Erasmus School of Economics, Erasmus University Rotterdam
Business Analytics and Quantative Marketing

Supervisor
O. Karabağ PhD
External supervisors
I. Lansdorp-Vogelaar PhD
R.G.S. Meester PhD
L. de Jonge MSc

Author
Linde Pronk (547815)
Master's Thesis
Econometrics
2019-2020
2020-2021

Bayesian calibration of the MISCAN-Colon microsimulation model

A comparison of Bayesian algorithms for simultaneous estimation of test characteristics and cancer dwelling times

Abstract

In 2014, the Netherlands initiated a national screening program for colorectal cancer (van Hees et al., 2015). Microsimulation models such as MISCAN-Colon simulate the effects of screening interventions on population cancer incidence and mortality (Rutter et al., 2009). MISCAN-Colon's parameters estimated under the current Dutch policy are outdated and possibly inaccurate. The availability of new data on the outcomes of the Dutch CRC screening program over the years 2014 to 2017 and the employment of better-performing algorithms require a novel calibration. It is additionally examined how modeling test characteristics as a continuous function of age affects the performance of the estimation methods. The ABC-SMC algorithm and NUTS, both Bayesian methods, are compared for the simultaneous calibration of MISCAN-Colon's FIT characteristics and preclinical cancer dwelling times using new data on outcomes of the Dutch national CRC screening program. Test characteristics are modeled according to a sigmoid and Richard functional form which are subsequently compared as well across both algorithms. Based on the reported statistics, the NUTS algorithm shows more consistent and promising convergence behavior. However, although the fit is similar across the functional forms and targets, the ABC algorithm with the Richard form leads to the best fit. Regardless, the Richard form leads to implausible shapes for the algorithms' test sensitivities. Due to high computational time, only up to 25 iterations are considered in this study. Ideally, more iterations for a large population size are run to facilitate convergence of the algorithms and minimize stochastic error (Ozik et al., 2016).

Statement of originality

I, Linde Michelle Pronk, hereby declare that this document is written by me and that I take full responsibility for its contents.
I declare that the text and the work presented in this document are original and that no other sources than those mentioned in the text and its references have been used in creating it.
The Faculty of Economics and Business is responsible solely for the supervision of completion of the work, not for the contents.

Acknowledgements

Hereby I would like to thank the Erasmus University Rotterdam for the supervision of my studies. In particular, I would like to thank my supervisors, Oktay, Iris, Reinier, and Lucie for their extensive guidance, providing me with all the knowledge and tools required for writing an Econometrics master's thesis in the health sciences field, but above all for their encouragement and support. An additional word of thanks goes out to my colleagues and their willingness to aid me in all matters, small or big. My final word of thanks goes out to my family and friends who have warmheartedly supported and reassured me over the past year.

Contents

Table 1: TABLE OF ABBREVIATIONS

Abbreviation	**Definition**
CRC	colorectal cancer
NAAD	non-advanced adenoma
AAD	advanced adenoma
IC	interval cancers
EMC	Erasmus Medical Center
CISNET	Cancer Intervention and Surveillance Modeling Network
IKNL	Integraal Kanker Instituut Nederland
MISCAN	Microsimulation Screening Analysis
MSM	microsimulation model
FIT	fecal immunochemical test
Hb	hemoglobin
ABC	Approximate Bayesian Computation
MCMC	Markov Chain Monte Carlo
ABC-PMC	Approximate Bayesian Computation Population Monte Carlo
ABC-SMC	Approximate Bayesian Computation Sequential Monte Carlo
GOF	Goodness Of Fit
CI	confidence interval
ESS	effective sample size
MH	Metropolis-Hastings
RW-MH	random walk Metropolis-Hastings
HMC	Hamiltonian Monte Carlo
NUTS	No-U-Turn Sampler
TP	true positive
TN	true negative
FN	false negative

1 Introduction

Colorectal cancer (CRC) is one of the most common cancers worldwide (Bray et al., 2018). Bray et al. (2018) report that CRC took the third position for cancer incidence out of all cancers, in 2018. Additionally, it took the second leading position out of cancer casualties. Out of all countries, the Netherlands showed one of the highest incidence rates for CRC (Bray et al., 2018). However, CRC took a top-three position both concerning cancer incidence and cancer mortality already in 2000 as well as in 2012 (Parkin, 2001; Ferlay et al., 2013). This phenomenon urged the Netherlands to establish a national screening program which was initiated in 2014 (van Hees et al., 2015). Before elaborating further upon the Dutch CRC screening program, the definition of CRC is shortly introduced.

1.1 Colorectal cancer

CRC[1] is identified as a malignant tumor subsiding in either the colon or rectum. Such tumors do not emerge suddenly but generally start with the development of a polyp; a lump in the intestinal wall. Most polyps are benign and will remain so. However, some benign polyps may evolve into pre-malignant tumors, named adenomas. These adenomas may turn malignant. The probability of an adenoma turning malignant increases with age and other risk factors (RIVM, 2013). When turned malignant, the tumor is named cancer. At this stage, the tumor may spread to other parts of the body through the lymph system or bloodstream. This is called metastasis.

Table 2: Colorectal cancer stages

Stage	Grouping[a]	Description
0	T0, N0, M0	The cancer has not spread beyond the colon or rectum yet
I	T1-T2 N0, M0	The cancer has penetrated the intestinal wall but has not yet spread to any lymph nodes (N0) or other sites (M0)
II	T3-T4, N0, M0	The cancer has grown through the colon or rectum wall or has even penetrated nearby tissues and or organs but has not yet spread to any lymph nodes (N0) or other sites (M0)
III	T1-T4, N1-N2, M0	The cancer has spread to 1 up to 3 nearby lymph nodes (N1) or to 4 up to 6 nearby lymph nodes (N2a) or to more than 7 nearby lymph nodes (N2b) but has not yet spread to other sites (M0)
IV	T1-T4, N1-N4, M1	The cancer has grown through the colon or rectum walls, has affected nearby lymph nodes and has reached distant sites

[a] The TNM classification as specified by American Cancer Society (2018); T indicates the invasiveness of the cancer, N indicates the number of nearby lymph nodes the cancer has spread to and M indicates the state of metastasis to distant sites.

[1] Colon cancer, bowel cancer, intestinal cancer.

In order to determine if, how much and where cancer has spread, CRC is divided into four cancer stages; I through IV. This is done according to the TNM classification as described by the American Joint Committee on Cancer Staging (Table 2). Generally, the lower the stage, the less metastasis has occurred. An elaboration is found in American Cancer Society (2018) and American Cancer Society (2020).

When metastasis occurs and cancer spreads to other parts of the body, treatment prospects and overall prognosis deteriorate significantly. Therefore, early detection of cancer and removal of adenomas is essential for the survival of the patient. Detection and treatment as part of cancer screening lead to an improvement in quality of life and an increase in the life expectancy. This is because early detection should nullify the necessity of more drastic procedures in a later stage, and subsequent early treatment has a better prognosis, leading to a longer expected lifespan of the treated individual. Therefore, early detection of precursor lesions and early-stage cancers was one of the primary reasons for establishing the Dutch screening program. An elaboration on the setup of the Dutch screening program and how it was informed is, thus, essential.

1.2 MISCAN-Colon and the Dutch screening program

In order to quantify the possible harms and benefits of implementing a nationwide screening program, researchers at the Erasmus Medical Center (EMC) in Rotterdam were enquired to develop a model which could assess the requirements needed for such a large-scale operation. Hence, the Microsimulation Screening Analysis model for CRC (MISCAN-Colon) was developed. Microsimulation models (MSMs), such as MISCAN-Colon, simulate the effects of screening interventions on population cancer incidence and mortality (Rutter et al., 2009). By simulating these effects for different screening strategies, proper screening policy decisions could be made for the national screening program that was to be implemented. Based on the model results, it was decided that the Dutch screening program would include biennial screening with as test modality the fecal immunochemical test (FIT). The screening ages would lie in the interval of 55 to 75. However, the program was originally estimated to start in 2013 and eligible participants who would turn 75 in that year had already been invited. As a result, for the first year of the program, participants aged 76 and 77 were invited as well. According to a phased roll-out, participants were invited to the screening program by cohort (Figure 1). From the figure, it is seen that indeed the first year of the program included cohorts for individuals that would turn 76 that year.

The FIT test that was chosen as test modality measures hemoglobin (Hb) levels in the stool. In case this level exceeds the fixed cutoff value, the test is considered positive. A positive FIT is an indicator of the possible presence of a lesion, an abnormality in the tissue. When a FIT returns positive, the individual is referred for a diagnostic follow-up

test. In the case of the Netherlands, it was decided this would be a colonoscopy. During the colonoscopy, present lesions are verified. After the lesion(s) have been detected, an appointment is made to have it surgically removed. In case the lesion is still very small, it may be removed during the diagnostic colonoscopy. After confirming the presence of one or more lesions during a diagnostic colonoscopy, the patient enters a surveillance track. Therein, depending on the number and stage of the detected lesions, the patient will remain for a short or long interval respectively. This interval is set at 5 or 10 years. The follow-up test for surveillance is, again, a colonoscopy. If the surveillance period ends before the patient has reached the age of 75, they will re-enter the regular population screening program.

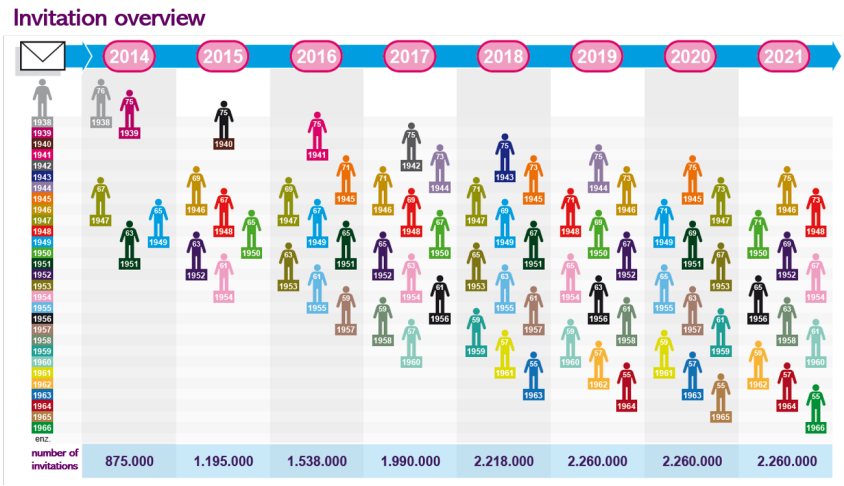


Figure 1: Invitation overview per cohort

Based on the analysis done with the MISCAN-Colon model, it was decided the initial cut-off value for the FIT test would be 15 μg Hb/g of feces. However, participation levels were higher than anticipated, and a shortage in colonoscopy capacity occurred (van Hees et al., 2014). Again, with the use of the MISCAN-Colon model, an analysis was performed which revealed the most cost-effective option was to elevate the FIT cut-off for referral to colonoscopy. Cost-effectiveness is measured in terms of life-years gained relative to the cost of colonoscopy and the loss of life-years due to CRC mortality. This decision lead to an increase from 15 μg Hb/g feces to 47 μg Hb/g feces in the second half of 2014. The current cut-off remains at this level. From their first screening onwards, individuals are biennially re-invited to participate in screening, given they did not attend the previous round or they had a negative FIT. Individuals that are not born in the same year will not attend their first screening in the same year. Hence, detection rates by round are not ordered by year but depend on the individuals' screening participation behavior. As a result, data on the outcomes of the national screening program in the Netherlands from 2014 to 2017 covers only round 1 and round 2 outcomes. Round 2 outcomes are at the earliest reported for 2016 since no individuals can have their second screening round before that year. Additionally, depending on the invitation overview (Figure 1), some

individuals will have their first screening later than 2014. This depends on the cohort they belong to.

