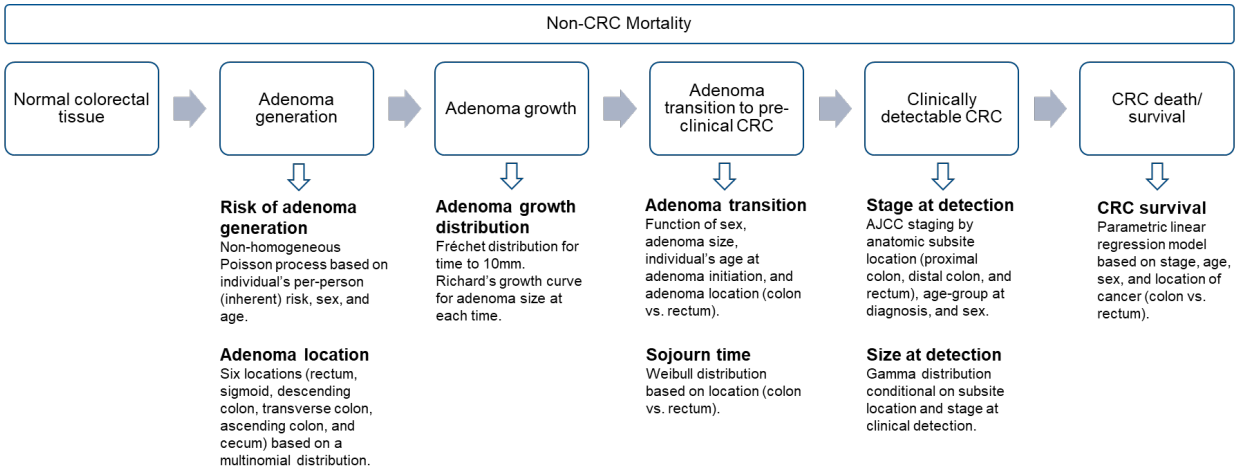
CRC-AIM model structure

Supplemental Figure 1. Overview of the CRC-AIM natural history model (adapted from Vahdat et al.[1])



Abbreviations: AJCC, American Joint Committee on Cancer; CRC, colorectal cancer; CRC-AIM, Colorectal Cancer-Adenoma Incidence and Mortality model.

CRC-AIM models the natural history of colorectal cancer (CRC) using an adenoma-carcinoma sequence. The natural history of CRC is comprised of five components: (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 (**Supplemental Figure 1**). Individuals may develop one or more adenomas, which may transition into a preclinical cancer as it grows. A preclinical cancer may ultimately transition into a clinically detectable cancer, and then lead to CRC-specific mortality. CRC-AIM does not model CRCs that occur through sessile serrated pathway (SSP), which is a major limitation of our model. Approximately 14% - 30% [2-4] of CRCs are estimated to arise from sessile serrated lesions and polyps, which develop mainly via the CpG island methylation pathway.[5, 6] In addition, the CRC screening modalities vary substantially in the detection of SSPs, since serrated polyps are less likely to bleed compared to adenoma.[2] Also, our model does not represent the CRC-related events experienced by individuals who are at high risk such as patients with inflammatory bowel disease (e.g., Crohn's disease, ulcerative colitis) and those with a personal or family history of CRC. This is because such patients are likely to have a different natural history of CRC and require different screening/surveillance strategies than average-risk individuals.

The ColoRectal Cancer Simulated Population Incidence and Natural history (CRC-SPIN) model (one of three CISNET CRC models) inspired the development of CRC-AIM, and the models share features (as obtained or derived from publicly available sources).[7, 8]

However, in addition to some structural differences, almost all the parameter values differ between CRC-AIM and CRC-SPIN as explained in detail in "Differences between CRC-AIM and CRC-SPIN."

In the remainder of this section, we provide details of each five subcomponents of CRC-AIM.

Adenoma generation

The risk of developing adenoma is assumed to depend on individuals' sex, age and baseline risk, with individuals aged ≤20 years not at risk for adenoma development.[9] The risk of generation an adenoma is governed by a non-homogenous Poisson process based on baseline risk (varying by individuals), sex, and age. We introduce the following notation:

- $r_i(a)$: risk of developing an adenoma for individual i at age a;
- $I_F(i)$:indicator function, which is equal to 1 when individual i is female and 0 otherwise;
- $I_{[a_0,\infty)}(a)$: indicator function that is equal 1 if the invidual's age $a \in [a_0, \infty)$;
- β_j:regression coefficients of the log-linear model.

