MDM
Medical Decision Making

Medical Decision Making
2023, Vol. 43(6) 719–736

DOI: 10.1177/0272989X231184175
journals.sagepub.com/home/mdm
Sage

Calibration and Validation of the Colorectal Cancer and Adenoma Incidence and Mortality (CRC-AIM) Microsimulation Model Using Deep Neural Networks

Vahab Vahdat, Oguzhan Alagoz, Jing Voon Chen, Leila Saoud, Bijan J. Borah, and Paul J. Limburg

Objectives. Machine learning (ML)–based emulators improve the calibration of decision-analytical models, but their performance in complex microsimulation models is yet to be determined. **Methods.** We demonstrated the use of an ML-based emulator with the Colorectal Cancer (CRC)-Adenoma Incidence and Mortality (CRC-AIM) model, which includes 23 unknown natural history input parameters to replicate the CRC epidemiology in the United States. We first generated 15,000 input combinations and ran the CRC-AIM model to evaluate CRC incidence, adenoma size distribution, and the percentage of small adenoma detected by colonoscopy. We then used this data set to train several ML algorithms, including deep neural network (DNN), random forest, and several gradient boosting variants (i.e., XGBoost, LightGBM, CatBoost) and compared their performance. We evaluated 10 million potential input combinations using the selected emulator and examined input combinations that best estimated observed calibration targets. Furthermore, we cross-validated outcomes generated by the CRC-AIM model with those made by CISNET models. The calibrated CRC-AIM model was externally validated using the United Kingdom Flexible Sigmoidoscopy Screening Trial (UKFSST). **Results.** The DNN with proper preprocessing outperformed other tested ML algorithms and successfully predicted all 8 outcomes for different input combinations. It took 473 s for the trained DNN to predict outcomes for 10 million inputs, which would have required 190 CPU-years without our DNN. The overall calibration process took 104 CPU-days, which included building the data set, training, selecting, and hyperparameter tuning of the ML algorithms. While 7 input combinations had acceptable fit to the targets, a combination that best fits all outcomes was selected as the best vector. Almost all of the predictions made by the best vector laid within those from the CISNET models, demonstrating CRC-AIM's cross-model validity. Similarly, CRC-AIM accurately predicted the hazard ratios of CRC incidence and mortality as reported by UKFSST, demonstrating its external validity. Examination of the impact of calibration targets suggested that the selection of the calibration target had a substantial impact on model outcomes in terms of life-year gains with screening. **Conclusions.** Emulators such as a DNN that is meticulously selected and trained can substantially reduce the computational burden of calibrating complex microsimulation models.

Corresponding Author
Vahab Vahdat, Health Economics and Outcome Research, Exact Sciences Corporation, 5505 Endeavor Ln, Madison, WI 53719, USA; (vvahdatzad@exactsciences.com; crcaim@exactsciences.com).

Highlights

- Calibrating a microsimulation model, a process to find unobservable parameters so that the model fits observed data, is computationally complex.
- We used a deep neural network model, a popular machine learning algorithm, to calibrate the Colorectal Cancer Adenoma Incidence and Mortality (CRC-AIM) model.
- We demonstrated that our approach provides an efficient and accurate method to significantly speed up calibration in microsimulation models.
- The calibration process successfully provided cross-model validation of CRC-AIM against 3 established CISNET models and also externally validated against a randomized controlled trial.

Keywords
microsimulation, calibration, colorectal cancer, deep neural networks

Date received: August 30, 2022; accepted: June 5, 2023

A crucial component of developing cancer microsimulation models is calibration, which involves estimating the directly unobservable natural history parameters from repeated simulation experiments.[1] Conventional approaches to calibration require running the microsimulation model with a large number of input combinations to identify a parameter set that best fits calibration targets such as observed cancer incidence and mortality.[2] There are 2 important challenges in calibration. First, running a complex simulation model with many input combinations is computationally prohibitive. Second, there is little knowledge of how different targets for calibration may affect model outcomes.

Previous efforts to improve the calibration of simulation models with heuristic or statistical engines such as simulated annealing[3–5] and Bayesian calibration[6–9] are powerful yet timely and complex. Alternatively, machine learning (ML) and statistical methods are simpler to implement, do not require optimization knowledge, and can be used to accelerate the calibration process compared with conventional calibration methods.

To address these challenges, we compared several ML algorithms and selected a deep neural network (DNN) framework as an emulator to facilitate microsimulation model calibration. Emulators or surrogate models have recently received attention for calibration of simulation models[2,10]; however, previous studies either used only 1 ML algorithm or calibrated using only a few targets. Furthermore, we incorporated multiple calibration targets into our framework and showed that the heterogeneity in the estimated unknown parameters can be achieved.

We demonstrate the effectiveness and validity of our approach using the Colorectal Cancer-Adenoma Incidence and Mortality (CRC-AIM) model, which is designed to answer questions related to colorectal cancer (CRC) progression and screening. CRC is the second leading cause of cancer deaths in the United States,[11] and early detection through screening reduces CRC incidence and mortality.[12] While screening is recommended by major medical organizations including the US Preventive Services Task Force (USPSTF),[13] American Cancer Society (ACS),[14]

Health Economics and Outcome Research, Exact Sciences Corporation, Madison, WI, USA (VV, JVC, LS, PJL); Departments of Industrial & Systems Engineering and Population Health Sciences, University of Wisconsin–Madison, Madison, WI, USA (OA); Division of Health Care Delivery Research, Mayo Clinic, Rochester, MN, USA (BJB). The majority of this work was presented as an oral presentation at SMDM 2022. The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: O. Alagoz has been a paid consultant for Exact Sciences. O. Alagoz has also been the owner of Innovo Analytics LLC as well as served as a consultant to Johnson & Johnson and Bristol Myers Squibb, outside of the submitted work. B. J. Borah is a consultant to Exact Sciences and Boehringer Ingelheim on projects unrelated to the submitted work. V. Vahdat, J. V. Chen, and L. Saoud are employees of Exact Sciences Corporation. P. J. Limburg serves as chief medical officer for screening at Exact Sciences. The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Financial support for this study was provided entirely by Exact Sciences Corporation. The following authors are employed by the sponsor: V. Vahdat, J. V. Chen, L. Saoud, and P. J. Limburg.

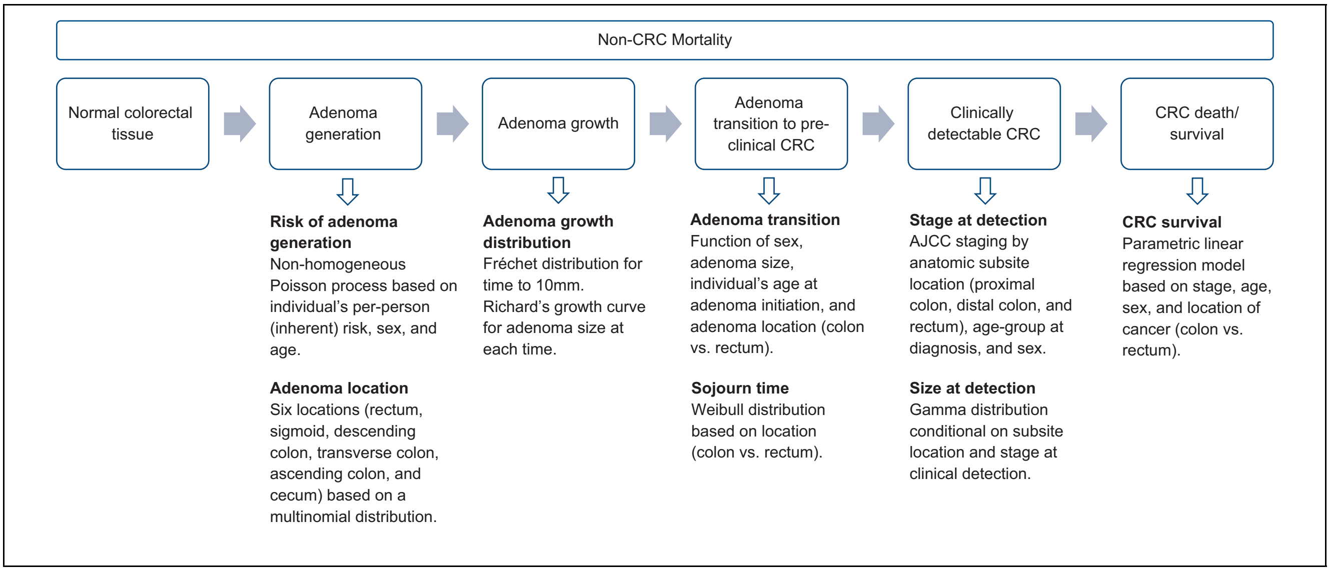
Non-CRC Mortality
Normal colorectal tissue
Adenoma generation
Adenoma growth
Adenoma transition to pre-clinical CRC
Clinically detectable CRC
CRC death/ survival
Risk of adenoma generation
Non-homogeneous Poisson process based on individual's per-person (inherent) risk, sex, and age.
Adenoma location
Six locations (rectum, sigmoid, descending colon, transverse colon, ascending colon, and cecum) based on a multinomial distribution.
Adenoma growth distribution
Fréchet distribution for time to 10mm. Richard's growth curve for adenoma size at each time.
Adenoma transition
Function of sex, adenoma size, individual's age at adenoma initiation, and adenoma location (colon vs. rectum).
Sojourn time
Weibull distribution based on location (colon vs. rectum).
Stage at detection
AJCC staging by anatomic subsite location (proximal colon, distal colon, and rectum), age-group at diagnosis, and sex.
Size at detection
Gamma distribution conditional on subsite location and stage at clinical detection.
CRC survival
Parametric linear regression model based on stage, age, sex, and location of cancer (colon vs. rectum).

Figure 1 Overview of the CRC-AIM natural history model.
AJCC,American Joint Committee on Cancer; CRC, colorectal cancer; CRC-AIM, Colorectal Cancer-Adenoma Incidence and Mortality model.

and American College of Gastroenterology,[15] full consensus has not been achieved with respect to some key considerations, such as the optimal frequency, age range for screening, and so forth. In addition, as new CRC screening modalities with different accuracies are developed, comparative effectiveness must be carefully reassessed. For this purpose, microsimulation models are increasingly used by policy makers to address comparative effectiveness and other questions related to CRC screening.[16]

In this study, we first show how we develop, train, and tune the hyperparameters of several ML algorithms and select the best emulator for the calibration. Then, we illustrate how our DNN-based emulator efficiently identifies multiple sets of unknown natural history–related parameters of CRC-AIM that fit well to primary calibration targets. We then demonstrate the validity of the calibrated CRC-AIM model using cross-model validation and external validation. For cross-model validation, we compare CRC-AIM's outcomes for CRC incidence, mortality, and life-years gained (LYG) from screening to the 3 established microsimulation models of the National Cancer Institute's (NCI's) Cancer Intervention and Surveillance Modeling Network (CISNET), which were used to inform USPSTF and ACS CRC screening guidelines.[14,16,17] For external validation, we replicate a large randomized controlled trial on CRC screening using CRC-AIM and compare the model's outcomes against the trial's findings. Finally, we demonstrate how calibration targets used for CRC-AIM affect the predicted CRC mortality reduction and LYG by screening.

