Health Care Manag Sci (2011) 14:1–21
DOI 10.1007/s10729-010-9138-3

Probability model for estimating colorectal polyp progression rates

Chaitra Gopalappa · Selen Aydogan-Cremaschi · Tapas K. Das · Seza Orcun

Received: 14 November 2009 / Accepted: 13 September 2010 / Published online: 5 October 2010

Abstract According to the American Cancer Society, colorectal cancer (CRC) is the third most common cause of cancer related deaths in the United States. Experts estimate that about 85% of CRCs begin as precancerous polyps, early detection and treatment of which can significantly reduce the risk of CRC. Hence, it is imperative to develop population-wide intervention strategies for early detection of polyps. Development of such strategies requires precise values of population-specific *rates* of incidence of polyp and its progression to cancerous stage. There has been a considerable amount of research in recent years on developing screening based CRC intervention strategies. However, these are not supported by population-specific mathematical estimates of progression rates. This paper addresses this need by developing a probability model that estimates polyp progression rates considering race and family history of CRC; note that, it is ethically infeasible to obtain polyp progression rates through clinical trials. We use the estimated rates to simulate the progression of polyps in the population of the State of Indiana, and also the population of a clinical trial conducted in the State of Minnesota, which was obtained from literature. The results from the simulations are used to validate the probability model.

Keywords Colorectal cancer · Disease progression · CRC intervention · Polyp progression · CRC simulation · Applied probability

This research is part of the Cancer Care Engineering project (http://ccehub.org/) and was partly supported by a grant from Regenstrief Foundation, Indianapolis, Indiana.

C. Gopalappa (✉) · T. K. Das
Department of Industrial and Management Systems Engineering, University of South Florida, 4202 E. Fowler Avenue, ENB 118, Tampa, FL 33620, USA
e-mail: chaitrag@gmail.com

T. K. Das
e-mail: das@usf.edu

S. Aydogan-Cremaschi
Department of Chemical Engineering, University of Tulsa, 800 S. Tucker Drive, KEP U309, Tulsa, OK 74104, USA
e-mail: selen-cremaschi@utulsa.edu

S. Orcun
Discovery Park, Purdue University, 203 South Martin Jischke Dr., West Lafayette, IN 47907, USA
e-mail: sorcun@purdue.edu

1 Introduction

Colorectal cancer (CRC) is the third most common cause of cancer related deaths in the U.S. Most CRCs begin as a precancerous polyp [1, 2] and is referred to as a adenoma-carcinoma sequence [3]. Employing effective population-wide strategies for early detection and treatment at precancerous stages can lead to a significant reduction in CRC mortalities. The literature presents a considerable amount of research on developing CRC screening strategies with varying tests and time lines, and examining their influence on mortality rates. However, developing feasible intervention strategies requires a system-based model of cancer care that must consider, in addition to screening alternatives, various other interacting elements of the system including the social-behavioral traits of the people and the physicians, and the parameters of the insurance

policies. Two important processes of the system-based cancer care model are: *polyp incidence* and *polyp progression*. Polyps follow a natural incidence and progression and, upon diagnosis, drive the behavior and interaction of the system elements. Thus, precise models portraying the incidence and the progression processes are fundamental to developing effective intervention strategies. In this paper, our attention is focused on developing a probability model to estimate the progression rates of colorectal polyps.

The literature contains a considerable number of simulation and mathematical models for CRC screening strategies ([4–12] and CISNET models [13]). All of these models have a natural history component for the incidence and progression of polyps, most of which are modeled using variants of Markovian techniques. The main inputs required for these models are the *incidence rates* of polyps and the *progression rates* between stages, e.g., the inverse of the time that polyps take to progress from adenoma (pre-malignant) to carcinoma. The incidence rates have been estimated based on case study results involving randomized screening and follow-up. However, for estimating progression rates, a case study approach is not feasible as it is unethical to keep a diagnosed polyp under observation without treatment. As a result, most of the cited models use progression rates that are derived based on expert opinion obtained either by convening a panel or by utilizing the available data in [14] and [15].

Mathematical models, in contrast to the models based on expert opinion, can incorporate characteristics like race and family history of CRC, and hence are capable of estimating population-specific progression rates. The National Academy of Engineering, under one of the Grand Challenges for the 21st century—*Engineering Better Medicine*, noted the need for engineering personalized rather than standardized medicine since "people differ in susceptibility to disease and response to medicine." The progression rates in the pre-diagnosis phase may also vary between populations, which underscores the need for population-specific progression rates. Although literature presents numerous models and cost based analysis of CRC screening strategies and numerous models and case studies on incidence rates, mathematical models to estimate polyp progression rates have been limited [16, 17]. The study presented in [16] uses the data from the national colonoscopy screening database of Germany to develop a statistical approach for obtaining annual transition rate and the ten-year cumulative risk of CRC specific to sex and age groups. The transition rates have been obtained only for after the onset of advanced adenoma (polyp $\geq$1 cm). A Markov model to estimate the incidence of cancer and the progression rates for stages after the onset of cancer is presented in [17]. The model was built based on the results of a case study conducted on a population with high-risk of CRC.

In this paper we present a probability model for estimating polyp progression rates, specific to race and family history status, from the incidence of polyp to carcinoma and between stages of carcinoma. Note that, estimating progression rates from incidence of polyp to carcinoma are of vital importance for developing early pre-cancer intervention strategies. The polyp progression pathways considered in our model (Fig. 1) are described as follows.

CRC pathways—polyp incidence and stages of progression Most CRCs originate as visible precancerous polyps and only a small percentage begin as flat carci-

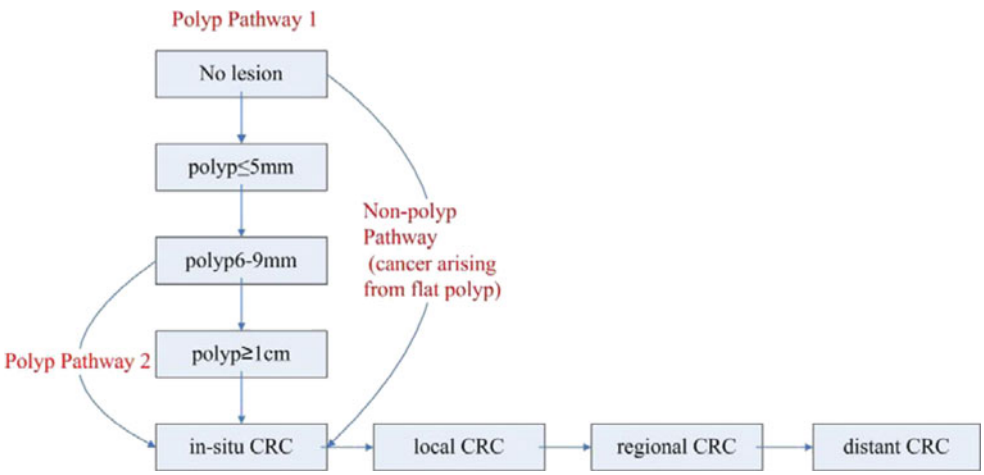


Fig. 1 CRC pathways—polyp incidence and stages of progression

noma. Not all polyps are pre-malignant and hence only some progress to carcinoma. While CRCs generally develop from polyps greater than 1 cm, carcinoma has also been diagnosed in polyps between 6 and 9 mm [15]. In this research, as depicted in Fig. 1, we consider three possible pathways for polyp progression before the onset of cancer and four stages of cancer. Polyp-pathways 1 and 2 refer to the progression types that begin with a visible adenoma polyp before progressing to cancer; these were adopted from the pathways presented in [15]. Non-polyp pathway refers to the cancers arising from flat polyps [18]. After the in-situ stage, the polyps progress through the three stages of invasive cancer: local, regional, and distant, which, based on Dukes classification of cancer [19], correspond to stages A+B, C, and D, respectively. While most cancers begin as pre-cancerous polyp, not all polyps progress to cancer. Hence, polyps can be categorized into progressive and non-progressive [15]. In what follows, we present our probability model, results estimated from the model, the model validation, and concluding remarks.

2 Probability model to estimate progression rates

One of the main inputs required to develop a progression rate model is the *incidence rate* of polyps specific to patient's age, race, and family history status of CRC. In this section, we first present our method for estimating incidence rates, followed by the probability model for estimating the progression rates. The probability model was developed based on the assumptions that non-polyp pathway accounts for 15% of the cancers and 70% of the cancers with polyps arise from pathway 1 (based on expert opinions cited in [15] and [6]). All notation used in the model presented in this section are summarized in Table 1.

2.1 Polyp incidence rates

We estimated the rate of incidence of polyp ≤ 5 mm, which is the first visible stage of a polyp, using the case studies presented in [20] and [21]. The study in [20] presents data on the number of positive sigmoidoscopy results (indicating polyps) in a population that had tested negative (indicating normal) three years back. Using this repeated test data, we estimated the probability of incidence of polyp ≤ 5 mm per year. Studies have shown that the rate of polyp incidence varies with age, however, [20] did not contain enough data to estimate the incidence probabilities based on age groups. Hence, we used the statistics in [21], which presents age-based data of patients undergoing polyp detection and removal at the Rochester Methodist Hospital, Rochester, Minnesota. It has also been observed in the literature that a family history of CRC increases the risk of developing CRC [21, 22]. Therefore, we considered population specific incidence rates, which we estimated as follows.

Let $p_{5_{t_1}}$ denote an event of incidence of polyp $\leq$ 5 mm at time t_1, and $A_{t_1} \in Z^+$ be the random variable denoting the age at t_1, where Z^+ denotes the set of positive integers. Let R and F denote the random variables of race and the status of family history of CRC, respectively, of a randomly selected individual. We consider $R \in \{Caucasian, African\ American,$

Table 1 Notation used in the probability model

Symbol	Description
$p_5, \mathcal{S}, \mathcal{L}, \mathcal{R}, \mathcal{D}$	Events of incidence of polyp ≤ 5 mm, in-situ CRC, local CRC, regional CRC, and distant CRC
$p_{5_t}, \mathcal{S}_t, \mathcal{L}_t, \mathcal{R}_t, \mathcal{D}_t$	Events of incidence of polyp ≤ 5 mm, in-situ CRC, local CRC, regional CRC, and distant CRC, at time t
$\tilde{\mathcal{S}}_t, \tilde{\mathcal{L}}_t, \tilde{\mathcal{R}}_t, \tilde{\mathcal{D}}_t$	Events of prevalence of in-situ CRC, local CRC, regional CRC, and distant CRC, at time t
$\tilde{\mathcal{C}}_t$	Event of prevalence of CRC (in-situ, local, regional, or distant), at time t
$\tilde{\mathcal{IC}}_t$	Event of prevalence of *invasive* CRC (local, regional, or distant), at time t
F	Random variable denoting family history of CRC of a randomly selected individual
R	Random variable denoting race of a randomly selected individual
A_t	Random variable denoting age (in years) of an individual at time t
L	Random variable denoting length of life in years
Z^+	Set of positive integers
N_S	Number of stages from p_5 to $\mathcal{S}$
N_L	Number of stages from $\mathcal{S}$ to $\mathcal{L}$
N_D	Number of stages from initial event (p_5 or $\mathcal{S}$) to distant CRC ($\mathcal{D}$)
T_i^j	Time to progress from event i to j
$\lambda_{i\|rf}^j$	Progression rate from event i to event j given $R = r$ and $F = f$

Other} and $F \in \{0, >0\}$, where $F = 0$ indicates no family history of CRC and $F > 0$ indicates at least one case of CRC in the family. We computed the joint probability of $p_{5_{t_1}}$ and A_{t_1} in age interval $[a, b]$ for given events of $F = f$ and $R = r$ as follows. Let the probability of $p_{5_{t_1}}$ given $R = r$ and $F = f$ (i.e., $P\{p_{5_{t_1}} | R = r \cap F = f\}$) be denoted by $P_{rf}\{p_{5_{t_1}}\}$. Applying the definition of conditional probability, and since $p_{5_{t_1}}$ and $R = r \cap F = f$ are dependent events, we can write that

$$P_{rf}\{p_{5_{t_1}}\} = \frac{P\{p_{5_{t_1}} \cap R = r \cap F = f\}}{P\{R = r \cap F = f\}}. \quad (1)$$

Also since $P_{rf}\{p_{5_{t_1}}\}$ and $a \le A_{t_1} \le b$ are dependent events, we can write that

$$P_{rf}\left\{(a \le A_{t_1} \le b) \cap p_{5_{t_1}}\right\} = P_{rf}\left\{(a \le A_{t_1} \le b) | p_{5_{t_1}}\right\} P_{rf}\left\{p_{5_{t_1}}\right\}. \quad (2)$$

The probability values for the elements on the right hand side of Eqs. 1 and 2 are estimated using data from: [20] for probabilities of $p_{5_{t_1}}$, [21] for probability distributions on A_{t_1} and F, and [23] for probability distributions on R. See Appendix A for a description of the estimation of the right hand side elements of Eqs. 1 and 2. Note that, we consider the minimum age for developing a polyp as 40 years, since risk of cancer below 40 is low based on the discussions presented in [24] and [25]. The report by the American Cancer Society in [24] notes that 90% of CRCs are diagnosed in individuals above the age of 50. Also, the expert panel from the U.S. Multisociety Task Force on Colorectal Cancer suggests a starting screening age of 50 years (40 years) for individuals without (with) a family history of colorectal polyps [25].

