it is important to conduct a thorough comparison among various GOF measures. This is crucial because each measure has its unique strengths and limitations, and its performance varies in different scenario. For instance, MSE is sensitive to outliers, leading to a heavy penalty for larger errors. Secondly, MSE may not allow the capturing the temporal trends in observed outcomes such as incidence, which may be significant for accurate representation. For example, a recent thyroid cancer simulation model reported that successful replication of thyroid cancer epidemiology in the US requires modeling the drastic increase in thyroid cancer incidence between 1990s and 2010s, and the use of MSE may not allow choosing parameters combinations to reflect this temporal trend [17].

Given the advancement of machine learning algorithms in the recent decade, it is surprising that only a small number of studies, specifically five articles in our review, have utilized machine learning-based algorithms. This observation suggests a potential underutilization of these powerful tools in the field. Modelers and researchers should be aware that machine learning is highly accessible. In-depth expertise in machine learning is no longer a prerequisite for implementation, thanks to the proliferation of user-friendly libraries and frameworks. These resources efficiently manage the intricate aspects of algorithmic processing, so the users only need to know coding

to implement such algorithms. However, users still need machine learning knowledge to pick the suitable algorithms specific to the model.

We identified only one prior systematic review that conducted systematic reviews of calibration process [6]. This review, however, is limited to articles published before 2006 and does not concentrate on parameter search algorithms. Only a small proportion of the articles they included have clearly documented their search algorithm during calibration process, therefore the comparison of our present study's findings to that article is not possible.

Our review also highlighted the need for reporting of the calibration methods in simulation modeling papers. Many of the studies included in this review lack a comprehensive description of their calibration process, including how they selected the GOF metric and parameter search algorithm, as well as the absence of acceptance criteria and stopping rules. To enhance transparency and understanding, we advise authors to include a detailed account of the calibration procedure in the main body of the text or in a supplement. This addition would significantly help readers in understanding the full extent and robustness of the calibration process.

In addition to providing a detailed account of the calibration procedure, we strongly encourage modelers to consider the possibility of making their source code publicly available, given that the nature of their project permits it. Future researchers who are interested in the model can gain deeper understanding of the model and can potentially build upon the current model.

4.1 Future Research Directions for Calibration of Simulation Models

The strong need for computationally efficient simulation calibration methods and growing interest from the cancer

research community on simulation calibration provide an opportunity for future research. There is a noticeable lack of articles that perform a comparative analysis of multiple parameter search algorithms used for calibration. This omission results in ambiguity regarding the selection of the most suitable parameter algorithms. A more thorough comparison in future studies could provide valuable insights into the efficacy and applicability of different algorithms. A comparison would also make the calibration process more robust. Furthermore, considering the significant advancements in machine learning, its current application in the field appears underutilized. We encourage more studies to explore the incorporation of machine learning-based algorithms, especially given the ease of implementation afforded by contemporary libraries. The integration of these advanced techniques could lead to more refined and efficient modeling approaches.

4.2 Limitations

Our study has several limitations. First, while numerous parameter search algorithms are categorized under random search, making it the most commonly used algorithm, this classification may be misleading due to a lack of detailed information. In cases where only the use of randomness is mentioned without further details, the classification may not accurately reflect the actual algorithm used in the model. Secondly, it is possible that some authors utilized multiple parameter search algorithms in their papers but only reported the most successful one in the paper and omitted the details for others. Similarly, studies may have used a combination of search algorithms.

5 Conclusion

This study summarizes the calibration methods used by cancer simulation models. We found that the most common used parameter search algorithm is random search and the most commonly used GOF metric is MSE.

Given the recent advancement of machine learning techniques, we found fewer than expected number of models adopting this method. The findings also signal a critical need for enhanced transparency and standardization in reporting calibration processes. Detailed documentation of the calibration methods is essential for replicability and further methodological advancements.

References

1. Trentham-Dietz A, Chapman C, Jayasekera J. Breast Cancer Screening with Mammography: An Updated Decision Analysis for the US Preventive Services Task Force. Publication; 2023.

2. Mandelblatt JS, Stout NK, Schechter CB, van den Broek JJ, Miglioretti DL, Krapcho M, et al. Collaborative Modeling of the Benefits and Harms Associated With Different U.S. Breast Cancer Screening Strategies. Ann Intern Med. 2016 Feb 16;164(4):215-25, doi:10.7326/M15-1536.

