Validation of Models Used to Inform Colorectal Cancer Screening Guidelines: Accuracy and Implications

Carolyn M. Rutter, PhD, Amy B. Knudsen, PhD, Tracey L. Marsh, MS, V. Paul Doria-Rose, DVM, PhD, Eric Johnson, MS, Chester Pabiniak, MS, Karen M. Kuntz, ScD, Marjolein van Ballegooijen, MD, PhD, Ann G. Zauber, PhD, Iris Lansdorp-Vogelaar, PhD

Background. *Microsimulation models synthesize evidence about disease processes and interventions, providing a method for predicting long-term benefits and harms of prevention, screening, and treatment strategies. Because models often require assumptions about unobservable processes, assessing a model's predictive accuracy is important.* ***Methods.*** *We validated 3 colorectal cancer (CRC) microsimulation models against outcomes from the United Kingdom Flexible Sigmoidoscopy Screening (UKFSS) Trial, a randomized controlled trial that examined the effectiveness of one-time flexible sigmoidoscopy screening to reduce CRC mortality. The models incorporate different assumptions about the time from adenoma initiation to development of preclinical and symptomatic CRC. Analyses compare model predictions to study estimates across a range of outcomes to provide insight into the accuracy of model assumptions.* ***Results.*** *All 3 models accurately predicted the relative reduction in CRC mortality 10 years after screening (predicted hazard ratios, with 95% percentile intervals: 0.56 [0.44, 0.71], 0.63 [0.51, 0.75], 0.68 [0.53, 0.83]; estimated with 95% confidence interval: 0.56 [0.45, 0.69]). Two models with longer average preclinical duration accurately predicted the relative reduction in 10-year CRC incidence. Two models with longer mean sojourn time accurately predicted the number of screen-detected cancers. All 3 models predicted too many proximal adenomas among patients referred to colonoscopy.* ***Conclusion.*** *Model accuracy can only be established through external validation. Analyses such as these are therefore essential for any decision model. Results supported the assumptions that the average time from adenoma initiation to development of preclinical cancer is long (up to 25 years), and mean sojourn time is close to 4 years, suggesting the window for early detection and intervention by screening is relatively long. Variation in dwell time remains uncertain and could have important clinical and policy implications.* ***Key words:*** *colorectal cancer; preventive medicine; gastroenterology; discrete event simulation; simulation methods.* ***(Med Decis Making XXXX;XX:xx–xx)***

The Cancer Intervention and Surveillance Modeling Network (CISNET) consortium focuses on using models to guide public health research and priorities by improving our understanding of cancer control interventions in prevention, screening, and treatment and their effects on population trends in incidence and mortality. As part of CISNET, 3 independent groups developed microsimulation models for colorectal cancer (CRC) that have been used to inform policy decisions, including cost-effectiveness analyses for the Centers for Medicare & Medicaid Services[1–3] and a decision analysis for the US Preventive Services Task Force.[4,5]

Understanding the natural history of disease is critical to designing optimal prevention strategies. Microsimulation models synthesize evidence about disease processes and interventions to prevent or treat disease and are used to predict the impact of interventions that cannot otherwise be evaluated in an affordable, timely, or ethical manner.[6] The 3 CISNET-CRC models included in this validation were informed by studies describing observable disease processes, such as the prevalence of adenomas,[7] the incidence of CRC,[8] and systematic review of the

DOI: 10.1177/0272989X15622642

available evidence on the effectiveness of screening, including the sensitivity and specificity of available screening tests and the harms of screening tests.[9]

Models simulate the natural history of disease using best available evidence, but this evidence is imperfect. Furthermore, some disease processes simulated by models are unobservable. Model calibration is the process of selecting parameters, including those describing unobservable processes, to produce a model that yields predictions that match known or observed outcomes.[10,11] Models are often complex, and many different parameter combinations can result in good-fitting predictions. Therefore, model validation is a critical component of model development. Eddy and colleagues[12] outlined 5 types of validity: face validity, internal validity, cross validity, external validity, and predictive validity. Face validity refers to whether the model ''makes sense.'' Internal validity refers to coding accuracy. Cross-validity, also known as comparative modeling, refers to comparison of predictions from different models. External validity refers to how well the model is able to predict (or ''fit'') data that were not used for model calibration. Predictive validity takes this idea a step further and refers to how well the model is able to predict study outcomes before they are observed. External and predictive validity provide the most direct assessment of model accuracy, especially in terms of the ability to predict outcomes under new scenarios, which is the main purpose of microsimulation models. Such validation is relatively uncommon, largely because when models are developed, they are informed by all available data.[13]

This article presents an external validation of 3 CISNET-CRC models. All 3 models have been evaluated for face validity, reviewed for internal validity, and calibrated so that predictions reproduce the data used for model development.[14,15] Cross-validation of the 3 models demonstrated that they predict similar adenoma prevalence and lifetime incidence of CRC but different preclinical durations, that is, the time from adenoma formation to clinically detected CRC.[16,17] Preclinical duration is closely tied to screening effectiveness and is an important factor to consider when designing observational studies of screening.[18] External validation offers an opportunity to evaluate the impact of model assumptions, including assumptions about preclinical duration, on the models' ability to predict study outcomes.

We validated the 3 CISNET-CRC models to the United Kingdom Flexible Sigmoidoscopy Screening (UKFSS) Trial of once-only screening for CRC with flexible sigmoidoscopy. The UKFSS Trial was conducted in a population that was not yet routinely screened for CRC, so that published trial results, from screening outcomes[19] through 10-year incidence and mortality,[20] provide unique information about the risk of CRC after screening that can be used to shed light on the preclinical duration of CRC.