2.2 A probability model for progression rates

Based on expert opinion presented in [16] and [4], we consider that the progression times are exponentially distributed with event dependent parameters (progression rates) for the following events of progressions: polyp ≤ 5 mm to in-situ CRC, in-situ to local CRC, local to regional CRC, and regional to distant CRC. It may be noted that, accurate estimation of the progression rate from the incidence of pre-cancerous polyp (polyp ≤ 5 mm) to carcinoma (in-situ) is crucial for developing effective pre-cancer intervention strategies.

Let $\lambda^{j}_{i|rf}$ denote the progression rate from event i to event j given $R = r$ and $F = f$. In what follows, we present models for estimating $\lambda^{j}_{i|rf}$ for pre-cancer events (from incidence of polyp ≤ 5 mm to incidence of in-situ) considering polyp pathways 1 and 2 (see Fig. 1) and post-cancer events (between different CRC stages) considering all pathways.

2.2.1 Estimating pre-cancer progression rate

Pre-cancer progression rate, which we will denote as $\lambda^{\mathcal{S}}_{p_5|rf}$, refers to the inverse of the expected time to progress from incidence of the first stage of visible polyp ≤ 5 mm (p_5) to incidence of in-situ CRC ($\mathcal{S}$) given $R = r$ and $F = f$.

Note: Not all p_5 progress to S. Those that do progress are called progressive polyps and the rest non-progressive. In this research, $\lambda^{\mathcal{S}}_{p_5|rf}$ is estimated by including cases of both progressive and non-progressive polyps. Meaning, if the random value for the time to progress to in-situ, drawn from the exponential($\lambda^{\mathcal{S}}_{p_5|rf}$) distribution, exceeds the natural life, then the polyp is considered non-progressive.

We now present the model for estimating $\lambda^{\mathcal{S}}_{p_5|rf}$. Let $\mathcal{S}_{t_2}$ denote the event of incidence of in-situ CRC at time t_2. Recall that $p_{5_{t_1}}$ denotes p_5 at time t_1 and A_t denotes age at time t, then, following the polyp pathways in Fig. 1, $t_2 > t_1$ (see Fig. 2). For a population with $\mathcal{S}_{t_2}$, since events of $p_{5_{t_1}} \cap A_{t_1} = \alpha \cap A_{t_2} = \beta$ are mutually exclusive (i.e., $\sum_{\alpha} \sum_{\beta > \alpha} (p_{5_{t_1}} \cap A_{t_1} = \alpha \cap A_{t_2} = \beta) = 1$) and exhaustive, we can apply the total probability rule and write that

$$P(\mathcal{S}_{t_2}) = \sum_{\alpha} \sum_{\beta > \alpha} P\left\{\mathcal{S}_{t_2} | (p_{5_{t_1}} \cap A_{t_1} = \alpha \cap A_{t_2} = \beta)\right\} \times P\left\{p_{5_{t_1}} \cap A_{t_1} = \alpha \cap A_{t_2} = \beta\right\}. \quad (3)$$

Let $T^{\mathcal{S}}_{p_5}$ be a random variable denoting the time to progress from p_5 to $\mathcal{S}$. Referring to Fig. 2, $P\{\mathcal{S}_{t_2} | (p_{5_{t_1}} \cap A_{t_1} = \alpha \cap A_{t_2} = \beta)\}$ is equivalent to $P\{T^{\mathcal{S}}_{p_5} = \beta - \alpha |$

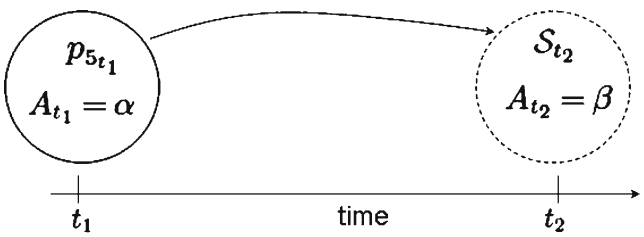


Fig. 2 Event of incidence of polyp ≤ 5 mm at time t_1 ($p_{5_{t_1}}$) and its progression to an event of incidence in-situ CRC at time t_2 ($\mathcal{S}_{t_2}$), with age at t_1 and t_2 as α and β, respectively

$(p_{5_{t_1}} \cap A_{t_1} = \alpha \cap A_{t_2} = \beta)\}$. Therefore, we can rewrite Eq. 3 as

$$P(\mathcal{S}_{t_2}) = \sum_{\alpha}\sum_{\beta>\alpha} P\left\{T^{\mathcal{S}}_{p_5} = \beta-\alpha | (p_{5_{t_1}} \cap A_{t_1} = \alpha \cap A_{t_2} = \beta)\right\} \times P\left\{p_{5_{t_1}} \cap A_{t_1} = \alpha \cap A_{t_2} = \beta\right\}. \quad (4)$$

Applying conditional probability, Eq. 4 can be written as

$$P(\mathcal{S}_{t_2}) = \sum_{\alpha}\sum_{\beta>\alpha} P\left\{T^{\mathcal{S}}_{p_5} = \beta-\alpha | (p_{5_{t_1}} \cap A_{t_1} = \alpha \cap A_{t_2} = \beta)\right\} \times P\left\{p_{5_{t_1}} | (A_{t_1} = \alpha \cap A_{t_2} = \beta)\right\} \times P\left\{A_{t_1} = \alpha \cap A_{t_2} = \beta\right\}, \quad (5)$$

where, $P\{p_{5_{t_1}} | (A_{t_1} = \alpha \cap A_{t_2} = \beta)\}$ can simply be written as $P\{p_{5_{t_1}} | A_{t_1} = \alpha\}$, since the incidence of polyp $\leq$ 5 mm at time t_1 is only dependent on age at t_1 and not on age at any future time t_2. Using the estimate from Eq. 2, $P\{p_{5_{t_1}} | A_{t_1} = \alpha\}$ can be computed as $\frac{P\{p_{5_{t_1}} \cap (a \leq A_{t_1} \leq b)\}}{b-a+1} \frac{1}{P(A_{t_1}=\alpha)}$, $a \leq \alpha \leq b$, i.e., by applying conditional probability and assuming constant rate of incidence within each age interval. Note that the assumption is in accordance with that in the microsimulation model MISCAN-colon that evaluates CRC screening policies [4] and whose input parameter values were based on expert estimates presented in meetings at the National Cancer Institute. Further, in Eq. 5, $P\{(T^{\mathcal{S}}_{p_5} = \beta - \alpha) | (p_{5_{t_1}} \cap A_{t_1} = \alpha \cap A_{t_2} = \beta)\}$ can be substituted as $\lambda^{\mathcal{S}}_{p_5} e^{-\lambda^{\mathcal{S}}_{p_5}(\beta-\alpha)}$, where $\lambda^{\mathcal{S}}_{p_5}$ denotes progression rate from p_5 to $\mathcal{S}$. The remaining term on the right hand side of Eq. 5 can be estimated using population demographics from U.S. census data. Hence, $\lambda^{\mathcal{S}}_{p_5}$ can be estimated using Eq. 5 if the probability on the left hand side is available. However, $P\{\mathcal{S}_{t_2}\}$ is unknown and is infeasible to estimate with the currently available data as explained below.

Let us divide the in-situ CRC stage as a series of sequential events $\{s_0, s_1, s_2, \cdots, s_n\}$, where s_i is an event indicating i time units of cancer progression in the in-situ stage. To relate $\mathcal{S}$ to s_i, we need to consider small time units, e.g., day, in which case $\mathcal{S}$ is equivalent to s_0 thus denoting the event of *epoch* of *incidence* of in-situ. Note that, for a diagnosed case of in-situ CRC, it is not possible to determine the value of i, and hence we cannot obtain data related to the occurrence of each event s_i. Therefore, it is not feasible to estimate $P\{\mathcal{S}\}$. However, it is possible to estimate the probability of event of *prevalence* of in-situ CRC, i.e., $P\{\cup_{i=1}^{n} s_i\}$, as equal to the proportion of people in stage in-situ CRC in a randomized screening trial (we will denote $P\{\cup_{i=1}^{n} s_i\}$ at an arbitrary time t_2 as $P\{\tilde{\mathcal{S}}_{t_2}\}$). Due to the unavailability of a suitable randomized study that can be used to estimate $P\{\tilde{\mathcal{S}}_{t_2}\}$, we estimated $P\{\tilde{\mathcal{C}}_{t_2}\}$, which is the probability of prevalence of CRC at t_2. That is, $P\{\tilde{\mathcal{C}}_{t_2}\} = P\{\tilde{\mathcal{S}}_{t_2} \cup \tilde{\mathcal{L}}_{t_2} \cup \tilde{\mathcal{R}}_{t_2} \cup \tilde{\mathcal{D}}_{t_2}\}$, where, the events in the probability term on the right hand side of the equation denote the prevalences of in-situ CRC, local CRC, regional CRC, and distant CRC, respectively, at time t_2. Therefore, Fig. 2 is modified to include the above changes and is presented as Fig. 3. As illustrated by Scenarios 1–4 in the figure, for a randomly chosen individual at t_2, $\tilde{\mathcal{C}}_{t_2}$ corresponds to an event of prevalence of one of the CRC stages.

To reflect the above changes we modify Eq. 5 as follows,

$$P\{\tilde{\mathcal{C}}_{t_2}\} = \sum_{\alpha}\sum_{\beta>\alpha} P\left\{T^{\mathcal{S}}_{p_5} \leq \beta-\alpha | (p_{5_{t_1}} \cap A_{t_1} = \alpha \cap A_{t_2} = \beta)\right\} \times P\left\{p_{5_{t_1}} | A_{t_1} = \alpha\right\} P\left\{A_{t_1} = \alpha \cap A_{t_2} = \beta\right\} \quad (6)$$

where the left hand side has been replaced with $P\{\tilde{\mathcal{C}}_{t_2}\}$. Also, we have $T^{\mathcal{S}}_{p_5} \leq \beta - \alpha$ instead of $T^{\mathcal{S}}_{p_5} = \beta - \alpha$, as an occurrence of $\tilde{\mathcal{C}}$ at t_2 does not necessarily mean occurrence of s_0 (i.e., $s_0 \subsetneq \tilde{\mathcal{C}}$). Also, since $\mathcal{S}$ (or s_0) is the first of the set of chronological events that make $\tilde{\mathcal{C}}$, age at $\mathcal{S}$ is $\leq \beta$ implying that $T^{\mathcal{S}}_{p_5} \leq \beta - \alpha$. $P\{\tilde{\mathcal{C}}_{t_2}\}$ was estimated using data from [26], which presents screen results from CRC counseling and screening conducted by ten health departments in 15 diverse counties in the state of North Carolina as part of a pilot study on cancer coordination and control.