The functional represention of the adenoma risk model may then be expressed as:

$$\ln(r_i(a)) = \beta_{0i} + \beta_1 I_F(i) + \beta_2 I_{[20,\infty)}(a)\min\{a-20, 30\} + \beta_3 I_{[50,\infty)}(a)\min\{a-50, 10\} + \beta_4 I_{[60,\infty)}(a)\min\{a-60, 10\} + \beta_5 I_{[70,\infty)}(a)(a-70)$$

Upon the generation of an adenoma, its location is determined via a multinomial distribution informed by autopsy studies[10-18] as derived in Rutter et al.[8] (**Supplemental Table 1**).

Supplemental Table 1. Multinomial distribution used for assigning the location of new adenomas

Site	Rectum	Distal colon		Proximal colon		
		Sigmoid	Descending	Transverse	Ascending	Cecum
Probability	0.09	0.24	0.12	0.24	0.23	0.08

Adenoma growth

The diameter in millimeters (mm) of an adenoma j at time t, measured after its initiation, in individual i is modeled using the Richard's growth model[19]:

$$d_{ij}(t) = d_{max}\left[1 + \left(\left(\frac{d_{min}}{d_{max}}\right)^{\frac{1}{p}} - 1\right)e^{-\lambda_{ij}t}\right]^p$$

where p represents the unknown parameter of the growth model, λ_{ij} represents the growth rate of adenoma j in individual i, d_{min} and d_{max} represent the minimum and maximum diameter. The diameter of an adenoma is confined between 1 mm and 50 mm, so that $d_{min} = 1\ mm$ and $d_{max} = 50\ mm$.

The growth rate λ_{ij} is determined stochastically by first sampling the time to reach 10 mm in diameter, denoted as t_{10mm}, which is assumed to follow the Fréchet distribution with scale parameter s_l and shape parameter α_l ($l \in$ {colon, rectum}). The cumulative distribution function (CDF) at time t after the initiation of an adenoma located at l is given by

$$F_l(t) = \exp\left[-\left(\frac{t}{s_l}\right)^{-\alpha_l}\right]$$

Once t_{10mm} is sampled from the CDF, the growth rate can be determined by solving the growth model using $d_{ij}(t_{10mm}) = 10$, leading to:

$$\lambda_{ij} = -\frac{1}{t_{10mm}} \ln \frac{\left(\frac{10}{d_{max}}\right)^{\frac{1}{p}} - 1}{\left(\frac{d_{min}}{d_{max}}\right)^{\frac{1}{p}} - 1}$$

Transition from adenoma to preclinical cancer

The cumulative transition probability of progressing from adenoma to preclinical cancer was modeled using a lognormal cumulative distribution function based on sex, size and age at adenoma initiation. [20, 21] For an adenoma that is initiated at age a located in $l \in \{\text{colon, rectum}\}$ of an individual whose sex is $s \in \{\text{male, female}\}$, the probability of transition to preclinical cancer at or before diameter d is given by:

$$P_{ls}(d, a) = \Phi\left(\frac{\ln(\gamma_{1ls} d) + \gamma_{2ls}(a - 50)}{\gamma_3}\right)$$

where $\Phi(\cdot)$ denotes the standard normal CDF. Following Rutter et al.,[8] γ_3 is set to be 0.5. The annual transition probability from year t to $t + 1$ may be expressed as

$$\frac{P_{ls}(d_{t+1}, a) - P_{ls}(d_t, a)}{1 - P_{ls}(d_t)}$$

where d_t represents the diameter of the adenoma at year t.