METHODS

The 3 models we evaluated are part of CISNET: Colorectal Cancer Simulated Population model for Incidence and Natural history (CRC-SPIN),[14] Simulation Model of Colorectal Cancer (SimCRC),[21] and Microsimulation Screening Analysis (MISCAN).[22] Detailed model descriptions can be found in cited articles and on the CISNET website, although this online resource includes updates to the MISCAN model made after this validation work.[23]

The models are similar in their overarching assumptions about the natural history of CRC: simulated individuals begin in a disease-free state and may progressively transition to an adenoma state, a preclinical CRC state, and a clinically detected CRC state, from which they may die of CRC (Figure

Received 23 March 2015 from RAND Corporation, Santa Monica, CA, USA (CMR); Institute for Technology Assessment, Massachusetts General Hospital, Boston, MA, USA (ABK); Department of Biostatistics, University of Washington, Seattle, WA, USA (TLM); National Cancer Institute, Health Systems & Intervention Research Branch, Bethesda, MD, USA (VPD); Group Health Research Institute, Seattle, WA, USA (EJ, CP); Department of Health Policy and Management, School of Public Health, University of Minnesota, Minneapolis, MN, USA (KMK); Department of Public Health, Erasmus MC, Rotterdam, Netherlands (MVB, IL); and Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, NY, USA (AGZ). This publication was made possible by financial support for this study by a grant from the National Cancer Institute (U01-CA-52959) as part of the Cancer Intervention and Surveillance Modeling Network (CISNET). The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of the National Cancer Institute. For all authors other than Zauber, the entirety of this work was supported by U01-CA-52959. Zauber was also supported by a Center Core Grant (P30-CA-008748). Revision accepted for publication 20 October 2015.

Supplementary material for this article is available on the *Medical Decision Making* Web site at http://mdm.sagepub.com/supplemental.

Address correspondence to Carolyn M. Rutter, PhD, RAND Corporation, 1776 Main St, Santa Monica, CA 90401-2138, USA; e-mail: crutter@rand.org.

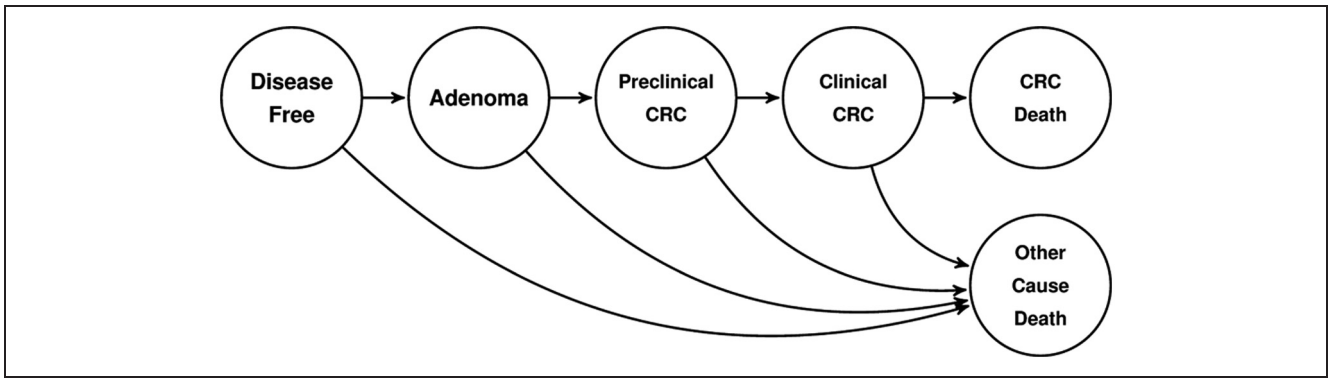

Figure 1 General model structure: all 3 models assume a progressive disease model, with all colorectal cancers (CRCs) arising from adenomas.

1). Individuals may die of other causes at any time. Each model simulates adenoma risk that varies systematically by sex and age and varies randomly across individuals. Individuals accumulate adenomas based on this risk. Each simulated adenoma is assigned a location in the colon or rectum. The SimCRC and CRC-SPIN models specify adenoma location distributions based on autopsy and (for CRC-SPIN) endoscopy studies. The MISCAN model specifies an adenoma location distribution based on the location of detected CRC. As a result, the MISCAN model assigns relatively more adenomas in the distal colon and rectum than the SimCRC or CRC-SPIN models.

Each model describes adenoma growth and transition to CRC, and none allow adenoma regression. Each model allows multiple adenomas and preclinical cancers within individuals, as well as variability in the duration of the adenoma state (the time to progression from adenoma to preclinical disease) across both individuals and adenomas within individuals.

Each model uses the same set of inputs describing overall and CRC-specific mortality. Age and sex-specific all-cause survival probabilities are based on estimates from the National Center for Health Statistics Databases.[24] CRC survival is based on analysis of data from the Surveillance, Epidemiology, and End Results (SEER) Program.[25] Models are governed by parameters that are selected (or calibrated) so that predictions match a set of targets. Each model is calibrated to match SEER CRC incidence rates in 1975–1979, a period when there was little or no CRC screening. The number of parameters varies across models: MISCAN calibrated 64 parameters, SimCRC calibrated 82 parameters (41 per sex), and CRC-SPIN calibrated 23 parameters.

Each of the models simulates variability in the ''dwell time'' of adenomas. Dwell time is defined for adenomas that transition to clinically detectable CRC as the time it takes for the adenoma to become a symptomatically detected cancer. The models do not include specific parameters that describe dwell time. Instead, dwell times are an output, something predicted by each model. Dwell time distributions are driven by assumptions about adenoma growth and progression. For example, the CRC-SPIN and SimCRC models allow all adenomas to progress to cancer (although most do not within a person's lifetime) and have longer dwell times, while the MISCAN model allows only some adenomas to progress, but these progressive adenomas tend to have shorter dwell times. Prior model comparisons found that the average and interquartile range (IQR) of predicted dwell times were 25.8 years (IQR: 17–33 years) for CRC-SPIN, 25.2 years (IQR: 15–33 years) for SimCRC, and 10.6 years (IQR: 5–14 years) for MISCAN and that among patients with CRC diagnosed at age 55 years, the predicted proportion arising from adenomas formed within the prior 10 years was 4% for CRC-SPIN, 10% for SimCRC, and 72% for MISCAN.[16]

Validation Study

We used our models to predict published results from the UKFSS Trial.[19,20] Participants in the UKFSS Trial were between the ages of 55 and 64 years and were randomized to receive either usual care (n = 113,178) or a one-time flexible sigmoidoscopy (n = 57,258).