1.3 Calibration of MISCAN-Colon

The employment of MISCAN-Colon for informing the Dutch screening policy is one of the many applications of MSMs. Similarly, cancer screening interventions for breast, cervical and esophageal cancer have previously been simulated with the use of MISCAN-Fadia, MISCAN-STDSIM, and MISCAN-ESO respectively. Based on the estimated effects, an optimal screening strategy is advised. Optimal is then defined as that strategy that will maximize the program effects on the health of the population while minimizing adverse effects of screening and the accompanying costs (Meester, 2017). Consequently, an optimal screening strategy will lead to the highest number of life-years gained while maintaining as low as possible costs of treatment and screening.

However, the current national screening program is established using data of old randomized controlled trials (RCT). These RCTs are trials in which treatment is randomly assigned to the experiment population. In this case, screening of individuals' stools was randomly decided. Moreover, model parameters that are difficult to determine through clinical expertise or previous research have been estimated using the Nelder-Mead estimation method. This method proves to be inferior compared to other algorithms with regards to multiple aspects. This concerns aspects such as accuracy, speed, the optimal solution found, and possibilities for constructing summary statistics and performance statistics. It has been shown that other sophisticated algorithms such as genetic algorithms (de Jonge, 2019) and Bayesian approaches (de Weerdt, 2019; Rutter et al., 2009) are superior due to the limitations of Nelder-Mead. As a result, the model parameters estimated under the current Dutch policy are outdated and possibly inaccurate. This may lead to an inaccurate prediction of future outcomes such as incidence and mortality rates and the number of life-years gained by screening. In turn, this may have an impact on the prognosis of many individuals in the Dutch population as well as introduce unexpected costs of the screening program for the Dutch government.

The availability of new data on the outcomes of the Dutch screening program and the employment of better-performing algorithms allows for a novel calibration. Through calibration, model parameters are estimated in such a way that they ensure that the predicted outcomes are as close as possible to the actual screening outcomes that have been observed. Better-performing estimation methods for determining MISCAN-Colon's model parameters will provide the Department of Public Health of the EMC with a model that remedies the limitations the current estimation method has.

The main objective of this research is, therefore, to find a proper estimation method for calibrating MISCAN-Colon's parameters using new data on screening. The realm of

candidate algorithms that is explored is reduced to that of Bayesian algorithms. Based on the studies that are presented in the Literature, two different methods are chosen and the differences in their performance examined. Specifically, an Approximate Bayesian Computation (ABC) approach (de Weerdt, 2019) is compared to a Markov Chain Monte Carlo (MCMC) approach that has not previously been used on models such as MISCAN-Colon. A discussion about the details on the ABC and MCMC approaches is found in the Methodology section.

While a calibration using a novel estimation method and newly available data is necessary, other extensions of MISCAN-Colon may be interesting as well. Specifically, we are interested in a model that can distinguish between age-specific differences within some of MISCAN-Colon's parameters on the FIT test's characteristics. This interest stems from the suspicion that test sensitivity is not constant but changes over different age levels. For younger individuals, test sensitivity is expected to be lower, such that fewer lesions are detected in younger individuals. This appears to be a rational consideration knowing that adenoma prevalence is lower at lower ages (Corley et al., 2013). Conversely, test sensitivity is expected to be higher in older individuals, such that more lesions are detected in older individuals. A similar effect is expected to hold for other test characteristics, such as test specificity and systematic components as well. This distinction between age-specific effects could be important and allow us to improve the model's performance by modeling a flexible relationship between the parameters and screening age. Possibly facilitating the inference of these parameters. Additionally, no study has previously addressed these effects directly. Here, we fill this gap in the literature. The research into age-specific dependencies can thus be considered as part of the contribution of this study to the literature. Correspondingly, this thesis will additionally examine how modeling the dependency of test characteristics as a continuous function of the screening age affects the estimation methods' performance.

As a result of the discussion provided above, this thesis will focus on the following key points.

- *"What will be the effective methodology for determining MISCAN-Colon's parameters?"*
- *How does modeling test characteristics as a continuous function of screening age affect the performance of the selected estimation methods?"*

By addressing these research questions, the current model is improved and the aforementioned parameters updated. This could lead to more accurately predicted future outcomes in terms of mortality and quality of life. As a result, this research will provide both the government and the EMC with the following benefits. The government will be provided with more accurate predictions of expected CRC mortality and expected benefits

of CRC screening in terms of quality of life metrics, avoiding the possibility of unprecedented costs and mortality rates. Whereas the EMC is provided with an updated model that can make predictions more accurately and more efficiently.

2 Literature

To answer the research questions posed in Section 1, it is of importance to elaborate upon the definition of calibration and to investigate what calibration practices are generally employed for microsimulation models such as MISCAN-Colon.

Stout et al. (2009) provide a comparison of multiple studies on cancer screening simulation models. The overarching definition of calibration handled by Stout et al. (2009) to find articles of interest is the following. "Formally, model calibration is the process of determining parameter values so that model output replicates empirical data." As a result, this indicates that the parameters that result from calibration are those parameters for which the model outcomes represent the observed outcomes as close as possible. Here, closeness is quantified using an error measure that is inherent to the calibration approach, also called the goodness of fit (GOF).

The reasons for performing a calibration are vast. Vanni et al. (2011) argue that one of the most common reasons for performing a calibration is because it allows us to estimate parameters that are generally unobservable. Since these parameters are not generally observable, we cannot determine them from regular estimation procedures, nor can they be informed by researchers' expertise or previous literature. Through calibration, we can use empirical data to infer them. An example of such non-observable parameters in a cancer screening setting is given by de Jonge (2019). One of the parameters in the MISCAN-EAC model for esophageal cancer is the dwelling time of preclinical cancer. Since preclinical cancer does not show any symptoms, its dwelling time cannot be directly determined. Therefore, other methods such as calibration are useful tools to infer such non-observable parameters.

Literature on the proper practices of calibration is scarce. Nonetheless, the next section discusses studies that do aim to provide a clear framework. The proposed practices largely coincide and lead to an approach consisting of seven steps. The subsequent sections follow the order of this framework and highlight the most important details of each presented calibration step. The selection of a proper algorithm for calibration is included in the steps of the proposed approach. The final section of this chapter justifies the selection of the ABC-SMC algorithm and the NUTS sampler for calibration of MISCAN-Colon based on those findings.

2.1 Calibration approach

In Stout et al. (2009), the authors attempt to find a unanimous approach to calibration for cancer-screening simulation models such as MISCAN-EAC and MISCAN-Colon. Their findings suggest a step-wise approach to calibration that shares many resemblance with the seven-step approach to calibration by Vanni et al. (2011) (Table 3). Although the approach suggested by Stout et al. (2009) is specifically constructed for cancer screening simulation models, it is a subset of the approach by Vanni et al. (2011) and only covers steps 2-6 of their approach. The most remarkable difference is that Stout et al. (2009) do not regard the identification of the calibration parameters of interest to be inherent to the calibration procedure, but more so to be inherent to the problem at hand. Regardless, both studies agree on the majority of essential steps needed for calibration.

Table 3: The seven-step approach to calibration

1.	Identification of parameters to vary in the calibration process
2.	Selection of calibration targets
3.	Selection of an appropriate GOF
4.	Selection of an appropriate estimation method
5.	Convergence/acceptance criteria of the selected method
6.	Selecting a proper stopping rule (typically convergence of observed outcomes or max number of parameter searches/iterations)
7.	Integration of calibration results into the model

The methodological decisions made for each of the steps in the approach are dependent on the calibration problem at hand. The following sections, therefore, provide a comparison of modeling decisions for each of the calibration steps based on the currently available literature.

2.2 Selection of calibration parameters and targets

Steps 1 and 2 (Table 3) urge the specification of the parameters to be calibrated and the calibration targets. Vanni et al. (2011) argue that the parameters to be calibrated are generally those parameters that are unobservable due to the unknown process that produces them. However, they additionally state that observable parameters can be included for calibration as well. The advantage of this practice is that it allows for modeling the correlation between input parameters. Despite this, they add that when correlation patterns are expected, this relationship should be explicitly modeled either way. Multiple parameters in cancer screening models are generally unobservable. Preclinical cancer

dwelling times are such parameters (de Jonge, 2019). Other uncertain parameters include test characteristics, costs of screening, and the transition probabilities for the progression of adenomas and preclinical cancers (Alarid-Escudero et al., 2019).

Necessarily, calibration parameters are calibrated with the use of calibration targets. Those are defined by Stout et al. (2009) as the empirical outcomes which are to be replicated by the model. Targets are chosen as those outcomes whom we expect to be influenced by the calibration parameters of choice. Hence, the chosen targets depend on the parameters of interest. In de Weerdt (2019), the author argues that when calibrating the parameter 'cancer risk', a proper calibration target would be the number of clinical cancers. This is because the number of people who are diagnosed with cancer is expected to be of influence on the risk of cancer an individual has. Alarid-Escudero et al. (2019) calibrate transition probabilities of adenomas and preclinical cancers on adenoma prevalence and CRC incidence for early and late stages. They argue that prevalence and incidence data are often used as calibration targets in cancer screening MSMs. Besides selecting the calibration targets based on their dependency on the calibration parameters, other selection criteria are possible. Stout et al. (2009) explain that the choice and number of calibration targets may also both depend on data availability as well as on the quality of the data. Besides, the authors argue that model complexity may influence the decision as well. However, Hazelbag et al. (2020) find that many studies select a number of calibration parameters that far exceed the number of selected targets. Regardless, neither Stout et al. (2009) nor Hazelbag et al. (2020) identify an issue with the number of calibration parameters exceeding the number of calibration targets. It is therefore unclear how model complexity influences the number of selected calibration targets.

To determine which parameter values lead to a proper fit of the model outcomes to the calibration target set and what is considered a proper fit, measures for determining the goodness of fit (GOF) have to be identified. Both Stout et al. (2009) and Vanni et al. (2011) argue that identifying a proper GOF for each calibration target is an essential step in the calibration procedure. This is supported by multiple other studies such as de Jonge (2019), van der Steen et al. (2016), Sai et al. (2019), and Karnon and Vanni (2011). The next section will provide a discussion about good practices in choosing a GOF metric and how this decision depends on the calibration problem at hand.