Referring to Fig. 3, the upper bound on $T^{\mathcal{S}}_{p_5}$, i.e., $T^{\mathcal{S}}_{p_5} = \beta - \alpha$, is represented by Scenario 1, while a lower bound would be represented by Scenario 4. We can quantify the lower bound as one year, however, this value may not be realistic and can be explained with the following examples. For a combination of values for $\{A_{t_1} = \alpha, A_{t_2} = \beta\}$ consider an example of $\{A_{t_1} = 40, A_{t_2} = 43\}$. Referring to Scenario 4, if $T^{\mathcal{S}}_{p_5} = 1$, it will imply that age at $\tilde{\mathcal{S}}_{t_2}$ is 41 and age at $\tilde{\mathcal{D}}_{t_2}$ is 43, i.e., it takes two years to progress from in-situ to distant CRC. Similarly, if instead, we consider an example of $\{A_{t_1} = 40, A_{t_2} = 70\}$, if $T^{\mathcal{S}}_{p_5} = 1$, it will imply that age at

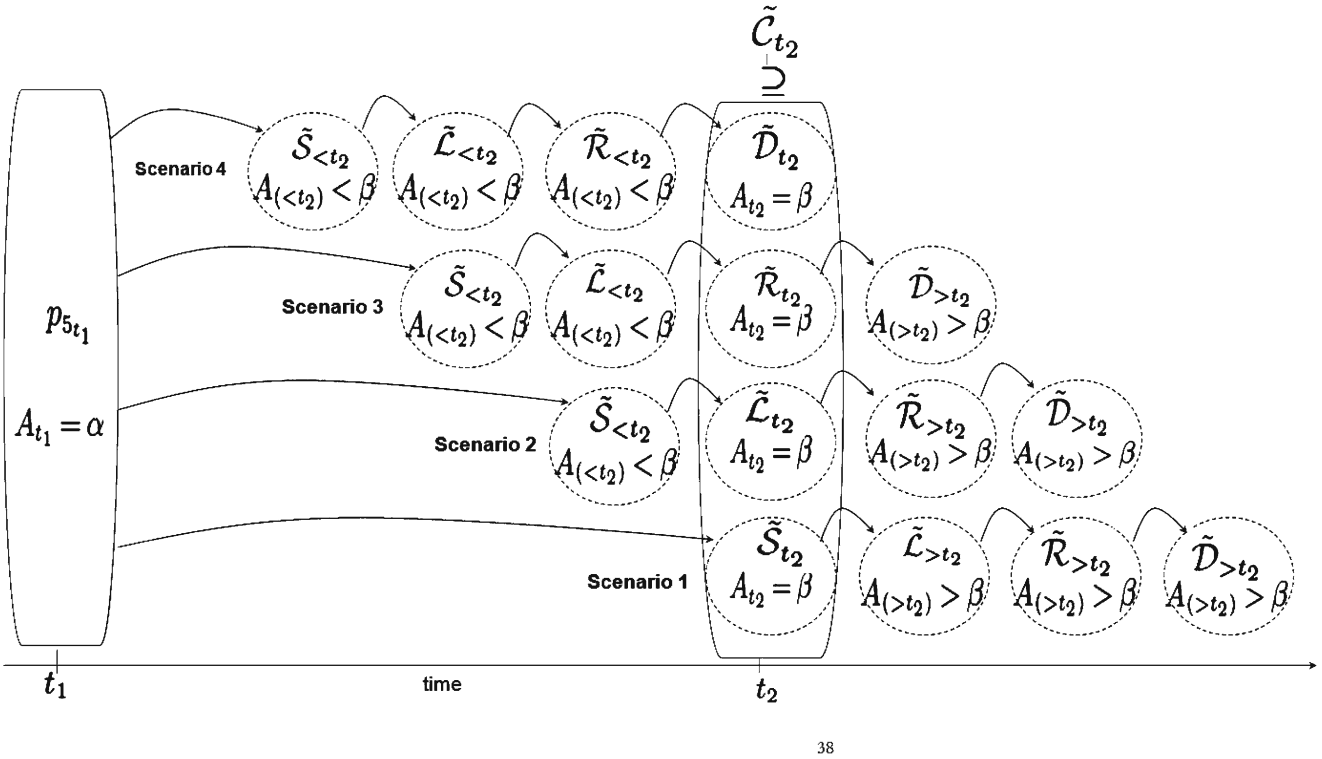


Fig. 3 Event of incidence of polyp ≤ 5 mm at time t_1 ($p_{5_{t_1}}$) and its progression to an event of prevalence of CRC at time t_2 ($\tilde{\mathcal{C}}_{t_2}$) (i.e., either $\tilde{\mathcal{S}}_{t_2}$, $\tilde{\mathcal{L}}_{t_2}$, $\tilde{\mathcal{R}}_{t_2}$, or $\tilde{\mathcal{D}}_{t_2}$), with age at t_1 (A_{t_1}) and t_2 (A_{t_2}) as α and β, respectively

$\tilde{\mathcal{S}}_{t_2}$ is 41 and age at $\tilde{\mathcal{D}}_{t_2}$ is 70, i.e., it takes 29 years to progress from in-situ to distant CRC while it took only one year to progress from polyp ≤ 5 mm to in-situ CRC. This is highly unlikely since the progression between cancer stages is faster compared to precancer stages. Therefore, in order to place a more realistic lower bound, we consider equal time to progress between stages, and referring to Scenario 4, we write $T^{\mathcal{S}}_{p_5} \geq \frac{(\beta-\alpha)N_S}{N_D}$, where N_S and N_D are the number of stages from polyp ≤ 5 mm to in-situ CRC and polyp ≤ 5 mm to distant CRC, respectively. As an example, for polyp pathway 1 in Fig. 1, $N_S = 3$ and $N_D = 7$. Note that, the equal time between stages was an assumption made only for obtaining a more realistic lower bound, than using an arbitrary value of one year, and was not an assumption on the progression rate estimation.

The modified equation can be written as,

$$P(\tilde{\mathcal{C}}_{t_2}) = \sum_{\alpha}\sum_{\beta>\alpha} P\left\{\frac{N_S(\beta-\alpha)}{N_D} \leq T^{\mathcal{S}}_{p_5} \leq (\beta-\alpha) | (p_{5_{t_1}} \cap A_{t_1}=\alpha \cap A_{t_2}=\beta)\right\} \times P\left\{p_{5_{t_1}} | A_{t_1}=\alpha\right\} P\{A_{t_1}=\alpha \cap A_{t_2}=\beta\} \quad (7)$$

where,

$$P\left\{\frac{N_S(\beta-\alpha)}{N_D} \leq T^{\mathcal{S}}_{p_5} \leq (\beta-\alpha) | (p_{5_{t_1}} \cap A_{t_1}=\alpha \cap A_{t_2}=\beta)\right\} = \left[1-e^{-\lambda^{\mathcal{S}}_{p_5}(\beta-\alpha)}\right] - \left[1-e^{-\lambda^{\mathcal{S}}_{p_5} N_S\left(\frac{\beta-\alpha}{N_D}\right)}\right]$$

since the bounds on the progression time is equivalent to the event $T^{\mathcal{S}}_{p_5} \leq (\beta-\alpha) \bigcap T^{\mathcal{S}}_{p_5} \geq \frac{N_S(\beta-\alpha)}{N_D}$. Note that, the formulation in Eq. 7 will imply that every individual with p_5 at t_1 will live through to time t_2, however, in reality this is not the case. Therefore, denoting $L \in Z^+$ as a random variable for length of life, we write Eq. 7 as,

$$P(\tilde{\mathcal{C}}_{t_2}) = \sum_{\alpha}\sum_{\beta>\alpha} P\left\{\frac{N_S(\beta-\alpha)}{N_D} \leq T^{\mathcal{S}}_{p_5} \leq (\beta-\alpha) | (p_{5_{t_1}} \cap A_{t_1}=\alpha \cap A_{t_2}=\beta)\right\} \times P\left\{p_{5_{t_1}} | A_{t_1}=\alpha\right\} P\left\{A_{t_1}=\alpha \cap A_{t_2}=\beta\right\} \times P\{L>\beta\}. \quad (8)$$

In Eq. 8, the only unknown value is the progression rate $\lambda^{\mathcal{S}}_{p_5}$, which can be computed by iteratively incrementing its value to where it best fits (Eq. 8). How-

ever, before doing so, in order to obtain population-specific progression rates, we estimate $\lambda^{\mathcal{S}}_{p_5|rf}$, which denotes $\lambda^{\mathcal{S}}_{p_5}$ given race $(R=r)$ and family history status $(F=f)$, as follows.

Since events of $R=r \cap F=f$ are mutually exclusive (i.e., $\sum_r \sum_f P\{R=r \cap F=f\}=1$) and exhaustive, applying the total probability rule, we can write,

$$\begin{aligned} P(\tilde{\mathcal{C}}_{t_2}) &= \sum_r \sum_f P\left\{\tilde{\mathcal{C}}_{t_2}|(R=r \cap F=f)\right\} \\ &\quad \times P\{R=r \cap F=f\} \\ &= \sum_r \sum_f P\left\{\tilde{\mathcal{C}}_{t_2} \cap R=r \cap F=f\right\}. \end{aligned} \tag{9}$$

Note that, Eq. 8 for a given $R=r \cap F=f$ is equivalent to $P\left\{\tilde{\mathcal{C}}_{t_2}|(R=r \cap F=f)\right\}$ of Eq. 9. Therefore, we can write

$$\begin{aligned} &P\left\{\tilde{\mathcal{C}}_{t_2} \cap R=r \cap F=f\right\} \\ &= \Bigg[\sum_{\alpha}\sum_{\beta>\alpha} P_{rf}\left\{\frac{N_S(\beta-\alpha)}{N_D} \le T^{\mathcal{S}}_{p_5} \right. \\ &\qquad \left. \le (\beta-\alpha)|(p_{5_{t_1}} \cap A_{t_1}=\alpha \cap A_{t_2}=\beta)\right\} \\ &\quad \times P_{rf}\left\{p_{5_{t_1}}|A_{t_1}=\alpha\right\} P_r\left\{A_{t_1}=\alpha \cap A_{t_2}=\beta\right\} \\ &\quad \times P_r\{L>\beta\}\Bigg] P\{R=r \cap F=f\} \quad \forall r \forall f, \end{aligned} \tag{10}$$

where,

$$\begin{aligned} &P_{rf}\left\{\frac{N_S(\beta-\alpha)}{N_D} \le T^{\mathcal{S}}_{p_5} \le (\beta-\alpha)|(p_{5_{t1}} \cap A_{t_1}=\alpha \cap A_{t_2}=\beta)\right\} \\ &= \left[1-e^{-\left(\lambda^{\mathcal{S}}_{p_5|rf}\right)(\beta-\alpha)}\right] - \left[1-e^{-\left(\lambda^{\mathcal{S}}_{p_5|rf}\right)N_S\left(\frac{\beta-\alpha}{N_D}\right)}\right]. \end{aligned}$$

Note that, occurrences of $P_{rf}\{(.)\}$ in Eq. 10 (and henceforth) represent $P\{(.)|(R=r \cap F=f)\}$, and have been written as such for notational convenience. Applying conditional probability, we can write $P\{\tilde{\mathcal{C}}_{t_2} \cap R=r \cap F=f\}=P\{(R=r \cap F=f)|\tilde{\mathcal{C}}_{t_2}\}P\{\tilde{\mathcal{C}}_{t_2}\}$. Since not enough data is available to determine the dependence of events $F=f$ and $R=r$ when given $\tilde{\mathcal{C}}_{t_2}$, we assume independence and write $P\{\tilde{\mathcal{C}}_{t_2} \cap R=r \cap F=f\}=P\{R=r|\tilde{\mathcal{C}}_{t_2}\}P\{F=f|\tilde{\mathcal{C}}_{t_2}\}P\{\tilde{\mathcal{C}}_{t_2}\}$. As mentioned earlier, $P\{\tilde{\mathcal{C}}_{t_2}\}$ is estimated using data presented in [26]. We compute $P\{(R=r)|\tilde{\mathcal{C}}_{t_2}\}=\frac{\text{Number of CRC cases in race } r}{\text{Total number of CRC cases}}$, where the required numbers are obtained from the Indiana State Department of Health database [27]. We consider $P\{(F=f)|\tilde{\mathcal{C}}_{t_2}\}=0.2$ based on the observations reported by the American Cancer Society in [28]. Note that, estimates of $P\{R=r \cap F=f\}$ were earlier obtained for Eq. 1, the details of which are described in Appendix A. Therefore, the only unknown element in Eq. 10 is $\lambda^{\mathcal{S}}_{p_5|rf}$, and hence, it can be easily estimated for all values of r and f.