Simulation of study results

To predict both average outcomes and variability in model predictions, each model simulated 2000 UKFSS ''trials'' that matched the study on sample

size, average age, age range, and percent male. Models specify that the only screening that occurred in study patients was the one-time flexible sigmoidoscopy in the intervention group; because there was no screening program in place in the United Kingdom during the period under study, we assumed no screening in the control group. Models assume complete follow-up of study participants who remain alive throughout the trial. Life-table methods were used to calculate model-predicted cumulative incidence and mortality that account for other-cause death.[26]

Assumptions about endoscopic examinations among intervention participants

The probabilities of undergoing flexible sigmoidoscopy and, among those screening positive, undergoing subsequent colonoscopy were based on published results.[19,27] Based on a previous study of UK endoscopists,[28] we assumed that flexible sigmoidoscopy examinations completely visualized the rectum for all individuals, completely visualized the sigmoid colon in 88% of individuals, and completely visualized the descending colon in 6% of individuals but never visualized the colon proximal to the descending colon. We assumed that colonoscopy examinations completely visualized the rectum for all individuals and completely visualized the entire colon for 95% of individuals.

Within the reach of the endoscope, we applied a lesion-specific sensitivity based roughly on a meta-analysis of colonoscopy miss rates.[29] MISCAN and SimCRC used sensitivities of 0.75 for adenomas 1 to 5 mm, 0.85 for adenomas 6 to 9 mm, and 0.95 for adenomas 10 mm and larger and for any preclinical cancer. CRC-SPIN used sensitivity (Se) that was a function of continuous size, s, with $\mathrm{Se}(s) = 0.66 + 0.0349s - 0.0009s^2$; when $s < 20$ mm (e.g., Se(3) = 0.76, Se(7.5) = 0.87, Se(12) = 0.95), $\mathrm{Se}(s) = 1$ when $s \geq 20$ mm and sensitivity equal to $\max(\mathrm{Se}(s),0.95)$ for preclinical cancer.

Assumptions about referral to colonoscopy

Study participants were referred to colonoscopy if any of 5 conditions were found at sigmoidoscopy: 1) a polyp ≥10 mm, 2) 3 or more adenomas, 3) an adenoma with high-grade dysplasia or villous components, 4) 20 or more hyperplastic polyps proximal to the distal rectum, or 5) preclinical cancer. The models simulate adenomas but not hyperplastic polyps, and so we assumed referral to colonoscopy when any of 3 conditions were found at sigmoidoscopy: 1) an adenoma ≥10 mm, 2) 3 or more adenomas, or 3) preclinical cancer. To simulate referral for trial participants with multiple hyperplastic polyps, large nonadenomatous polyps, or smaller adenomas with high-grade dysplasia or villous components, we estimated referral rates for participants with other findings using published results.[19,20,27] We estimated that participants with 1 or 2 small adenomas detected were referred to colonoscopy with probability 0.126, and participants with no adenomas detected were referred to colonoscopy with probability 0.006.

Surveillance after screening

We made assumptions about surveillance after screening that were consistent with recommendations in place in the United Kingdom during the trial[30]: 1) surveillance in 1 year when a very large (>2 cm) adenoma was detected or when 5 or more adenomas were detected and 2) surveillance in 3 years when a large (1–2 cm) adenoma or 3 to 4 adenomas were detected. Surveillance was simulated to end after 2 consecutive examinations with 2 or fewer small (<10 mm) adenomas detected. We arbitrarily assumed that 80% of individuals were completely adherent to adenoma surveillance and that 20% never underwent surveillance colonoscopy.

Sensitivity analysis

Sensitivity analysis was used to address uncertainty in assumptions about endoscopy reach and adherence to surveillance. We examined the impact of these model assumptions by considering runs with no surveillance colonoscopy and runs with less complete examinations (complete colonoscopy in 89% of participants and complete flexible sigmoidoscopy, to the sigmoid colon, in 86% of participants, with the descending colon completely visualized in only 1% of participants).

Outcomes

Our primary prediction targets were study-estimated hazard ratios of CRC incidence and mortality 10 years after screening in intervention v. control participants, based on per-protocol analysis comparing screened patients to controls, which are reported with confidence limits.[20] We selected CRC incidence and mortality because these were the targets of the trial and because these end points are the focus of screening. Furthermore, these outcomes constitute the end of the adenoma-carcinoma sequence and therefore provide information about the entire CRC

Table 1 Outcomes at 10-Year Follow-Up: Estimated and Predicted Hazard Ratios, and 10-year CRC Incidence and Mortality, Reported with 95% Confidence Intervals for Trial Estimates and with 95% Percentile Intervals for Model Predictions

					10-Year Rate per 100,000 Person Years	
Outcome	**Source**	**HR**	**95% Interval**	**Interval Width**	**Control**	**Screened**
CRC mortality	**Estimated**	**0.56**	**(0.45, 0.69)**	**0.24**	**44 (40, 48)**	**25 (21, 30)**
	CRC-SPIN	0.57	(0.44, 0.71)	0.27	38 (34, 42)	21 (17, 26)
	SimCRC	0.63	(0.52, 0.76)	0.24	52 (48, 57)	33 (28, 39)
	MISCAN	0.68	(0.53, 0.83)	0.30	37 (34, 41)	25 (21, 30)
CRC	**Estimated**	**0.68**	**(0.60, 0.76)**	**0.16**	**149 (143, 156)**	**100 (191, 110)**
Incidence	CRC-SPIN	0.62	(0.54, 0.69)	0.15	135 (129, 142)	84 (175, 193)
	SimCRC	0.74	(0.66, 0.82)	0.16	167 (160, 175)	127 (116, 139)
	MISCAN	0.86	(0.78, 0.94)	0.16	183 (175, 191)	160 (147, 173)

CRC, colorectal cancer; CRC-SPIN, CISNET: Colorectal Cancer Simulated Population model for Incidence and Natural history; HR, hazard ratio; MISCAN, Microsimulation Screening Analysis; SimCRC, Simulation Model of Colorectal Cancer.