2.3 Selection of GOF measures

GOF metrics quantify the model's accuracy by measuring how well the model outcomes fit the observed outcomes. There are good and less good practices when deciding on appropriate GOF metrics. Vanni et al. (2011) propose a selection of various GOF metrics that are appropriate and which depend on the selected calibration targets. Popular choices are the least-squares measure, chi-squared measure, and likelihood-based measure (Vanni

et al., 2011). de Jonge (2019) and van der Steen et al. (2016) argue that in a disease-modeling setting, most calibration targets are binomially and Poisson distributed. As a result, for the likelihood-based measure, binomial deviance and Poisson deviance are popular deviances.

Table 4: Various common GOF statistics

Statistic	**Formula**
Sum-of-squared error	$\sum_x (y(x) - f(x\|\theta))^2$
Pearson's Chi-squared	$\sum_x (\frac{y(x)-f(x\|\theta)}{\sigma_x})^2$
Poisson	$2\left[y(x)ln\left(\frac{y(x)}{f(x\|\theta)}\right) - (y(x) - f(x\|\theta))\right]$
Binomial	$2[y(x)\left(ln(\frac{y(x)}{n_1}) - ln(\frac{f(x\|\theta)}{n_2})\right) +$ $(n_1 - y(x))\left(ln(\frac{n_1-y(x)}{n_1}) - ln(\frac{n_2-f(x\|\theta)}{n_2})\right)]$
Overall sum	$\sum_{i\ inK} W_i GOF_i$[c]

[a] $y(x)$ indicate the empirical outcomes. $f(x|\theta)$ indicate the simulated model outcomes conditional on the model parameters θ

[b] n_1 indicates the empirical dataset size. n_2 indicates the model simulated sample size.

[c] $K = 1, \ldots, k$ outcomes

While multiple GOF metrics are commonly used in practice, a clear guideline for which metric to use and when is missing (Vanni et al., 2011; Karnon and Vanni, 2011). However, both Karnon and Vanni (2011) and van der Steen et al. (2016) provide a comparison of the selection of GOF metrics mentioned earlier (Table 4).

In contrast to the findings of van der Steen et al. (2016), Karnon and Vanni (2011) find that a chi-squared GOF measure performs better compared to a likelihood-based approach. The chi-squared measure better distinguishes between the accuracy of different parameter sets as compared to the likelihood-based measure. However, they only test this hypothesis for a breast cancer model. They conclude other models may prefer the likelihood-based measure and that it is best practice to choose a GOF measure that will produce mean model outcomes that are closest to the observed outcomes.

Where Karnon and Vanni (2011) compare the chi-squared and the likelihood-based GOF measures based on the mean output being closest to the outputs associated with the best model fit, van der Steen et al. (2016) argue that each GOF criterion leads to the best model fit according to its incremental definition. According to them, comparing different GOF measures directly is not insightful. Instead, they provide a comparison of GOF measures based on three statistical measures. Specifically, a calibration procedure

on MISCAN-Colon using the currently practiced Nelder-Mead optimization method and a sample size of 100 million individuals is performed. They subsequently compare the sum of squared errors, Pearson chi-square, a Poisson-deviance based likelihood, and a binomial-deviance based likelihood as GOF measures. These four measures are evaluated based on the root mean squared prediction error of the selected parameters, the computation time of the procedure, and the impact on the estimated cost-effectiveness ratios. van der Steen et al. (2016) find that likelihood-based GOF measures work best for microsimulation disease models such as the MISCAN models, showing good performance in all calibration scenarios tried. While a comparison of various GOF measures is their main goal, van der Steen et al. (2016) additionally compare a grid search algorithm to the Nelder-Mead parameter search method and find similar results. This indicates that the likelihood-based GOF measures are expected to have consistent performance amongst other estimation methods as well.

Besides choosing the appropriate metric for each calibration target, both van der Steen et al. (2016) and Vanni et al. (2011), as well as Stout et al. (2009) and Kong et al. (2009) argue that for multi-objective optimization, where multiple calibration targets are used, a combined GOF metric can be computed. This allows for weighing the distinct targets differently. A simple idea is to use a weighted-sum approach, where the weight of each target outcome indicates the relative importance of the target (Kong et al., 2009). This leads to an approach where the most important targets may be assigned a weight of 1 and the targets which are less informative or contaminated in some way may be down-weighted, being assigned a weight smaller than 1. As Kong et al. (2009) argue, this may be because these targets characterize some rare event or because they are subject to measurement errors.

Since the study by van der Steen et al. (2016) illustrates that likelihood-based measures are expected to perform consistently amongst estimation methods, the estimation method can be chosen irrespective of the GOF metrics selected, and thus irrespective of the targets and parameters selected. A search of the literature stream concerning microsimulation and individual-based models serves to provide insight into the realm of currently employed estimation methods in a calibration setting. This should facilitate the decision for selecting a proper estimation method.

2.4 Selection of estimation methods

In order to determine which estimation methods could be appropriate for calibration of MISCAN-Colon, we've searched the literature for previous calibration procedures performed on microsimulation models such as MISCAN-Colon and the other MISCAN models. A handful of methodological estimation techniques have been explored by public health researchers with the goal of properly calibrating their model's parameters.

One such studies is performed by de Jonge (2019). This study focuses on the calibration of the MISCAN-EAC model for esophageal cancer. Calibration is performed on a subset of seven parameters that influence the calibration target parameter quantifying EAC incidence. Following the same GOF measure selection as is proposed by van der Steen et al. (2016), a genetic algorithm is employed for calibrating EAC incidence and Barrett's Esophagus prevalence. Subsequently, the performance of the genetic algorithm is compared to that of the Nelder-Mead simplex algorithm. In de Jonge (2019), the author elaborates on the limitations of the current Nelder-Mead Simplex estimation method. They summarize that it performs poorly and returns locally optimal solutions. Moreover, a single solution is provided instead of a set of solutions. Lastly, in the MISCAN models, calibration targets are often affected by multiple parameters. Correlation between parameters should be modeled to obtain unbiased results. The author in de Jonge (2019) continues that simulated annealing and genetic algorithms are used most often and have been suggested for calibration of other microsimulation models. Where simulated annealing performs better than grid search algorithms, it does not guarantee a globally optimal solution as opposed to a genetic algorithm.

2.4.1 Bayesian methods

Additionally, de Jonge (2019) states that the Bayesian method employed by Rutter et al. (2009) appears to be a suitable estimation method for calibrating microsimulation models as well. Bayesian-based approaches are attractive since they allow for modeling uncertainty through incorporating preliminary knowledge of the user. The Bayesian approach is frequently compared to the frequentist approach in statistical literature. It finds its basis in Bayes' rule (Equation 1).

$$\begin{aligned} p(\boldsymbol{\theta}|y) &= \frac{p(\boldsymbol{\theta}, y)}{p(y)} \\ &= \frac{p(\boldsymbol{\theta}) \cdot p(y|\boldsymbol{\theta})}{p(y)} \\ &\propto p(\boldsymbol{\theta}) \cdot p(y|\boldsymbol{\theta}) \end{aligned} \tag{1}$$

$$\Rightarrow \pi(\boldsymbol{\theta}|y) \propto \pi(\boldsymbol{\theta}) \cdot f(y|\boldsymbol{\theta})^{2} \tag{2}$$

The most prevalent difference between the Bayesian and frequentist approaches is the estimate that is found for the parameter(s) of interest. Where frequentist approaches aim to find point estimates of the parameter(s), Bayesian approaches aim to find a whole distribution of the parameters. This is called the posterior distribution $p(\boldsymbol{\theta}|y)$, from now on

[2]Notation as described in (Greenberg, 2014).

denoted as $\pi(\boldsymbol{\theta}|y)$ (Equation 2). In the Bayesian framework, $\boldsymbol{\theta}$ is regarded as an unknown quantity, and can therefore be considered a random variable with its own probability distribution (Greenberg, 2014; Minter and Retkute, 2019). From Bayes' rule, it is seen that the posterior distribution of the parameter(s) of interest $\boldsymbol{\theta}$ is proportional to its prior $\pi(\boldsymbol{\theta})$ multiplied by the theoretical empirical distribution $f(y|\boldsymbol{\theta})$. The multiplication of the two terms is alternatively called the posterior kernel. In accordance, Turner and van Zandt (2012) state that Bayesian methods assume two requirements for computing the posterior distribution of the parameters in $\boldsymbol{\theta}$. The first being the presence of the theoretical probability distribution of y conditional on $\boldsymbol{\theta}$ and the second being the prior distribution in which the preliminary knowledge of the user is contained.

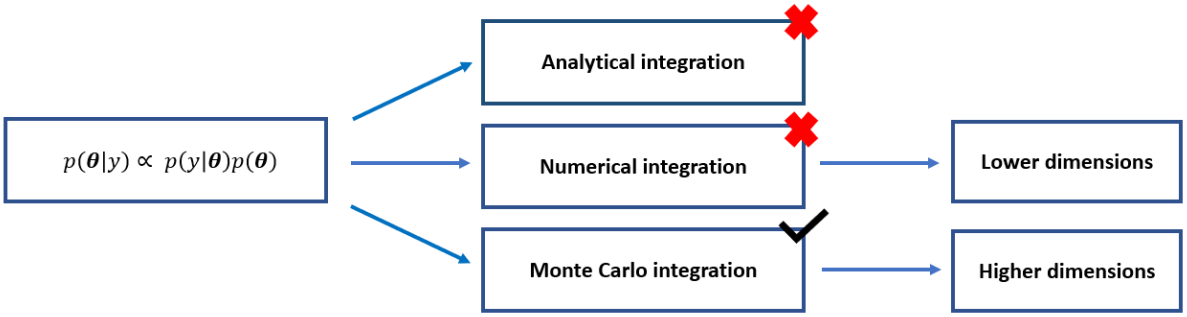


Figure 2: Infographic on distribution integration

As is the case for many frequentist approaches, the theoretical probability distribution is oftentimes intractable and cannot be solved analytically. This is emphasized by Hazelbag et al. (2020), who provide a systematic review of calibration of epidemiological models in infectious disease modeling. They explain that parameter calibration of individual-based models is often impeded because of analytical intractability of the likelihood. This is caused by the high complexity of the models. As a result, numerical methods have to be used to obtain posterior results (Paap, 2019). However, Hazelbag et al. (2020) additionally argue that deterministic numerical calculation of the likelihood is often prevented because of the complexity of the systems as well. Turner and van Zandt (2012) argue that this problem arises frequently in simulation-based models as well. Because of this, simulation-based methods that avoid a closed form of the likelihood have been developed. Out of these methods, MCMC methods belong to the chain of Monte Carlo integration techniques often used for sampling from the posterior distribution. Figure 2 shows a general overview of the consideration of the different integration methods.