Methods

Overview of CRC-AIM

CRC-AIM was inspired by the ColoRectal Cancer Simulated Population Incidence and Natural history (CRC-SPIN) model, 1 of 3 CISNET CRC models, and therefore shares many of this model's features (as obtained or derived from publicly available sources).[18,19] In this section, we provide a brief description of the CRC-AIM and include full details of the model in Supplementary Section A. We also describe the key differences between CRC-AIM and CRC-SPIN models in Supplementary Section B.

CRC-AIM simulates CRC-related events for individuals at average risk of developing CRC. The natural history of CRC is based on an adenoma-carcinoma sequence and consists of 5 subcomponents: 1) adenoma generation, 2) adenoma growth, 3) transition from adenoma to preclinical cancer, 4) transition from preclinical cancer to clinically detectable cancer (i.e., sojourn time), and 5) survival (Figure 1). CRC-AIM included stylistic probability distributions to model CRC progression. We used these probability distributions as they have been reported to accurately represent CRC natural history.[19]

1. *Adenoma generation*. CRC-AIM assumes that the risk of developing an adenoma depends on an individual's sex, age, and baseline risk, in which individuals younger than age 20 are not at risk of developing adenomas.[20] After an adenoma is

created, it is assigned to 1 of the 6 locations according to a multinomial distribution derived from several autopsy studies: rectum, sigmoid colon, descending colon, transverse colon, ascending colon, and cecum (Supplementary Table 1).

2. *Adenoma growth*. The size (i.e., diameter) of an adenoma is determined using the Richard's growth model,[21] in which its growth rate is calculated by the time required to reach 10 mm in diameter, sampled from a Fréchet distribution.[22] The parameters of the adenoma growth model are differentiated by colon and rectum.
3. *Transition from adenoma to preclinical cancer*. CRC-AIM models the cumulative transition probability of progressing from adenoma to preclinical cancer using a log-normal cumulative distribution function that is based on sex, size, and age at initiation of adenoma.[23,24] Adenoma to preclinical cancer transitions differ between colon and rectum.
4. *Transition from preclinical cancer to clinically detectable cancer (sojourn time)*. A Weibull distribution is used to model the time between the transition from preclinical cancer to when the preclinical cancer becomes clinically symptomatic (also known as the sojourn time) for colon cancers. A proportional hazards model is assumed between colon and rectal cancers, and consequently, the sojourn times for both locations follow the Weibull distribution.
5. *Survival*. Upon clinical detection of cancer, the stage at clinical detection is sampled using NCI's Surveillance, Epidemiology, and End Results (SEER) Program 1975–1979 data[25] and is found to be a function of age, sex, and location (rectum, proximal colon, and distal colon). The size at clinical detection, conditional on location and stage at clinical detection, is modeled as a gamma distribution (Supplementary Table 2) using SEER 2010–2015 data that are confined within cases diagnosed at ages 20 to 50 y (prior to eligibility for CRC screening in the United States). SEER 2010–2015 data for CRC size generation are preferred to SEER 1975–1979 data due to 1) the uncertainty regarding American Joint Committee on Cancer (AJCC) staging estimate within the older era and 2) notable differences in cancer sizes between the 2 time periods (Supplementary Figure 2). Survival from CRC is sampled from parametric models, with age at diagnosis and sex as covariates for each stage and location (colon v. rectum) fitted to cause-specific survival from SEER (see Supplementary Section A.5). We applied a 7% reduction in hazard, estimated using the 5-y cause-specific relative survival between periods 2000–2003 and 2010–2019 from SEER, for cases diagnosed after 2000 to reflect the improvement in CRC-specific survival in the recent years.[26] All-cause mortality by age were based on the 2017 U.S. life table.[27]

List of Calibrated Natural History Parameters

CRC-AIM includes 23 directly unobservable parameters governing the natural history of CRC (Table 1), which need to be estimated using calibration. To calibrate these parameters, we first identified a plausible range for each parameter, which was informed by CRC-SPIN.[18,19] We then supplemented the initial plausible range using our calibration process.

Calibration Targets

To estimate the natural history parameters, we used several calibration targets. Our primary targets included SEER 1975–1979 CRC incidence per 100,000, which encompass the most comprehensive population-based nationwide CRC data prior to widespread CRC screening in the United States, hence providing a crucial input for natural history model development. These data have also been used by several other CRC models focusing on the United States, including CISNET CRC models.[28–31] Because AJCC staging was not recorded in SEER data prior to 1988, stage-specific CRC incidence was not available to be used as a calibration target. To overcome this limitation, in addition to using overall CRC incidence by age as a calibration target, we also included the CRC incidence by location (colon and rectum) and gender (male and female) among our calibration targets. While SEER data include very useful data, they do not provide sufficient details such as average adenoma size, which are needed for precise natural history model development. For this purpose, we supplemented the primary calibration targets by including studies by Corley et al.[32] and Pickhardt et al.,[33] 2 high-impact studies reporting the adenoma prevalence and distribution by size based on a large sample of asymptomatic patients.[32,33]

In addition, we used 3 studies as secondary calibration targets to verify preclinical cancer prevalence and size distribution.[34–36] Since preclinical cancer prevalence is highly attributed with prior screening history and removal of adenomas, these studies were unique in identifying participants without a history of screening. The chances of detecting precancerous lesions are low; thus, for each of the secondary calibration targets, we generated a tolerance interval based on confidence intervals to determine whether model predictions fall within the

Table 1 Unknown Parameters of CRC-AIM Natural History Model

Unknown Parameter	Plausible Range	Best Parameter Value Selected by Calibration
Adenoma generation		
Baseline log risk, β_0	$\beta_0 \sim TN_{[-7,-5]}(-6.3,\ 0.4)$	−5.661
Standard deviation of baseline log-risk, σ_0	$\sigma_0 \sim TN_{[1,2]}\ (1.1,\ 0.2)$	1.270
Sex effect, β_1	$\beta_1 \sim TN_{[-0.5,-0.1]}(-0.5,\ 0.1)$	−0.384
Age effect (ages 20–<50), β_2	$\beta_2 \sim TN_{[0.03,0.07]}(0.045,\ 0.007)$	0.039
Age effect (ages 50–<60), β_3	$\beta_3 \sim TN_{[0.01,0.05]}(0.03,\ 0.01)$	0.023
Age effect (ages 60–<70), β_4	$\beta_4 \sim TN_{[-0.01,0.05]}(0.03,\ 0.01)$	0.020
Age effect (ages ≥70), β_5	$\beta_5 \sim TN_{[-0.02,0.03]}(0.03,\ 0.03)$	−0.018
Adenoma growth (time to 10 mm)		
Scale (colon), s_c	$s_c \sim U(10.7, 40)$	24.364
Shape (colon), α_c	$\alpha_c \sim U(0.5, 4)$	1.388
Scale (rectum), s_r	$s_r \sim U(5, 20)$	6.734
Shape (rectrum), α_r	$\alpha_r \sim U(2, 5)$	3.601
Adenoma growth (Richard's growth model)		
Shape parameter, p	$p \sim TN_{[0.5,3.2]}(1.0,\ 0.5)$	0.710
Transition from adenoma to cancer		
Size (male, colon), γ_{1cm}	$\gamma_{1cm} \sim U(0.02, 0.06)$	0.040
Age at initiation (male, colon), γ_{2cm}	$\gamma_{2cm} \sim U(0.0, 0.02)$	0.016
Size (male, rectum), γ_{1rm}	$\gamma_{1rm} \sim U(0.02, 0.07)$	0.039
Age at initiation (male, rectum), γ_{2rm}	$\gamma_{2rm} \sim U(0.0, 0.02)$	0.004
Size (female, colon), γ_{1cf}	$\gamma_{1cf} \sim U(0.02, 0.05)$	0.043
Age at initiation (female, colon), γ_{2cf}	$\gamma_{2cf} \sim U(0.0, 0.02)$	0.014
Size (female, rectum), γ_{1rf}	$\gamma_{1rf} \sim U(0.02, 0.055)$	0.035
Age at initiation (female, rectum), γ_{2rf}	$\gamma_{2rf} \sim U(0.0, 0.02)$	0.010
Sojourn time		
Scale (colon), λ_c	$\lambda_c \sim U(3.0, 5.0)$	4.683
Shape (colon and rectum), k	$k \sim U(2.0, 5.0)$	3.620
Log-hazard ratio, α	$\alpha \sim U(-1.0, 1.0)$	−0.018

$TN_{[a,b]}(\mu,\ \sigma)$ represents a truncated normal distribution with mean μ and standard deviation σ over the domain [a,b]. $U(a,b)$ represents a uniform distribution with domain (a,b). CRC-AIM, Colorectal Cancer-Adenoma Incidence and Mortality Model.

reported values (Supplementary Table 7b). The use of secondary targets required adapting CRC-AIM to replicate the study settings in terms of the age and sex distribution of the study population (Supplementary Table 6).

Model Calibration

Because a single run of CRC-AIM simulation takes approximately 30 min in a standalone desktop PC with population size of 500K, it is computationally infeasible to evaluate all possible combinations of the parameters listed in Table 1 to identify the best combination. To speed up this process, we evaluated several ML algorithms as an emulator, which approximates the CRC-AIM model outcomes based on inputs and has substantially shorter computational times compared with CRC-AIM.[37] Figure 2 shows a schematic flowchart of our calibration framework.