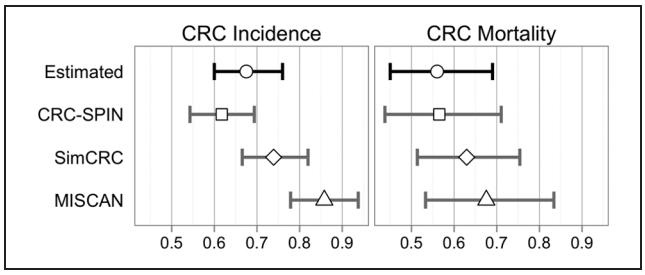

Figure 2 Hazard ratios: study-estimated and model-predicted intervention effects, with 95% percentile intervals for model predictions and 95% confidence intervals for estimates. Both estimated and predicted hazard ratios compare screening-adherent intervention participants to control participants who underwent no screening. CRC, colorectal cancer; CRC-SPIN, CISNET: Colorectal Cancer Simulated Population model for Incidence and Natural history; MISCAN, Microsimulation Screening Analysis; SimCRC, Simulation Model of Colorectal Cancer.

disease trajectory. We also described the model-predicted and study-estimated 10-year cumulative incidence and mortality in control and intervention groups.

Three secondary prediction targets were study-estimated screen-detected cancer and adenoma detection rates by location in the colon and rectum and stage at screen detection.[19] The rectum and sigmoid colon were defined as distal locations, and the descending colon was defined as a proximal location, consistent with reporting of UKFSS Trial results.

Model predictions were based on the average across 2000 simulated trials, with 95% percentile intervals estimated by the 2.5th and 97.5th percentiles across the 2000 simulated trials. We evaluated whether model predictions fell within the study's 95% confidence limit.

RESULTS

Effects of Intervention on CRC Mortality

All 3 models predicted relative hazard ratios for CRC-specific mortality in the intervention group relative to the control group that were within the study-estimated 95% confidence interval (Figure 2, Table 1). Predicted 10-year cumulative CRC mortality per 100,000 in the intervention group (Figure 3, Table 1) was lower than study-estimated CRC mortality for CRC-SPIN and MISCAN and higher than estimated for SimCRC.

Effects of Intervention on CRC Incidence

CRC-SPIN and SimCRC models predicted hazard ratios for CRC incidence in the intervention group relative to the control group that were within the study-estimated 95% confidence interval, and MISCAN predicted a weaker effect than estimated (Figure 2, Table 1). The predicted 10-year cumulative CRC incidence among the intervention group (Figure 4, Table 1) was lower than the study estimate for CRC-SPIN and higher for SimCRC and MISCAN models.

Disease Detection at Screening

Screening in the UKFSS Trial detected 140 cancers, CRC-SPIN predicted 47 screen-detected

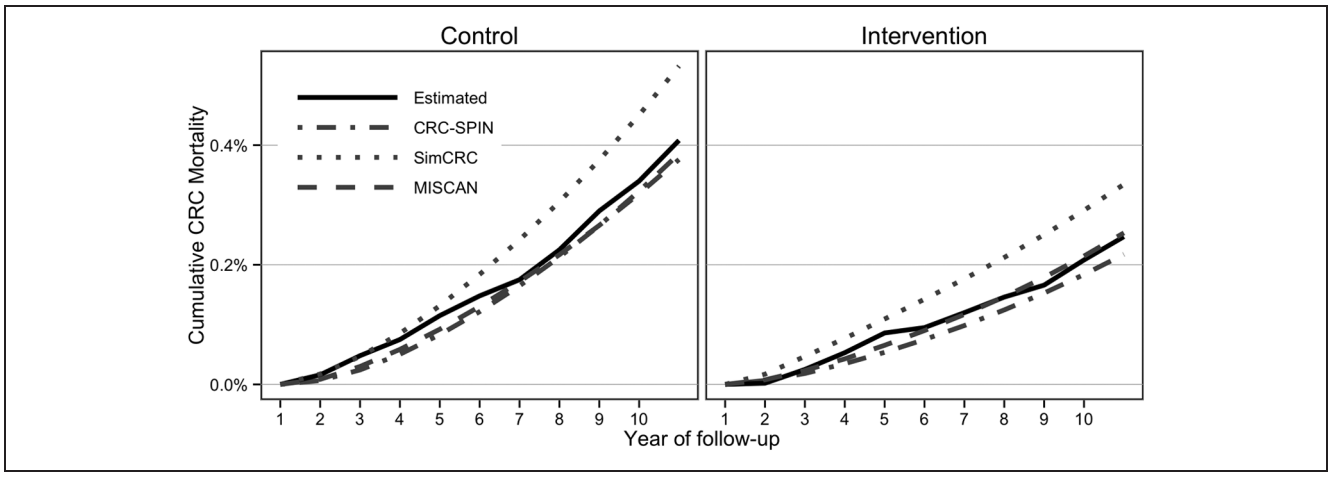

Figure 3 Cumulative mortality: study-estimated and model-predicted colorectal cancer mortality over the 10 years after randomization in intervention and control groups. CRC, colorectal cancer; CRC-SPIN, CISNET: Colorectal Cancer Simulated Population model for Incidence and Natural history; MISCAN, Microsimulation Screening Analysis; SimCRC, Simulation Model of Colorectal Cancer.