MCMC methods are built upon Markov theory where the limiting distribution of the Markov chain equals the posterior distribution. Monte Carlo simulations are used for sampling from the Markov chain until it has converged. The resulting limiting distribution approximates the posterior distribution. Subsequently, the resulting posterior distribution can be used to calculate summary statistics such as the posterior means, variances, or marginal densities of the parameters.

Metropolis-Hastings sampler

One of the most frequently used MCMC samplers is the Metropolis-Hastings (MH) sampler. Rutter et al. (2009) employ the random-walk MH sampler for a simulated calibration study on US data using their own colorectal cancer model. They decide to reparameterize the likelihood such that $p(y|\theta)$ becomes $p(y|g(\theta))$. The reparameterization is justified by the same reason other integration methods for MSMs and individual-based models have to be proposed; the likelihood is intractable due to the unknown relationship of the MSM's parameters. Rutter et al. (2009) apply random-permutation to improve the mixing of the sampler. This facilitates the sampler's ability to move away from the starting point and explore the full support of the posterior distribution. As a result, they report the chain appears to mix slowly but does eventually seem to converge based on the reported Gelman-Rubin statistics computed from two chains. Additionally, priors and posteriors are visually examined to assess whether the posteriors had shifted away from the prior distributions. This was the case for most parameters, though there are some exceptions. Remarkably, duration parameters appeared to be largely informed by their priors, though this was not the case for (colon) cancer durations. Lastly, Rutter et al. (2009) report prediction intervals which indicate a good prediction of calibration data. However, some prediction intervals are rather wide, indicating uncertainty in some of the parameters. Rutter et al. (2009) conclude that although MCMC calibration is computationally expensive, its advantages justify their use over previously proposed calibration methods.

Approximate Bayesian Computation

The MH sampler by Rutter et al. (2009) shows promising performance. However, the MH sampler itself is rather outdated and over time, alternative methods have been proposed to solve the issue of the intractability of the likelihood. One such solution comes from Turner and van Zandt (2012) who propose a Bayesian method that does not require the theoretical probability distribution, called ABC. They particularly propose this method because it can be applied to simulation models as well, which are widely used in the social sciences. Turner and van Zandt (2012) explain that ABC is an estimation method in which the estimation of the likelihood function $p(y|\theta)$ is replaced by simulating a dataset X for some θ^*. Subsequently, the simulated dataset is compared to an empirical dataset Y using a pre-specified distance function, $\rho(X, Y)$. If the distance between the estimated and observed datasets is smaller than some tolerance level ε then the parameters θ^* that lead to the simulated dataset X are assumed to be reasonable. Turner and van Zandt (2012) describe the evolution of the simple rejection ABC algorithm, arguing that embedding MH computations into the algorithm may help when priors are uninformative and the

posteriors hence vary far from the priors. This yields an algorithm similar to the rejection algorithm, called ABC-MCMC. The exception comes from the acceptance ratio. Instead of accepting candidate parameter values by comparing distances with the tolerance directly, the candidate is accepted according to the acceptance ratio of the Metropolis-Hastings sampler. This probability is larger than 0 but not necessarily 1 when the distance function is smaller than the tolerance level, while the candidate is rejected indefinitely when the distance function is larger than the tolerance level. A downside to this embedment of the Metropolis-Hastings acceptance ratio is that when the proposal distribution is chosen poorly, the algorithm returns highly correlated samples (Turner and van Zandt, 2012). As a result, no unbiased summary statistics can be calculated. Moreover, the rejection rate can be very high when simulated data resulting from inferred parameters are required to be sufficiently close to the observed data.

Bergqvist (2020) illustrate the shortcomings of the ABC-MCMC algorithm with their study into the calibration of a random-effects model for breast cancer. In their approach, the researchers evaluate the ABC-MCMC algorithm using synthetic data that results from a simulation study on their random-effects model. They find that ABC cannot be used in conjunction with their model in the presence of screening. As a result, constraints need to be set. They, therefore, adopt the assumption that screening attendance is homogeneous amongst all participants. As such, individuals are screened at the same age and for the same number of rounds. A Euclidean-based distance and an uninformative flat prior are chosen. Three scenarios are compared, of which the first two only include symptomatic detection of tumors in absence of screening. The third scenario includes homogeneous screening attendance. In the absence of screening, the Metropolis-Hastings-based ABC-MCMC algorithm appears to show proper mixing and convergence after 5,000 iterations. However, reported acceptance rates of 0.0354 and 0.0762 for the first respectively second scenario lie lower than the required acceptance rate of regular MH. Bergqvist (2020) argue the inclusion of one-time breast cancer screening between ages 42 to 48 results in proper mixing and decent performance as well. However, for none of the parameters, the chain fully converges, even after 20,000 iterations. Another limitation outside of the downsides of the ABC-MCMC algorithm is that the study by Bergqvist (2020) is performed using a simulation study. The authors argue that the statistical properties of the algorithm should be further examined using information on observed outcomes.

Conclusively, ABC-MCMC shows limitations for its use in both MSMs and random-effects models such as those described in Bergqvist (2020). de Weerdt (2019) therefore employs a subsequent modified version of the ABC algorithm, the ABC-sequential Monte Carlo (ABC-SMC) algorithm, for calibration of MISCAN-Colon. In the algorithm, the posterior is approximated through moving from the prior through a set of intermediate distribution, consisting of multiple parameter sets called particles. Particles are weighted and perturbed samples from the previous intermediate distribution and each leads to a

separate set of simulated outcomes. In particular, de Weerdt (2019) aims to compare the performance of ABC-SMC as a method for calibrating CRC FIT sensitivities for large adenomas and cancers[3] to the Nelder-Mead Simplex method as calibration method. The latter being the main model for calibrating MISCAN-Colon at the time. She finds that for all comparable analyses, the average relative absolute error is lower for the ABC-SMC algorithm. The same observation is done for the deviances of the algorithms, though the deviance measures differ across the two. Additionally, all but one of the analyses terminates at their final tolerance level before reaching the maximum iteration level for the ABC-SMC algorithm. In contrast, Nelder-Mead appears to preliminary get stuck before it can reach the final tolerance level. Besides these findings, ABC-SMC shows better consistency than Nelder-Mead, indicated by lower average coefficients of variation. It is concluded that Nelder-Mead requires a considerably higher number of model runs than ABC-SMC. As a result, it is proven that ABC-SMC outperforms Nelder-Mead with respect to accuracy, consistency, and efficiency. A limitation of the study performed by de Weerdt (2019) is that ABC-SMC could only be calibrated on simulated data. Hence, proper FIT test sensitivities could not be estimated that would represent draws found for real-life data as we have for the national screening program in the Netherlands. This is the case for the study performed by Rutter et al. (2009) and Bergqvist (2020) as well, who largely base their results on simulated calibration data.

No-U-Turn sampler

Though the MH sampler is rather outdated, it seemed to perform well on some parameters in the calibration performed by Rutter et al. (2009). Furthermore, the ABC-SMC algorithm appeared to hold promising performance as well, at least compared to the previous benchmark Nelder-Mead algorithm. However, besides the MH sampler, other samplers within the MCMC scheme may prove to be promising as well. One of such samplers is the Hamiltonian Monte Carlo sampler (HMC). This sampler appears to be a proper substitute to the MH sampler by Rutter et al. (2009). They argue that the covariance matrix of the proposal distribution determines the direction the sampler takes. As a result, to efficiently and fully explore the support space, they state a reasonable estimate of the covariance matrix is required. However, Betancourt (2017) argues that regardless of how this covariance matrix is tuned, the random-walk MH sampler will explore the support space particularly slow in high-dimensional spaces. This is an issue when the calibration problem is large and many parameters have to be calibrated. Betancourt (2017) continues that in the worst case this leads to highly biased estimators. In less bad cases, we still end up with large autocorrelations and imprecise estimators.

The HMC sampler remedies the limited exploration of higher-dimensional spaces by

[3]Sensitivity $= \frac{TP}{Tp+FN}$; the probability of correctly finding a lesion in an ill individual.

being able to make larger jumps away from the starting point. This is done through exploitation of the differential structure by means of the gradient (Betancourt, 2017). Hoffman and Gelman (2014) explain this goes through first-order gradient information, which should lead the sampler away from parameterized focused neighborhoods and further towards what is called the typical set of the target distribution, in this case, the posterior distribution. In this way, HMC avoids the random-walk behavior of the MH sampler and facilitates convergence to the target distribution in high-dimensional problems (Chong and Lam, 2017).

Yet, Hoffman and Gelman (2014) argue there is a downside to HMC's efficiency. The step-size ε and the number of steps L are user-specified and as such complicate proper tuning of the sampler. Hoffman and Gelman (2014) extend the MH sampler, removing the need for hand-tuning these parameters. They call this no-tuning implementation of the HMC sampler the No-U-Turn Sampler (NUTS). They provide a comparison of NUTS with the Gibbs and random-walk MH samplers. They conclude that while the cost per iteration of the random-walk MH is equal to the cost per gradient evaluation of NUTS, it has barely explored the space, where NUTS appears to have generated various independent samples.

In a building energy models setting, Chong and Lam (2017) provide a comparison of the random-walk MH sampler, the Gibbs sampler, and NUTS as well. They find similar results as Hoffman and Gelman (2014), where NUTS is found to be more effective as compared to its counterparts. They compare the three algorithms by examining convergence and mixing abilities of the samplers based on Gelman-Rubin (GR) statistics and trace plots. To fairly compare the algorithms, they are run on an equal number of iterations. Initially, MH is faster than NUTS. However, after tuning the acceptance ratio of the MH sampler such that it lies between 20% and 25%, NUTS is significantly faster. The trace plots of the MH sampler indicate poor mixing. Additionally, the high values of its GR statistic indicate large variances between chains. This suggests its convergence is poor and more iterations are needed. In contrast, Chong and Lam (2017) argue the trace plots for the NUTS sampler show good mixing, and its GR statistics are close to 1, indicating proper convergence as well. Chong and Lam (2017) conclude that based on the trace plots and GR statistics NUTS is able to converge to the posterior distribution faster than MH, requiring significantly fewer numbers of iterations and no hand-tuning.

The Bayesian algorithms proposed here as candidates for calibrating MISCAN-Colon show promising performance in previous studies. While they are generally computationally expensive, researchers argue that the advantages outweigh the high computational time that is needed (Rutter et al., 2009; de Jonge, 2019). Such advantages include solving the intractability problem of the likelihood by means of sampling, construction of point and interval estimates, the inclusion of primary knowledge about the underlying process, and the ability to simultaneously calibrate multiple parameters that cannot be observed

directly. The latter is of particular interest in cancer screening problems where many parameters cannot be directly informed and the parameter space is vast.