Emulator Selection

We first generated 15,000 different combinations of the unknown parameters from the plausible ranges using Latin hypercube sampling (LHS)[38] and ran CRC-AIM to evaluate the corresponding target values (shown as D_1 in Figure 2). To select the best population size for generating outcomes, the precision of CRC-AIM in predicting CRC with different population sizes was evaluated. We found that the modeled incidence remained relatively stable when the population size was at least 500K (Supplementary Figure 4). Hence, we simulated 500K individuals in each run, and we used the following aggregated calibration targets to select ML algorithms: CRC incidence by location and gender from SEER (4 outcomes), adenoma size distribution for the age groups of 50 to 59, 60 to 69, and 70+ y based on Corley et al.[32] (3 outcomes), and the percentage of small adenoma detected by same-day virtual and optical colonoscopy from

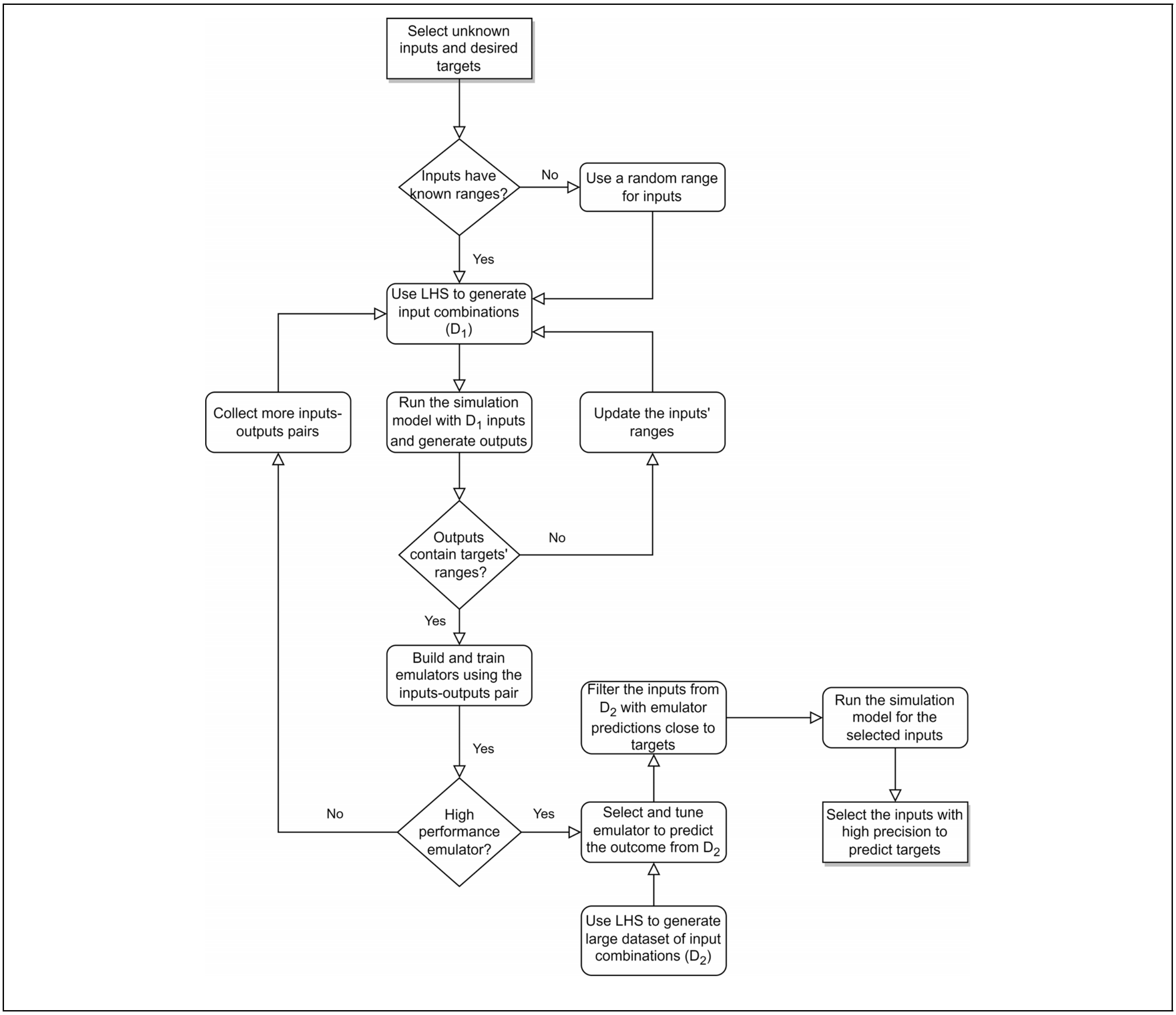
Select unknown inputs and desired targets
Inputs have known ranges?
No
Use a random range for inputs
Yes
Use LHS to generate input combinations (D_1)
Collect more inputs-outputs pairs
Run the simulation model with D_1 inputs and generate outputs
Update the inputs' ranges
Outputs contain targets' ranges?
No
Yes
Build and train emulators using the inputs-outputs pair
Yes
Filter the inputs from D_2 with emulator predictions close to targets
Run the simulation model for the selected inputs
No
High performance emulator?
Yes
Select and tune emulator to predict the outcome from D_2
Select the inputs with high precision to predict targets
Use LHS to generate large dataset of input combinations (D_2)

Figure 2 Calibration framework using an emulator as a surrogate for actual microsimulation model.
LHS, Latin hypercube sampling.

Pickhardt et al.[33] (1 outcome). Since the confidence intervals around the mean for secondary targets were wide (Supplementary Table 7b), the additional value of adding them as outcomes to our ML algorithms was minimal; hence, they were excluded in selecting the best emulator but included in calibration validation.

Using the 15,000 input-output combination pairs, we evaluated several ML algorithms, such as DNN,[39] random forest,[40] and several gradient boosting methods including conventional gradient boosting,[41,42] eXtreme Gradient Boosting (XGBoost) with advanced L1 and L2 regularization,[43] LightGBM[44] (light gradient boosting machine), and CatBoost (categorical boosting)[45] and compared their performance. For this purpose, we divided the simulation runs into training and testing data sets with the ratio of 3:1 and trained each ML algorithm with the training data. To preprocess the data set, we evaluated 2 scaling methods, standardization (mean of zero and standard deviation of unity) and normalization (min-max scaler). The goodness-of-fit (GOF) metrics (i.e., mean squared error [MSE], mean absolute error, mean absolute percentage error, and mean squared log error) for the training and testing data sets were calculated. To investigate and tune hyperparameters of each

ML model, we used k-fold cross-validation (k = 5) and a random search with 100 hyperparameter combinations for finding the best set that maximizes GOF.

Calibration Process with the Trained Emulator

Once the best ML algorithm for the emulator was identified, we compared emulator-based predictions against CRC-AIM–generated outcomes with the testing data set and confirmed the predictive accuracy of the emulator. We then used the emulator to evaluate 10 million input vector combinations generated from LHS (denoted as D_2 in Figure 2) to identify the most promising input vector combinations that are within 5% difference to targets. We used this 5% deviation from the calibration target to ensure that all potentially acceptable inputs suggested by the emulator would be further analyzed. The selected input vector combinations were simulated by CRC-AIM, and their simulated outcomes were compared with the calibration targets.

Since we had several primary and secondary calibration targets, among vectors with simulated outcomes fall within targets' ranges, a rank-ordered hierarchical process of eliminating implausible models based on their fit to the calibration targets was employed. The list of target priorities and the scoring framework are provided in Supplementary Table 14. For the natural history modeling, our highest-rank target was CRC incidence by age, location, and sex, followed by adenoma prevalence and sojourn time. Adenoma size distribution for different age groups, dwell time, and secondary calibration targets were weighted less. The input vectors with acceptable natural history fit to primary and secondary calibration targets were further examined by cross-model validation experiments.

Cross-Model Validation

After the calibration was completed and all of the input vector combinations with high precision to primary and secondary calibration targets were selected, CRC-AIM was cross-validated against the 3 CISNET models, CRC-SPIN, MISCAN-COLON, and SimCRC, which reported extensive results as part of the 2021 USPSTF CRC screening guideline update.[46] We compared several outcomes such as LYG with screening, CRC incidence and deaths (in the presence and absence of screening), and total number of colonoscopies conducted by screening modality. Three screening strategies (at their recommended screening intervals) for individuals aged 45 to 75 y were compared: colonoscopy every 10 y, annual fecal immunochemical test (FIT), and triennial multitarget stool DNA (mt-sDNA) test. Screening test sensitivity for CRC and adenomas (by size) and specificity are provided in Supplementary Table 10. Consistent with the USPSTF modeling approach,[46] sensitivity for stool tests was calibrated to match the overall nonadvanced adenoma sensitivity (Supplementary Section F). Perfect adherence to screening (100%) was assumed, and an incidence rate ratio was applied to reflect the increasing underlying risk of developing CRC since 1970.[47] Previous analysis showed that the CRC incidence for adults younger than 50 y who were not eligible to receive national screening has substantially increased for both men and women in colon and rectum.[47] Based on prior analysis reported in USPSTF,[46] the incidence rate ratio was set to 1.19. The incidence rate ratio was assumed to be driven by an increase in the baseline log risk, (β_0) in adenoma generation (the full equation is provided in Supplementary Section A.1), and is applied throughout each simulated individuals' life span.

Similar to natural history model selection, we used a hierarchical process to rank our cross-model validation experiments since multiple outputs were compared against other models. The criteria specified that the differences of model-predicted outcomes compared with those from the 3 CISNET models should be sufficiently small. The outcomes, sorted from most important to least, included CRC incidence, LYG due to screening, CRC deaths averted with screening, total number of colonoscopies, and CRC cases and deaths without screening. We prioritized LYG due to screening since USPSTF "focused on estimated LYG (compared with no screening) as the primary measure of the benefit of screening" in their 2021 CRC screening guideline update.[48]

External Validation Using the United Kingdom Flexible Sigmoidoscopy Screening Trial (UKFSST)

External validation was performed by comparing modeled outcomes from CRC-AIM against those reported by UKFSST, a randomized controlled trial that examined CRC incidence and mortality outcomes following a 1-time flexible sigmoidoscopy.[49–51] UKFSST was conducted in a population that was not yet routinely screened for CRC; therefore, published trial results provided unique information on the preclinical duration of CRC and the screening impact on the risk of CRC. As a result, many simulation models including CSNET CRC models used UKFSST as an external validation target.[52–55]

Briefly, in UKFSST, participants aged 55 to 64 y from 14 centers were randomized into a control group and a sigmoidoscopy screening group. As the UKFSST was a UK-based trial, 1996–1998 UK life tables were used to modify all-cause mortality in CRC-AIM.[56] No other modifications to the natural history of the model were made. The trial was simulated 500 times, each time generating a cohort with size, age, and sex distributions similar to the observed data from the trial. Details regarding sensitivity and specificity of sigmoidoscopy and colonoscopy, referral to colonoscopy, and surveillance with colonoscopy are presented in Supplementary Section G. Primary outcomes included hazard ratios of CRC incidence and mortality, whereas secondary outcomes included long-term cumulative incidence and mortality over 17 y of follow-up.

For each of the input combinations that demonstrated successful fit to natural history targets and screening cross-validation, external validity against UKFSST was also examined. A vector that partially fails to meet this criterion (i.e., only 1 of incidence or mortality hazard ratio is within confidence interval range) was regarded as acceptable, since it is likely for that vector to demonstrate external validity against other trials. At the end of selecting final input vectors, we identified 1 single input vector that has the best performance in terms of calibration targets. However, we also included multiple input vectors with acceptable performance in our final model, to reflect the heterogeneity of the CRC natural history.