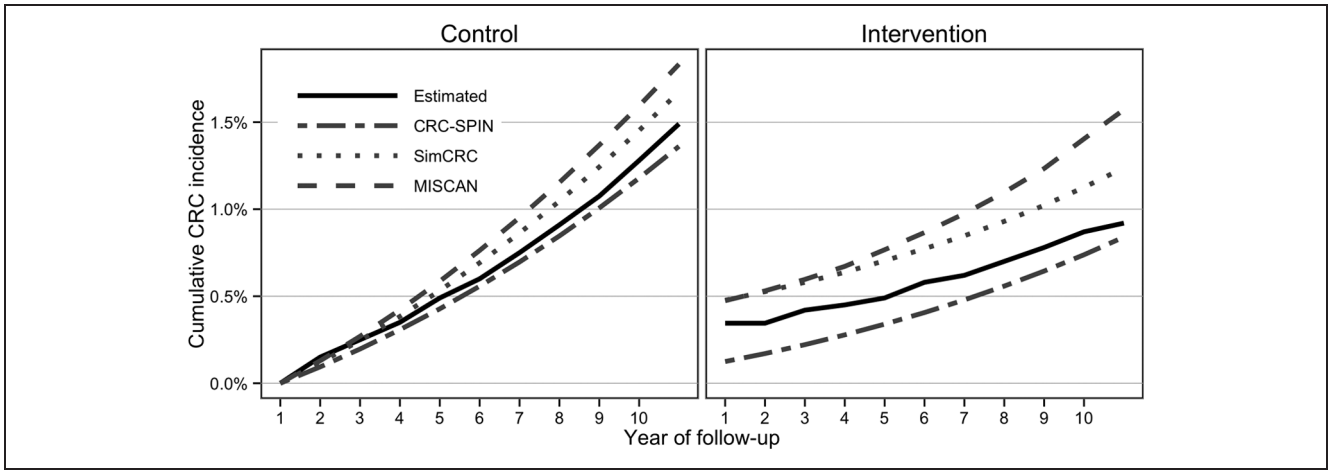

Figure 4 Cumulative incidence: study-estimated and model-predicted colorectal cancer incidence over the 10 years after randomization in intervention and control groups. CRC, colorectal cancer; CRC-SPIN, CISNET: Colorectal Cancer Simulated Population model for Incidence and Natural history; MISCAN, Microsimulation Screening Analysis; SimCRC, Simulation Model of Colorectal Cancer.

cancers, SimCRC predicted 180, and MISCAN predicted 181. Seventy-three percent of screen-detected cancers were early stage (I or II). The models predicted 90% (CRC-SPIN), 81% (SimCRC), and 77% (MISCAN) of screen-detected cancers at early stages.

Predicted adenoma detection rates at baseline flexible sigmoidoscopy (''distal'' location) were low for CRC-SPIN and SimCRC and high for MISCAN (Table 2). Colonoscopy referral rates follow similar patterns. Among those referred to colonoscopy, predicted adenoma detection rates (''proximal'' location) were high for all models.

Incidence in the Unscreened (Control) Group

The overall 10-year CRC incidence in the control group was 149 cancers per 100,000; the CRC-SPIN model underpredicted 10-year incidence (135 per 100,000); SimCRC and MISCAN overpredicted incidence (167 and 183 per 100,000, respectively) (Table 1). The UKFSS Trial reported nearly twice as many distal cancers than proximal cancers (98 and 51 per 100,000) (Table 3). The CRC-SPIN model predicted too few distal and too many proximal cancers. The SimCRC model predicted too many distal cancers

Table 2 Outcomes at Screening: Estimated and Predicted Percentage of Patients with Adenomas Detected, Referral to Colonoscopy, and Cancer Detected, Reported with 95% Confidence Intervals for Trial Estimates and with 95% Percentile Intervals for Model Predictions

Outcome	Source	Percent	95% Interval	Interval Width
Adenomas detected at flexible sigmoidoscopy	**Estimated**	**12.1**	**(11.8, 12.4)**	**0.6**
	CRC-SPIN	9.4	(9.2, 9.7)	0.6
	SimCRC	8.8	(8.5, 9.0)	0.6
	MISCAN	22.7	(22.4, 23.1)	0.8
Referred to colonoscopy	**Estimated**	**5.2**	**(4.3, 6.2)**	**1.9**
	CRC-SPIN	3.6	(3.4, 3.8)	0.4
	SimCRC	4.0	(3.8, 4.2)	0.4
	MISCAN	7.2	(7.0, 7.5)	0.5
Adenomas detected at colonoscopy	**Estimated**	**18.8**	**(17.1, 20.5)**	**3.4**
	CRC-SPIN	39.1	(36.7, 41.7)	5.0
	SimCRC	35.6	(33.4, 37.9)	4.5
	MISCAN	45.6	(43.8, 47.3)	3.5
CRC detected at screening	**Estimated**	**0.34**	**(0.29, 0.40)**	**0.11**
	CRC-SPIN	0.11	(0.08, 0.15)	0.07
	SimCRC	0.44	(0.38, 0.51)	0.13
	MISCAN	0.43	(0.37, 0.49)	0.12

CRC, colorectal cancer; CRC-SPIN, CISNET: Colorectal Cancer Simulated Population model for Incidence and Natural history; MISCAN, Microsimulation Screening Analysis; SimCRC, Simulation Model of Colorectal Cancer.

but matched the number of proximal cancers. The MISCAN model matched the number of distal cancers but predicted too many proximal cancers.

Sensitivity Analysis

Predictions from sensitivity analyses were very similar to primary results (data not shown). The modified assumptions only affected predictions in the intervention group and resulted in more pessimistic predictions of the effect of screening. For example, for a scenario with pessimistic reach assumptions and no participation in surveillance, the predicted hazard ratios for CRC incidence in the screened intervention v. control patients were 0.64 for CRC-SPIN, 0.75 for SimCRC, and 0.87 for MISCAN, and predicted hazard ratios for CRC mortality were 0.59 for CRC-SPIN, 0.64 for SimCRC, and 0.71 for MISCAN.

Model Changes

As a result of validation findings, the MISCAN model was recalibrated using the UKFSS Trial data, resulting in a longer average adenoma dwell times.[31] The updated MISCAN model-predicted hazard ratios for both 10-year CRC incidence and mortality were within the study error bounds (results in the online appendix).

DISCUSSION

Microsimulation models synthesize information from multiple sources, including randomized trials, observational data, and expert opinion, and provide decision makers with a method for carrying out in silico experiments comparing outcomes under different policy recommendations to inform policy decisions. Model validation, which evaluates the predictive accuracy of models, is critical because of the intended use of models, combined with the inherent uncertainty in model structure. We evaluated the accuracy of 3 models for CRC that have been used to inform policy decisions, comparing model-predicted outcomes to results from the UKFSS Trial.[19,20] This study is uniquely well suited for validation because it examines a one-time intervention, and contamination in the control group is unlikely because the study took place before a screening program was in place. The UKFFS Trial data provide information about the time it takes for new CRC to develop after detection and removal of adenomas detected by flexible sigmoidoscopy and subsequent colonoscopy among

Table 3 Ten-Year Incidence per Person Year in the Unscreened (Control) Group, by Location: Estimated and Predicted CRC Incidence per 100,000 Person Years among the Control Population, Reported with 95% Confidence Intervals for Trial Estimates and with 95% Percentile Intervals for Model Predictions