2.2.2 Estimating progression rates between post-cancer stages

This section discusses the estimation of progression rates between CRC stages, i.e., between stages in-situ, local, regional, and distant, with events of incidences denoted as $\mathcal{S}$, $\mathcal{L}$, $\mathcal{R}$, and $\mathcal{D}$, respectively. A similar model as that developed for estimating $\lambda^{\mathcal{S}}_{p_5|rf}$ in Eq. 10 can be used in estimation of the progression rates between the CRC events. For example, as represented by Fig. 4, consider event of incidence of in-situ at time t_2 $(\mathcal{S}_{t_2})$ and consider $\tilde{\mathcal{IC}}_{t_3}$, the event of prevalence of *invasive* CRC at time t_3 (i.e., $P\{\tilde{\mathcal{IC}}_{t_3}\}=P\{\tilde{\mathcal{L}}_{t_3} \cup \tilde{\mathcal{R}}_{t_3} \cup \tilde{\mathcal{D}}_{t_3}\}$). We can estimate the progression rate from $\mathcal{S}$ to $\mathcal{L}$ given $R=r$ and $F=f$, denoted as $\lambda^{\mathcal{L}}_{\mathcal{S}|rf}$, by the following equation

$$\begin{aligned} P_{rf}\{\tilde{\mathcal{IC}}_{t_3}\} &= \sum_{\alpha}\sum_{\beta>\alpha} P_{rf}\left\{\frac{N_L(\beta-\alpha)}{N_D} \le T^{\mathcal{L}}_{\mathcal{S}} \right. \\ &\qquad \left. \le (\beta-\alpha)|(\mathcal{S}_{t_2} \cap A_{t_2}=\alpha \cap A_{t_3}=\beta)\right\} \\ &\quad \times P_{rf}\left\{\mathcal{S}_{t_2}|A_{t_2}=\alpha\right\} P_r\left\{A_{t_2}=\alpha \cap A_{t_3}=\beta\right\} \\ &\quad \times P_r\{L>\beta\} \quad \forall r, \forall f \end{aligned} \tag{11}$$

where,

$$\begin{aligned} &P_{rf}\left\{\frac{N_S(\beta-\alpha)}{N_D} \le T^{\mathcal{L}}_{\mathcal{S}} \le (\beta-\alpha)|(\mathcal{S}_{t_2} \cap A_{t_2}=\alpha \cap A_{t_3}=\beta)\right\} \\ &= \left[1-e^{-\left(\lambda^{\mathcal{L}}_{\mathcal{S}|rf}\right)(\beta-\alpha)}\right] - \left[1-e^{-\left(\lambda^{\mathcal{L}}_{\mathcal{S}|rf}\right)N_L\left(\frac{\beta-\alpha}{N_D}\right)}\right], \end{aligned}$$

$T^{\mathcal{L}}_{\mathcal{S}}$ denotes time to progress from $\mathcal{S}$ to $\mathcal{L}$, and N_L and N_D in the lower bound of $T^{\mathcal{L}}_{\mathcal{S}}$ now represent the number of stages from $\mathcal{S}$ to $\mathcal{L}$ and $\mathcal{S}$ to $\mathcal{D}$, respectively. $P_{rf}\{\tilde{\mathcal{IC}}\}$ at an arbitrary time t_3 can be estimated as $P_{rf}\{\tilde{\mathcal{IC}}\}=P_{rf}\{\tilde{\mathcal{IC}} \cap \tilde{\mathcal{C}}\}=P_{rf}\{\tilde{\mathcal{C}}\}P_{rf}\{\tilde{\mathcal{IC}}|\tilde{\mathcal{C}}\}$. $P_{rf}\{\tilde{\mathcal{C}}\}$ is obtained from [26] and $P_{rf}\{\tilde{\mathcal{IC}}|\tilde{\mathcal{C}}\}$ is obtained using CRC diagnosed data from the Indiana database [29]. Therefore, we can estimate $\lambda^{\mathcal{L}}_{\mathcal{S}|rf}$ from Eq. 11 if the only unknown $P_{rf}\{\mathcal{S}_{t_2}|A_{t_2}=\alpha\}$ can be computed, since the rest of the terms can be obtained similar to the equivalent terms in Eq. 10. However, similar to the infeasibility in estimating $P\{\mathcal{S}\}$ using results from randomized screening trials as explained in Section 2.2.1, it is not feasible to estimate $P_{rf}\{\mathcal{S}_{t_2}|A_{t_2}=\alpha\}$ using screening trials. Note however that, at this point in the model, we have estimated $\lambda^{\mathcal{S}}_{p_5|rf}$, the progression rate

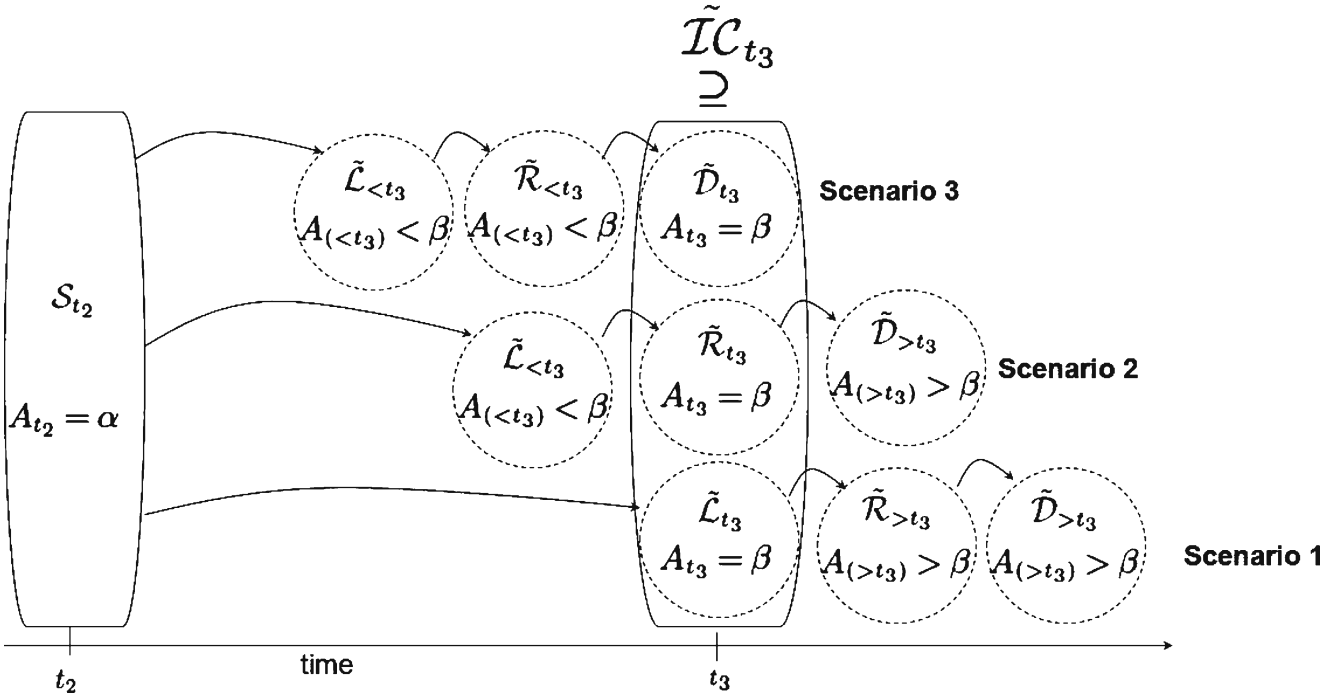


Fig. 4 Event of incidence of in-situ CRC at time t_2 ($\mathcal{S}_{t_2}$) and its progression to an event of prevalence of invasive CRC at time t_3 ($\tilde{\mathcal{IC}}_{t_3}$) (i.e., either $\tilde{\mathcal{L}}_{t_3}$, $\tilde{\mathcal{R}}_{t_3}$, or $\tilde{\mathcal{D}}_{t_3}$), with age at t_2 (A_{t_2}) and t_3 (A_{t_3}) as α and β, respectively

from p_5 to $\mathcal{S}$. Therefore, using the estimated values of $Pr_f\{p_{5_{t_1}} \cap (a \le A_{t_1} \le b)\}$ from Section 2.1 and $\lambda^{\mathcal{S}}_{p5|rf}$ from Section 2.2.1, we developed a model to estimate $Pr_f\{\mathcal{S}_{t_2}|A_{t_2} = \alpha\}$, which is explained below.

The schematic of the probability model developed for estimating $Pr_f\{\mathcal{S}_{t_2}|A_{t_2} = \alpha\}$ is presented in Fig. 5, and can be interpreted as follows. Consider p_5 at time t_1 and its progression to $\mathcal{S}$ at t_2, $t_1 < t_2$. Now considering age at t_1 and t_2, if $c \le A_{t_2} \le d$ and $a \le A_{t_1} \le b$, then $\{\forall a, b, c, d : [a, b] \le [c, d]\}$. For example, considering age intervals [40, 49], [50, 64], [65, 74], and [75,], if $A_{t_2} = [40, 49]$ then $A_{t_1} = [40, 49]$, and if $A_{t_2} = [50, 64]$ then $A_{t_1} = [40, 49]$ or $A_{t_1} = [50, 64]$. Accordingly, $Pr_f\{\mathcal{S}_{t_2} \cap (c \le A_{t_2} \le d)\}$ can be estimated by using $Pr_f\{p_{5_{t_1}} \cap (a \le A_{t_1} \le b)\}$ and $\lambda^{\mathcal{S}}_{p5|rf}$, $\forall a, b, c$, and d, as follows,

$$
\begin{aligned}
&Pr_f\left\{\mathcal{S}_{t_2} \cap (c \le A_{t_2} \le d)\right\} \\
&= \sum_{[a,b]\le[c,d]} Pr_f\left\{p_{5_{t_1}} \cap (a \le A_{t_1} \le b)\right\} \\
&\times \left[\sum_{k=a}^{b-1} \sum_{\substack{m= \\ max(k+1,c)-k}}^{d-k} Pr_f\{T^{\mathcal{S}}_{p_5} = m | (p_{5_{t_1}} \cap A_{t_1} = k)\}\right. \\
&\left. \times\ P\{L > m + k\}\right] \qquad (12)
\end{aligned}
$$

where, $Pr_f\{T^{\mathcal{S}}_{p_5} = m\} = \left(\lambda^{\mathcal{S}}_{p5|rf}\right) e^{-m\left(\lambda^{\mathcal{S}}_{p5|rf}\right)}$. Note that, Eq. 12 was derived by a simple application of the total probability rule. For example, considering one specific

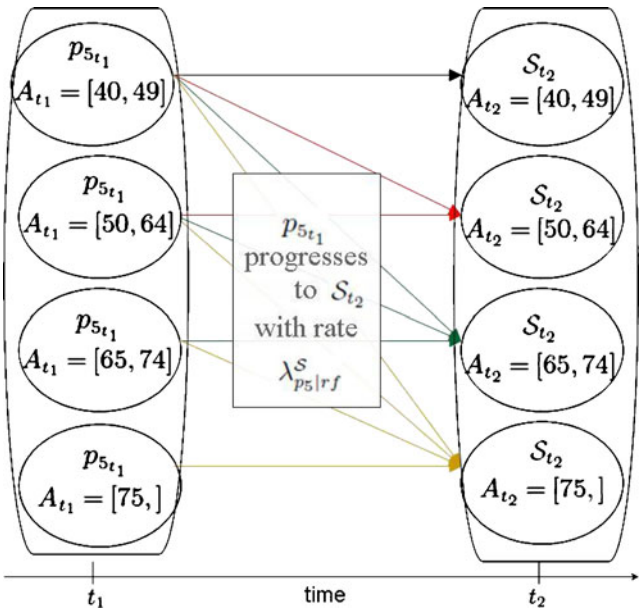


Fig. 5 Event of incidence of polyp ≤ 5 mm at t_1 ($p_{5_{t_1}}$) and its progression to an event of incidence of in-situ CRC at t_2 ($\mathcal{S}_{t_2}$), with age $a \le A_{t_1} \le b$ and $c \le A_{t_2} \le d$ such that $[a, b] \le [c, d]$

age at A_{t_2} as 50, we can write $P_{rf}\{\mathcal{S}_{t_2} \cap A_{t_2} = 50\} = \sum_{\alpha<50} P_{rf}\{\mathcal{S}_{t_2} \cap A_{t_2} = 50 | (p_{5_{t_1}} \cap A_{t_1} = \alpha)\} P_{rf}\{p_{5_{t_1}} \cap A_{t_1} = \alpha)\} = P_{rf}\{T^{\mathcal{S}}_{p_5} = 50 - \alpha \,|\, (p_{5_{t_1}} \cap A_{t_1} = \alpha)\} P_{rf}\{p_{5_{t_1}} \cap A_{t_1} = \alpha)\}$. Equation 12 however has been written for an age interval and also contains the variable L to include the length of life of an individual.

$P_{rf}\{\mathcal{S}_{t_2} | A_{t_2} = \alpha\}$, required for Eq. 11, can now be estimated from $P_{rf}\{\mathcal{S}_{t_2} \cap (c \le A_{t_2} \le d)\}$ by applying conditional probability and by considering a constant rate of incidence within each age interval, the latter of which is in accordance to the literature as explained earlier. Equation 11 can now be solved to obtain $\lambda^{\mathcal{L}}_{\mathcal{S}|rf}$. The progression rates from local to regional CRC and regional to distant CRC can similarly be estimated by cyclically computing the probability of event of incidence (similar to computation of $P_{rf}\{\mathcal{S}_{t_2} | A_{t_2} = \alpha\}$ using Eq. 12) followed by estimation of the progression rates (similar to estimation of $\lambda^{\mathcal{L}}_{\mathcal{S}|rf}$ using Eq. 11). Note that, for stages past in-situ, L also includes survival based on stage of cancer in addition to the natural life of an individual.

3 Results: estimated incidence and progression rates

In Tables 2 and 3, we present rates of p_5 for polyp pathways and rates of $\mathcal{S}$ for non-polyp pathway, respectively, for different combinations of age, race, and family history status. For example, in Table 2, the percentage of incidence of polyp ≤ 5 mm at age $50 \le A \le 64$ and $R = Caucasian$ and $F > 0$ is 4.25. The mean times in years to progress from event i to event j given $R = r$ and $F = f$, i.e., $\frac{1}{\lambda^{j}_{i|rf}}$, are presented in Tables 4 and 5 for polyp pathways and non-polyp pathway, respectively. The values in Tables 2–5 will serve as input for simulating polyp progression in a multi-agent simulation model that will be built for obtaining CRC intervention strategies.