Impact of Calibration Targets on Outcomes

To test the importance of calibration target selection, we examined the LYG from screening for 4 input combinations that were regarded as unacceptable for 1 of the primary targets: 1 that fit SEER incidence and the calibration target from the study by Pickhardt et al. but not that from the study by Corley et al. (model U1), 1 that fit the calibration target from the studies by Corley et al. and Pickhardt et al. well but not SEER incidence (model U2), 1 that fit SEER incidence and the calibration target from the study by Corley et al. but not that from the study by Pickhardt et al. (model U3), and 1 that fit SEER incidence but not the studies by Corley et al. and Pickhardt et al. well (model U4). We also performed the external validation experiments for these models.

Results

Selection and Fine-Tuning of the Emulator

Among all ML algorithms, DNN had the best GOF when standardization was used (Supplementary Table 8). Hence, DNN was selected to build the emulator for the calibration. For the hyperparameter tuning of the DNN, we explored its performance with different GOF measures, the number of nodes in hidden layers, activation functions, optimization algorithms, learning rates, epochs, and batch sizes. We used 5-fold cross-validation and random search with 100 hyperparameter combinations that took 4.16 h to complete. We further verified several parameter combinations with trial and error, to ensure the best hyperparameters were selected. The final DNN had an input layer with 23 nodes, an output layer with 8 nodes, and 4 dense hidden layers with 128, 64, 64, and 64 nodes, respectively (Figure 3). The activation function used in the first and third layers was the sigmoid function, whereas rectified linear units were used in other layers. We selected the Adam optimization algorithm,[57] a first-order gradient-based optimization of stochastic objective functions with a learning rate of 0.001 to train the model. Since the outputs are continuous and MSE provides a combination measurement of bias and variance of the prediction, it was used to quantify the GOF between the predicted and observed values in the test data set and as the loss function.

Performance of DNN

Using AWS (p3.2xlarge EC2 instance), it took 10 min to run 1 replication of CRC-AIM, whereas the DNN model was trained in 28.4 s with a training MSE of 0.014. On the test data set, predicted outcomes were comparable with actual outcomes in most cases, with an MSE less than 0.016. DNN-predicted versus CRC-AIM–predicted outcomes for the first 100 testing input combinations are shown in Figure 4. While the outcomes may substantially differ between each combination of input vector (shown in the x-axis), the DNN model was also able to predict these differences accurately (shown by the red line). For instance, the input vector used for run 20 of the testing data set (i.e., x-axis value equal to 20) led to a high adenoma prevalence and low CRC incidence, indicating a slow-growing adenoma scenario in which the proportion of small adenomas is high and they would not transition to cancer. The red and black points at $x = 20$ represent the DNN and CRC-AIM estimates, respectively. In most cases, red and black points associated with input vectors are very close to each other, indicating that the DNN successfully predicted the CRC incidence for each location and sex, adenoma prevalence by age, and proportion of small adenoma detected.

The trained DNN was used to predict outcomes for 10 million newly generated inputs in 473.16 s. Considering the computation times for input-output pair data

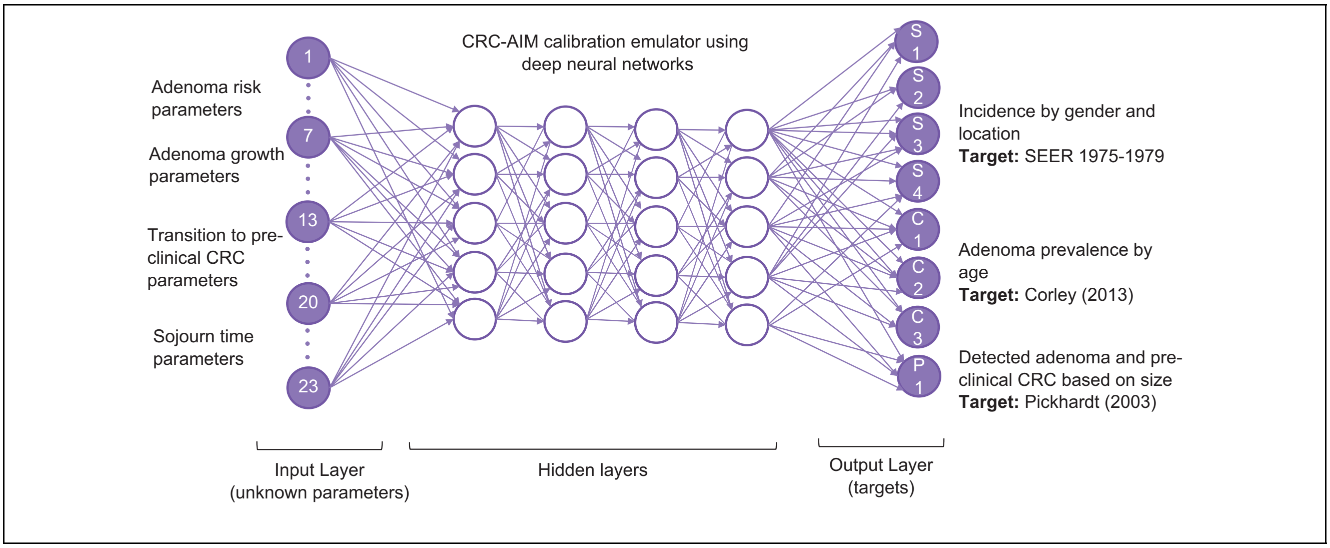
CRC-AIM calibration emulator using deep neural networks
1
7
13
20
23
Adenoma risk parameters
Adenoma growth parameters
Transition to pre-clinical CRC parameters
Sojourn time parameters
S 1
S 2
S 3
S 4
C 1
C 2
C 3
P 1
Incidence by gender and location
Target: SEER 1975-1979
Adenoma prevalence by age
Target: Corley (2013)
Detected adenoma and pre-clinical CRC based on size
Target: Pickhardt (2003)
Input Layer (unknown parameters)
Hidden layers
Output Layer (targets)

Figure 3 Graphical representation of the DNN emulator for CRC-AIM. The DNN consists of 23 nodes in the input layer representing the unobservable parameters, 4 hidden layers, and an output layer with 8 nodes representing the primary calibration targets, which include CRC incidence by location and gender from SEER 1975-1979, adenoma prevalence by age, and percentage of detected adenomas $\leq$ 5 mm).
CRC, colorectal cancer; CRC-AIM, Colorectal Cancer-Adenoma Incidence and Mortality model; DNN, Deep neural networks; SEER, Surveillance, Epidemiology, and End Results.

generation (2,500 h), the generation, storage, and retrieving of 10 million LHS combinations for prediction (7 min), Emulator model selection (5 h), training, testing, hyperparameter tuning of selected emulator (4.5 h), and filtering the predicted outcomes (10 min), the calibration process took approximately about 104 CPU days. In contrast, conventional calibration would have required a total of 190 CPU y with CRC-AIM.

Of 10 million inputs, only 101 input combinations that were within a deviation of 5% from the point estimates of primary calibration targets and were considered well-fitting and selected for further investigation. We then used CRC-AIM to evaluate these 101 input combinations for primary and secondary calibration targets. As shown in Supplementary Table 9, the overall difference between CRC-AIM's actual outcomes and the predicted outcomes by DNN were 4.4% (CI: 3.9%–4.7%), and the largest margin between predicted and actual outcomes was seen in CRC incidence of colon in females and rectal in males.

Selection of a Calibrated CRC-AIM

In total, 56 of 101 input vectors showed acceptable natural history outcomes and were further examined for cross-model validation. Among them, 16 of the 56 input parameters resulted in outcomes that were consistent with those reported by the CISNET models. We then used the UKFSST to test the external validity of our best input vectors. Seven input vectors with acceptable fit to the calibration targets and cross-model validity were selected as our final input vector combinations. These inputs with corresponding values, reflecting the heterogeneity of CRC natural history, are presented in Supplementary Table 12. Model predictions for all outcomes considered in cross-validation and external validation are presented in Supplementary Table 13. The score of each input vector for targets is summarized in Supplementary Table 14. One vector that best fits all outcomes was selected as the representative input vector (Table 1). The difference between predicted outcomes from DNN and actual outcomes from CRC-AIM for the representative input vector was 1.9%. The selected input vector matched age-specific CRC incidence as reported by SEER's 1975–1979 data (Figure 5) as well as adenoma prevalence reported by the autopsy studies[46] (Supplementary Figure 7). Distributions of adenomas by location (Supplementary Figure 8), adenoma size by age group (Supplementary Figure 9), and cancer stage at diagnosis (Supplementary Figure 11) estimated by CRC-AIM compared well against our estimates with SEER data[58] and CISNET models. The dwell time and sojourn time estimated by CRC-AIM were 20.3 y and 4.1 y, respectively, both of which fall within the estimated values from the literature[59–61] and CISNET models (Supplementary Figure 10).