3. Alagoz O, Berry DA, de Koning HJ, Feuer EJ, Lee SJ, Plevritis SK, et al. Introduction to the Cancer Intervention and Surveillance Modeling Network (CISNET) Breast Cancer Models. Med Decis Making. 2018 Apr;38(1_suppl):3S-8S, doi:10.1177/0272989X17737507.

4. Knudsen AB, Rutter CM, Peterse EFP, Lietz AP, Seguin CL, Meester RGS, et al. Colorectal Cancer Screening: An Updated Modeling Study for the US Preventive Services Task Force. Jama-J Am Med Assoc. 2021 May 18;325(19):1998-2011, doi:10.1001/jama.2021.5746.

5. Meza R. Evaluation of the benefits and harms of lung cancer screening with low-dose computed tomography : a collaborative modeling study for the U.S. Preventive Services Task Force. Rockville, MD: Agency for Healthcare Research and Quality; 2021.

6. Stout NK, Knudsen AB, Kong CY, Mcmahon PM, Gazelle GS. Calibration Methods Used in Cancer Simulation Models and Suggested Reporting Guidelines. PharmacoEconomics. 2009 2009-07-01;27(7):533-45, doi:10.2165/11314830-000000000-00000.

7. Cevik M, Ergun MA, Stout NK, Trentham-Dietz A, Craven M, Alagoz O. Using Active Learning for Speeding up Calibration in Simulation Models. Medical Decision Making. 2016 2016-07-01;36(5):581-93, doi:10.1177/0272989x15611359.

8. Rutter CM, Savarino JE. An Evidence-Based Microsimulation Model for Colorectal Cancer: Validation and Application. Cancer Epidemiology Biomarkers & Prevention. 2010 2010-08-01;19(8):1992-2002, doi:10.1158/1055-9965.epi-09-0954.

9. Bae S, Karnon J, Crane G, Bessen T, Desai J, Crowe P, et al. Cost-effectiveness analysis of imaging surveillance in stage II and III extremity soft tissue sarcoma: an Australian perspective. Cost Effectiveness and Resource Allocation. 2020 2020-12-01;18(1), doi:10.1186/s12962-020-0202-7.

10. Kim JJ, Kuntz KM, Stout NK, Mahmud S, Villa LL, Franco EL, et al. Multiparameter Calibration of a Natural History Model of Cervical Cancer. American Journal of Epidemiology. 2007 2007-06-07;166(2):137-50, doi:10.1093/aje/kwm086.

11. Hur C, Hayeck TJ, Yeh JM, Richards EB, Spechler SJ, Gazelle GS, et al. Development, Calibration, and Validation of a U.S. White Male Population-Based Simulation Model of Esophageal Adenocarcinoma. PLoS ONE. 2010 2010-03-01;5(3):e9483, doi:10.1371/journal.pone.0009483.

12. Taylor DCA, Pawar V, Kruzikas DT, Gilmore KE, Sanon M, Weinstein MC. Incorporating Calibrated Model Parameters into Sensitivity Analyses. PharmacoEconomics. 2012 2012-02-01;30(2):119-26, doi:10.2165/11593360-000000000-00000.

13. Hayes-Larson E, Shaw C, Ackley SF, Zimmerman SC, Glymour MM, Graff RE, et al. The Role of Dementia Diagnostic Delay in the Inverse Cancer-Dementia Association. J Gerontol A Biol Sci Med Sci. 2022 Jun 1;77(6):1254-60, doi:10.1093/gerona/glab341.

14. Chen Y, Watson TR, Criss SD, Eckel A, Palazzo L, Sheehan DF, et al. A simulation study of the effect of lung cancer screening in China, Japan, Singapore, and South Korea. PLOS ONE. 2019 2019-07-30;14(7):e0220610, doi:10.1371/journal.pone.0220610.

15. Wettstein MS, Naimark D, Hermanns T, Herrera‐Caceres JO, Ahmad A, Jewett MAS, et al. Required efficacy for novel therapies in BCG‐unresponsive non‐muscle invasive bladder cancer: Do current

recommendations really reflect clinically meaningful outcomes? Cancer Medicine. 2020 2020-05-01;9(10):3287-96, doi:10.1002/cam4.2980.

16. Hammer MM, Palazzo LL, Eckel AL, Barbosa EM, Kong CY. A Decision Analysis of Follow-up and Treatment Algorithms for Nonsolid Pulmonary Nodules. Radiology. 2019 2019-02-01;290(2):506-13, doi:10.1148/radiol.2018180867.