Outcome	Source	CRC per 100,000	95% Interval	Interval Width
Overall CRC	**Estimated**	**149**	**(143, 156)**	**13**
	CRC-SPIN	135	(129, 142)	14
	SimCRC	167	(160, 175)	15
	MISCAN	188	(180, 196)	16
Distal CRC	**Estimated**	**98**	**(92, 103)**	**11**
	CRC-SPIN	64	(59, 69)	10
	SimCRC	116	(109, 122)	13
	MISCAN	98	(92, 103)	11
Proximal CRC	**Estimated**	**51**	**(48, 56)**	**8**
	CRC-SPIN	71	(66, 77)	10
	SimCRC	51	(47, 56)	9
	MISCAN	85	(80, 90)	10

CRC, colorectal cancer; CRC-SPIN, CISNET: Colorectal Cancer Simulated Population model for Incidence and Natural history; MISCAN, Microsimulation Screening Analysis; SimCRC, Simulation Model of Colorectal Cancer.

referred patients. Before the UKFSS Trial, no other studies provided such information.

Each model describes the unobservable process of transition from benign adenoma to malignant tumor. The models are anchored by adenoma prevalence from autopsy studies and by CRC incidence in the prescreening era, but there are many ways to describe the transition between these 2 anchoring states. In past applications, these models have drawn similar but not identical conclusions about the relative benefit of different screening recommendations.[2] This validation exercise provided an opportunity to examine the impact of different transition assumptions on model predictions. All 3 models accurately predicted the relative effect of one-time flexible sigmoidoscopy on CRC mortality 10 years after screening. However, the models predicted the effect of screening on disease incidence and cumulative 10-year incidence and CRC mortality with varying degrees of success. When models missed validation targets, they missed in different ways. The accuracy of model predictions provides insight into the plausibility of different model assumptions.

A major difference between the models is the implied adenoma dwell time, with a more than twofold difference in dwell times between the MISCAN model compared to the CRC-SPIN and SimCRC models.[16] All 3 models accurately predicted relative effects of screening intervention on CRC mortality, but only the models with longer dwell times accurately predicted relative effects on CRC incidence. As shown in a prior model comparison,[17] the shorter dwell times implied by the MISCAN model allow CRC incidence in a screened group to return quickly to background incidence rates. These results suggest the average time from adenoma initiation to presentation with clinical CRC, at least in the distal colon, is longer than 10.6 years and may be closer to 25 years. This suggests that, on average, there is a long period of time when adenomas and preclinical cancers can be detected and removed before symptomatic detection.

Average dwell time is a simple summary of the dwell time distribution. The same average dwell time can arise from very different distributions, and the dwell time distribution underlying simulation models has implications for the predicted population-level effectiveness of different screening intervals. For example, consider 2 models with the same average dwell time: one that simulates a narrow range of dwell times and a second with a skewed dwell time distribution, resulting in wider variability in dwell times and a larger fraction of individuals with short dwell times who could benefit from shorter screening intervals. The model with dwell times that skew short is more likely to find benefit from shifting from a longer to a shorter screening interval. It is difficult to directly examine the effect of dwell time on model predictions because for all 3 models included in this validation dwell time is an output rather than a direct input or assumption. Models would need to be reconfigured to match different dwell time

distributions, and each model configuration would need to be calibrated to match target data. Therefore, we do not know how sensitive model predictions are to dwell time distributions. Future sensitivity analyses could explore the sensitivity of predictions to different dwell time distributions. Future validation exercises using longer term follow-up of the UKFSS Trial or randomized endoscopy studies with repeated examinations might also provide information about dwell time distributions.

Another difference between the models is mean sojourn time, the time from initiation of preclinical cancer to presentation with clinical CRC. Mean sojourn time is simulated as 1.6 years for CRC-SPIN, 4.0 years for SimCRC, and 4.7 years for MISCAN.[16] The mean sojourn time implied by the CRC-SPIN model may be too short, resulting in a low number of screen-detected cancers. In contrast, the SimCRC and MISCAN models more closely predicted the number of screen-detected cancers but detected too many. Unpublished model comparisons found that the lower rate of preclinical cancers predicted by CRC-SPIN more closely matched autopsy findings.[32] Preclinical CRC is a rare outcome and therefore difficult to calibrate. Our findings suggests that mean sojourn time lies between 1.6 and 4.0 years and may be closer to the upper end of this range.

All 3 models predicted detection of too many proximal adenomas in participants referred to colonoscopy following flexible sigmoidoscopy. This likely results from assumptions that are common across the models, because all models make the same error. One possible explanation is related to adenoma risk assumptions: all 3 models include a person-level adenoma risk component, which implies that people with distal lesions (preclinical cancer and adenomas) are more likely to have proximal lesions. The difference between model predictions and study estimates could occur because the correlation between proximal and distal lesions is not as strong as the models imply because proximal and distal colon cancers arise through somewhat different underlying processes.[33,34] An alternative (or additional) explanation is related to assumptions about colonoscopy accuracy: all 3 models also assume that colonoscopy is equally sensitive throughout the large intestine, but colonoscopy may be less sensitive in the proximal colon than in the distal colon and rectum. Observational studies have suggested reduced effectiveness of colonoscopy for prevention of proximal colon cancer.[35–38] Variation in the sensitivity of colonoscopy may be related to the occurrence of sessile serrated polyps, which are more common in the proximal colon, more difficult to detect than adenomas, and estimated to account for one-third of CRCs,[39] or to the quality of bowel cleanliness in the proximal colon relative to more distal segments.[40,41]

The models we validated produced similar results in terms of the relative effect of one-time flexible sigmoidoscopy on CRC mortality, but each model missed some targets and the specific targets missed varied across models. This analysis focused on validating each model independently and making comparisons across models to gain insights into how model assumptions influence model predictions. Our findings demonstrate the value of using multiple models to inform policy guidelines. Model predictions could be combined with the goal of obtaining more robust predictions that are closer to validation targets. Bayesian model averaging or ''ensemble modeling'' combines results from multiple models into a single set of predictions, with measures of uncertainty, and is potentially more robust than single-model prediction.[42] We know of only one example when model averaging was used to combine predictions across models used for guideline development.[43] Methods for handling the unique challenges faced in this context, such as the relatively limited data available to assess goodness of fit, are not yet fully developed.[44] The role of model averaging to inform policy is an important area for future research.