2.5 Selection of convergence criteria and stopping rules

In order to determine the quality of the final solutions, the convergence of the estimation methods needs to be assessed. For MCMC algorithms, the Geweke test for single Markov chains and the GR statistic for multiple chains can be computed as formal tests of convergence (Geweke, 1991; Gelman and Rubin, 1992). Additionally, trace plots can be constructed to visually inspect the convergence of the samplers such as Rutter et al. (2009) suggest. Visual inspection of shifting of the posterior from its prior is another way to assess convergence, which is suggested by Rutter et al. (2009) as well. In addition, the effective sample size (ESS) and acceptance rates of the individual methods inform us about their convergence as well.

In addition to convergence criteria, it is desired to set stopping rules in case the algorithms do not converge or take a long time to convergence. The selection of possible stopping rules for algorithm termination is vast as well. The main stopping criteria that might facilitate comparison between algorithms is the maximum number of iterations. While, preferably, this is set such that all algorithms receive an equal chance of converging. Setting the maximum number of iterations equal across algorithms allows for comparison of the computational time needed for the algorithms to terminate. Particularly for the ABC-SMC algorithm, the tolerance level ε can be set to a desirable level such that the algorithm terminates before the maximum number of iterations is reached.

2.6 Determining calibration performance

Setting convergence criteria and identifying proper stopping rules facilitates the assessment of individual calibrations as well as facilitates comparing multiple calibration algorithms. Here, the stopping rules ensure the algorithms are compared under the same criteria. The convergence statistics subsequently can be used as performance metrics with which to compare the algorithms.

Indeed, in the comparison of three MCMC algorithms by Chong and Lam (2017), the performance of the algorithms is assessed by comparing their trace plots and Gelman-Rubin convergence statistics. In this way, it is determined how well each of the algorithms converges to the posterior distribution and, thus, which algorithm converges best. However, convergence alone does not implicate that the results found provide a good fit to the data. It is, therefore, straightforward to subsequently compare the algorithms both on their fit to the data. This can be done by plotting the simulated outcomes alongside the observed data, including confidence intervals (CIs) in order to compare model

fit across the algorithms. However, comparing model fit is possible through other means as well. van der Steen et al. (2016) find that for MSM such as the MISCAN models, likelihood-based GOF measures worked best. Additionally, these measures are dependent on the problem at hand and perform consistently across estimation methods. As a result, another way to compare different algorithms for the calibration of MISCAN-Colon is to compare their GOF. Continuing on with comparing algorithmic performance, ESSs and acceptance rates can be found both for MCMC algorithms as well as for ABC-SMC. In conjunction with other convergence criteria, these measures help indicate how well the algorithms perform and how their values stroke with the ESS and acceptance ratios we would expect for the different algorithms. Similar to the practices of Chong and Lam (2017), for the MH sampler an acceptance ratio between 20% and 25% is preferred. For the NUTS sampler, this ratio is much higher and lies somewhere around 80% (Paap, 2019). Lastly, setting equal stopping rules allows for comparison of computational time needed for each of the algorithms to reach the maximum number of iterations. MISCAN-Colon is a high-dimensional model and if the simulated population is large, model simulation alone can take up significant time. Therefore it is crucial that an upper bound on the number of iterations is set.

Based on the discussion posed in this chapter, the subset of the NUTS MCMC sampler and the ABC-SMC algorithm show to be promising. ABC-SMC proves to perform well in a simulation calibration study on MISCAN-Colon. The NUTS sampler, though not yet evaluated using any MSM, performs better than the MH sampler on multiple occasions, as proven by the studies performed by Hoffman and Gelman (2014) and Chong and Lam (2017). As a result, the ABC-SMC algorithm and the NUTS sampler will be used for the simultaneous calibration of MISCAN-Colon's test characteristics and dwelling times using new data on outcomes of the Dutch screening program. Considering both are Bayesian algorithms, which are proven to perform better than Nelder-Mead, their performance will be compared. This way, it is decided which of the two is most suited for calibrating parameters of MISCAN-Colon.

3 MISCAN-Colon simulation

The MISCAN-Colon model is the MSM developed by the Department of Public Health, EMC in Rotterdam in collaboration with the National Cancer Institute's Cancer Intervention and Surveillance Modeling Network (CISNET) which informs the current Dutch screening program. It aims to simulate the life histories of individuals and quantify subsequent effects of the intervention of screening. Habbema et al. (1985) provide a concise description of the model workings.

"Basically, the MISCAN program first simulates a large number of individual

life histories, according to assumptions (input specifications) concerning the epidemiology and the natural history of the disease under consideration. Then, these life histories are subjected to screening, according to assumptions (input specifications) on screening policy, attendance, characteristics of the screening test, and prognostic consequences of early detection. Some of the life histories will be changed by this simulated screening experience. These changes, be it in a favourable or in an unfavourable sense, constitute the simulated effect of screening. (Habbema et al., 1985)"

In our own words, for each individual with their own risk of developing CRC, the model simulates the onset of adenomas and development to CRC, screening participation, screening outcomes, and possible follow-up and surveillance. A graphical representation of the simulation of a single individual life-history is displayed in Figure 3[4].

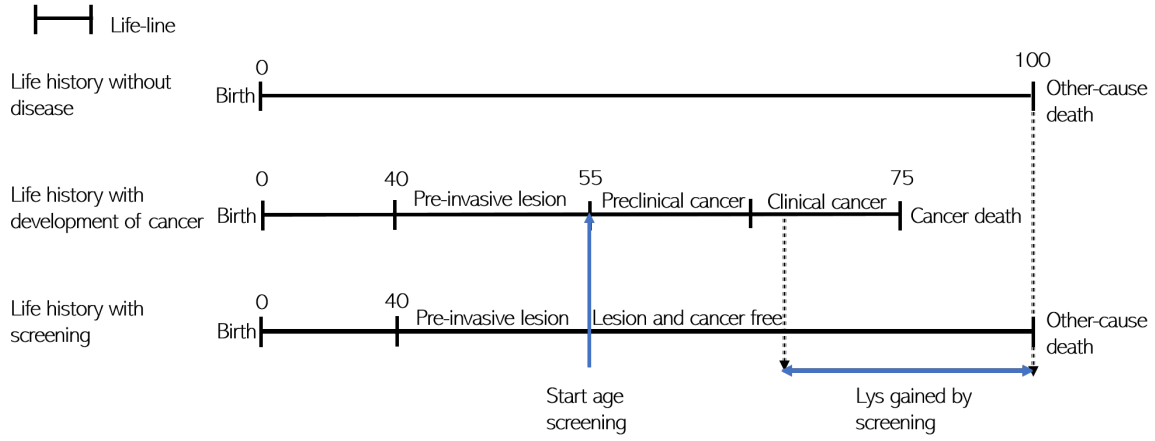


Figure 3: Example of a life-history as simulated in MISCAN-Colon

Individuals' life histories, the onset and development of lesions, and the intervention of screening are modeled according to three modules. These are the demography module, the natural-life history module, and the screening module.

3.1 Demography module

The demography module simulates an individual's life-history without disease. This is displayed in the top lifeline of Figure 3. This means a date of birth is simulated for the individual as well as a date of death from other causes than cancer, which results in a life-history where cancer is absent. For the Dutch setting, dates of birth are based on the cohorts eligible for screening (Figure 1). The date of death in turn is drawn based on a table containing probabilities to die at a certain age. Such tables are freely available from sources such as the Centraal Bureau voor de Statistiek.

[4]Figure replicated from Naber (2017).

3.2 Natural history module

The natural history module simulates the onset of small adenomas and their progression into different CRC stages. While not all individuals will develop a lesion in their lifetime, some may develop multiple, depending on individual characteristics such as age and CRC risk. The lesions that we differentiate between in MISCAN-Colon are adenomas, preclinical cancers, CRCs, and interval cancers. Adenomas progress from small ($\leq 5mm$) into medium adenomas ($6-9mm$), medium adenomas into large adenomas ($\geq 10mm$), and large adenomas progress into preclinical cancers. Preclinical cancers are broken down into four stages. Each stage indicates the transition to one of the four cancer stages characterized in Table 2.

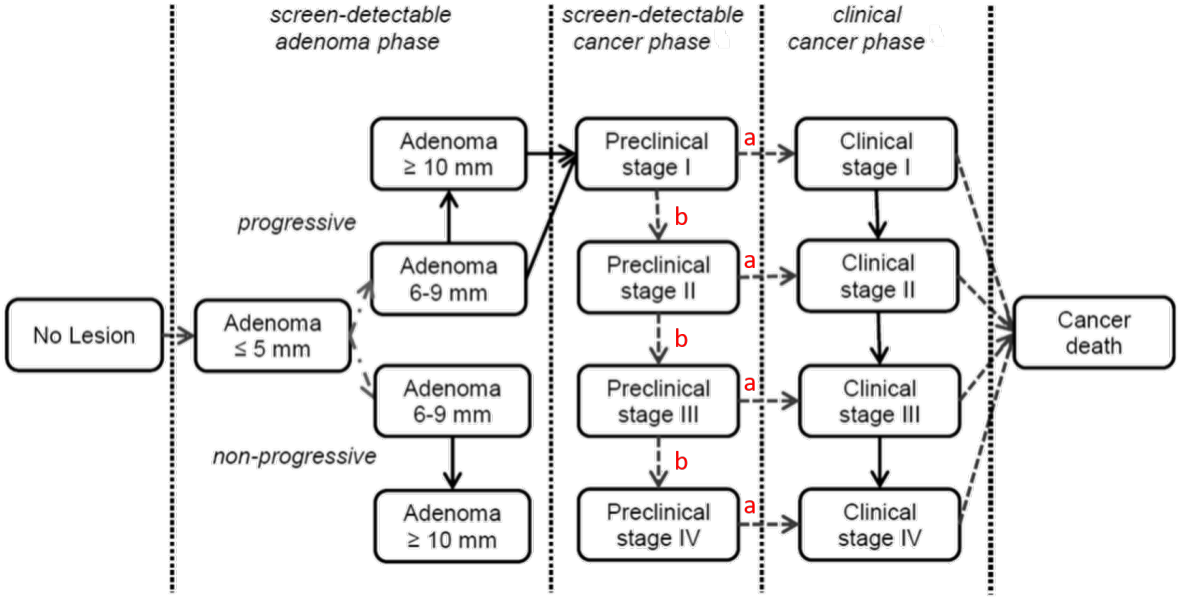


Figure 4: Diagram of the transition of states in MISCAN-Colon

Transitions between states in the MISCAN-Colon model follow a Markov process (Figure 4[5]). Except stage IV preclinical cancer, each of the four preclinical stages can be divided into a short before clinical diagnosis state, denoted by subscript a and a long before clinical diagnosis state, denoted by subscript b. Stage IV preclinical cancers can only progress into clinical cancer and therefore belong to the short before clinical diagnosis states denoted by subscript a. We assume lesions can only progress and thus cannot return to previous, less malignant states.

Transitions between states depend on the dwelling time of the current state. This dwelling time indicates the time spent in the current state, independently of age and risk. For each state, dwelling times are modeled according to a dwelling time distribution. For

[5]Figure adapted from Cancer Intervention and Surveillance Modeling Network (2015).

the transitions shown in Figure 4, this is an exponential distribution with parameter the mean dwelling time of the corresponding state (Loeve et al., 2000). The model requires mean dwelling times as input for these distributions. Dwelling times of CRC cannot be observed in the data and therefore they need to be calibrated. Current input values for dwelling times result from previous calibrations and previously performed clinical trials.