17. Alagoz O, Zhang Y, Arroyo N, Fernandes-Taylor S, Yang DY, Krebsbach C, et al. Modeling Thyroid Cancer Epidemiology in the United States Using Papillary Thyroid Carcinoma Microsimulation Model. Value Health. 2023 Dec 22, doi:10.1016/j.jval.2023.12.007.

18. Dinh T, Ladabaum U, Alperin P, Caldwell C, Smith R, Levin TR. Health benefits and cost-effectiveness of a hybrid screening strategy for colorectal cancer. Clin Gastroenterol Hepatol. 2013 Sep;11(9):1158-66, doi:10.1016/j.cgh.2013.03.013.

19. Berry DA, Inoue L, Shen Y, Venier J, Cohen D, Bondy M, et al. Chapter 6: Modeling the Impact of Treatment and Screening on U.S. Breast Cancer Mortality: A Bayesian Approach. JNCI Monographs. 2006 2006-10-01;2006(36):30-6, doi:10.1093/jncimonographs/lgj006.

20. Plevritis SK, Sigal BM, Salzman P, Rosenberg J, Glynn P. Chapter 12: A Stochastic Simulation Model of U.S. Breast Cancer Mortality Trends From 1975 to 2000. JNCI Monographs. 2006 2006-10-01;2006(36):86-95, doi:10.1093/jncimonographs/lgj012.

21. Eminaga O, Shkolyar E, Breil B, Semjonow A, Boegemann M, Xing L, et al. Artificial Intelligence-Based Prognostic Model for Urologic Cancers: A SEER-Based Study. Cancers. 2022 2022-06-26;14(13):3135, doi:10.3390/cancers14133135.

22. Seigneurin A, Labarere J, Francois O, Exbrayat C, Dupouy M, Filippi M, et al. Overdiagnosis and overtreatment associated with breast cancer mammography screening: A simulation study with calibration to population-based data. Breast. 2016 Aug;28:60-6, doi:10.1016/j.breast.2016.04.013.

23. Ward ZJ, Yeh JM, Bhakta N, Frazier AL, Girardi F, Atun R. Global childhood cancer survival estimates and priority-setting: a simulation-based analysis. The Lancet Oncology. 2019 2019-07-01;20(7):972-83, doi:10.1016/s1470-2045(19)30273-6.

24. Cheng CY, Calderazzo S, Schramm C, Schlander M. Modeling the Natural History and Screening Effects of Colorectal Cancer Using Both Adenoma and Serrated Neoplasia Pathways: The Development, Calibration, and Validation of a Discrete Event Simulation Model. MDM Policy Pract. 2023 Jan-Jun;8(1):23814683221145701, doi:10.1177/23814683221145701.

25. Karlsson A, Jauhiainen A, Gulati R, Eklund M, Grönberg H, Etzioni R, et al. A natural history model for planning prostate cancer testing: Calibration and validation using Swedish registry data. PLOS ONE. 2019 2019-02-14;14(2):e0211918, doi:10.1371/journal.pone.0211918.

26. Sai A, Vivas-Valencia C, Imperiale TF, Kong N. Multiobjective Calibration of Disease Simulation Models Using Gaussian Processes. Medical Decision Making. 2019 2019-07-01;39(5):540-52, doi:10.1177/0272989x19862560.

27. Alagoz O, Ergun MA, Cevik M, Sprague BL, Fryback DG, Gangnon RE, et al. The University of Wisconsin Breast Cancer Epidemiology Simulation Model: An Update. Medical Decision Making. 2018 2018-04-01;38(1_suppl):99S-111S, doi:10.1177/0272989x17711927.

28. Ward ZJ, Scott AM, Hricak H, Abdel-Wahab M, Paez D, Lette MM, et al. Estimating the impact of treatment and imaging modalities on 5-year net survival of 11 cancers in 200 countries: a simulation-based analysis. The Lancet Oncology. 2020 2020-08-01;21(8):1077-88, doi:10.1016/s1470-2045(20)30317-x.

29. Burger EA, Dyer MA, Sy S, Palefsky JM, de Pokomandy A, Coutlee F, et al. Development and Calibration of a Mathematical Model of Anal Carcinogenesis for High-Risk HIV-Infected Men. J Acquir Immune Defic Syndr.