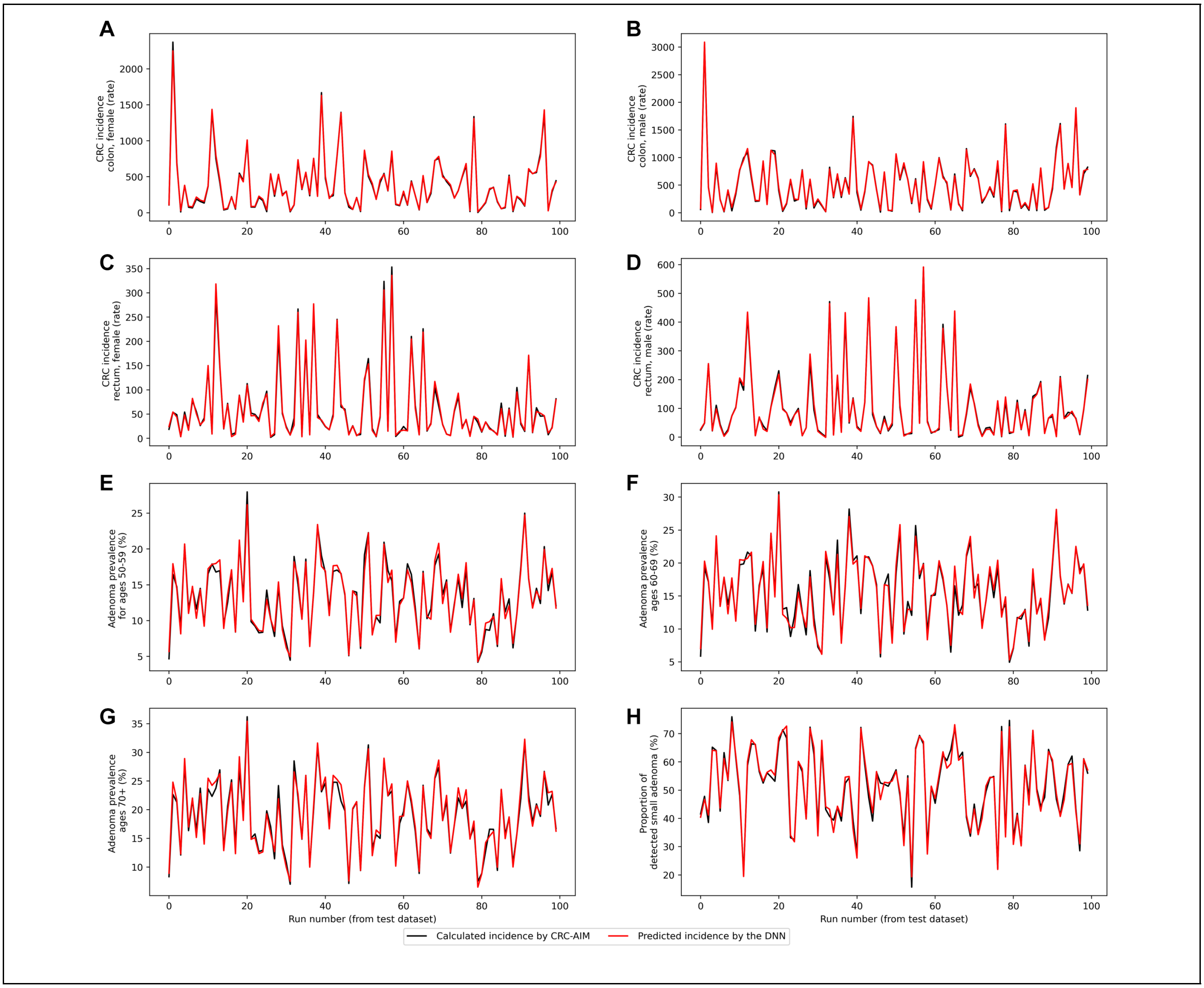

Figure 4 DNN- and CRC-AIM–predicted outcomes for the first 100 testing input combinations. CRC incidence rate by sex and location is represented in panels A through D, followed by adenoma prevalence by age groups in panels E through G and percentage of small adenoma (≤ 5 mm) in panel H. Note that the red lines and black lines perfectly overlap for most of the instances; therefore, black lines are often invisible.

CRC, colorectal cancer; CRC-AIM, Colorectal Cancer-Adenoma Incidence and Mortality model; DNN, deep neural networks.

Cross-Model Validation with CISNET Models and External Validation with UKFSST

Screening-related outcomes estimated by CRC-AIM, including LYG, incidence and mortality reductions associated with colonoscopy, FIT, and mt-sDNA screening strategies (Figure 6) as well as the associated numbers of colonoscopies and stool-based tests (Supplementary Figures 12–15), were comparable with CISNET model predictions.

The hazard ratios of CRC incidence and CRC mortality at 17-y follow-up (Figure 7) and the cumulative probabilities of CRC incidence and mortality (Supplementary Figure 16) estimated by CRC-AIM were consistent with the reported outcomes from UKFSST,[50] demonstrating the external validity of CRC-AIM.

Impact of Calibration Targets on Outcomes

The LYG from screening colonoscopy for individuals aged 45 to 75 y was 338, 283, 344, and 414 per 1,000 people screened for models U1, U2, U4, and U4, respectively (Supplementary Table 13). Models U2 and U4

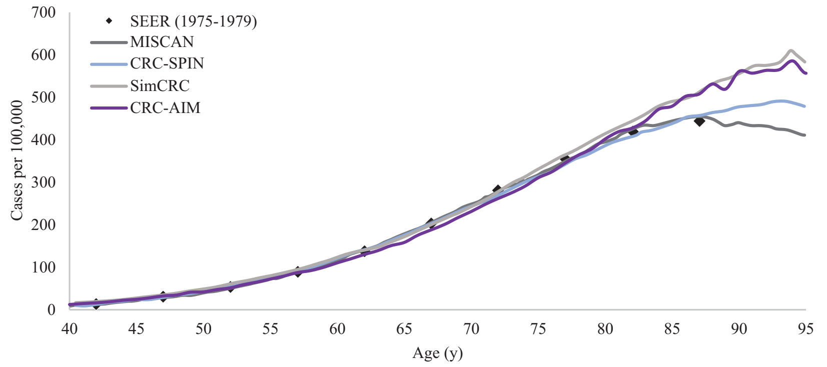

Figure 5 CRC-AIM and CISNET predictions of colorectal cancer cases per 100,000 people by age (adapted from Knudsen et al.[46]).
CRC-AIM, Colorectal Cancer-Adenoma Incidence and Mortality model; CRC-SPIN, ColoRectal Cancer Simulated Population Incidence and Natural history model; MISCAN, MIcrosimulation SCreening Analysis; SEER, Surveillance, Epidemiology, and End Results; SimCRC, Simulation Model of Colorectal Cancer.

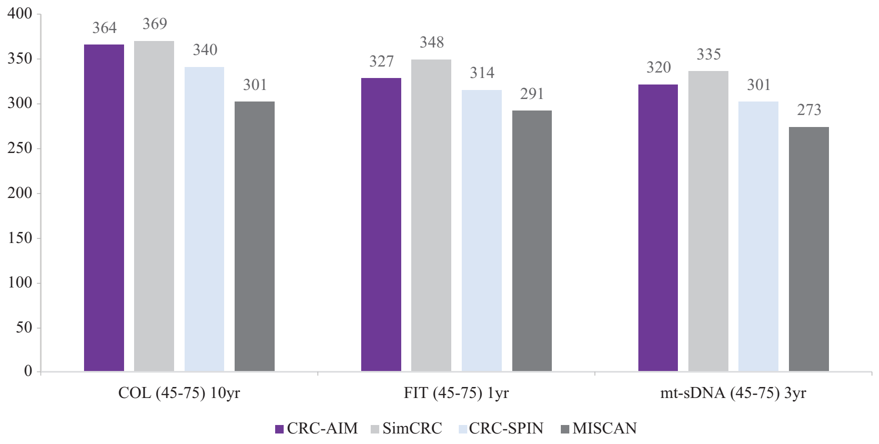

Figure 6 CRC-AIM and CISNET[46] estimates of life-years gained from screening by modality (10 y colonoscopy, annual FIT, and triennial mt-sDNA).
COL, colonoscopy; CRC-AIM, Colorectal Cancer-Adenoma Incidence and Mortality model; CRC-SPIN, ColoRectal Cancer Simulated Population Incidence and Natural history model; FIT, fecal immunochemical test; MISCAN, MIcrosimulation SCreening Analysis; mt-sDNA, multi-target stool DNA; SimCRC, Simulation Model of Colorectal Cancer; yr, year.

were unable to predict LYG within expected range, and models U2 and U3 failed external validation. Therefore, fitting to SEER incidence and calibration targets from screening studies are both critical for model validation.

Discussion

Extensive computational needs for calibration, a crucial step in the development of cancer microsimulation models, require methods to accelerate this lengthy process. In this study, we used an ML-based framework to increase the efficiency of calibration for simulation models without compromising the quality of the calibrated parameters. We demonstrated our framework's utility using CRC-AIM, a microsimulation model representing the CRC epidemiology in the United States. We found that using a DNN as an emulator substantially reduced calibration time from 190 CPU-years to 104 CPU-days.

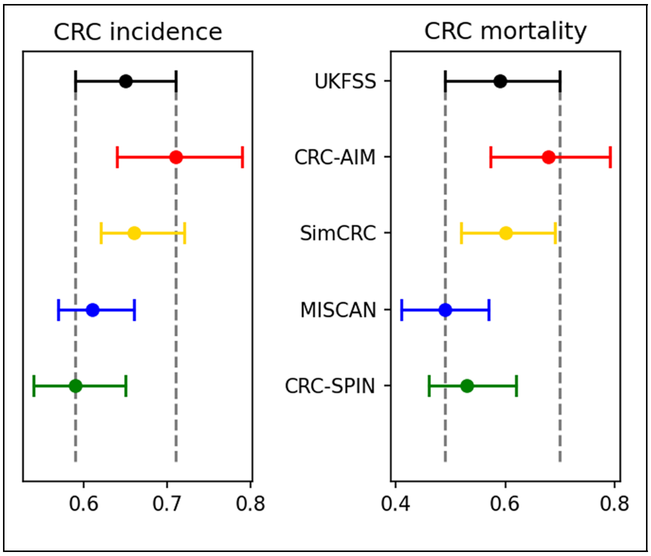

Figure 7 External validation with UKFSST: hazard ratios of colorectal cancer incidence and mortality between screening and control groups over the 17-y follow-up (adapted from Knudsen et al.[46])

CRC, colorectal cancer; CRC-AIM, Colorectal Cancer-Adenoma Incidence and Mortality model; CRC-SPIN, Colorectal Cancer Simulated Population Incidence and Natural history model; MISCAN, MIcrosimulation SCreening Analysis; SimCRC, simulation model of colorectal cancer; UKFSST, United Kingdom Flexible Sigmoidoscopy Screening Trial.

We demonstrated the validity of calibrated parameters by comparing model-predicted outcomes such as CRC incidence, LYG, and CRC mortality reduction due to screening to those reported by the 3 established CISNET CRC models. Model-predicted LYG and CRC incidence and mortality reduction resulting from the evaluated CRC screening strategies were within CISNET models' predictions. In addition, we showed that the calibrated CRC-AIM estimated a reduction in CRC incidence and mortality from 1-time sigmoidoscopy screening similar to that reported by the UKFSS trial, representing the external validity of the model.