Transition from preclinical cancer to clinically detectable cancer (sojourn time)

For colon adenomas, the sojourn time is modeled using a Weibull distribution with shape parameter k and location-specific scale parameter λ_c where the survival function is given by:

$$S_c(t) = \exp\left(-\left(\frac{t}{\lambda_c}\right)^k\right)$$

Assuming proportional hazards between rectal and colon adenomas with the following hazard ratio (HR),

$$\text{HR} = \exp(\alpha)$$

Rectal adenomas also follow Weibull distribution with the following survival distribution

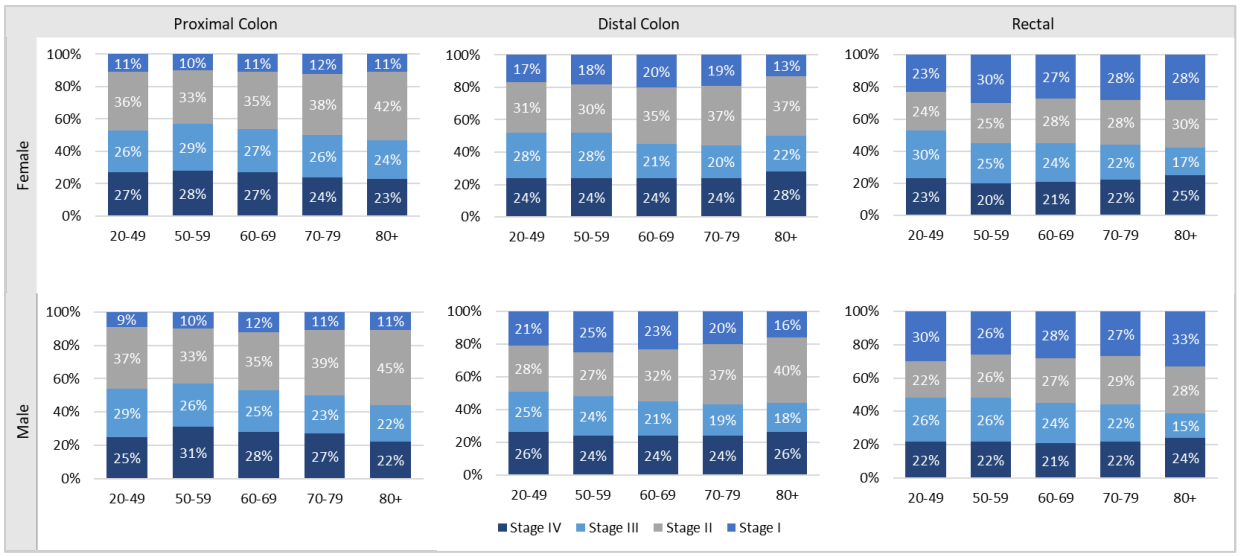
$$S_r(t) = \exp\left(-\left(\frac{t}{\lambda_c \, exp(\alpha)^{-\frac{1}{k}}}\right)^k\right)$$

Here, $\lambda_r = \lambda_c \exp(\alpha)^{-\frac{1}{k}}$. In summary, the sojourn time model for both colon and rectal adenomas has a total of three parameters: λ_c, k, and α.

CRC stage at clinical detection

When a preclinical cancer becomes clinically detectable (i.e., at the end of sojourn time), CRC-AIM first attributes the AJCC stage (5th edition) at clinical detection using a multinomial distribution, stratified by anatomic subsite location (proximal colon, distal colon, and rectum), age-group at diagnosis, and sex. Proximal colon anatomic subsite includes cecum, ascending, hepatic flexure, transverse colon, and splenic flexure. Distal colon consists of descending colon and sigmoid colon. Finally, rectosigmoid colon and rectum are included in rectum. The distributions were derived from SEER 1975-1979 data. Since AJCC staging was not recorded prior to 1988, we adopted the methodology introduced by Schrag[22] for estimating the cancer stages by combining site-specific surgery codes and the staging codes (**Supplemental Figure 2**).