2018 Sep 1;79(1):10-9, doi:10.1097/QAI.0000000000001727.

30. Jahn B, Sroczynski G, Bundo M, Mühlberger N, Puntscher S, Todorovic J, et al. Effectiveness, benefit harm and cost effectiveness of colorectal cancer screening in Austria. BMC Gastroenterology. 2019 2019-12-01;19(1), doi:10.1186/s12876-019-1121-y.

31. Campos NG, Burger EA, Sy S, Sharma M, Schiffman M, Rodriguez AC, et al. An updated natural history model of cervical cancer: derivation of model parameters. Am J Epidemiol. 2014 Sep 1;180(5):545-55, doi:10.1093/aje/kwu159.

32. Lubitz C, Ali A, Zhan T, Heberle C, White C, Ito Y, et al. The thyroid cancer policy model: A mathematical simulation model of papillary thyroid carcinoma in The U.S. population. PLOS ONE. 2017 2017-05-08;12(5):e0177068, doi:10.1371/journal.pone.0177068.

33. Ward ZJ, Walbaum M, Walbaum B, Guzman MJ, Jimenez de la Jara J, Nervi B, et al. Estimating the impact of the COVID-19 pandemic on diagnosis and survival of five cancers in Chile from 2020 to 2030: a simulation-based analysis. Lancet Oncol. 2021 Oct;22(10):1427-37, doi:10.1016/S1470-2045(21)00426-5.

34. Mühlberger N, Kurzthaler C, Iskandar R, Krahn MD, Bremner KE, Oberaigner W, et al. The ONCOTYROL Prostate Cancer Outcome and Policy Model. Medical Decision Making. 2015 2015-08-01;35(6):758-72, doi:10.1177/0272989x15585114.

35. Goldhaber-Fiebert JD, Stout NK, Ortendahl J, Kuntz KM, Goldie SJ, Salomon JA. Modeling human papillomavirus and cervical cancer in the United States for analyses of screening and vaccination. Population Health Metrics. 2007 2007-12-01;5(1):11, doi:10.1186/1478-7954-5-11.

36. Kong CY, Mcmahon PM, Gazelle GS. Calibration of Disease Simulation Model Using an Engineering Approach. Value in Health. 2009 2009-06-01;12(4):521-9, doi:10.1111/j.1524-4733.2008.00484.x.

37. Coldman AJ, Phillips N, Brisson J, Flanagan W, Wolfson M, Nadeau C, et al. Using the Cancer Risk Management Model to evaluate colorectal cancer screening options for Canada. Curr Oncol. 2015 Apr;22(2):e41-50, doi:10.3747/co.22.2013.

38. Mandelblatt J, Schechter CB, Lawrence W, Yi B, Cullen J. Chapter 8: The SPECTRUM Population Model of the Impact of Screening and Treatment on U.S. Breast Cancer Trends From 1975 to 2000: Principles and Practice of the Model Methods. JNCI Monographs. 2006 2006-10-01;2006(36):47-55, doi:10.1093/jncimonographs/lgj008.

39. Bailey SL, Sigal BM, Plevritis SK. A Simulation Model Investigating the Impact of Tumor Volume Doubling Time and Mammographic Tumor Detectability on Screening Outcomes in Women Aged 40–49 Years. JNCI: Journal of the National Cancer Institute. 2010 2010-08-18;102(16):1263-71, doi:10.1093/jnci/djq271.

40. Bessen T, Karnon J. A Patient-Level Calibration Framework for Evaluating Surveillance Strategies: A Case Study of Mammographic Follow-Up After Early Breast Cancer. Value in Health. 2014 2014-09-01;17(6):669-78, doi:10.1016/j.jval.2014.07.002.

41. Peters MLB, Eckel A, Mueller PP, Tramontano AC, Weaver DT, Lietz A, et al. Progression to pancreatic ductal adenocarcinoma from pancreatic intraepithelial neoplasia: Results of a simulation model. Pancreatology. 2018 2018-12-01;18(8):928-34, doi:10.1016/j.pan.2018.07.009.

42. Pandharipande PV, Heberle C, Dowling EC, Kong CY, Tramontano A, Perzan KE, et al. Targeted Screening of Individuals at High Risk for Pancreatic Cancer: Results of a Simulation Model. Radiology. 2015 2015-04-01;275(1):177-87, doi:10.1148/radiol.14141282.

43. Eisemann N, Waldmann A, Garbe C, Katalinic A. Development of a Microsimulation of Melanoma Mortality for Evaluating the Effectiveness of Population-Based Skin Cancer Screening. Medical Decision Making. 2015 2015-02-01;35(2):243-54, doi:10.1177/0272989x14543106.