None of the models accurately predicted absolute CRC incidence rates in the intervention and control groups, none accurately predicted absolute CRC mortality rates in the control group, and one model did not accurately predict absolute CRC mortality rates in the intervention group. The failure of models to predict absolute rates could be the result of differences between the US population used for model calibration and the UK population used for validation. This suggests that it is important to calibrate models to target populations when they are used to predict absolute risks and differences. Overall mortality rates are similar in the United States and England up to age 65 years and are somewhat higher in England after age 65 years.[45] It is difficult to compare CRC risk, observed as CRC incidence and mortality rates, across the 2 countries because of differences in screening rates. We know of no studies that have undertaken this comparison. Finally, while studies such as the UKFSS Trial are key to our understanding of screening interventions, they are imperfect. Study participants can act in unexpected ways that can bias study results. For example, healthy volunteer effects can manifest in randomized trials if healthier patients, who are at lower risk for CRC, are more likely to enroll in a trial.[46]

Some might argue that our findings, with well-established models missing some validation targets, provide evidence that models should not be used to inform policy decisions. However, without modeling to inform policy, we are left to expert opinion, which is inherently subjective. In addition, our findings demonstrate the value of using multiple models to inform policy guidelines and the importance of validation as way to better understand model assumptions and to improve model accuracy.

Model validation to new data and to findings describing a range of natural history outcomes is uncommon. The International Society for Pharmacoeconomics and Outcomes Research and the Society for Medical Decision Making published guidelines for carrying out external validation,[12] but further thought needs to be given to how models are compared to study estimates and the relative importance of different validation targets (e.g., absolute rates and relative intervention effects, as well as the impact of interventions on precursor lesions). Furthermore, it is unclear how closely a model should match validation targets. Given the difficult task of validation, is it possible for a model to fit too well? What types of disagreements between predicted and study-estimated outcomes are acceptable for models to be useful and credible to clinicians and policy makers, given the variability of estimates?

We present the first validation of 3 CRC models used to inform screening guidelines. This validation is based on a single study that reported results spanning multiple steps in the CRC disease process. We found that our models accurately predicted the relative effect of screening on mortality, the primary outcome driving decisions about screening modalities. This example demonstrates how concurrent validation to multiple outcomes can uniquely inform the structure and accuracy of microsimulation models, ultimately providing insight into the disease process and enabling improved models with which to inform screening policy. Model accuracy, and thus reliability, can only be established through external validation analyses such as these and are therefore essential for any decision model.

REFERENCES

1. Lansdorp-Vogelaar I, Kuntz KM, Knudsen AB, Wilschut JA, Zauber AG, van Ballegooijen M. Stool DNA testing to screen for colorectal cancer in the Medicare population: a cost-effectiveness analysis. Ann Intern Med. 2010;153(6):368–77.

2. Knudsen AB, Lansdorp-Vogelaar I, Rutter CM, et al. Cost-effectiveness of computed tomographic colonography screening for colorectal cancer in the Medicare population. J Natl Cancer Inst. 2010;102(16):1238–52.

3. van Ballegooijen M, Habbema JDF, Boer R, Zauber AG, Brown ML. Report to the Agency for Healthcare Research and Quality: a comparison of the cost-effectiveness of fecal occult blood tests with different test characteristics in the context of annual screening in the Medicare population. August, 2003. Available from: http://www.cms.gov/medicare-coverage-database/details/technology-assessments-details.aspx?TAId=20

4. Zauber AG, Lansdorp-Vogelaar I, Knudsen AB, Wilschut J, van Ballegooijen M, Kuntz KM. Evaluating test strategies for colorectal cancer screening: a decision analysis for the U.S. Preventive Services Task Force. Ann Intern Med. 2008;149(9):659–69.

5. USPSTF. Screening for colorectal cancer: recommendation and rationale. Ann Intern Med. 2002;137(2):129-31.

6. Rutter CM, Zaslavsky AM, Feuer EJ. Dynamic microsimulation models for health outcomes: a review. Med Decis Making. 2010; 31(1):10–8.

7. Rutter CM, Yu O, Miglioretti DL. A hierarchical non-homogenous Poisson model for meta-analysis of adenoma counts. Stat Med. 2007;26(1):98–109.

8. Ries LAG, Kosary CL, Hankey BF, Miller BA, Edwards BK, eds. SEER Cancer Statistics Review, 1973–1995. Bethesda, MD: National Cancer Institute; 1998.

9. Whitlock EP, Lin JS, Liles E, Beil TL, Fu R. Screening for colorectal cancer: a targeted, updated systematic review for the U.S. Preventive Services Task Force. Ann Intern Med. 2008;149(9):638–58.

10. Stout NK, Knudsen AB, Kong CY, McMahon PM, Gazelle GS. Calibration methods used in cancer simulation models and suggested reporting guidelines. Pharmacoeconomics. 2009;27(7): 533–45.

11. Jackson CH, Jit M, Sharples LD, De Angelis D. Calibration of complex models through bayesian evidence synthesis: a demonstration and tutorial. Med Decis Making. 2015;35(2):148–61.

12. Eddy DM, Hollingworth W, Caro JJ, et al. Model transparency and validation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7. Med Decis Making. 2012;32(5):733–43.

13. Afzali HH, Gray J, Karnon J. Model performance evaluation (validation and calibration) in model-based studies of therapeutic interventions for cardiovascular diseases: a review and suggested reporting framework. Appl Health Econ Health Policy. 2013; 11(2):85–93.

14. Rutter CM, Miglioretti DL, Savarino JE. Bayesian calibration of microsimulation models. J Am Stat Assoc. 2009;104(488):1338–50.

15. Rutter CM, Savarino JE. An evidence-based microsimulation model for colorectal cancer. Cancer Epidemiol Biomarkers Prev. 2010;19(8):1992–2002.

16. Kuntz KM, Lansdorp-Vogelaar I, Rutter CM, et al. A systematic comparison of microsimulation models of colorectal cancer: the role of assumptions about adenoma progression. Med Decis Making. 2011;31(4):530–9.