3.3 Screening module

Finally, the screening module adds the intervention of screening to the life-history of an individual in which the onset of adenomas and development of CRC has been simulated by the natural history module. The screening module introduces the practice of screening for CRC. This is done by specifying a test modality, in the Dutch setting the FIT, for which we specify test characteristics. These test characteristics describe how well the test is able to detect lesions in an individual. In MISCAN-Colon we specify test sensitivity, lack of specificity, and systematic lack of sensitivity. Sensitivity is the possibility of a true positive (TP) and hence describes the probability of finding a lesion in an ill individual. Specificity is the probability of a true negative (TN) and hence describes the probability of finding nothing in a healthy individual. Lastly, lack of sensitivity is the probability we do not find something in an ill individual, a false negative (FN). The systematic component then indicates we systematically always find nothing in an ill individual. One scenario in which adenomas are missed is for the case of adenomas that do not bleed for several years (Van der Meulen et al., 2016). Similarly as for the dwelling times, it is impossible to observe whether we correctly found a true negative nor is it possible to know whether we consistently miss a lesion in an ill individual if nothing is found. These parameters are therefore specifically of interest for calibration.

Before closing the explanation of MISCAN-Colon, we go more in depth about how the test characteristics are currently modeled with respect to age. In chapter 1 it was argued that it is suspected that test sensitivity may differ across individuals of different ages. As a result, test characteristics are specified separately for different age groups. That means that within the model different test sensitivities, lack of specificities, and systematic lack of sensitivities are specified for individuals aged [55,59), [60, 64), [65, 69), [70, 74), and [75+). Since sensitivities and systematic lack of sensitivities are specified for each adenoma size group and preclinical cancer, a large set of parameters results for each age group. It is therefore not optimal to calibrate all test characteristics while they are specified separately by state for each of the four age categories.

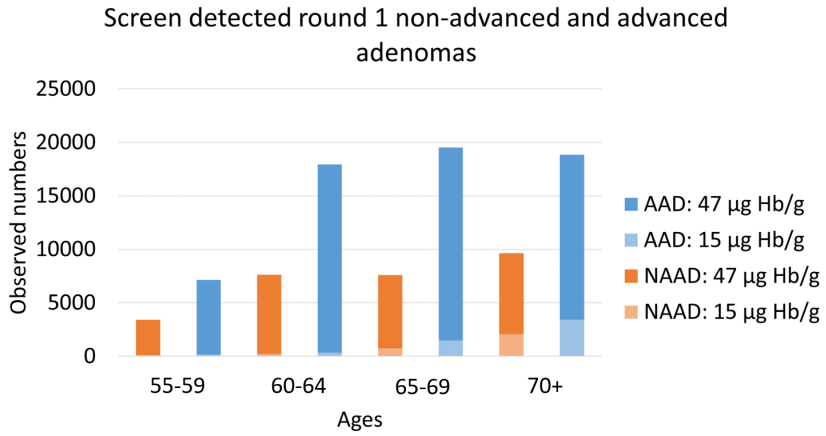
As a result, modeling the test characteristics as a continuous function of age not only allows for more flexibility for estimating them, but it reduces the number of parameters needed to be estimated as well. Before elaborating upon the specification of the functional form of the test characteristics as a function of age, the data used for calibration is discussed.

4 Data

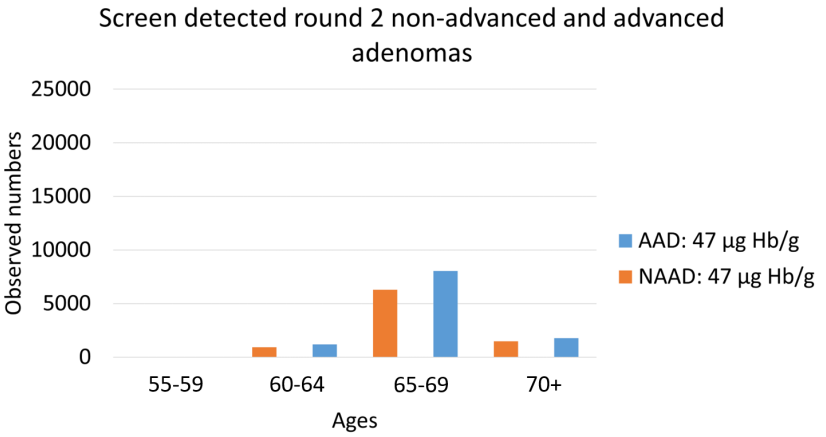
For the calibration of the MISCAN-Colon model, data is needed on the observed outcomes of the screening program. Using this data, the aim is to simultaneously calibrate test characteristics and cancer dwelling times. The former defines input parameters for the test characteristics used in the screening module, and the latter defining the time until the transition to the next state, used in the natural life-history module. Data on the calibration targets stems from the results of the Dutch screening program over the period 2014-2017, which covers individuals who have attended at most two screening rounds considering a two-year follow-up screening. This data is obtained from the national CRC screening database (ScreenIT) and the National Cancer Registry (IKNL). Descriptive statistics on age and sex distribution are found in Table 11. Outcomes of the national screening program are presented for the first and second screening round separately. Information about screen-detected stage I, II, III, and IV CRCs, the number of screen-detected non-advanced (NAAD) adenomas of size (5, 9) mm, the number of screen-detected advanced adenomas (AAD) of sizes [10, 10+) mm, and the total number of interval cancers, aggregated over all four cancer stages is available. The aggregation of adenomas is in accordance with the definition of non-advanced and advanced adenomas from Toes-Zoutendijk et al. (2018). For the Dutch biennial FIT screening program, interval cancers are identified as clinically detected cancers found within two years after the last attended screening, following the general definition of interval cancers provided in the previous chapter.

Figure 5[6] displays the number of the first (a) respectively the second screening round (b) screen-detected (non-) advanced adenomas. The lighter bars display the numbers reported for the lower cutoff of 15 μg Hb/g feces, used in the first half of 2014, and the darker bars indicate the numbers reported for the current cutoff of 47 μg Hb/g feces. Due to the lower number of individuals screened with the lower cutoff, fewer lesions were detected. Few individuals were screened with the lower cutoff, because the cutoff was elevated already six months after the start of the program, on July 1st of 2014. Consequently, no lesions are reported for the lower cutoff for the second screening round. Since the first screening round is a prevalence round, the number of screen-detected adenomas in the second screening round is significantly lower. No screen-detected adenomas were found in the second screening round for individuals aged between 55 and 59 years old because none of them had had their second screening round yet, because of the graduate roll-out of the program displayed in Figure 1. In addition to the differences between the first and second screening rounds, the number of advanced adenomas is higher than the number of non-advanced adenomas in both screening rounds. This difference can be ex-

[6]NAAD: non-advanced adenomas, AAD: advanced adenomas (Toes-Zoutendijk et al., 2018)



(a)



(b)

Figure 5: Screen-detected adenomas by size group in the first (a) and second (b) screening rounds of the national CRC screening program in the Netherlands from 2014 to 2017

plained by the fact that non-advanced adenomas are smaller and thus can indicate new lesions. Advanced adenomas in the second screening round are then more likely to be the result of false-negative lesions from the first screening round that have developed into more advanced adenomas.

Screen-detected cancers are displayed according to the four stages identified in Table 2. Correspondingly to the findings for screen-detected adenomas, the number of screen-detected round 1 CRCs is higher for the higher cutoff of 47 μg Hb/g feces as well, compared to the lower cutoff of 15 μg Hb/g feces (Figure 6). The reason for this is the same as was given for the screen-detected adenomas. Most screen-detected CRCs of the first round are found for stages I-III. Since CRCs in later stadia are expected to show symptoms, these lesions are generally clinically detected before screening has taken place, which indeed results in lower detection numbers for higher stage CRCs. The same phenomenon is observed for cancers detected in the second round (Figure 7). This consistent finding

of detection of early-stage cancers for the first and second rounds is what we require a proper screening program to result in (Toes-Zoutendijk et al., 2018).

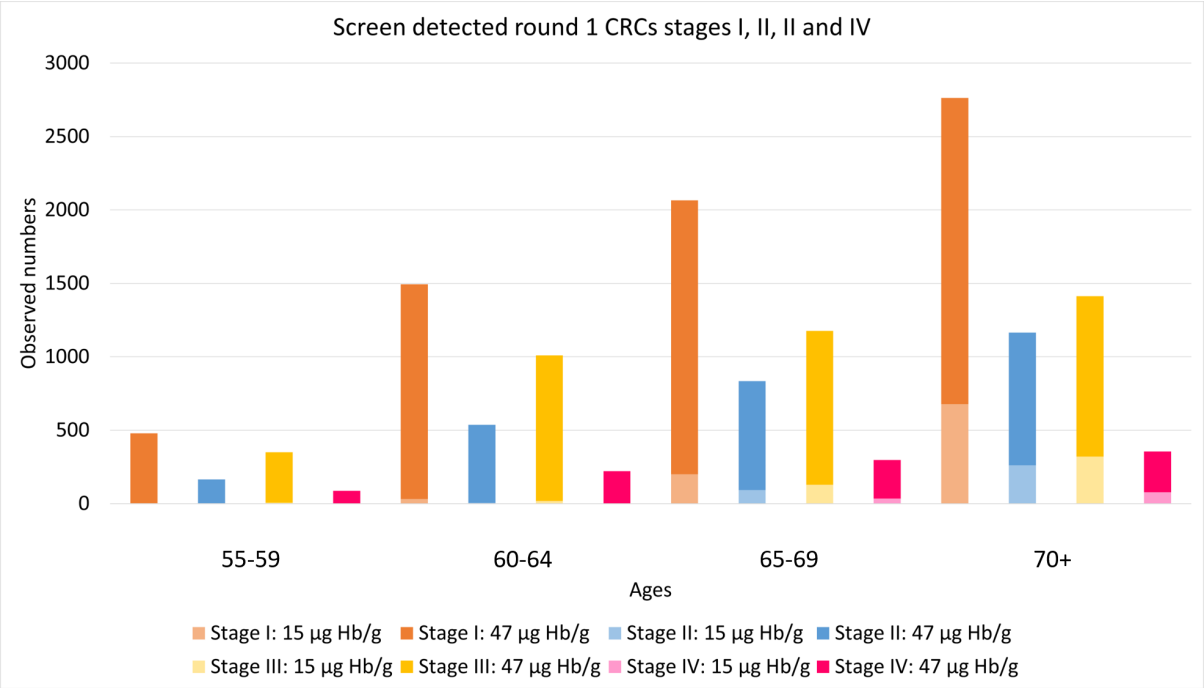


Figure 6: screen-detected CRCs by stage distribution in the first screening round of the national CRC screening program in the Netherlands from 2014 to 2017