44. Lee JM, Mcmahon PM, Kong CY, Kopans DB, Ryan PD, Ozanne EM, et al. Cost-effectiveness of Breast

MR Imaging and Screen-Film Mammography for ScreeningBRCA1Gene Mutation Carriers. Radiology. 2010 2010-03-01;254(3):793-800, doi:10.1148/radiol.09091086.

45. de Gelder R, Bulliard JL, de Wolf C, Fracheboud J, Draisma G, Schopper D, et al. Cost-effectiveness of opportunistic versus organised mammography screening in Switzerland. Eur J Cancer. 2009 Jan;45(1):127-38, doi:10.1016/j.ejca.2008.09.015.

46. Shi L, Tian H, McCarthy WJ, Berman B, Wu S, Boer R. Exploring the uncertainties of early detection results: model-based interpretation of mayo lung project. BMC Cancer. 2011 Mar 7;11:92, doi:10.1186/1471-2407-11-92.

47. Kroep S, Lansdorp-Vogelaar I, Rubenstein JH, De Koning HJ, Meester R, Inadomi JM, et al. An Accurate Cancer Incidence in Barrett's Esophagus: A Best Estimate Using Published Data and Modeling. Gastroenterology. 2015 2015-09-01;149(3):577-85.e4, doi:10.1053/j.gastro.2015.04.045.

48. Tappenden P, Chilcott J, Brennan A, Squires H, Glynne-Jones R, Tappenden J. Using Whole Disease Modeling to Inform Resource Allocation Decisions: Economic Evaluation of a Clinical Guideline for Colorectal Cancer Using a Single Model. Value in Health. 2013 2013-06-01;16(4):542-53, doi:10.1016/j.jval.2013.02.012.

49. Keeney E, Sanghera S, Martin RM, Gulati R, Wiklund F, Walsh EI, et al. Cost-Effectiveness Analysis of Prostate Cancer Screening in the UK: A Decision Model Analysis Based on the CAP Trial. Pharmacoeconomics. 2022 Dec;40(12):1207-20, doi:10.1007/s40273-022-01191-1.

50. Zhong C, Xu L, Peng HL, Tam S, Xu L, Dahlstrom KR, et al. An economic and disease transmission model of human papillomavirus and oropharyngeal cancer in Texas. Sci Rep. 2021 Jan 19;11(1):1802, doi:10.1038/s41598-021-81375-5.

51. Lin RS, Plevritis SK. Comparing the benefits of screening for breast cancer and lung cancer using a novel natural history model. Cancer Causes Control. 2012 Jan;23(1):175-85, doi:10.1007/s10552-011-9866-9.

52. Schultz FW, Boer R, de Koning HJ. Chapter 7: Description of MISCAN-lung, the Erasmus MC Lung Cancer microsimulation model for evaluating cancer control interventions. Risk Anal. 2012 Jul;32 Suppl 1(Suppl 1):S85-98, doi:10.1111/j.1539-6924.2011.01752.x.

53. Spencer JC, Brewer NT, Coyne-Beasley T, Trogdon JG, Weinberger M, Wheeler SB. Reducing Poverty-Related Disparities in Cervical Cancer: The Role of HPV Vaccination. Cancer Epidemiol Biomarkers Prev. 2021 Oct;30(10):1895-903, doi:10.1158/1055-9965.EPI-21-0307.

54. Berkhof J, Bogaards JA, Demirel E, Diaz M, Sharma M, Kim JJ. Cost-effectiveness of cervical cancer prevention in Central and Eastern Europe and Central Asia. Vaccine. 2013 Dec 31;31 Suppl 7:H71-9, doi:10.1016/j.vaccine.2013.04.086.

55. Wang Z, Zhang Q, Wu B. Development of an Empirically Calibrated Model of Esophageal Squamous Cell Carcinoma in High-Risk Regions. BioMed Research International. 2019 2019-05-22;2019:1-9, doi:10.1155/2019/2741598.

56. Bieri U, Hübel K, Seeger H, Kulkarni GS, Sulser T, Hermanns T, et al. Management of Active Surveillance-Eligible Prostate Cancer during Pretransplantation Workup of Patients with Kidney Failure: A Simulation Study. Clinical Journal of the American Society of Nephrology. 2020 2020-06-08;15(6):822-9, doi:10.2215/cjn.14041119.