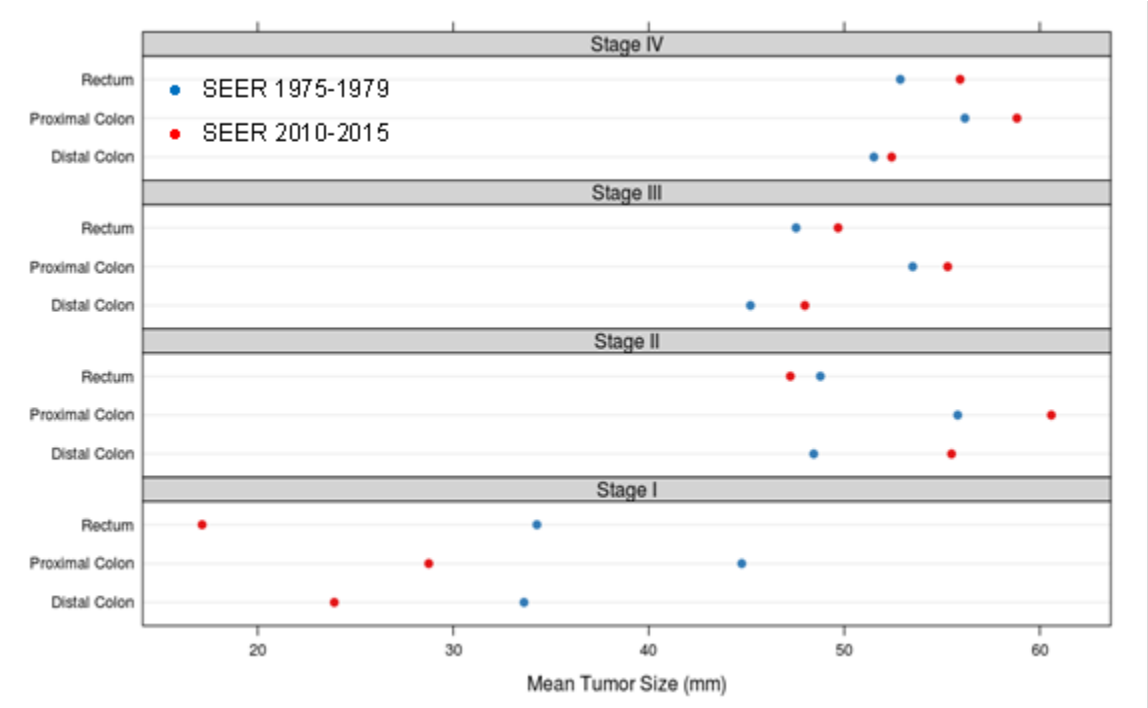
Supplemental Figure 2. Distribution of stage at clinical detection by age, sex, and location according to SEER 1975-1979 data



CRC size at clinical detection

The size at clinical detection, conditional on location and stage at clinical detection, is assigned using SEER 2010-2015 data, limited to cases diagnosed at ages 20-50. The comparison between SEER 2010-2015 data and SEER 1975-1979 data for the same stage (AJCC 5th edition), age group, and anatomic location is presented in **Supplemental Figure 3**. Depending on location, stage I tumors in the 2010-2015 data were 10-15 mm smaller, on average, than the stage I tumors in the 1975-1979 data (**Supplemental Figure 3**). There was no effect of age on tumor size at diagnosis in the 1975-1979 data, nor was there a difference in stage I tumor size between patients younger than 50 years compared with patients 50 years or older. This suggests that in modern times, even without screening, stage I tumors may be diagnosed when they are smaller.

Supplemental Figure 3. Mean cancer size by stage and location among cases diagnosed between ages 20 and 50 according to SEER (SEER 1975-1979 vs. SEER 2010-2015)