Comparison of results The expected progression time estimates (i.e., Tables 4 and 5) can be used to compute one-step transition probabilities needed to build Markov models (as in [11]) for polyp progression. For example, probabilities for $R = Caucasian$ and $F = 0$ case are depicted in Fig. 6. We compared the transition probabilities derived from our model with those compiled by [11], who analyze the cost-effectiveness of screening for a population without a family history of cancer. Though the study in [11] is based on the population of Israel, the reason for our comparison is to only check if our estimates are within commonly observed ranges, and is not meant as a validation. Polyp stages considered in [11] are low risk polyps (<1 cm), high risk (≥1 cm) polyps, local CRC, regional CRC, and distant CRC, which, as seen in Fig. 6, are slightly different from that in our model. Therefore, to obtain a rough comparison, we assumed equal progression time between stages p_5 and $\mathcal{S}$ (see pathway 1 in Fig. 1), and computed the transition probability from p_5 to polyp ≥ 1 cm. As shown in Table 6, for similar stages (i.e., rows 2, 3, and 4), the transition probabilities obtained from our model are comparable to that assumed in [11]. Using our mathematical modeling approach of progression rate estimation, we can further compute population-specific transition probabilities to build Markov models for developing effective CRC intervention strategies.

4 Validation of progression rates estimated from probability model

In order to validate the progression rates estimated in Section 2, we used a simulation based approach as follows. A simulation model was constructed such that it initially generates a population based on user-input demographics data of specific populations. For validation purpose, we considered two different populations: i) population of the State of Indiana, and ii) population of the clinical trial described in [30–32] conducted in the State of Minnesota. Further, the simulation model was built such that it executes the following three events every year for each person in the population: *event 1*) updating age of each person, and creating new births and generating mortalities in the population; *event 2*) the natural incidence and progression of polyps using values presented in Tables 2–5; and *event 3*) screening based on the actual compliance rates of the corresponding population. Note that, based on change in age (through event 1), event 2 generates a polyp and

Table 2 $P_{rf}(p_5 \cap a \le A \le b)100$: percentage incidence of polyp ≤ 5 mm at age $[a, b]$, given $R = r$ and $F = f$, for polyp pathways

Age group	All race		R = Caucasian		R = African American	
$[a, b]$	$F > 0$	$F = 0$	$F > 0$	$F = 0$	$F > 0$	$F = 0$
[40, 49]	0.74	0.12	0.73	0.13	0.90	0.10
[50, 64]	4.33	0.70	4.25	0.74	5.28	0.58
[65, 74]	4.54	0.73	4.46	0.78	5.53	0.61
[75,]	2.13	0.34	2.09	0.36	2.59	0.29

Table 3 $P_{rf}(\mathcal{S} \cap a \leq A \leq b)100$: percentage incidence of in-situ CRC at age $[a, b]$, given $R = r$ and $F = f$, for non-polyp pathway

Age group	All race		R = Caucasian		R = African American	
$[a, b]$	$F > 0$	$F = 0$	$F > 0$	$F = 0$	$F > 0$	$F = 0$
[40, 49]	0.025	0.002	0.026	0.003	0.024	0.002
[50, 64]	0.257	0.026	0.265	0.029	0.252	0.018
[65, 74]	0.330	0.034	0.339	0.038	0.318	0.023
[75,]	0.344	0.040	0.346	0.044	0.332	0.026

Table 4 Mean times (in years) to progress from event i to event j, given $R = r$ and $F = f\left(\frac{1}{\lambda^{j}_{i|rf}}\right)$ on polyp pathways

Event $i \rightarrow$ event j	All race		R = Caucasian		R = African American	
	$F > 0$	$F = 0$	$F > 0$	$F = 0$	$F > 0$	$F = 0$
$p_5 \rightarrow$ in-situ[a]	23	41.6	21.5	39	29	50
In-situ $\rightarrow$ local	3.4	3.4	3.5	3.5	3.1	3.1
Local $\rightarrow$ regional	5	5	4.5	4.5	3.5	4
Regional $\rightarrow$ distant	0.95	0.95	0.95	0.95	0.88	0.9

[a] $\lambda^{\mathcal{S}}_{p_5|rf}$ was estimated considering progressive and non-progressive polyp

Table 5 Mean times (in years) to progress from event i to event j, given $R = r$ and $F = f\left(\frac{1}{\lambda^{j}_{i|rf}}\right)$ on non-polyp pathway

Event $i \rightarrow$ event j	All race		R = Caucasian		R = African American	
	$F > 0$	$F = 0$	$F > 0$	$F = 0$	$F > 0$	$F = 0$
In-situ $\rightarrow$ local	3.4	3.3	3.4	3.4	3.1	3.1
Local $\rightarrow$ regional	3.5	4.8	4	4.5	3.5	3.5
Regional $\rightarrow$ distant	0.9	0.95	0.9	0.95	0.9	0.9

Fig. 6 One-step transition probabilities for polyp pathway 1 ($R = Caucasian$, $F = 0$)

Table 6 One-step transition probabilities between stages: comparing results presented in this paper to the literature presented in [11]

Literature [11]		This research	
From $\rightarrow$ to stages	Transition probability	From $\rightarrow$ to stages	Transition probability
Low-risk polyp $\rightarrow$ high risk polyp	0.02	$p_5 \rightarrow$ polyp$\geq$1cm	0.035
High-risk polyp $\rightarrow$ local CRC	0.05 (0.02–0.10)	Polyp ≥ 1 cm $\rightarrow$ local CRC	0.06
Local CRC $\rightarrow$ regional CRC	0.28 (0.20–0.35)	Local CRC $\rightarrow$ regional CRC	0.20
Regional CRC $\rightarrow$ distant CRC	0.63 (0.50–0.70)	Regional CRC $\rightarrow$ distant CRC	0.65

Table 7 Simulated vs. actual Indiana CRC counts per 100,000 of population

Stages	Race	Simulated 95% CI		Actual Indiana
		Lower CI	Upper CI	counts
Local + regional + distant	All race	48.83	57.55	56.02
	Caucasian	52.50	61.97	57.68
	African American	31.97	56.53	47.93
In-situ + local + regional + distant	All race	52.98	61.90	60.70

updates the natural polyp progression until a successful screen (through event 3) leads to the polyp's diagnosis. The number and stage of the new cases of polyps that are diagnosed each year are recorded. For validation, the simulated statistics on diagnosed cases of CRCs are compared with the actual statistics of the corresponding population. The simulation model was constructed in *Repast* [33], a java agent-based modeling framework. The reason for using an agent-based approach is for ease of including the behavior and interaction between the system entities (including physician and insurance policy), which is a part of our future research for obtaining cancer intervention strategies. Simulation events 2 and 3, which were mentioned above, are described using flowcharts in Appendix B. We present below the details of our validation study on the two populations.

4.1 Simulation of the Indiana population

Actual proportions of the population belonging to different race, sex, and age groups, as estimated from census data, were used to generate an initial sample population for the State of Indiana. The mortality and birth rates required for event 1 were also obtained from the census data. The incidence and progression rates required for event 2 were obtained from Tables 2–5. The screening rates required for event 3 were computed using the data from a survey that was conducted in year 2001 by the Indiana State Department of Health as part of the Behavioral Risk Factor Surveillance System [34]. The survey considered three screening options: FOBT (fecal occult blood test), colonoscopy, and sigmoidoscopy. The sensitivity and specificity of the screening tests were taken from the values used in an intervention study conducted in year 2006 [35] and also used in the MISCAN-Colon microsimulation model [4]. The values of the screening parameters used in the simulation are summarized in Appendix C.

The simulation was run with a sample population of 10,000 for 30 trials. As presented in Tables 7 and 8, the confidence interval (CI) of the simulated new cases of CRCs diagnosed over five years were compared with the actual 2000–2004 diagnosed cases of CRC, which was available on the Indiana State Department of Health website [23, 29]. Table 7 presents CRC counts per 100,000 of population. The large range in CI for the African American population can be attributed to its small proportion (8%) of the total population of Indiana. Table 8 presents the percentage distribution of CRCs among various stages at the time of diagnosis. As can be seen in both tables, the actual values lie in between the simulated CI. Note that, some of the actual values in Table 8 fall on the boundary of the simulated CI, which can be attributed to the fact that about 6.5% of actual CRC cases did not have a stage identifier (un-staged). It may be noted that, since the percentage distribution of diagnosed CRC in different stages was initially used in the probability model, Table 8 serves as a verification. However, Table 7 serves as a validation, as the simulated CRC results presented in the table are not equivalent to the cancer prevalence probabilities ($P\{\tilde{C}_{t_2}\}$ estimated using [26]) that were initially used in the probability model. The difference between the two is as follows:

- The cancer prevalence probabilities used in the probability model were estimated using data from *clinical trial* studies (not from Indiana database). However, we compare the simulated diagnosed cancer counts with the *actual diagnosed counts* in Indiana.
- The second difference is inherent in the terms *clinical trial rates* and *actual diagnosed counts*. The former statistic includes all cases of cancer in the population, since, in a clinical trial, all participants get screened (except for false negatives, but screen

Table 8 Simulated vs. actual Indiana values for stage at time of diagnosis as percentage of total CRC counts

CRC stage	Simulated 95% CI		Actual Indiana
	Lower CI	Upper CI	values
In-situ	6.02	9.04	7.70
Local	34.99	41.88	34.94
Regional	29.58	36.75	33.75
Distant	17.99	23.76	17.12
Un-staged	NA	NA	6.50

NA not applicable

sensitivity is the same in both cases). Whereas, the latter statistic does not include all cases, since, in an actual population, not everyone is compliant to screening.

Therefore, the accuracy of the simulated diagnosed cancer counts are dependent on the population's screening compliance rates as well as the polyp natural incidence rates and the expected progression times estimated from the probability model. Since the compliance rates were computed from the Indiana population database, the results in Table 7 serve as a validation of the probability model.

Results from the simulation model The above simulation model was developed to validate the probability model. However, the combination of the probability model followed by the simulation, as constructed in this paper, serves as a model in itself for obtaining certain polyp related estimates of interest. One such set of estimates is related to the progressive polyps. The simulated CIs on the maximum likelihood estimate of the exponential distribution parameter, i.e., on the mean *time to progress* (in years) from p_5 to $\mathcal{S}$ given $R = Caucasian$, are presented in Table 9. It may be seen that, our estimated value for the progression from polyp ≤ 5 mm to in-situ CRC (row 1 of the table) compares well with expert opinion in [14], where an average time of approximately ten years to progress from adenomatous polyp (mainly <1 cm) to invasive CRC (local CRC and beyond) is suggested. Use of a mathematical modeling approach for estimating population-specific values, as in this research, allows us to quantify any variations across populations. For example, see Table 9, which shows shorter progression time to cancer for a population with $F > 0$. We also obtained results for the proportion of polyp ≤ 5 mm progressing to in-situ CRC (i.e., proportion of progressive polyps), which are presented in Table 10 for $R = Caucasian$. Note that, the proportion of progressive polyps in a population with $F > 0$ is approximately 1.8 times as much as that in a population with $F = 0$. Though it is known that a family history of CRC increases the life-time chances of cancer, such mathematical model based quantifications of polyp progression, as presented in Tables 9 and 10, could not be found in the literature.