57. Elliott TM, Lord A, Simms LA, Radford-Smith G, Valery PC, Gordon LG. Evaluating a risk assessment tool to improve triaging of patients to colonoscopies. Intern Med J. 2019 Oct;49(10):1292-9, doi:10.1111/imj.14267.

58. Campos NG, Castle PE, Wright TC, Jr., Kim JJ. Cervical cancer screening in low-resource settings: A cost-effectiveness framework for valuing tradeoffs between test performance and program coverage. Int J Cancer. 2015 Nov 1;137(9):2208-19, doi:10.1002/ijc.29594.

59. Lansdorp-Vogelaar I, Goede SL, Bosch LJW, Melotte V, Carvalho B, van Engeland M, et al. Cost-effectiveness of High-performance Biomarker Tests vs Fecal Immunochemical Test for Noninvasive Colorectal

Cancer Screening. Clin Gastroenterol Hepatol. 2018 Apr;16(4):504-12 e11, doi:10.1016/j.cgh.2017.07.011.

60. Foy M, Deng L, Spitz M, Gorlova O, Kimmel M. Chapter 11: Rice-MD Anderson Lung Cancer Model. Risk Analysis. 2012 2012-08-01;32:S142-S50, doi:10.1111/j.1539-6924.2011.01741.x.

61. Hazelton WD, Goodman G, Rom WN, Tockman M, Thornquist M, Moolgavkar S, et al. Longitudinal multistage model for lung cancer incidence, mortality, and CT detected indolent and aggressive cancers. Math Biosci. 2012 Nov;240(1):20-34, doi:10.1016/j.mbs.2012.05.008.

62. Snowsill T, Yang H, Griffin E, Long L, Varley-Campbell J, Coelho H, et al. Low-dose computed tomography for lung cancer screening in high-risk populations: a systematic review and economic evaluation. Health Technology Assessment. 2018 2018-11-01;22(69):1-276, doi:10.3310/hta22690.

63. Van Oortmarssen GJ, Habbema JDF, Lubbe JTHN, Van Der Maas PJ. A model-based analysis of the hip project for breast cancer screening. International Journal of Cancer. 1990 1990-08-15;46(2):207-13, doi:10.1002/ijc.2910460211.

64. Draisma G, Boer R, Otto SJ, Van Der Cruijsen IW, Damhuis RAM, Schroder FH, et al. Lead Times and Overdetection Due to Prostate-Specific Antigen Screening: Estimates From the European Randomized Study of Screening for Prostate Cancer. JNCI Journal of the National Cancer Institute. 2003 2003-06-18;95(12):868-78, doi:10.1093/jnci/95.12.868.

65. Retsky M, Demicheli R. Multimodal Hazard Rate for Relapse in Breast Cancer: Quality of Data and Calibration of Computer Simulation. Cancers. 2014 2014-11-27;6(4):2343-55, doi:10.3390/cancers6042343.

66. Rose J, Augestad KM, Kong CY, Meropol NJ, Kattan MW, Hong Q, et al. A simulation model of colorectal cancer surveillance and recurrence. BMC Medical Informatics and Decision Making. 2014 2014-12-01;14(1):29, doi:10.1186/1472-6947-14-29.

67. Berkhof J, Coupé VM, Bogaards JA, Van Kemenade FJ, Helmerhorst TJ, Snijders PJ, et al. The health and economic effects of HPV DNA screening in The Netherlands. International Journal of Cancer. 2010 2010-11-01;127(9):2147-58, doi:10.1002/ijc.25211.

68. Wallis CJD, Morton G, Jerath A, Satkunasviam R, Szumacher E, Herschorn S, et al. Adjuvant Versus Salvage Radiotherapy for Patients With Adverse Pathological Findings Following Radical Prostatectomy: A Decision Analysis. MDM Policy & Practice. 2017 2017-01-01;2(1):238146831770947, doi:10.1177/2381468317709476.

69. Landy R, Windridge P, Gillman MS, Sasieni PD. What cervical screening is appropriate for women who have been vaccinated against high risk HPV? A simulation study. International Journal of Cancer. 2018 2018-02-15;142(4):709-18, doi:10.1002/ijc.31094.

70. Cockrell C, Axelrod DE. Combination Chemotherapy of Multidrug-resistant Early-stage Colon Cancer: Determining Optimal Dose Schedules by High-performance Computer Simulation. Cancer Res Commun. 2023 Jan 3;3(1):21-30, doi:10.1158/2767-9764.crc-22-0271.