17. van Ballegooijen M, Rutter CM, Knudsen AB, et al. Clarifying differences in natural history between models of screening: the case of colorectal cancer. Med Decis Making. 2011;31(4):540–9.

18. Weiss NS. Case-control studies of the efficacy of screening tests designed to prevent the incidence of cancer. Am J Epidemiol. 1999;149(1):1–4.

19. Atkin WS, Cook CF, Cuzick J, et al. Single flexible sigmoidoscopy screening to prevent colorectal cancer: baseline findings of a UK multicentre randomised trial. Lancet. 2002;359(9314):1291–300.

20. Atkin WS, Edwards R, Kralj-Hans I, et al. Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: a multicentre randomised controlled trial. Lancet. 2010;375(9726):1624–33.

21. Frazier AL, Colditz GA, Fuchs CS, Kuntz KM. Cost-effectiveness of screening for colorectal cancer in the general population. JAMA. 2000;284(15):1954–61.

22. Loeve F, Boer R, van Oortmarssen GJ, van Ballegooijen M, Habbema JD. The MISCAN-COLON simulation model for the evaluation of colorectal cancer screening. Comput Biomed Res. 1999;32(1):13–33.

23. CISNET. 2014 [cited 2014 April 30]. Available from: http://cisnet.cancer.gov

24. National Center for Health Statistics. US Life Tables 2000. Available from: http://www.cdc.gov/nchs/products/pubs/pubd/lftbls/life/1966.htm

25. Rutter CM, Johnson EA, Feuer EJ, Knudsen AB, Kuntz KM, Schrag D. Secular trends in colon and rectal cancer relative survival. J Natl Cancer Inst. 2013;105(23):1806–13.

26. Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. New York: John Wiley; 1980.

27. Wardle J, Miles A, Atkin W. Gender differences in utilization of colorectal cancer screening. J Med Screen. 2005;12:20–7.

28. Painter J, Saunders DB, Bell GD, Williams CB, Pitt R, Bladen J. Depth of insertion at flexible sigmoidoscopy: implications for colorectal cancer screening and instrument design. Endoscopy. 1999;31(3):227–31.

29. van Rijn JC, Reitsma JB, Stoker J, Bossuyt PM, van Deventer SJ, Dekker E. Polyp miss rate determined by tandem colonoscopy: a systematic review. Am J Gastroenterol. 2006;101(2):343–50.

30. Atkin WS, Valori R, Kuipers EJ, et al. European guidelines for quality assurance in colorectal cancer screening and diagnosis. First edition—colonoscopic surveillance following adenoma removal. Endoscopy. 2012;44(Suppl 3):SE151–63.

31. van Hees D, Habbemma J, Meester R, Lansdorp-Vogelaar I, van Ballegooijen M, Zauber AG. Should colorectal cancer screening be considered in elderly persons without previous screening? a cost-effectiveness analysis. Ann Intern Med. 2014;160(11):750–9.

32. Berg JW, Downing A, Lukes RJ. Prevalence of undiagnosed cancer of the large bowel found at autopsy in different races. Cancer. 1970;25(5):1076–80.

33. Carethers JM. One colon lumen but two organs. Gastroenterology. 2011;141(2):411–2.

34. Burnett-Hartman AN, Newcomb PA, Phipps AI, et al. Colorectal endoscopy, advanced adenomas, and sessile serrated polyps: implications for proximal colon cancer. Am J Gastroenterol. 2012;107(8):1213–9.

35. Brenner H, Chang-Claude J, Jansen L, Seiler CM, Hoffmeister M. Role of colonoscopy and polyp characteristics in colorectal cancer after colonoscopic polyp detection: a population-based case-control study. Ann Intern Med. 2012;157(4):225–32.

36. Baxter NN, Warren JL, Barrett MJ, Stukel TA, Doria-Rose VP. Association between colonoscopy and colorectal cancer mortality in a US cohort according to site of cancer and colonoscopist specialty. J Clin Oncol. 2012;30(21):2664–9.

37. Doubeni CA, Weinmann S, Adams K, et al. Screening colonoscopy and risk for incident late-stage colorectal cancer diagnosis in average-risk adults: a nested case-control study. Ann Intern Med. 2013;158(5 Pt 1):312–20.

38. Brenner H, Hoffmeister M, Volker A, Stegmaier C, Altenhofen L, Haug U. Protection from right- and left-sided colorectal neoplasms after colonoscopy: population-based study. J Natl Cancer Inst. 2010;102(2):89–95.

39. Rex DK, Ahnen DJ, Baron JA, et al. Serrated lesions of the colorectum: review and recommendations from an expert panel. Am J Gastroenterol. 2012;107(9):1315–29; quiz 4, 30.

40. Chokshi RV, Hovis CE, Hollander T, Early DS, Wang JS. Prevalence of missed adenomas in patients with inadequate bowel preparation on screening colonoscopy. Gastrointest Endosc. 2012;75(6):1197–203.

41. Chiu HM, Lin JT, Lee YC, et al. Different bowel preparation schedule leads to different diagnostic yield of proximal and nonpolypoid colorectal neoplasm at screening colonoscopy in average-risk population. Dis Colon Rectum. 2011;54(12):1570–7.

42. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial (vol 14, pg 382, 1999). Stat Sci. 2000;15(3):193–5.

43. de Koning HJ, Meza R, Plevritis SK, et al. Benefits and harms of computed tomography lung cancer screening strategies: a comparative modeling study for the U.S. Preventive Services Task Force. Ann Intern Med. 2014;160(5):311–20.

44. Lindstrom T, Tildesley M, Webb C. A Bayesian ensemble approach for epidemiological projections. PLoS Comput Biol. 2015;11(4):e1004187.

45. Banks J, Muriel A, Smith JP. Disease prevalence, disease incidence, and mortality in the United States and in England. Demography. 2010;47(Suppl):S211–31.

46. Pinsky PF, Miller A, Kramer BS, et al. Evidence of a healthy volunteer effect in the prostate, lung, colorectal, and ovarian cancer screening trial. Am J Epidemiol. 2007;165(8):874–81.