Table 9 Estimated confidence interval on mean time to progress (in years) from polyp ≤ 5 mm to in-situ CRC according to family history

Family history (F)	Upper 95% CI	Lower 95% CI
All[a]	10.7	8.3
$F = 0$	12.1	9.4
$F > 0$	9.9	7.7

[a]Includes $F = 0$ and $F > 0$

Table 10 Estimated proportion of p_5's progressing to $\mathcal{S}$ for $R = Caucasian$

Family history (F)	Proportion (in %)
All[a]	20.9
$F = 0$	19.5
$F > 0$	34.2

[a]Includes $F = 0$ and $F > 0$

4.2 Simulation of Minnesota study

The study in [30–32] present a clinical trial conducted in Minnesota, where a sample population in the age group of 50–80 years with no history of cancer was recruited and randomly divided into three groups. Groups 1 and 2 were subject to annual and biennial FOBT screening, respectively, and group 3 was a control group. The objective of the study was to identify the difference in CRC related mortality rates among the three groups, and hence analyze the effect of annual and biennial FOBT screening on mortalities. Phase I of the study was conducted from 1978 to 1982, and continued to Phase II from February 1986 to February 1992. Study groups 1 and 2 were simulated, separately, by utilizing the values for proportions of people in age ranges 50–59, 60–69, and 70–80 that were given by [30], as follows. The simulation first generated people between age 0–30 years, with proportions of people in age ranges 0–9, 10–19, and 20–30 equal to the proportions of people in age ranges 50–59, 60–69, and 70–80, respectively, of the actual study. The simulation was first run for 50 years so that the population is now between age group 50–80 years, and then later run for a period of 14 years representing the timeline of the actual clinical trial. The mortality and birth rates (event 1) were kept at zero during the first 50 years. Event 2, i.e., polyp incidence and progression was run during the entire period, the rates for which were obtained from Tables 2–5. During the first 50 years, any symptomatic cases of CRC were removed from the simulation in order to remove existing diagnosed cases of cancer, and the proportion of population in the three age groups were adjusted. The time to symptomatic was taken as per the times (preclinical to clinical) considered in the MISCAN-colon model [5], whose parameters, as mentioned earlier, were based on expert estimates presented in meetings at the National Cancer Institute. During the 14 year simulation run that mimicked the actual study, groups

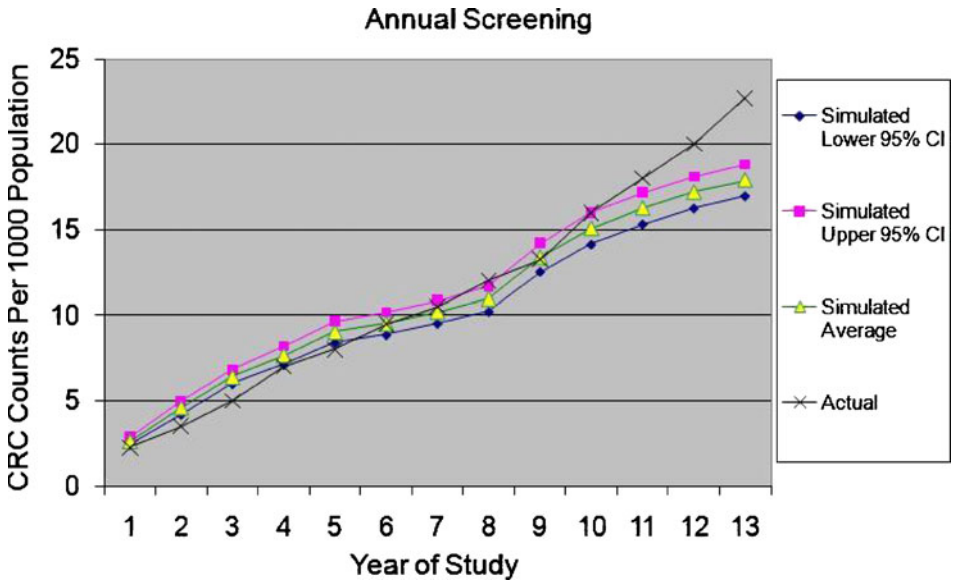


Fig. 7 Comparing simulated versus actual CRC cases per 1,000 population—annual group of the Minnesota study

1 and 2 were subject to annual and biennial screening, respectively, as per screening details and test sensitivity as provided by [30].

The simulated cases of CRC during the 13 years of the study were compared with the actual cases given by [30] and the results (represented as CRC cases per 1,000 population) are presented in Figs. 7 and 8 for annual and biennial screening groups, respectively. Since the simulated screening intervals matched that in the clinical trial, the accuracy of the simulated CRC cases is dependent on the natural polyp progression, whose rates were estimated from the probability model. Moreover, tracking a population and comparing results over a 13 year period is a stronger analysis of the polyp progression, and therefore the results presented in Figs. 7 and 8 can be used to validate the probability model.

While the actual values for the biennial group fall within the simulated 95% confidence interval in most years, the actual values for annual group are lower during Phase 1 of the study and higher during Phase II of study. In addition to acknowledging that there are other population-specific factors that need to be identified and considered to increase accuracy of the estimates, some of the trends in the aforementioned deviations can be attributed to the uncontrollable factors as follows. It was not possible to simulate the exact study since we had to make several assumptions

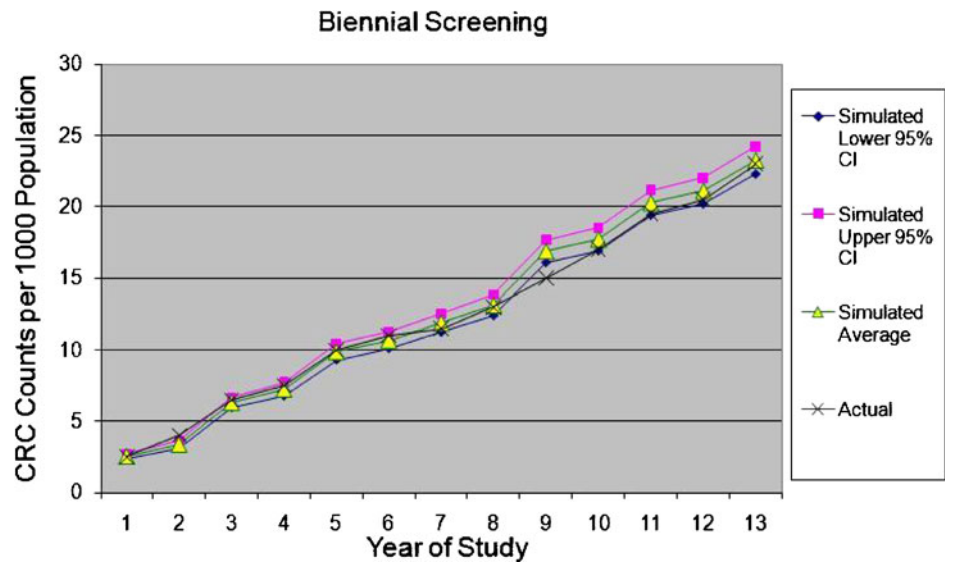


Fig. 8 Comparing simulated versus actual CRC cases per 1,000 population—biennial group of the Minnesota study

Table 11 CRC counts per 1,000 population and stage at diagnosis—for annually screened group of Minnesota study

Simulated 95% CI			Actual counts	
CRC stage	Lower CI	Upper CI	Dukes CRC staging	Actual values
Local	8.83	9.76	A+B	13.3
Regional	4.64	5.36	C	5.6
Distant	2.87	3.56	D	2.3
Un-staged	NA	NA	Un-staged	2.1
Total CRC	16.34	18.68	Total CRC	23.3

for the unknown data. We discuss below some of the assumptions and their relation to the noticed trends in deviation between the actual and simulated results.

- **Lack of data on the number of screenings** It is noted in [30] that not everyone participated in the scheduled number of screens, which was 11 for annual group and 6 for biennial group [31] over the entire duration of study. For each person, let $X =$ number of screens obtained during study duration. Using the information provided in [30] and [31] data could only be extracted for the following probabilities: $P(X \geq 1)$, $P(X \geq 6)$, $P(X \geq 9)$, and $P(X = 11)$ for the annual group; and $P(X \geq 1)$, $P(X \geq 3)$, $P(X \geq 5)$, and $P(X = 6)$ for the biennial group. However, since we need $P(X = x)$ for all $x = 1, 2, .., 11$ for the annual group and for all $x = 1, 2, .., 6$ for the biennial group, the probabilities had be estimated by assuming uniform distribution. For example, $P(X = 1) = P(X = 2) = \frac{P(1 \leq X \leq 2)}{2}$ for the biennial group, and $P(X = 1) = P(X = 2) = .. = P(X = 5) = \frac{P(1 \leq X \leq 5)}{5}$ for the annual group. Also, under both groups, for each individual a random value of X was drawn, and e.g., if $X = 5$, it was assumed that the individual participated in the first five of the scheduled screens, while in reality it could be any five of the scheduled screens. That is, it would be 5 of any of the 6 required screens if it is a biennial group participant, while it would be 5 of any of the 11 required screens if it is an annual group participant. As observed from the above examples, the information available for the biennial group was more compared to the annual group. This may have contributed to the larger deviations in results for the annual group. Based on the above assumptions made on the annual group due to the lack of data, we can expect that there will be more than actual number of screens during the initial period of study and hence more CRCs get diagnosed in the simulation. Similarly, we can also expect that, along with causing lesser than actual number of screens during the latter period of study, increased screening in the beginning reduces the number of polyps that progress to CRC, and hence further reducing CRC incidence during the latter period. This is evident in Fig. 7.
- **Lack of information on screening outside of the study** As mentioned in the study, participants could have obtained screening from a source outside the study during and in-between the two phases, and yearly updates were obtained on any diagnosed cases of CRC. Since information on outside screening was not available, it was assumed that a person undergoes screening outside of study in symptomatic cases only. The time to symptomatic was extracted from the MISCAN-colon parameters presented in [5]. The above assumption together with the fact that advanced stages of CRC tend to be more symptomatic than earlier stages, we can expect that, while looking at the simulated minus actual CRC cases per stage at diagnosis, the value will be higher for distant stage compared to regional compared to local. This trend is evident in Tables 11 and 12 for the annual and biennial groups, respectively, which presents the CRC cases per 1,000 population according to stage at diagnosis over the 13 year period. Note that, in the clinical trial, staging of CRC was done by using Dukes classification. We considered Dukes stages A and B as equivalent to local, and Dukes stages C and D as equivalent to regional and distant, respectively.

Table 12 CRC counts per 1,000 population and stage at diagnosis—for biennially screened group of Minnesota study

Simulated 95% CI			Actual counts	
CRC stage	Lower CI	Upper CI	Dukes CRC staging	Actual values
Local	9.51	10.76	A+B	11.6
Regional	6.77	7.73	C	6.1
Distant	4.66	5.49	D	3
Un-staged	NA	NA	Un-staged	2.1
Total CRC	20.94	23.98	Total CRC	22.8

Table 13 Stage at diagnosis as percentage of total CRC counts—for annual group of Minnesota study

Simulated 95% CI					Actual values	
CRC stage	End of 5 years		End of 13 years		End of 13 years	
	Lower CI	Upper CI	Lower CI	Upper CI	Dukes CRC staging	Actual values
Local	58.57	64.63	51.20	55.24	A+B	57.08
Regional	20.99	26.60	27.07	29.86	C	24.03
Distant	12.43	16.78	16.62	20.00	D	9.87
Un-staged	NA	NA	NA	NA	Un-staged	9.01

Based on our assumptions: 1) that an individual undergoes the first n of the scheduled screens, and 2) that outside screening was done only in symptomatic cases, we can hypothesize that the percentages of diagnosed CRCs in stages regional and distant are more during the latter period of study compared to those in the initial period. The reasoning leading to the hypothesis can be explained as follows. Assumption 1 generates more than actual screenings during the initial few years, thus causing less cases to reach a symptomatic stage. Consequently, the latter duration has lesser than actual number of screens causing more cases to reach symptomatic stage. Since advanced stages of CRC are more symptomatic than earlier stages, we can thus infer the above mentioned hypothesize as a consequence of assumption 1 combined with assumption 2. This hypothesis can be verified by the numbers in Tables 13 and 14. As speculated, the percentage of CRCs in regional and distant stage is higher at the end of year 13 compared to that at the end of year 5.

5 Summary

Precise values of polyp incidence and progression rates are crucial for developing population-wide CRC intervention strategies. Polyp incidence rates for population groups characterized by age, race, and family history of CRC, were estimated by using data from the literature. The data sources included clinical CRC screening trials, population databases, and evidence-based reports from state health departments and national institutes. On the other hand, the natural progression timeline of polyps could not be directly estimated using observed data, since it is infeasible to allow a diagnosed polyp to progress naturally without treatment. Hence, we developed a probability model to estimate population-specific rates of polyp progression. The probability model was constructed based on known concepts of the natural progression of polyps. Thereafter, using the model, data from the above mentioned sources were synthesized to estimate progression rates. These rates are characterized by race and family history of CRC and correspond to both progressive and non-progressive polyps. The estimated incidence and progression rates were used to simulate the natural history of colorectal polyps for the population in the State of Indiana and a subset of the population in the State of Minnesota. The simulation results were used to validate the probability model. The simulation model also yielded 1) the expected time for progressive polyps to reach in-situ CRC from the polyp ≤ 5 mm stage, and 2) the proportion of polyps reaching in-situ CRC (i.e., progressive polyps).

The polyp progression related values available in the literature are mainly experience based approximations and are not population-specific. Though the literature contains several data sources related to the incidence of either precancerous polyp or carcinoma, these data had not been synthesized to mathematically estimate population-specific polyp progression times (as in Tables 4 and 5). Mathematical estimation helps to identify and quantify any variation across populations that are critical for developing early intervention strategies.

Table 14 Stage at diagnosis as percentage of total CRC counts—for biennial group of Minnesota study

Simulated 95% CI					Actual values	
CRC stage	End of 5 years		End of 13 years		End of 13 years	
	Lower CI	Upper CI	Lower CI	Upper CI	Dukes CRC staging	Actual values
Local	48.34	53.56	43.28	46.70	A+B	50.88
Regional	28.93	34.35	30.48	34.29	C	26.75
Distant	15.13	19.70	20.88	24.36	D	13.16
Un-staged	NA	NA	NA	NA	Un-staged	9.21

The probability model was developed to estimate rates considering both progressive and non-progressive polyps, and was subsequently used in the simulation model to obtain statistics of progressive polyps. Such an approach is significant, since, while it is known that a family history of CRC increases the risk of cancer, quantification of the increased risk based on proportion of progressive polyps (as in Table 10) has not been presented in the literature. Also, the risk based on progression time from polyp ≤ 5 mm to in-situ CRC had not been previously mathematically quantified (as in Table 9). Consideration of both progressive and non-progressive polyps also supports development of more comprehensive intervention strategies comprising resource needs and allocation.