There is a growing interest in improving the calibration process for microsimulation models.[62–66] Since the early 2000s, when the conventional approaches to calibration were trial and error,[67] maximum likelihood–based methods,[68–70] and grid or random search,[71] novel methods are increasingly being introduced. Hazelbag et al.[63] reviewed 84 publications that used calibration methods in simulation models. Only 40 of the models reported a search strategy that was further classified into optimization and sampling algorithms. Among optimization methods, grid search[72–74] and iterative optimization algorithms[75] were most commonly used. Examples of iterative optimization algorithms (e.g., meta-heuristic methods) that have been extensively used for calibration are genetic algorithms, simulated annealing, and particle swarm optimization. The methods have been suggested as a way to shorten the calibration time.[1,5,76,77] However, the complexity of the proposed methods has limited their use in real-world applications. Heuristic algorithms start from an initial input vector and sequentially update the vector by exploring the neighboring solutions at each iteration. The sequential nature of these algorithms and the likelihood of converging to a local optimum are the biggest limitations of these heuristic approaches. Compared with these methods, our approach enables parallel computations, thus achieving computational feasibility and time efficiency. We note that meta-models and emulators have also been used in simulation models for purposes other than calibration, such as conducting cost-effectiveness and value-of-information analyses or developing online decision support tools.[78,79]

Another class of search strategies for calibration involves statistical and sampling methods,[63] which includes Bayesian calibration with several variations such as Bayesian melding,[80] Sampling Importance Resampling, Rejection Approximate Bayesian Computation (ABC), and Incremental Mixture Importance Sampling (IMIS).[81] Volpatto et al.[9] and Wade et al.[82] used Bayesian calibration with Cascading Adaptive Transitional Metropolis in Parallel (CATMIP)[83] for parallel sampling. Ryckman et al.[64] considered 3 calibration techniques comparing random search to Bayesian calibration with the sampling-importance-resampling algorithm and IMIS to model the natural history of cholera and showed that Bayesian calibration with IMIS provided the best model fit while requiring the most computational resources.

Among Bayesian calibration methods, ABC received more attention, with Shewmaker et al.[7] and Niyukuri et al.[6] using ABC rejection sampling. ABC offers a likelihood-free method that provides an estimate from the posterior distribution by choosing parameters that closely fit the data. However, ABC can be inefficient when the number of unknown parameters for calibration is large or many calibration targets are involved. Also, these models are sensitive to the differences between prior and posterior distributions. Slipher and Carnegie[84] explored parameter calibration in epidemic network models using 2 search strategies: LHS and ABC. They found that parameter estimation with LHS is more dispersed and better covers the entire parameter space, while approximate Bayesian inference creates a focused distribution of values and is more computationally efficient. To overcome some of the shortcomings of ABC, the Bayesian Calibration using Artificial Neural Networks (BayCANN) framework was recently proposed.[10] BayCANN estimates the posterior joint distribution of calibrated parameters using neural networks and was 5 times faster than the IMIS algorithm.[10]

Compared with the present study, BayCANN included a smaller number of unknown parameters (9 inputs) while predicting a large number of outcomes (36). For a better comparison between our method and the Bayesian calibration with neural network emulators, we used the open-source code of BayCANN for our calibration experiment (Supplementary Section E.1). We observed that our method was more successful than BayCANN in matching calibration targets. Furthermore, our method generates a set of heterogeneous input vectors rather than clustered inputs, which may be more helpful when dealing with uncertainty. However, we also recognize that BayCANN was previously tested on multiple models and therefore has more promise for generalizability. Thus, while our approach appears to work well for our problem, its potential performance for other simulation models is unknown.

Recent calibration literature advocates for the use of ML algorithms and their efficiency. Chopra et al.[85] and Anirudh et al.[86] used neural networks to calibrate simulation models. These studies did not compare their calibration method with other ML algorithms but found neural network framework to be effective for calibration. Angione et al.[87] compared several ML algorithms (e.g., linear regression, support vector machines, neural networks) for an agent-based model of social care provision in the United Kingdom and found ML-based metamodels can facilitate robust sensitivity analyses while reducing computational time. However, this proof-of-concept study predicted only a single outcome of interest, rather than multiple outcomes concurrently. Sai et al.[88] and Reiker et al.[89] used Gaussian process for calibration. Reiker and colleagues proposed an optimization framework employing Gaussian process as a ML emulator function to calibrate a complex malaria transmission simulator.[89]

Similar to our study, Cevik et al.[2] demonstrated how an active learning-based algorithm could accelerate natural history calibration in microsimulation model, specifically a CISNET breast cancer model. However, that active learning algorithm required a feedback mechanism between the ML and microsimulation models, whereas our framework used the microsimulation model only to provide inputs to the ML algorithm. Therefore, unlike the study by Cevik et al.,[2] our framework does not require specifying a stopping condition to end the feedback mechanism between the ML and microsimulation models. Furthermore, our ML algorithm incorporated multiple calibration targets rather than a single calibration target, and such differences may have led to performance differences between the 2 studies.

We showed that the choice and number of calibration targets and the differential weights applied to them affected modeled outcomes. Because several input combinations generated outcomes close to the targets, adding cross-model targets for validation of our complex simulation model was crucial in identifying the final set of inputs. We demonstrated that if our study had relied on only 1 of the primary calibration targets sets instead of using all 3 of them, the model's LYG predictions would have been substantially different, demonstrating the impact of calibration target selection on model predictions. In fact, even the use of SEER data, the most comprehensive population-based calibration target for cancer modeling, was by itself insufficient to identify a model that passed the cross-model validation and external validation. The findings suggest the importance of establishing cross-model and external validity to obtain a robust set of input combinations that is best supported by all available evidence. To the best of our knowledge, no previous cancer simulation study has demonstrated the impact of choosing calibration targets on long-term model outcomes such as LYG, prohibiting direct comparison to these studies. Our findings suggest that modelers and policy makers may need to conduct a sensitivity analysis on the calibration targets to assess the robustness of the conclusions drawn from modeling studies and the uncertainties in the natural history of the disease. Such structural sensitivity analyses experiments and robust decision-making approaches could be useful for model development.

Unlike many of the calibration studies in the literature, we identified a set of input vectors that have acceptable performance in terms of calibration and validation targets. The selection of multiple input vectors as opposed to a single input vector provides an opportunity to evaluate the impact of heterogeneity and uncertainty in directly unobservable natural history parameters on final model outcomes. We identified an input vector with good fit to the outcomes explored in the natural history, but the LYG outcomes did not compare well with the CISNET model predictions (model U5 in Supplementary Table 14). While this vector may be regarded as unacceptable simply due to its failure to demonstrate cross-validity, we recognize that the choice of comparing well to CISNET models may appear arbitrary and suggest that models that predicted outcomes that are out of range must not be plausible. Excluding this vector may jeopardize the goal of obtaining robust conclusions. While we plan to use the best input vector for the base-case analyses, multiple input vectors will be used to conduct a structural and parametric sensitivity analysis of the natural history parameters in future experiments.

This study has several limitations. One of the challenges in calibrating complex cancer simulation models is overidentification. To assess the level of overidentification in our study, we plotted the distribution of the model parameters corresponding to the best-fitting 500K input combinations from our emulator, as shown in Supplementary Figure 5. While some parameters are tightly clustered (e.g., the mean and standard deviation of the baseline risk for developing adenoma and the impact of adenoma size in colon and rectum on transitioning of adenoma to preclinical cancer), no clustering was observed for other parameters (e.g., the impact of person's age at the time of adenoma initiation on transition of adenoma to preclinical cancer). The heterogeneity of our final selected sets of inputs also indicates that there might be multiple solutions for our calibration problem. Alarid-Escudero et al.[90] and Ryckman et al.[64] suggested using additional calibration targets, narrowing prior distribution ranges, and weighting the GOF function, as some methods to address nonidentifiability, which also applies to overidentification. Identifying and addressing the degree of overidentification solutions with metamodels is a topic of interest and recommended for future research.

While our approach has impressive empirical performance, it is not based on rigorous statistical methodology. Also, other sections of our calibration procedure such as input selection scoring system, targets' weights and importance, and selection of additional cross-validation target were based on empirical evidence rather than on a theoretical framework, which can be further investigated with more theoretical approaches. While empirical, we showed that having a sufficiently large sample for prediction, the difference between the emulator and statistical methods such as Bayesian calibration may be minimal. Although substantial time may need to be spent to fine-tune and tailor the DNN to a specific simulation model including hyperparameter tuning, the overall calibration time can be reduced. Unlike Bayesian calibration models, our method is not capable of producing conventional posterior distribution and uncertainty bounds around the estimates. However, our method generates a set of heterogeneous input vectors rather than clustered inputs, as discussed earlier. Furthermore, identifying the correlation between inputs and calibration targets is not trivial when using a DNN emulator. While some ML algorithms such as random forests generate the correlation of inputs to outputs, such a task is computationally burdensome for other models such as DNN, with many deep layers and thousands of hidden parameters.[91–93] There are methods that approximate the importance of inputs in DNN[94,95] but are beyond the scope of this research.

While we investigated the importance of calibration target selection in our research, we did not quantify uncertainty and incompatibility of calibration targets in cancer simulation modeling. Mandrik et al.[66] discussed methods for dealing with biased calibration targets, including adjustment of target means and standard errors to account for sampling uncertainty and data incompatibility. Further investigation is required to understand the efficiency of ML compared with Bayesian calibration when calibration data are incomplete or biased. Note that prior information about the input parameters from the CRC-SPIN model, which was used to design our model structure, may have helped us identify high-quality inputs relatively quickly. Therefore, our approach may not work efficiently on models in which there is no prior information. In terms of modeling the natural history of CRC, CRC-AIM does not consider the modeling of CRCs that occur through the sessile serrated pathway (SSP), which is a major limitation. Approximately 14% to 30%[96–98] of CRCs are estimated to arise from sessile serrated lesions and polyps, which develop mainly via the CpG island methylation pathway.[99,100] In fact, several CRC simulation models considered both adenoma-carcinoma and SSP pathways and have been extensively validated through several clinical trials.[101–103] Finally, further work is needed to demonstrate that CRC-AIM predictions approximate CISNET CRC model predictions for other screening scenarios and modalities.

In summary, this study showed that the use of powerful DNNs as an emulator could significantly speed up calibration for complex cancer microsimulation models with extensive computational requirements.

Acknowledgments

Will Johnson and Feyza Sancar of Exact Sciences provided writing assistance. We thank the following Exact Sciences employees: Burak A. Ozbay for providing supervision and technical support for this research and Darl Flake for conducting statistical analysis. The analytic methods and study materials may be available upon request. Any inquires should be submitted to crcaim@exactsciences.com.

ORCID iDs

Vahab Vahdat https://orcid.org/0000-0001-7831-7567
Oguzhan Alagoz https://orcid.org/0000-0002-5133-1382
Jing Voon Chen https://orcid.org/0000-0003-3632-6806

Supplemental Material

Supplementary material for this article is available online at https://doi.org/10.1177/0272989X231184175.