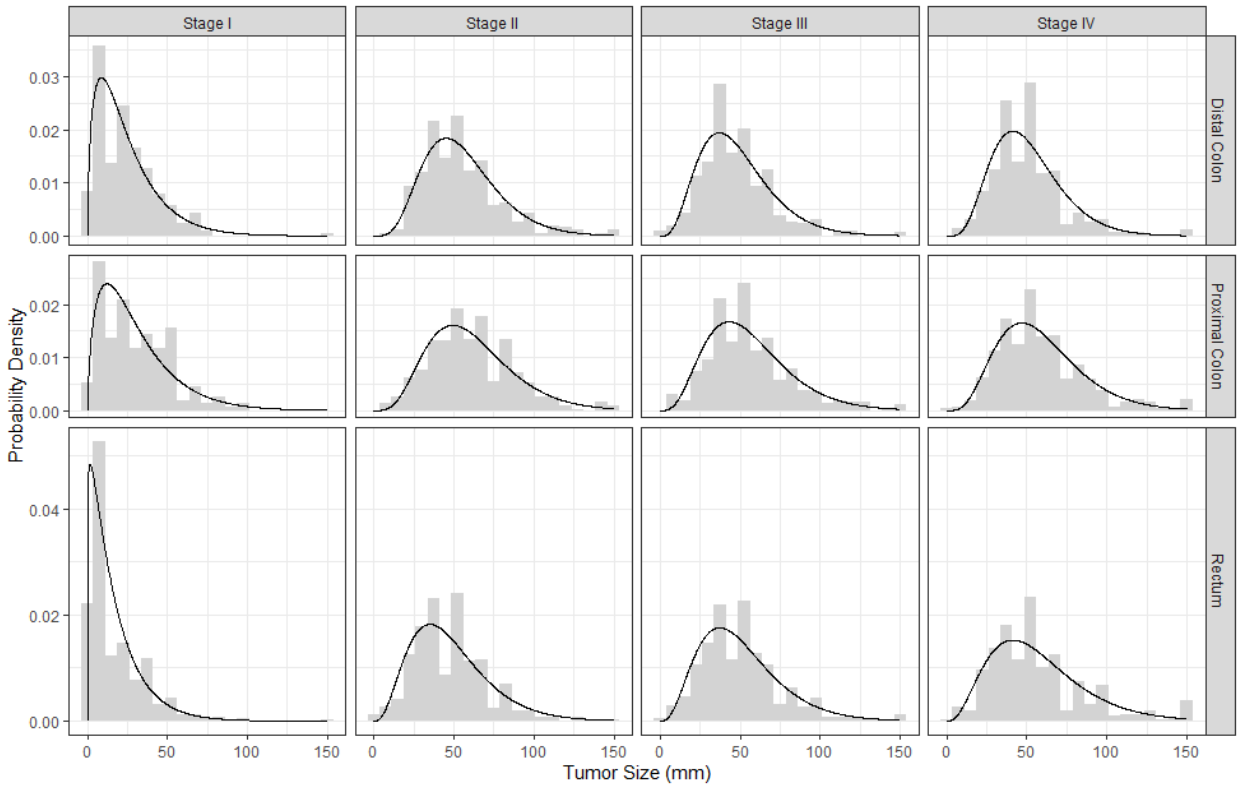
Abbreviations: mm, millimeter; SEER, Surveillance, Epidemiology, and End Results.

The true tumor size at diagnosis by stage and location was assumed to follow a gamma distribution. Within the 8,577 cases retrieved from SEER 2010-2015, tumors larger than 150 mm (n = 29, 0.5%) were set to equal to 150 mm and tumor sizes coded as more than 900 mm (n = 1,820) were excluded. Tumors in the appendix or unspecified locations of the large intestine were excluded (n = 992). After exclusions, 5,910 cases remained for modeling. Right censoring (i.e., sizes greater than 98 mm that are recorded as 98 mm as in the SEER 1975-1979 data) is more likely to skew distribution estimates than interval censoring (i.e., rounding of tumor size data). Because SEER 2010-2015 data were not right-censored, gamma distributions were fit to the data while ignoring the interval censoring using maximum likelihood (**Supplemental Table 2**). **Supplemental Figure 4** shows that gamma distribution fits the data well.

Supplemental Table 2. Estimated parameters of gamma distribution by stage and location that is used to assign tumor size in CRC-AIM

Location	Stage I		Stage II		Stage III		Stage IV	
	Shape	Rate	Shape	Rate	Shape	Rate	Shape	Rate
Distal Colon	1.5545	0.0639	5.5037	0.0998	4.3892	0.0919	5.3333	0.1050
Proximal Colon	1.6695	0.0552	5.1046	0.0833	4.4307	0.0796	4.9185	0.0836
Rectum	1.1040	0.0645	3.7173	0.0777	3.8285	0.0763	3.6448	0.0641

Supplemental Figure 4. Goodness of fit of gamma distribution used to assign tumor size in CRC-AIM



Abbreviations: mm, millimeter.