Accuracy of estimation Though our model estimates polyp progression rates specific to race and family history of CRC, for better accuracy of the estimates, it is essential that the model be expanded to consider other dependent factors. Examples of dependent factors include, number and histology of polyps, classification of first and second degree relatives with CRC, and personal history of other medical conditions. However, inclusion of these factors in the model would require significant additional data, which is currently unavailable. Also, estimating progression rates between each of the pre-in-situ stages would help to simulate a more comprehensive natural progression of polyps. As in any estimated value, the progression rates presented in this manuscript could contain some error. While synthesizing data from various sources is beneficial for estimating the rates, the variations in the data acquisition processes across these sources could induce some estimation error. For example, the value of $P\{F > 0|p_{5_{t_1}}\}$ was estimated based on diagnosed data at the Rochester Methodist Hospital [21]. Whereas, the value of $P\{F > 0|\tilde{\mathcal{C}}_{t_2}\}$ was estimated based on expert observation reported by the American Cancer Society. Since these values were based on either large amount of data or long-term observations, we expect the error to be relatively small. Due to the unavailability of required data, some of the values were based on assumptions, hence creating room for estimation errors. For example, based on expert opinion in the literature, the time to progress was assumed to follow exponential distribution. Though we cannot empirically ascertain this assumption, based on the validation results, we believe that the exponential distribution is a good alternative. It may be noted that, this paper presents a model framework that has the potential to estimate population-specific progression rates, and the accuracy of the estimates can be improved as more data becomes available.

Future research Obtaining effective cancer intervention strategies encompasses not only development of screening strategies, but also analyzing factors pertaining to the availability of resources such as the patient's access to physician and hospital, and effective dissemination of evidence based information to the population. For example, it would be useful to assess the population's compliance to screening guidelines based on available knowledge of cancer screening tests and cancer risk factors. This knowledge can be relayed to the patient through interaction with their physician and/or through other sources. The model can then be used for a cost-effectiveness analysis of programs to increase risk awareness and its impact on reduction in CRC cases. It would also be interesting to model the impact of insurance policies under different system settings. Therefore, in addition to simulating the population entity, we will need to include entities like physicians and insurance policies, and their interactions. Note that, the current simulation model has been constructed as an agent-based model for the convenience of developing such a system-based simulation, which is part of our future research plan. Such a systems approach will allow for a more realistic analysis of feasible intervention strategies.

In summary, the probability model in the current state can be considered a base model that presents potential for use in developing population-specific intervention strategies. The work presented here can be used to support the need for collection of specific data required for analyzing and identifying more population-specific factors of interest. While the model developed in this manuscript has been specifically applied to colorectal cancer, most cancer types follow a similar pattern, i.e., incidence and progression. Following a similar procedure, but with disease specific modeling details, the current framework could be utilized for estimating progression and developing intervention strategies for other types of cancers as well. In addition, by inclusion of a transmission model in the current framework, cost-effectiveness analysis of prevention programs for infectious diseases like HIV/AIDS could be developed.

Acknowledgements The authors gratefully acknowledge Drs. Brad Doebbeling and David Haggstrom, Veterans Affairs at Indiana University, for their medical guidance and expertise to the project. The support of Dr. Joseph F. Pekny, Purdue University,

and Lori Nola Losee, Regenstrief Institute, is also appreciated. This research is part of the *Cancer Care Engineering* project (http://ccehub.org/) and was partly supported by a grant from Regenstrief Foundation, Indianapolis, Indiana.

The authors also acknowledge the insightful comments and suggestions of the anonymous reviewers.

Appendix A: Estimation of elements required for obtaining polyp incidence rates

In this section, we describe the estimation of elements on the right hand side of Eqs. 1 and 2 of Section 2.1.

A.1 Estimating $P\{F = f \cap R = r\}$

Using the definition of conditional probability, since $F = f$ and $R = r$ are dependent events, we can write $P\{F = f \cap R = r\} = P\{F = f|R = r\}P\{R = r\}$. To compute $P\{F = f|R = r\}$, we consider that the number of CRCs per family (i.e., the family history status F) is Poisson distributed with mean μ_r for a given race. We estimate μ_r for each race as $\left\{\frac{\text{Number of CRC cases in the population (i.e., CRC prevalence count)}}{\text{Total population}} * \text{Average family size}\right\}$. The CRC prevalence count for year 2006 for each race was obtained from SEER (Surveillance Epidemiology and End Results) database [36], which presents CRC statistics of the U.S. population. The total population count in year 2006 for each race was obtained from the U.S. census data. The average family size of the U.S. population is 3.20, as reported by the census [37]. With the inclusion of second degree relatives, we assume that the average family size for all race is 7. Using the Poisson distribution probability density function, we compute $P\{F = f|R = r\} = \frac{(\mu_r)^f e^{-\mu_r}}{f!}$, and $P\{R = r\}$ can be easily computed using the U.S. census data. For Eq. 1, we compute $P\{F = 0 \cap R = r\}$ as above, and $P\{F > 0 \cap R = r\} = (1 - P\{F = 0|R = r\})P\{R = r\}$. We present below the estimates of μ_r and $P\{F > 0|R = r\}100$.

i. Estimates of the Poisson parameter μ_r

R = all race	R = Caucasian	R = African American
0.026	0.028	0.018

ii. Percentage of population with family history of CRC given race ($P\{F > 0|R = r\}$*100)

R = all race	R = Caucasian	R = African American
2.55	2.77	1.77

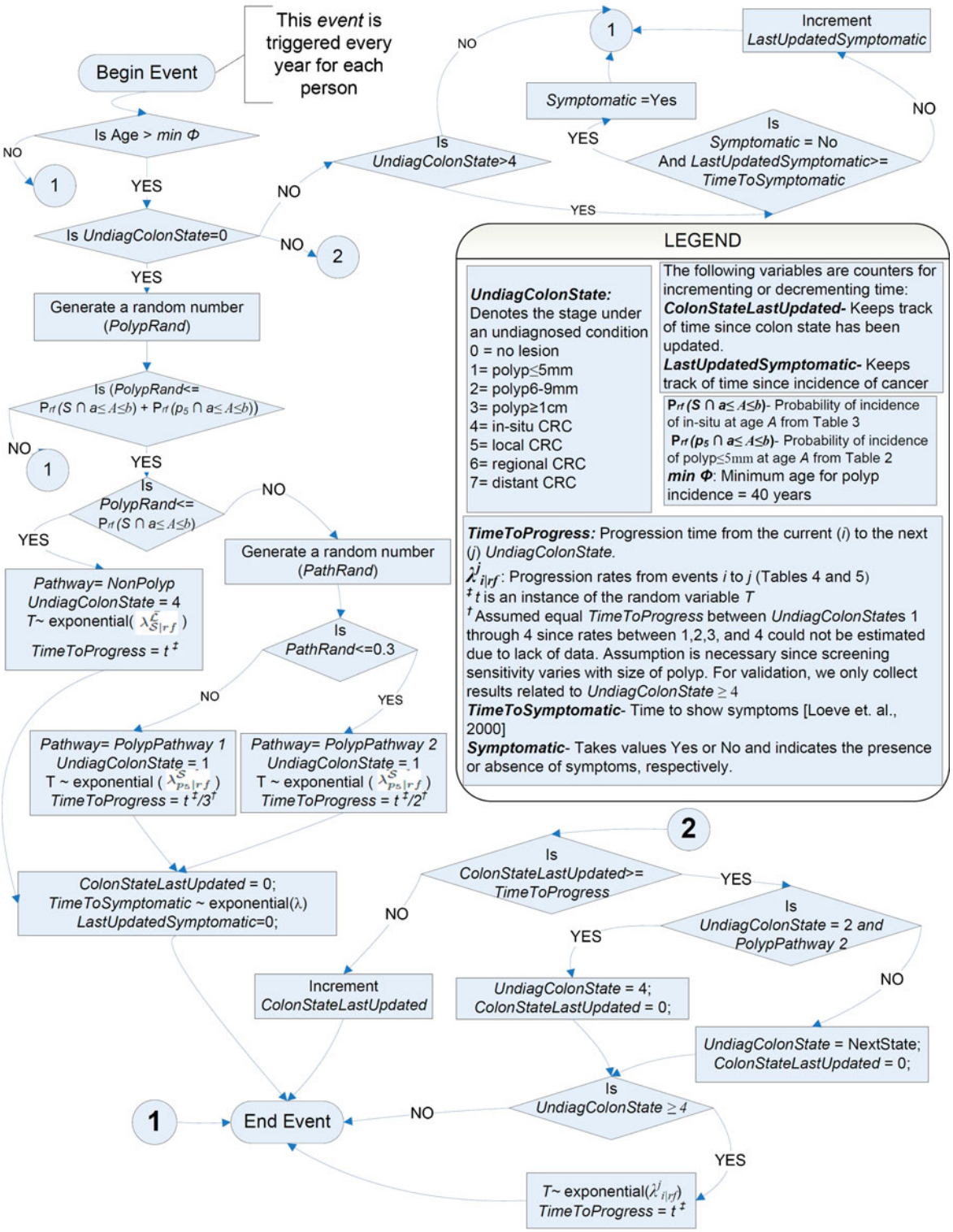
A.2 Estimating $P\{p_{5_{t_1}} \cap R = r \cap F = f\}$ and $P_{rf}\left\{(a \le A_{t_1} \le b)|p_{5_{t_1}}\right\}$

The article in [20] presents part of the results of a large-scale randomized Prostate, Lung, Colorectal, and Ovarian Screening Trial (PLCO) [38], that was conducted to test the effect of various screening tests on mortalities from the cancers. As part of the colorectal cancer study, initially, a population in the age group 55–74 was screened for colorectal polyps. On the population that tested negative (indicating normal), a repeated screening test was conducted three years after the initial screen test. The number of positive (indicating presence of polyp) screen results from the repeated test has been presented in [20]. The diagnosed polyps have been categorized into sizes <0.5, 0.5–0.9, and ≥1.0 cm. Since all polyps should have started as <0.5 cm with the event of incidence occurring during one of the three years between tests, we estimate the probability of incidence of polyp ≤ 5 mm at an arbitrary year t_1 ($P\{p_{5_{t_1}}\}$) as $\frac{1}{3} * \frac{\text{Number of people tested positive}}{\text{Total tested}}$, where, we multiply by $\frac{1}{3}$ by assuming that there were equal number of incidences in each of the three years. However, note that, age groups 40–54 and >74 were not part of the study population in [20], and hence, the above estimate of $P\{p_{5_{t_1}}\}$ will only apply to population in age 55–74. Therefore, to estimate $P\{p_{5_{t_1}}\}$ for the required population (i.e., age >40) we perform simple mathematical calculations using: i) the percentage distribution of polyps across age groups from [21], and ii) percentage distribution of U.S. population across age groups, taken from the U.S. census data.