71. Jalal H, Trikalinos TA, Alarid-Escudero F. BayCANN: Streamlining Bayesian Calibration With Artificial Neural Network Metamodeling. Front Physiol. 2021;12:662314, doi:10.3389/fphys.2021.662314.

72. Jackson CH, Jit M, Sharples LD, De Angelis D. Calibration of Complex Models through Bayesian Evidence Synthesis. Medical Decision Making. 2015 2015-02-01;35(2):148-61, doi:10.1177/0272989x13493143.

73. Lowry KP, Lee JM, Kong CY, McMahon PM, Gilmore ME, Cott Chubiz JE, et al. Annual screening strategies in BRCA1 and BRCA2 gene mutation carriers: a comparative effectiveness analysis. Cancer. 2012 Apr 15;118(8):2021-30, doi:10.1002/cncr.26424.

74. Erenay FS, Alagoz O, Banerjee R, Cima RR. Estimating the Unknown Parameters of the Natural History of Metachronous Colorectal Cancer Using Discrete-Event Simulation. Medical Decision Making. 2011 2011-07-01;31(4):611-24, doi:10.1177/0272989x10391809.

75. Etzioni R, Gulati R, Falcon S, Penson DF. Impact of PSA screening on the incidence of advanced stage

prostate cancer in the United States: a surveillance modeling approach. Med Decis Making. 2008 May-Jun;28(3):323-31, doi:10.1177/0272989X07312719.

TABLES AND FIGURES

Table 1. Data source types for the calibration targets used in the cancer simulation models

Data source type	Number of studies	Reference
Cancer registry	42	[8], [10], [18], [12], [19], [20], [21], [22], [23], [24], [25], [14], [13], [26], [7], [27], [11], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52]
Observational study	24	[9], [8], [10], [18], [12], [24], [14], [13], [26], [53], [54], [29], [55], [35], [36], [56], [57], [40], [41], [58], [59], [48], [60], [61]
Randomized controlled trial	13	[62], [16], [63], [64], [65], [66], [34], [56], [67], [68], [46], [59], [69]
Other	1	[70]
Unspecified	3	[71], [72], [73]

Table 2. Calibration targets used in the cancer simulation models

Calibration target	Number of studies	Reference
Incidence	36	[62], [9], [12], [63], [64], [20], [22], [25], [13], [71], [7], [27], [11], [54], [30], [32], [55], [34], [35], [36], [37], [38], [39], [67], [40], [41], [43], [44], [45], [46], [58], [47], [48], [60], [49], [52]
Mortality	18	[9], [12], [63], [19], [20], [23], [14], [27], [28], [33], [36], [37], [40], [41], [42], [68], [45], [60]
Prevalence	18	[18], [12], [24], [14], [26], [71], [53], [11], [29], [31], [55], [34], [35], [67], [41], [42], [48], [69]
Survival	8	[21], [25], [13], [65], [66], [37], [56], [48]
Stage distribution	5	[64], [30], [36], [38], [42]
Other	4	[70], [57], [59], [61]
Unspecified	5	[8], [72], [73], [50], [51]

Table 3. Cancer types represented in the cancer simulation models

Cancer type	Number of studies	Reference
Breast	13	[63], [19], [20], [22], [65], [7], [27], [38], [39], [73], [40], [44], [45]
Colorectal	13	[8], [18], [24], [26], [71], [66], [30], [37], [70], [57], [74], [59], [48]
Cervical	11	[10], [12], [53], [54], [31], [35], [67], [72], [58], [69], [50]
Lung	9	[62], [16], [14], [36], [46], [60], [51], [52], [61]
Prostate	7	[64], [25], [34], [56], [68], [75], [49]
Multiple	4	[23], [13], [28], [33]
Esophageal	3	[11], [55], [47]
Pancreatic	2	[41], [42]
Soft tissue	1	[9]
Urologic	1	[21]
Anal	1	[29]
Thyroid	1	[32]
Skin	1	[43]
Bladder	1	[15]

Table 4. Goodness-of-fitness measures used in the cancer simulation models

Goodness-of-fitness measure	**Total number**	**Reference**
Sum of square	15	[16], [18], [20], [21], [23], [24], [66], [28], [33], [36], [44], [68], [74], [60], [61]
Likelihood	10	[10], [71], [53], [29], [31], [35], [40], [41], [42], [58]
Weighted sum of square	10	[12], [13], [26], [30], [56], [39], [67], [45], [46], [15]
Chi square (weighted sum)	7	[63], [64], [11], [32], [55], [38], [47]
Envelope method	3	[19], [7], [27]
Other	7	[26], [65], [34], [72], [43], [51], [52]
Unspecified	17	[62], [9], [8], [22], [25], [14], [54], [37], [70], [57], [73], [75], [59], [48], [69], [49], [50]