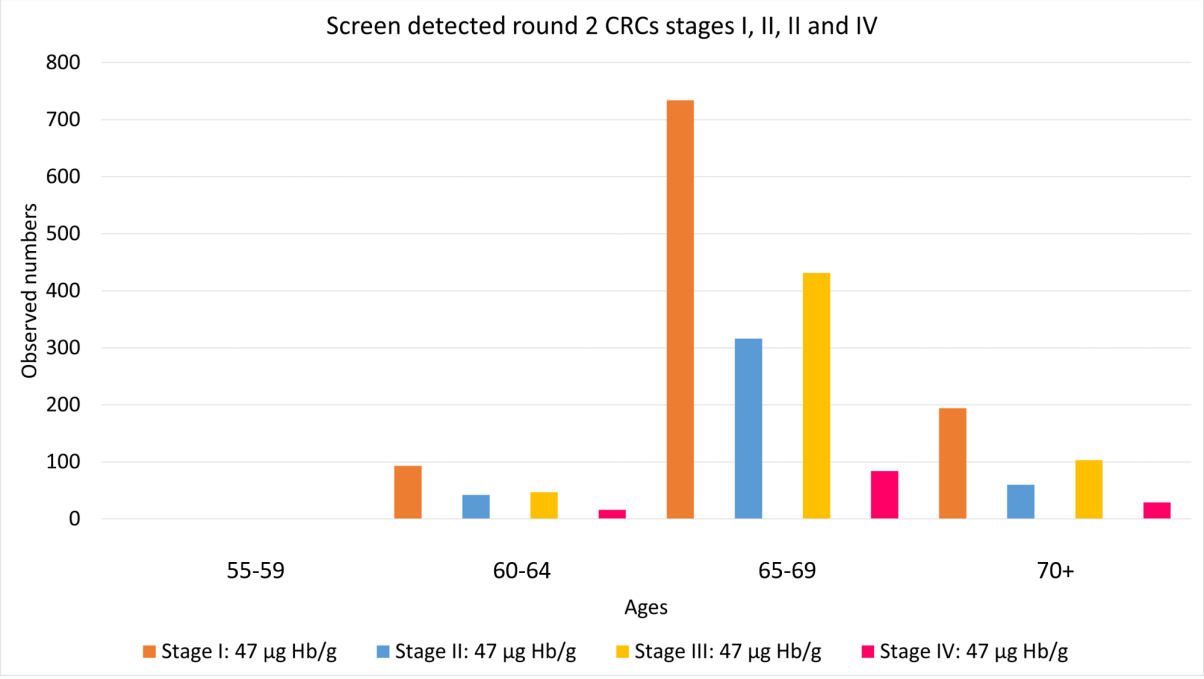


Figure 7: Screen-detected CRCs by stage distribution in the second screening round of the national CRC screening program in the Netherlands from 2014 to 2017

Simultaneously, older individuals are at higher risk for developing advanced neoplasia.

This may explain why more screen-detected round 1 CRCs are found for the higher age categories. Moreover, it is more likely that we find cancers in individuals that have their first round at a higher age. This can again be explained by the first round acting as a prevalence round. Surely, more lesions are found when individuals have never been screened before while their lesions have long been developing. Screen-detected CRCs of the second screening round show a unimodal shape where the highest incidence is found in individuals aged between 65 and 69. A possible reason for this is two-fold. On one hand, we argued older individuals are more at risk for developing advanced adenomas or CRCs. On the other hand, those individuals that had their second round were younger at the time of their first round. As a result, it is expected the detection numbers will behave this way across the age categories.

Remarkably, the reported total number of interval cancers for the first and second screening rounds are highest for the higher age categories (Figure 8). Possible reasons for this may be because elderly individuals may often bleed due to other reasons than the presence of a lesion. This in combination with a higher systematic lack of sensitivity for these individuals may lead to a higher number of interval cancers after a previous negative FIT. Moreover, elderly individuals belong to the higher-risk group for which cancers may develop more rapidly, resulting in a missed cancer during screening which subsequently develops into an interval cancer. Lastly, Toes-Zoutendijk et al. (2019) argue that interval cancers for the lower cutoff are oftentimes found in the rectum, in which the FIT test has poor reach. As a result, when these cancers bleed, the FIT test may miss them but the individual may notice blood in their stool and will consequently inquire their physician who will diagnose the cancer even though the individual had a negative FIT. Altogether, these may explain the higher number of interval cancers for the elderly.

MISCAN-Colon's current input parameters can be validated using the data on the outcomes of the Dutch national CRC screening program. Validation is examined by visualizing the outcomes simulated using the current parameter values (Figure 30). The figures indicate that the simulated outcomes under the current parameter values do not completely coincide with the observed outcomes over 2014 to 2017. Hence, this shows perspective for the selected calibration algorithms to infer the parameters such that they lead to a better model fit to the observed outcomes.

With the calibration parameters and targets identified and the current parameters validated, a subsequent step is to specify the functional form set for modeling the test characteristics as a continuous function of age. Thereafter, the workings of the selected Bayesian calibration algorithms are explained. Besides an explanation of the algorithms, the other methodological decisions that relate to the calibration approach introduced in Chapter 2 are set. We conclude the chapter with a selection of methods with which to compare the selected estimation methods.

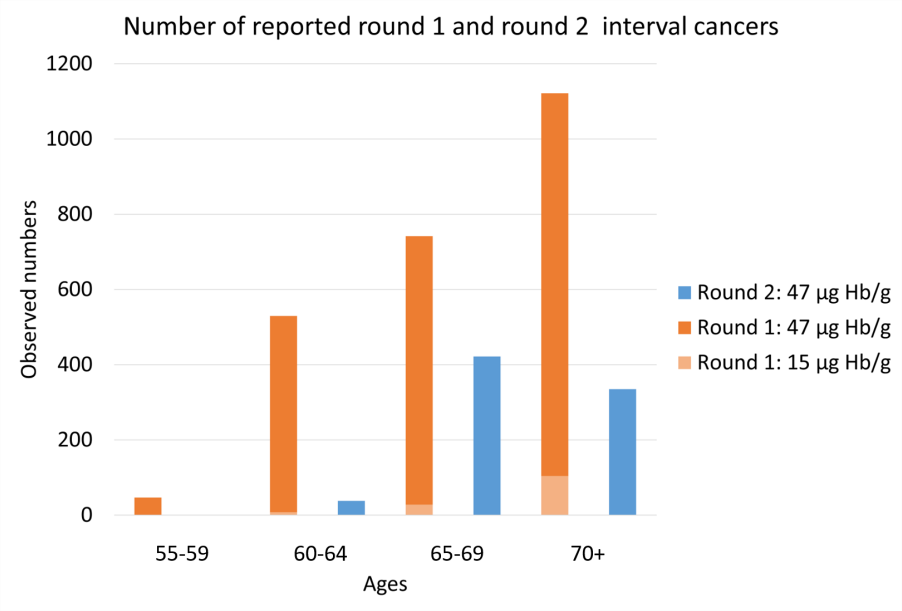


Figure 8: Total reported interval cancers in the first and second screening rounds of the national CRC screening program in the Netherlands from 2014 to 2017

5 Methodology

In order to provide a clear overview of the methodological decisions made, the chapter components will follow the calibration approach steps specified at the start of chapter 2.

5.1 Calibration parameters and targets

The justification of the calibration parameter selection and an overview are given first. Then, the selection and overview of calibration targets follow. With the calibration parameters and targets identified, the functional form for the dependency of the test characteristics calibration parameters on age is proposed.

The decision for the simultaneous calibration of test characteristics and cancer dwelling times results from the following reasoning. The proportion of cancers found by screening, the detection rate, is dependent both on how much disease and in which state is present in an individual, as well as how accurately the screening test can detect the disease. How much disease is present in an individual will depend on the dwelling time of the disease. Accordingly, how well the screening test performs in terms of being able to detect the disease is defined as the test sensitivity. Since multiple combinations of cancer dwelling times and test sensitivity may lead to the same cancer detection rate, it is important that test characteristics and cancer dwelling times are calibrated simultaneously. Additionally, it is suggested that it is of interest to calibrate the FIT systematic lack of sensitivities because we do not observe these parameters. Systematically missing small adenomas in ill individuals will not immediately lead to an increase in clinically detected cancers. This

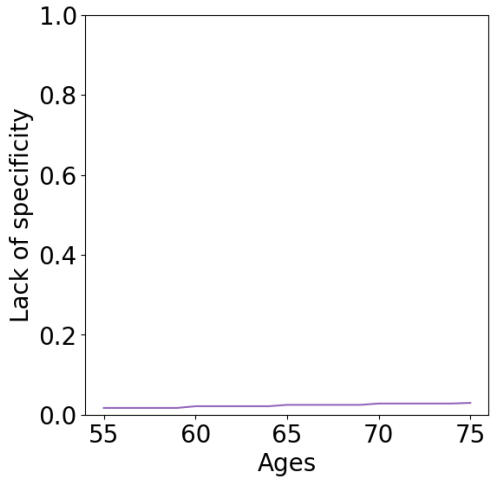
is because adenomas take a long time until they fully develop into CRC. However, for preclinical cancers, dwelling times are assumed to be much lower. As a result, systematically missing such cancers will lead to a higher number of reported interval cancers in expectation.

The final selection of calibration parameters comes down to a set of test characteristics, consisting of FIT sensitivities for medium and large adenomas and preclinical cancers stage I-IV, the lack of specificity, systematic lack of sensitivities for medium adenomas and large adenomas, and dwelling times of preclinical cancers. An overview of all calibration parameters is given in Table 12. The targets we calibrate on are model outputs that we expect to be influenced by the calibration parameters and are broken down into the first and second screening round targets in accordance with the presentation of the Dutch CRC screening outcomes. The final calibration is done on the number of screen-detected cancers by stage, the number of screen-detected (non-)advanced adenomas, and interval cancers. An overview of the calibration targets is given in Table 13.

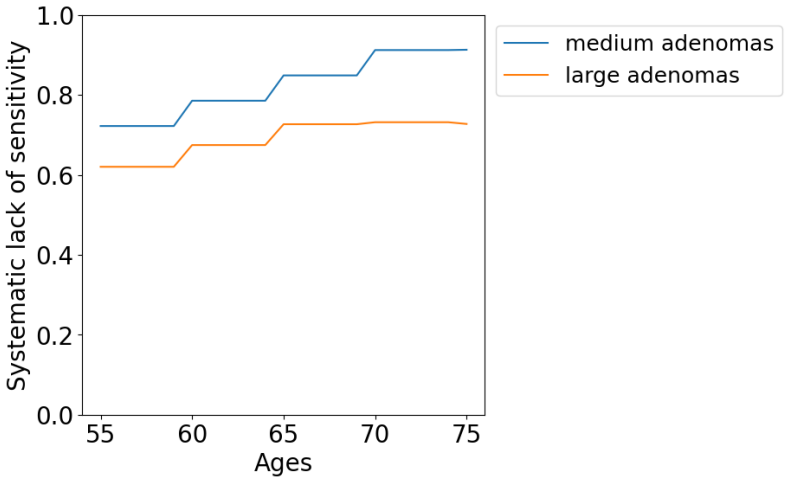
The final decisions that need to be made are those of the functional form for the parameters that are modeled as continuous functions of age and which priors to set for the calibration parameters. First, the functional form for the test characteristics is set. From the functional form, the decisions for the priors' specifications follow.