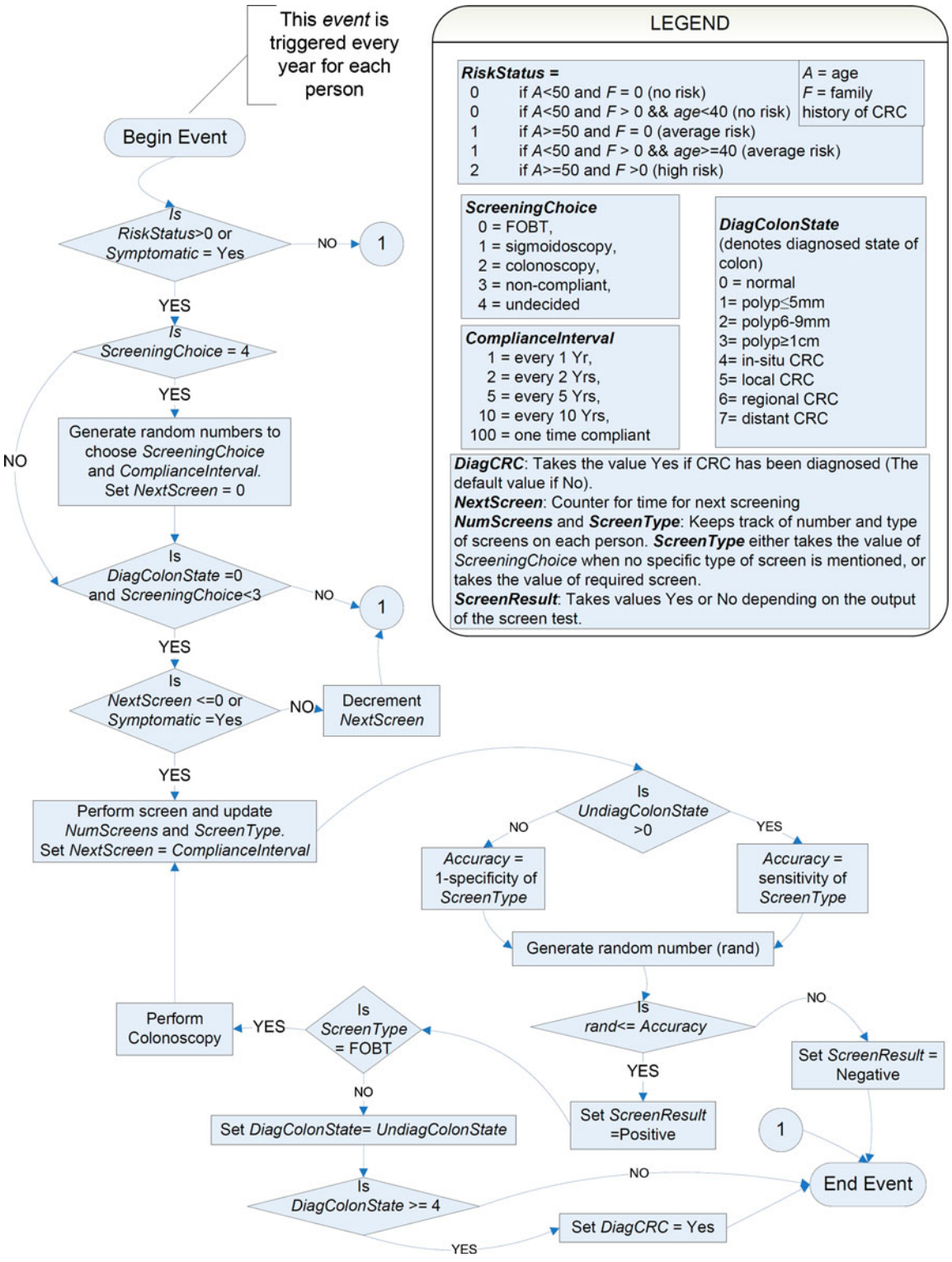
Applying the definition of conditional probability, we compute $P\{p_{5_{t_1}} \cap R = r \cap F = f\} = P\{(R = r \cap F = f)|p_{5_{t_1}}\}P\{p_{5_{t_1}}\}$. Since we do not have the data to determine dependence of events $F = f$ and $R = r$ when given event $p_{5_{t_1}}$, we assume independence and compute $P\{(R = r \cap F = f)|p_{5_{t_1}}\}P\{p_{5_{t_1}}\} = P\{R = r|p_{5_{t_1}}\}P\{F = f|p_{5_{t_1}}\}P\{p_{5_{t_1}}\}$. While we can estimate $P\{R = r|p_{5_{t_1}}\} = \frac{\text{Number of polyp cases in race } r}{\text{Total polyp cases}}$, we did not have suitable data to determine this proportion. Therefore, we approximated $P\{R = r|p_{5_{t_1}}\} = \frac{\text{Number of CRC cases in race } r}{\text{Total CRC cases}}$, data for which was obtained from the Indiana database [23]. We estimate $P\{F = 0|p_{5_{t_1}}\} = \frac{\text{Number of polyp cases with family history of CRC}}{\text{Total polyp cases}} = 0.14$ using data presented in [21], and $P\{F > 0|p_{5_{t_1}}\} = 0.86$. Also, $P_{rf}\{(a \le A_{t_1} \le b)|p_{5_{t_1}}\}$ is equated to the proportion of polyps in respective age groups as presented in [21].

Appendix B: Flowcharts of events in simulation model

B.1 Flowchart of simulation event 2: incidence and progression of polyps



B.2 Flowchart of simulation event 3: screening



Appendix C: Simulation parameters for the population of the state of Indiana

C.1 Screening parameters

i. Percentage of population compliant to screening [34]

Screening type	Percentage compliant
FOBT	43
Sigmoidoscopy	19
Colonoscopy	19
NeverCompliant	19

ii. Screening sensitivity [35]

Stage	FOBT	Sigmoidoscopy	Colonoscopy
Poly < 5 mm	0.02	0.75	0.80
Poly 6–9 mm	0.02	0.85	0.85
Poly > 1 cm	0.05	0.95	0.95
In-situ	0.05	0.95	0.95
Invasive CRC	0.60	0.95	0.95

iii. Screening specificity [35]

FOBT	Sigmoidoscopy	Colonoscopy
0.98	0.95	0.90

C.2 Assigning family history status for an individual

In determining the family history status for each person in the simulation, we assume that $F \sim Poisson(\mu_r)$ as described earlier in Appendix A. Note that, the CRC proportion (i.e., $\frac{\text{Number of CRC cases}}{\text{Total population}}$ in each race) for the State of Indiana is equivalent to the National proportion. Also, the average family size for the State of Indiana is 3.05 which is approximately equal to the National average (see Appendix A). Therefore, to generate random numbers in the simulation, we use μ_r as presented in Appendix A.

References

1. Morson B (1974) The polyp-cancer sequence in the large bowel. Proc R Soc Med 67:451–457
2. Neugut AI, Jacobson JS, DeVivo I (1993) Epidemiology of colorectal adenomatous polyps. Cancer Epidemiol Biomark Prev 2:159–176
3. Leslie A, Carey FA, Pratt NR, Steele RJC (2002) The colorectal adenoma–carcinoma sequence. Br J Surg 89:845–860
4. Loeve F, Boer R, Oortmarssen GJ, Ballegooijen MV, Habbema JDF (1999) The miscan–colon simulation model for the evaluation of colorectal cancer screening. Comput Biomed Res 32:13–33
5. Loeve F, Brown ML, Boer R, van Ballegooijen M, van Oortmarssen GJ, Habbema JDF (2000) Endoscopic colorectal cancer screening: a cost-saving analysis. J Natl Cancer Inst 92(7):557–563
6. Roberts S, Wang L, Klein R, Ness R, Dittus R (2007) Development of a simulation model of colorectal cancer. ACM Trans Model Comput Simul 18(1):1–30
7. Harper PR, Jones SK (2005) Mathematical models for the early detection and treatment of colorectal cancer. Health Care Manag Sci 8:101–109
8. Khandker RK, Dulski JD, Kilpatrick JB, Ellis RP, Mitchell JB, Baine WB (2000) A decision model and cost-effectiveness analysis of colorectal cancer screening and surveillance guidelines for average-risk adults. Int J Technol Assess Health Care 16(3):799–810
9. Wagner JL, Tunis S, Brown M, Ching A, Almeida R (1996) The cost effectiveness of colorectal cancer screening in average-risk adults. In: Young GP, Rozen P, Levin B (eds) Prevention and early detection of colorectal cancer. WB Saunders, Philadelphia, pp 321–356
10. Clemen RT, Lacke CJ (2001) Analysis of colorectal cancer screening regimens. Health Care Manag Sci 4:257–267
11. Leshno M, Halpern Z, Arber N (2003) Cost-effectiveness of colorectal cancer screening in the average risk population. Health Care Manag Sci 6:165–174
12. Vijan S, Hwang EW, Hofer TP, Hayward RA (2001) Which colon cancer screening test a comparison of costs, effectiveness, and compliance. Am J Med 111:593–601
13. National Cancer Institute NCI (2007) Cisnet-cancer intervention and surveillance modeling network. http://cisnet.cancer.gov/colorectal/profiles.html, http://cisnet.cancer.gov/publications/#Colorectal. Accessed on 3 February 2009
14. Winawer SJ, Fletcher RH, Miller L, Godlee F, Stolar MH, Mulrow CD, Woolf SH, Glick SN, Ganiats TG, Bond JH, Rosen L, Zapka JG, Olsen SJ, Giargiello FM, Sisk JE, Van Antwerp R, Brown-Davis C, Marciniak DA, Mayer RJ (1997) Colorectal cancer screening: clinical guidelines and rationale. Gastroenterology 112:594–642
15. Loeve F, Boer R, Zauber AG, van Ballegooijen M, van Oortmarssen GJ, Winawer SJ, Habbema JDF (2004) National polyp study data: evidence for regression of adenomas. Int J Cancer 111:633–639
16. Brenner H, Hoffmeister M, Stegmaier C, Brenner G, Altenhofen L, Haug U (2007) Risk of progression of advanced adenomas to colorectal cancer by age and sex: estimates based on 840,149 screening colonoscopies. Gut 56:1585–1589
17. Wong J-M, Yen M-F, Lai M-S, Duffy SW, Smith RA, Chen TH-H (2004) Progression rates of colorectal cancer by duke's stage in a high-risk group: analysis of selective colorectal cancer screening. Cancer J 10(3):160–169
18. Soetikno RM, Kaltenbach T, Rouse RV, Park W, Maheshwari A, Sato T, Matsui S, Friedland S (2008) Prevalence of nonpolypoid (flat and depressed) colorectal neoplasms in asymptomatic and symptomatic adults. J Am Med Assoc 299(9):1027–1035
19. Dukes CE (1932) The classification of cancer of the rectum. J Pathol Bacteriol 35(3):323–332
20. Schoen RL, Pinsky PF, Weissfeld JL, Bresalier RS, Church T, Prorok P, Gohagan JK (2003) Results of repeat sigmoidoscopy 3 years after a negative examination. J Am Med Assoc 290(1):41–48
21. Harewood GC, Lawlor GO (2005) Incident rates of colonic neoplasia according to age and gender. J Clin Gastroenterol 39(10):894–899
22. Noe M, Schroy P, Babayan R, Demierre M-F, Geller AC (2008) Increased cancer risk for individuals with a family

history of prostate cancer, colorectal cancer, and melanoma and their associated recommendations and practices. Cancer Causes Control 19:1–12
23. Epidemiology Resource Center Indiana State Department of Health and Division of Chronic/Communicable Disease (2004) Cancer incidence and mortality in Indiana. http://www.in.gov/isdh/reports/cancerinc/2004/section2c.htm. Accessed on 3 February 2009
24. American Cancer Society (2008) Cancer facts and figures 2008. American Cancer Society, Atlanta
25. Winawer SJ, Fletcher RH, Rex D, Bond J, Burt R, Ferrucci J, Ganiats T, Levin T, Woolf S, Johnson D, Kirk L, Litin S, Simmang C (2003) Colorectal cancer screening and surveillance: clinical guidelines and rationaleÚupdate based on new evidence. Gastroenterology 124:544–560
26. Jenkins L, Bradshaw D, Cannon P, Gierisch J, Freas W (2003) Colorectal cancer screening in local health departments—a pilot project of the North Carolina Advisory Committee on Cancer Coordination and Control and the North Carolina Division of Public Health
27. Epidemiology Resource Center Indiana State Department of Health and Division of Chronic/Communicable Disease (2004) Cancer incidence and mortality in Indiana. http://www.in.gov/isdh/reports/cancerinc/2004/section2r.htm. Accessed on 3 February 2009
28. American Cancer Society (2008) Cancer facts and figures 2008–2010. American Cancer Society, Atlanta
29. Epidemiology Resource Center Indiana State Department of Health and Division of Chronic/Communicable Disease (2004) Cancer incidence and mortality in Indiana. http://www.in.gov/isdh/reports/cancerinc/2004/section2s.htm. Accessed on 3 February 2009
30. Mandel JS, Bond JH, Church TR, Snover DC, Mary BG, Schuman LM, Ederer F (1993) Reducing mortality from colorectal cancer by screening for fecal occult blood. New Engl J Med 328(19):1365–1371
31. Mandel JS, Church TR, Ederer F, Bond JH (1999) Colorectal cancer mortality: effectiveness of biennial screening for fecal occult blood. J Natl Cancer Inst 91(5):434–437
32. Gilbertsen VA, McHugh R, Schuman L, Williams SE (1980) The earlier detection of colorectal cancers. Cancer 45:2899–2901
33. Argonne National Lab ANL (2007) Repast-recursive porous agent simulation toolkit. http://repast.sourceforge.net/. Accessed on 3 February 2009
34. Epidemiology Resource Center/Data Analysis Team Indiana State Department of Health (2002) Indiana health behavior risk factors 2001 state survey data. http://www.in.gov/isdh/reports/brfss/2001/index.htm. Accessed on 3 February 2009
35. Vogelaar I, van Ballegooijen M, Schrag D, Boer R, Winawer SJ, Habbema JDF, Zauber AG (2006) How much can current interventions reduce colorectal cancer mortality in the US. Cancer 107(7):1624–1633
36. National Cancer Institute (2006) Surveillance epidemiology and end results. http://seer.cancer.gov/faststats/selections.php#Output. Accessed on 3 February 2009
37. US Census Bureau (2009) 2006–2008 American community survey 3-year estimates. http://factfinder.census.gov/servlet/ACSSAFFFacts. Accessed on 3 February 2009
38. National Cancer Institute (2009) Prostate, lung, colorectal and ovarian cancer screening trial (plco). http://prevention.cancer.gov/programs-resources/groups/ed/programs/plco. Accessed on 3 February 2009