References

1. 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.
2. Cevik M, Ergun MA, Stout NK, Trentham-Dietz A, Craven M, Alagoz O. Using active learning for speeding up calibration in simulation models. *Med Decis Making*. 2016;36:581–93.
3. Lidbe AD, Hainen AM, Jones SL. Comparative study of simulated annealing, tabu search, and the genetic algorithm for calibration of the microsimulation model. *Simulation*. 2017;93:21–33.
4. Barbosa C, Dowd WN, Aldridge AP, Timko C, Zarkin GA. Estimating long-term drinking patterns for people with lifetime alcohol use disorder. *Med Decis Making*. 2019;39:765–80. DOI: 10.1177/0272989x19873627
5. Kong CY, McMahon PM, Gazelle GS. Calibration of disease simulation model using an engineering approach. *Value Health*. 2009;12:521–29.
6. Niyukuri D, Chibawara T, Nyasulu PS, Delva W. Inferring HIV transmission network determinants using agent-based models calibrated to multi-data sources. *Mathematics*. 2021;9:2645. DOI: 10.3390/math9212645
7. Shewmaker P, Chrysanthopoulou SA, Iskandar R, Lake D, Jutkowitz E. Microsimulation model calibration with approximate Bayesian computation in R: a tutorial. *Med Decis Making*. 2022;42:557–70. DOI: 10.1177/0272989x221085569
8. Menzies NA, Soeteman DI, Pandya A, Kim JJ. Bayesian methods for calibrating health policy models: a tutorial. *Pharmacoeconomics*. 2017;35:613–24. DOI: 10.1007/s40273-017-0494-4
9. Volpatto DT, Resende ACM, dos Anjos L, et al. A generalised SEIRD model with implicit social distancing mechanism: a Bayesian approach for the identification of the spread of COVID-19 with applications in Brazil and Rio de Janeiro state. *J Simul*. 2023;17:178–192. DOI: 10.1080/17477778.2021.1977731
10. Jalal H, Trikalinos TA, Alarid-Escudero F. BayCANN: streamlining Bayesian calibration with artificial neural network metamodeling. *Front Physiol*. 2021;12:662314.
11. American Cancer Society. Key statistics for colorectal cancer. 2022. Available from: https://www.cancer.org/cancer/colon-rectal-cancer/about/key-statistics.html [Accessed 7 June, 2022].
12. Edwards BK, Ward E, Kohler BA, et al. Annual report to the nation on the status of cancer, 1975-2006, featuring colorectal cancer trends and impact of interventions (risk factors, screening, and treatment) to reduce future rates. *Cancer*. 2010;116:544–73. DOI: 10.1002/cncr.24760
13. US Preventive Services Task Force. Screening for colorectal cancer: US Preventive Services Task Force recommendation statement. *JAMA*. 2021;325:1965–77. DOI: 10.1001/jama.2021.6238
14. Wolf AMD, Fontham ETH, Church TR, et al. Colorectal cancer screening for average-risk adults: 2018 guideline update from the American Cancer Society. *CA Cancer J Clin*. 2018;68:250–81. DOI: 10.3322/caac.21457
15. Shaukat A, Kahi CJ, Burke CA, Rabeneck L, Sauer BG, Rex DK. ACG clinical guidelines: colorectal cancer screening 2021. *Am J Gastroenterol*. 2021;116:458–79. DOI:10.14309/ajg.0000000000001122
16. Knudsen AB, Rutter CM, Peterse EF, et al. Colorectal cancer screening: an updated modeling study for the US Preventive Services Task Force. *JAMA*. 2021;325:1998–2011.
17. 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;315:2595–609. DOI: 10.1001/jama.2016.6828
18. Rutter CM, Miglioretti DL, Savarino JE. Bayesian calibration of microsimulation models. *J Am Stat Assoc*. 2009; 104:1338–50.
19. Rutter CM, Ozik J, DeYoreo M, Collier N. Microsimulation model calibration using incremental mixture approximate Bayesian computation. *Ann Appl Stat*. 2019;13:2189–212.
20. Koh K-J, Lin L-H, Huang S-H, Wong J-U. CARE—pediatric colon adenocarcinoma: a case report and literature review comparing differences in clinical features between children and adult patients. *Medicine (Baltimore)*. 2015;94: e503.
21. Tjørve E, Tjørve KM. A unified approach to the Richards-model family for use in growth analyses: why we need only two model forms. *J Theor Biol*. 2010;267:417–25.
22. Ramos PL, Louzada F, Ramos E, Dey S. The Fréchet distribution: estimation and application - An overview. *J Stat Manag Syst*. 2020;23:549–78. DOI: 10.1080/09720510.2019.1645400
23. Nusko G, Mansmann U, Altendorf-Hofmann A, Groitl H, Wittekind C, Hahn EG. Risk of invasive carcinoma in colorectal adenomas assessed by size and site. *Int J Colorectal Dis*. 1997;12:267–71.
24. Yamaji Y, Mitsushima T, Yoshida H, et al. The malignant potential of freshly developed colorectal polyps according to age. *Cancer Epidemiol Biomarkers Prev*. 2006;15:2418–21.
25. National Cancer Institute. Surveillance, Epidemiology, and End Results (SEER) Program. SEER*Stat Database: Incidence - SEER Research Data, 9 Registries, Nov 2020 Sub (1975-2018) - Linked To County Attributes - Time Dependent (1990-2018) Income/Rurality, 1969-2019 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, released April 2021, based on the November 2020 submission. 2021. Available from: www.seer.cancer.gov.
26. Rutter CM, Johnson EA, Feuer EJ, Knudsen AB, Kuntz KM, Schrag D. Secular trends in colon and rectal cancer relative survival. *J Natl Cancer Inst*. 2013;105:1806–13.

27. Arias E, Xu JQ. United States life tables, 2017. National Vital Statistics Reports; vol 68 no 7. Hyattsville, MD: National Center for Health Statistics; 2019.
28. Ladabaum U, Chopra CL, Huang G, Scheiman JM, Chernew ME, Fendrick AM. Aspirin as an adjunct to screening for prevention of sporadic colorectal cancer: a cost-effectiveness analysis. *Ann Intern Med*. 2001;135:769–81.
29. Ladabaum U, Song K. Projected national impact of colorectal cancer screening on clinical and economic outcomes and health services demand. *Gastroenterology*. 2005;129: 1151–62.
30. Telford JJ, Levy AR, Sambrook JC, Zou D, Enns RA. The cost-effectiveness of screening for colorectal cancer. *CMAJ*. 2010;182:1307–13.
31. Vijan S, Hwang I, Inadomi J, et al. The cost-effectiveness of CT colonography in screening for colorectal neoplasia. *Am J Gastroenterol*. 2007;102:380–90.
32. Corley DA, Jensen CD, Marks AR, et al. Variation of adenoma prevalence by age, sex, race, and colon location in a large population: implications for screening and quality programs. *Clin Gastroenterol Hepatol*. 2013;11:172–80.
33. Pickhardt PJ, Choi JR, Hwang I, et al. Computed tomographic virtual colonoscopy to screen for colorectal neoplasia in asymptomatic adults. *N Engl J Med*. 2003;349: 2191–200.
34. Imperiale TF, Wagner DR, Lin CY, Larkin GN, Rogge JD, Ransohoff DF. Risk of advanced proximal neoplasms in asymptomatic adults according to the distal colorectal findings. *N Engl J Med*. 2000;343:169–74.
35. Lieberman D, Moravec M, Holub J, Michaels L, Eisen G. Polyp size and advanced histology in patients undergoing colonoscopy screening: implications for CT colonography. *Gastroenterology*. 2008;135:1100–105.
36. Church JM. Clinical significance of small colorectal polyps. *Dis Colon Rectum*. 2004;47:481–5.
37. de Carvalho TM, van Rosmalen J, Wolff HB, Koffijberg H, Coupé VMH. Choosing a metamodel of a simulation model for uncertainty quantification. *Med Decis Making*. 2022;42(1):28–42.
38. Helton JC, Davis FJ. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliab Eng Syst Safe*. 2003;81:23–69.
39. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Cambridge (MA): MIT Press; 2016.
40. Breiman L. Random forests. *Mach Learn*. 2001;45:5–32. DOI: 10.1023/A:1010933404324
41. Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal*. 2002;38:367–78.
42. Natekin A, Knoll A. Gradient boosting machines, a tutorial. *Front Neurorobot*. 2013;7:21.
43. Chen T, Guestrin C.Xgboost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: Association for Computing Machinery; 2016. p 785–94.
44. Ke G, Meng Q, Finley T, et al. Lightgbm: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst*. 2017;30: 1–11. Available from: https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html.
45. Dorogush AV, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support. *arXiv Preprint arXiv:1810.11363*, 2018.
46. Knudsen AB, Rutter CM, Peterse EFP, et al. *Colorectal Cancer Screening: An Updated Decision Analysis for the US Preventive Services Task Force*. U.S. Preventive Services Task Force Evidence Syntheses, formerly Systematic Evidence Reviews. Rockville (MD): Agency for Healthcare Research and Quality; 2021.
47. Siegel RL, Fedewa SA, Anderson WF, et al. Colorectal cancer incidence patterns in the United States, 1974–2013. *J Natl Cancer Inst*. 2017;109: 1–6.
48. US Preventive Services Task Force. Final recommendation statement. Colorectal cancer: screening. 2021. Available from: https://www.uspreventiveservicestaskforce.org/uspstf/recommendation/colorectal-cancer-screening [Accessed 29 June, 2022].
49. Atkin WS, Edwards R, Kralj-Hans I, et al. Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: a multicentre randomised controlled trial. *Lancet*. 2010;375:1624–33. DOI: 10.1016/S0140-6736(10)60551-X
50. Atkin W, Wooldrage K, Parkin DM, et al. Long term effects of once-only flexible sigmoidoscopy screening after 17 years of follow-up: the UK Flexible Sigmoidoscopy Screening randomised controlled trial. *Lancet*. 2017;389: 1299–311. DOI: 10.1016/S0140-6736(17)30396-3
51. Atkin WS, Cook CF, Cuzick J, Edwards R, Northover JM, Wardle J; UK Flexible Sigmoidoscopy Screening Trial Investigators. Single flexible sigmoidoscopy screening to prevent colorectal cancer: baseline findings of a UK multicentre randomised trial. *Lancet*. 2002;359:1291–300. DOI: 10.1016/s0140-6736(02)08268-5
52. Lu B, Wang L, Lu M, et al. Microsimulation model for prevention and intervention of coloretal cancer in China (MIMIC-CRC): development, calibration, validation, and application. *Front Oncol*. 2022;12:883401.
53. Thomas C, Whyte S, Kearns B, Chilcott JB. External validation of a colorectal cancer model against screening trial long-term follow-up data. *Value Health*. 2019;22:1154–61.
54. Tappenden P, Chilcott J, Eggington S, Patnick J, Sakai H, Karnon J. Option appraisal of population-based colorectal cancer screening programmes in England. *Gut*. 2007;56: 677–84.
55. Areia M, Mori Y, Correale L, et al. Cost-effectiveness of artificial intelligence for screening colonoscopy: a modelling study. *Lancet Digit Health*. 2022;4:e436–44. DOI: 10.1016/s2589-7500(22)00042-5
56. Office of National Statistics. National life tables, United Kingdom, 1996-1998. 2020. Available from: https://www.ons.gov.uk/file?uri = /peoplepopulationandcommunity/birth

sdeathsandmarriages/lifeexpectancies/datasets/nationallifetablesunitedkingdomreferencetables/current/previous/v5/nationallifetables3yearuk.xls [Accessed 27 May, 2022].