Table 5. Parameter search algorithms used in the cancer simulation models

Search algorithm	Number of studies	Reference
Random search	16	[10], [18], [19], [71], [27], [53], [29], [31], [35], [56], [40], [43], [58], [15], [69], [16]
Grid search	9	[9], [63], [20], [13], [66], [38], [67], [68], [74]
Nelder-mead	8	[12], [64], [25], [30], [39], [47], [51], [52]
Bayesian	7	[8], [19], [22], [23], [24], [71], [72]
Simulated annealing	7	[11], [28], [32], [33], [36], [41], [44]
Genetic algorithm	3	[26], [55], [36]
Other machine learning	3	[26], [7], [70]
Neural network	2	[21], [71]
Visual, trial-and-error	2	[62], [65]
Other	4	[34], [42], [48], [60]
Unspecified	12	[14], [54], [37], [57], [73], [45], [46], [75], [59], [49], [50], [61]

Table 6. Acceptance criteria and stopping rule used in the cancer simulation models

Acceptance criteria/stopping rule	Number of studies	Reference
Neither acceptance criteria not stopping rule is used	21	[16], [64], [22], [7], [27], [11], [28], [29], [30], [31], [33], [38], [39], [41], [42], [43], [68], [58], [74], [47], [61]
Both acceptance criteria and stopping rule were used	14	[9], [10], [23], [24], [26], [53], [66], [32], [55], [35], [36], [67], [72], [40]
Only stopping rule was used	6	[18], [12], [71], [65], [37], [69]
Only acceptable criteria was used	3	[19], [21], [44]
Unspecified	23	[62], [8], [63], [20], [25], [14], [13], [54], [34], [56], [70], [57], [73], [45], [46], [75], [59], [69], [60], [49], [50], [51], [52]

Figure 1. PRISMA flow diagram for search strategy

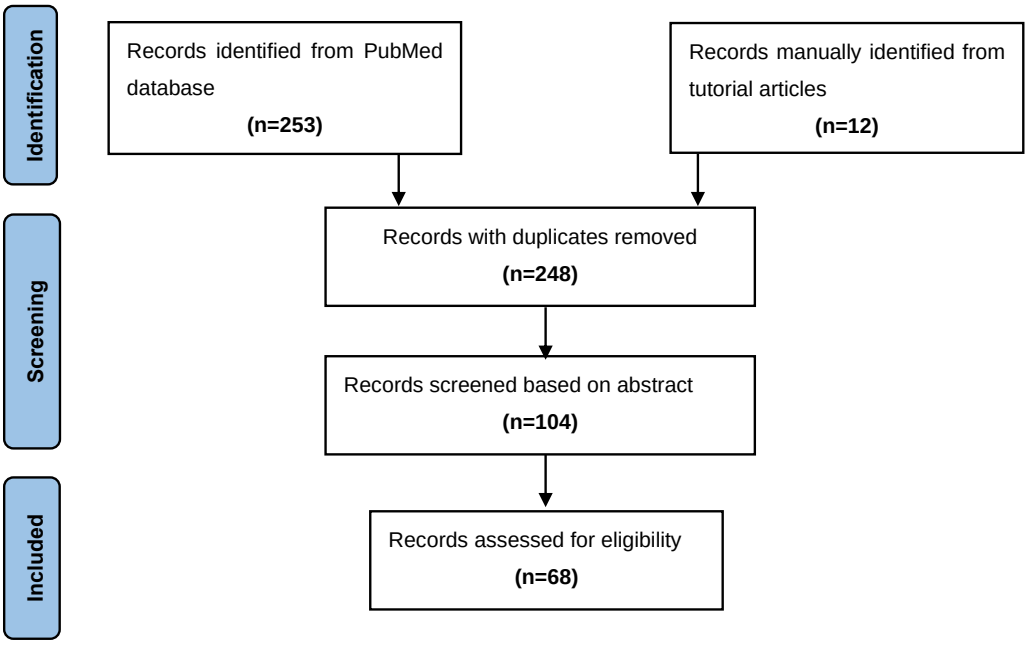

Figure 2. Number of cancer simulation models that utilized calibration by the year of publication