CRC survival

CRC survival is based on parametric models with sex and age at diagnosis as covariates for each stage and location (colon vs rectum) fitted to SEER-reported cause-specific survival. A 7% hazard reduction was applied based on 5-year cause-specific relative survival between periods 2000-2003 and 2010-2019 from SEER, for cases diagnosed after 2000 to replicate the recent improvement in CRC-specific survival.[23] Other-cause mortality by age and birth-year cohorts were based on the U.S. life tables.[24]

Once an individual is diagnosed with CRC, they are assigned a CRC-specific survival, which is estimates as follows. First, we fit five parametric regression models (Weibull, lognormal, exponential, Fréchet, and loglogistic) to the cause-specific survival data from SEER 2000-2003 for each cancer stage and location (colon vs rectum). Each regression model included two sets of covariates: age at diagnosis (20-49, 50-59, 60-69, 70-79, and ≥80 years) and sex (1 if female and -1 otherwise). We applied right-censoring as appropriate. Any subject with a time of death that was reliably recorded as 0 months was recoded to 0.5 months to prevent model-fitting issues. Among the five regression models, we chose the one based on the smallest corrected Akaike information criterion (AICc) value (**Supplemental Table 3**). Statistical significance of age and sex was based on the Wald test, and the final model (presented in **Supplemental Table 4**) was chosen by refitting the parametric regression models while excluding any covariate that was not statistically significant. Similar analyses were performed for the cause-specific CRC survival

distributions based on the SEER 1975-1979 data, which were used for comparing natural history outcomes against those from the CISNET models (data not shown).

Supplemental Table 3. Corrected Akaike information criterion (AICc) values associated with parametric regression models fit to SEER cause-specific survival

Location	Distribution	Stage I	Stage II	Stage III	Stage IV
Colon	Weibull	**6712.5586***	16392.3346	23078.6953	12638.3198
	Lognormal	6732.2328	**16356.1742***	**22728.0115***	**11934.9884***
	Exponential	6820.2702	16909.8138	23978.1744	14039.0074
	Fréchet	6778.3830	16423.1679	22747.3516	12219.7368
	Loglogistic	6714.0577	16372.8197	22845.1864	12048.4684
Rectum	Weibull	4427.7791	5542.9467	9203.2670	4760.9838
	Lognormal	4431.6023	**5520.6179***	**9061.5251***	**4604.2003***
	Exponential	4439.5690	5607.6837	9279.4864	4958.6266
	Fréchet	4462.4581	5566.6658	9086.8547	4801.3275
	Loglogistic	**4425.4890***	5524.4337	9103.2215	4614.0333

* Represents best fitting model based on smallest AICc.

Supplemental Table 4. Selected parametric regression model by stage and location based on SEER 2000-2003 cause-specific survival

Location	Parameters	Stage I	Stage II	Stage III	Stage IV
Colon	**Selected distribution**	**Weibull**	**Log-normal**	**Log-normal**	**Log-normal**
	Intercept	5.8797	4.5013	2.5361	-0.3766
	Age effect, age ∈ [20,50)	1.6097	0.6193	0.5416	0.6531
	Age effect, age ∈ [50,60)	0.6499	0.3778	0.4263	0.3190
	Age effect, age ∈ [60,70)	-0.0115	0.3243	0.2067	0.0861
	Age effect, age ∈ [70,80)	-0.4863	-0.2226	-0.1463	-0.2930
	Age effect, age ≥ 80	-1.7619	-1.0987	-1.0283	-0.7653
	Main sex effect	0.1321	0.1825	0	0
	Shape parameter	0.6991	2.8316	2.2040	1.5583
Rectum	**Selected distribution**	**Log-logistic**	**Log-normal**	**Log-normal**	**Log-normal**
	Intercept	4.2475	3.4680	2.3558	-0.1707
	Age effect, age ∈ [20,50)	0.4703	0.9608	0.4700	0.5883
	Age effect, age ∈ [50,60)	0.7033	0.4976	0.3899	0.5408
	Age effect, age ∈ [60,70)	0.1593	0.1146	0.3039	0.0489
	Age effect, age ∈ [70,80)	-0.2561	-0.3914	-0.1072	-0.3437
	Age effect, age ≥ 80	-1.0768	-1.1817	-1.0567	-0.8343
	Main sex effect	0	0	0	0
	Shape parameter	0.9026	2.1874	1.7519	1.4349

Abbreviations: SEER, Surveillance, Epidemiology, and End Results.

List of calibrated natural history parameters

Twenty-three directly unobservable parameters governing the natural history of CRC (**Supplemental Table 5)**, were estimated using calibration for CRC-AIM. Parameter calibration began with a plausible range for each parameter, informed by CRC-SPIN.[7, 8] The initial plausible range was then supplemented using our calibration process.

Supplemental Table 5. Unknown parameters of CRC-AIM natural history model (adapted from Vahdat et al.[1])

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).