57. Kingma DP, Ba J. Adam: a method for stochastic optimization. *arXiv Preprint arXiv:1412.6980*, 2014.
58. Schrag D. AJCC staging: staging colon cancer patients using the TNM system from 1975 to the present. 2007. Available from: https://www.mskcc.org/departments/epidemiology-biostatistics/epidemiology/ajcc-staging [Accessed 7 June, 2022].
59. Brenner H, Altenhofen L, Katalinic A, Lansdorp-Vogelaar I, Hoffmeister M. Sojourn time of preclinical colorectal cancer by sex and age: estimates from the German national screening colonoscopy database. *Am J Epidemiol*. 2011; 174:1140–6. DOI: 10.1093/aje/kwr188
60. Chen TH, Yen MF, Lai MS, et al. Evaluation of a selective screening for colorectal carcinoma: the Taiwan Multicenter Cancer Screening (TAMCAS) project. *Cancer*. 1999;86: 1116–28.
61. 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;36:604–14.
62. Reddy KP, Bulteel AJ, Levy DE, et al. Novel microsimulation model of tobacco use behaviours and outcomes: calibration and validation in a US population. *BMJ Open*. 2020;10:e032579.
63. Hazelbag CM, Dushoff J, Dominic EM, Mthombothi ZE, Delva W. Calibration of individual-based models to epidemiological data: a systematic review. *PLoS Comput Biol*. 2020;16:e1007893.
64. Ryckman T, Luby S, Owens DK, Bendavid E, Goldhaber-Fiebert JD. Methods for model calibration under high uncertainty: modeling cholera in Bangladesh. *Med Decis Making*. 2020;40:693–709.
65. DeYoreo M, Rutter CM, Ozik J, Collier N. Sequentially calibrating a Bayesian microsimulation model to incorporate new information and assumptions. *BMC Med Inform Decis Mak*. 2022;22:1–14.
66. Mandrik O, Thomas C, Whyte S, Chilcott J. Calibrating natural history of cancer models in the presence of data incompatibility: problems and solutions. *Pharmacoeconomics*. 2022;40:359–66.
67. Myers ER, McCrory DC, Nanda K, Bastian L, Matchar DB. Mathematical model for the natural history of human papillomavirus infection and cervical carcinogenesis. *Am J Epidemiol*. 2000;151:1158–71.
68. Loeve F, Boer R, Zauber AG, et al. National Polyp study data: evidence for regression of adenomas. *Int J Cancer*. 2004;111:633–9.
69. Salomon JA, Weinstein MC, Hammitt JK, Goldie SJ. Empirically calibrated model of hepatitis C virus infection in the United States. *Am J Epidemiol*. 2002;156:761–73.
70. Chia YL, Salzman P, Plevritis SK, Glynn PW. Simulation-based parameter estimation for complex models: a breast cancer natural history modelling illustration. *Stat Methods Med Res*. 2004;13:507–24.
71. Knudsen AB. *Explaining Secular Trends in Colorectal Cancer Incidence and Mortality with an Empirically-Calibrated Microsimulation Model*. Cambridge (MA): Harvard University; 2005.
72. Luo W, Katz DA, Hamilton DT, et al. Development of an agent-based model to investigate the impact of HIV self-testing programs on men who have sex with men in Atlanta and Seattle. *JMIR Public Health Surveill*. 2018;4:e58. DOI: 10.2196/publichealth.9357
73. Marshall BDL, Goedel WC, King MRF, et al. Potential effectiveness of long-acting injectable pre-exposure prophylaxis for HIV prevention in men who have sex with men: a modelling study. *Lancet HIV*. 2018;5:e498–505. DOI: 10.1016/s2352-3018(18)30097-3
74. Goedel WC, King MRF, Lurie MN, Nunn AS, Chan PA, Marshall BDL. Effect of racial inequities in pre-exposure prophylaxis use on racial disparities in HIV Incidence among men who have sex with men: a modeling study. *J Acquir Immune Defic Syndr*. 2018;79:323–9. DOI: 10.1097/qai.0000000000001817
75. Kasaie P, Berry SA, Shah MS, et al. Impact of providing preexposure prophylaxis for human immunodeficiency virus at clinics for sexually transmitted infections in Baltimore city: an agent-based model. *Sex Transm Dis*. 2018;45: 791–7. DOI: 10.1097/olq.0000000000000882
76. Lee JM, McMahon PM, Kong CY, et al. Cost-effectiveness of breast MR imaging and screen-film mammography for screening BRCA1 gene mutation carriers. *Radiology*. 2010; 254:793–800. DOI: 10.1148/radiol.09091086
77. Wang Z, Zhang Q, Wu B. Development of an empirically calibrated model of esophageal squamous cell carcinoma in high-risk regions. *Biomed Res Int*. 2019;2019:2741598. DOI: 10.1155/2019/2741598
78. McCandlish JA, Ayer T, Chhatwal J. Cost-effectiveness and value-of-information analysis using machine learning-based metamodeling: a case of hepatitis C treatment. *Med Decis Making*. 2023;43(1):68–77. DOI: 10.1177/0272989x221125418
79. Charles G, Wolock TM, Winskill P, Ghani A, Bhatt S, Flaxman S. Seq2Seq surrogates of epidemic models to facilitate Bayesian inference. *arXiv Preprint arXiv:2209.09617*, 2022.
80. McCormick AW, Abuelezam NN, Fussell T, Seage GR 3rd, Lipsitch M. Displacement of sexual partnerships in trials of sexual behavior interventions: a model-based assessment of consequences. *Epidemics*. 2017;20:94–101. DOI: 10.1016/j.epidem.2017.03.007
81. Suboi Z, Hladish TJ, Delva W, Hazelbag CM. Calibration of models to data: a comparison of methods. *bioRxiv*, 2020. DOI: 10.1101/2020.12.21.423763
82. Wade S, Weber MF, Sarich P, et al. Bayesian calibration of simulation models: a tutorial and an Australian smoking behaviour model. *arXiv Preprint arXiv:2202.02923*, 2022. DOI: 10.48550/ARXIV.2202.02923
83. Minson SE, Simons M, Beck JL. Bayesian inversion for finite fault earthquake source models I—theory and algorithm. *Geophys J Int*. 2013;194:1701–26. DOI: 10.1093/gji/ggt180

84. Slipher SK, Carnegie NB. Model calibration in network models of HIV. In: *epiDAMIK*, 2021; Virtual.
85. Chopra A, Rodríguez A, Subramanian J, Krishnamurthy B, Prakash BA, Raskar R. Differentiable agent-based epidemiological modeling for end-to-end learning [online]. In: *ICML 2022 Workshop AI for Agent-Based Modelling*, 2022.
86. Anirudh R, Thiagarajan JJ, Bremer P-T, Germann T, Del Valle S, Streitz F. Accurate calibration of agent-based epidemiological models with neural network surrogates. In: Xu P, Zhu T, Zhu P, Clifton DA, Belgrave D, Zhang Y, eds. *Proceedings of the 1st Workshop on Healthcare AI and COVID-19, ICML 2022*. Proceedings of Machine Learning Research (PMLR), PMLR, 2022, p 54–62. Available from: https://proceedings.mlr.press/v184/anirudh22a
87. Angione C, Silverman E, Yaneske E. Using machine learning to emulate agent-based simulations. *arXiv Preprint arXiv:2005.02077*, 2020.
88. Sai A, Vivas-Valencia C, Imperiale TF, Kong N. Multiobjective calibration of disease simulation models using Gaussian processes. *Med Decis Making*. 2019;39:540–52. DOI: 10.1177/0272989x19862560
89. Reiker T, Golumbeanu M, Shattock A, et al. Emulator-based Bayesian optimization for efficient multi-objective calibration of an individual-based model of malaria. *Nat Commun*. 2021;12:1–11.
90. 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;38:810–21.
91. Garson GD. Interpreting neural network connection weights. *Artif Intell Expert*. 1991;6:46–51.
92. Goh ATC. Back-propagation neural networks for modeling complex systems. *Artif Intell Eng*. 1995;9:143–51. DOI: 10.1016/0954-1810(94)00011-S
93. Rinke W. An algorithm to transform an artificial neural network into its open equation form and its potential applications. *Int J Neural Netw Adv Appl*. 2015;2:28–33.
94. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016; San Francisco.
95. Ahern I, Noack A, Guzman-Nateras L, Dou D, Li B, Huan J. NormLime: a new feature importance metric for explaining deep neural networks. *arXiv Preprint arXiv: 1909.04200*, 2019.
96. Imperiale TF, Ransohoff DF, Itzkowitz SH, et al. Multitarget stool DNA testing for colorectal-cancer screening. *N Engl J Med*. 2014;370:1287–97. DOI: 10.1056/NEJMoa1311194
97. Sweetser S, Smyrk TC, Sinicrope FA. Serrated colon polyps as precursors to colorectal cancer. *Clin Gastroenterol Hepatol*. 2013;11:760–7; quiz e754-765. DOI: 10.1016/j.cgh.2012.12.004
98. East JE, Vieth M, Rex DK. Serrated lesions in colorectal cancer screening: detection, resection, pathology and surveillance. *Gut*. 2015;64:991–1000. DOI: 10.1136/gutjnl-2014-309041
99. Hawkins N, Norrie M, Cheong K, et al. CpG island methylation in sporadic colorectal cancers and its relationship to microsatellite instability. *Gastroenterology*. 2002;122:1376–87. DOI: 10.1053/gast.2002.32997
100. Samowitz WS, Albertsen H, Herrick J, et al. Evaluation of a large, population-based sample supports a CpG island methylator phenotype in colon cancer. *Gastroenterology*. 2005;129:837–45. DOI: 10.1053/j.gastro.2005.06.020
101. Greuter MJ, Xu XM, Lew JB, et al. Modeling the adenoma and serrated pathway to colorectal cancer (ASCCA). *Risk Anal*. 2014;34:889–910. DOI: 10.1111/risa.12137
102. Lew J-B, Greuter MJ, Caruana M, et al. Validation of microsimulation models against alternative model predictions and long-term colorectal cancer incidence and mortality outcomes of randomized controlled trials. *Med Decis Making*. 2020;40:815–829.
103. Lew JB, St John DJB, Xu XM, et al. Long-term evaluation of benefits, harms, and cost-effectiveness of the National Bowel Cancer Screening Program in Australia: a modelling study. *Lancet Public Health*. 2017;2:e331–40. DOI: 10.1016/s2468-2667(17)30105-6