In the current implementation of MISCAN-Colon, test characteristics are specified for four different age groups. Since we estimate test sensitivities for all adenoma states and all preclinical cancers in addition to systematic lack of sensitivity for both medium and large adenomas and lack of specificity, estimating those separately for all four different age groups leads to a large calibration problem. Modeling test characteristics as a continuous function of age reduces the number of parameters that need to be estimated. Furthermore, such a specification allows for flexible estimation of the test characteristics.

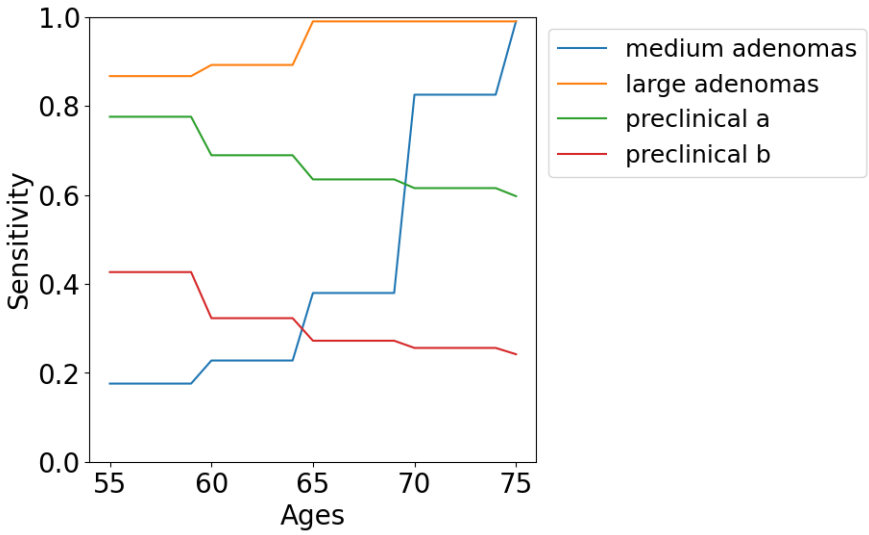
In order to make a well-informed decision for the functional form of the test characteristics, their values as they are currently set for MISCAN-Colon are visualized in Figure 9. The figure suggests for most parameters an upward sloping trend. We expect test sensitivity to be higher for higher ages since they are at higher risk of developing CRC. In addition, older individuals bleed more often, leading to a higher test sensitivity as well. Furthermore, lack of test specificity is expected to increase with age as well since those individuals generally have more blood in their stool due to reasons other than the presence of adenomas or cancers, resulting in more false negatives for these individuals. Lastly, systematic lack of sensitivity is expected to increase with age as well since the probability of systematically missing a lesion is higher in those individuals as well. Since older individuals more rapidly develop adenomas, it is possible they develop more non-bleeding adenomas than younger individuals (Van der Meulen et al., 2016). As a result, since those adenomas do not bleed, they are systematically missed during screening. Since the distinction between bleeding and non-bleeding lesions is most pronounced for



(a) Lack of specificity



(b) Systematic lack of sensitivity by state



(c) Sensitivity by state

Figure 9: Original lack of specificity (a), systematic lack of sensitivity per state (b) and original sensitivity per state (c)

adenomas, the systematic lack of sensitivity for those polyps are modeled.

Though the upward trend is most prominent for FIT sensitivity and systematic lack of sensitivity for medium and large adenomas, it is less visible for lack of specificity (Figure 9(a)). Figure 17(a) shows the upward sloping better. Conversely, Figures 17(b) and 17(c) suggest a downward sloping trend for preclinical cancer sensitivities. This would indicate that the FIT test is less capable of detecting preclinical cancers than adenomas for higher ages. Though the literature on this is sparse, preclinical cancers are modeled as an upward function of age as well. Their priors are subsequently set such that the function may reverse to a downward sloping function for preclinical cancer a and b.

All test characteristics denote a probability. Therefore, we require a functional form that is upward sloping and is bound between 0 and 1. As a result, we suggest a sigmoid-shaped functional form (Figures 18). More specifically, a standard logistic function with

two coefficients, a scale, and a shift parameter. The scale coefficient determines the steepness of the slope while the shift coefficient indicates the location of the midpoint. This ensures the test characteristics remain between 0 and 1. However, many original parameters show a very steep slope on small age intervals. As a result, it is suspected the standard sigmoid functional form may not be flexible enough for all test characteristics. It is alternatively decided to compare the standard sigmoid functional form with a generalized Richard functional form (Figure 19). These are generalized logistic functions for which more coefficients can be fixed, hence allowing for more flexibility of its shape. For the Richard functional form, we fix most coefficients but the Q, B, and ν coefficients. The Q coefficient relates to the shift coefficient and changes the midpoint location, the B coefficient determines the steepness of the slope and the ν coefficient determines near which of the two bounds the curve grows steepest. This introduces another set of extra calibration parameters. In order to keep the increase in the number of calibration parameters modest, it is therefore decided this functional form is only specified for test sensitivities of medium and large adenomas and preclinical cancer stages a and b. The other test characteristics in the Richard scenario keep their sigmoid functional form. In this way, the majority of the test characteristics are estimated by the Richard functional form. For the mathematical formulation of the sigmoid and Richard functional forms see Figure 20.

Due to the functional form specification, we calibrate the scale, shift, Q, B, and ν coefficients instead of the test characteristics directly. The scale and shift coefficients for the standard sigmoid-shaped functional form lead to a total number of 14 calibration parameters for test characteristics. Dwelling times are estimated for preclinical cancer stages I-IV, which leads to an additional 4 calibration parameters. The total calibration parameter count comes down to 18 parameters for the sigmoid functional form. For the Richard functional form for test sensitivities by state, we calibrate the Q, B, and ν coefficients. This leads to a total number of 12 test sensitivity calibration parameters. Within the Richard functional form, systematic lack of sensitivities by state and lack of specificity keep their sigmoid functional form which leads to an addition of 2 coefficients for lack of specificity and 4 for the systematic lack of sensitivities of adenomas by size group. Preclinical dwelling times again add an additional 4 parameters. The total calibration parameter count then comes down to 22 parameters for the Richard functional form.

Since dwelling times are generally modeled by an exponential distribution (Loeve et al., 2000), dwelling time priors for the preclinical cancer stages are set as the exponential distribution as well, with their λ set as their current dwelling time parameter. For the coefficients of the functional forms, it is less straightforward what a proper prior would be. In consensus with Bayesian Econometric literature and the study performed by de Weerdt (2019) and Bergqvist (2020), their priors are chosen as (continuous) uniform. Upper and lower bounds for the uniform priors are derived from approximating linear equations of

the original test characteristics' values.

5.2 GOF measures

In section 2.3, a discrepancy between the performances of different GOF measures is identified. Where Karnon and Vanni (2011) find that a chi-squared GOF measure performs best; van der Steen et al. (2016) find that likelihood-based GOF measures seem to perform best from a theoretical as well as from a practical point of view for cancer screening models such as MISCAN-Colon. As a result, we decide to evaluate the calibration results using these two likelihood-based GOF measures.

Additionally, van der Steen et al. (2016) state that calibration targets in a disease-modeling setting are typically binomially or Poisson distributed. Indeed most MISCAN-Colon outcomes are count data that statistically follow distributions such as the binomial, Poisson, or multinomial distributions. de Jonge (2019) confirms this, identifying interval cancers as Poisson distributed and screen-detected cancers as binomial distributed. Accordingly, we assume that screen-detected adenomas follow a binomial distribution as well. Therefore, binomial and Poisson deviances are used as GOF measures for the calibration algorithms we select. The formulas for the Poisson respectively binomial GOF measures are specified in Equations 3 and 4.[7] For each target, a GOF is calculated.

$$GOF_{poi} = 2 \cdot \left(obs \left[ln\left(\frac{obs}{sim}\right)\right] - (obs - sim)\right) \tag{3}$$

$$\begin{aligned} GOF_{bin} = 2 \cdot \Bigg(& obs \left[ln\left(\frac{obs}{n}\right) - ln\left(\frac{sim}{m}\right)\right] \\ & + (n - obs)\left[ln\left(\frac{n - obs}{n}\right) - ln\left(\frac{m - sim}{m}\right)\right] \end{aligned} \tag{4}$$

Here, *obs* indicates the number of screen-detected cancers, screen-detected adenomas, or interval cancers that result from the Dutch screening program. n indicates the intersection of the number of individuals with a negative FIT and the number of individuals with a positive FIT for the screen-detected cancers and adenomas, and the number of individuals with a negative FIT for the interval cancers. Both observed in the national screening program as well. *sim* indicates the simulated number of screen-detected cancers, screen-detected adenomas or interval cancers. Finally, m contains the number of simulated individuals from which *sim* is determined. Again, for interval cancers, these are individuals that had a previous negative FIT result and for screen-detected cancers and screen-detected adenomas, this is again the intersection of individuals with a positive FIT and individuals with a negative FIT. Hence we calculate the GOF of interval cancers with Equation 3 and the GOF of screen-detected cancers and screen-detected adenomas

[7] Notation as specified in van der Steen et al. (2016).

with Equation 4. This leads to three separate GOFs for each screening round. Within the GOF equations, we evaluate the observations and simulations relative to n and m. This ensures that the simulated output corresponds to the observed output for different population sizes (van der Steen et al., 2016).

In order to simultaneously compare multiple model outcomes, an overall GOF measure is required. This overall GOF is calculated as the weighted sum of all targets' individual GOF values as seen in Table 4. However, we see no reason to weigh the targets differently and therefore simply compute the aggregated sum over all individual GOFs.

5.3 Estimation methods

Chapter 2 discusses a selection of Bayesian algorithms previously used for calibrating CRC screening models such as the MISCAN-Colon model. Based on the findings described there, the ABC-SMC algorithm employed by de Weerdt (2019) and the NUTS sampler employed by Hoffman and Gelman (2014) are selected to simultaneously calibrate MISCAN-Colon's test characteristics and preclinical cancer dwelling times. The calibration results are used to compare the two algorithms and determine which of them is most suited for calibrating MISCAN-Colon's parameters. Both algorithms are implemented in Python, using the Panmodel; the Pythonic implementation of MISCAN-Colon.

5.3.1 ABC-SMC

The basic explanation of ABC-SMC from Chapter 2 is extended here in order to provide more intuition into the basics of the algorithm. A clear overview using diagrams is seen in Figure 10[8]. Accompanying the diagram, pseudo-code for the algorithm is found in Table 5.

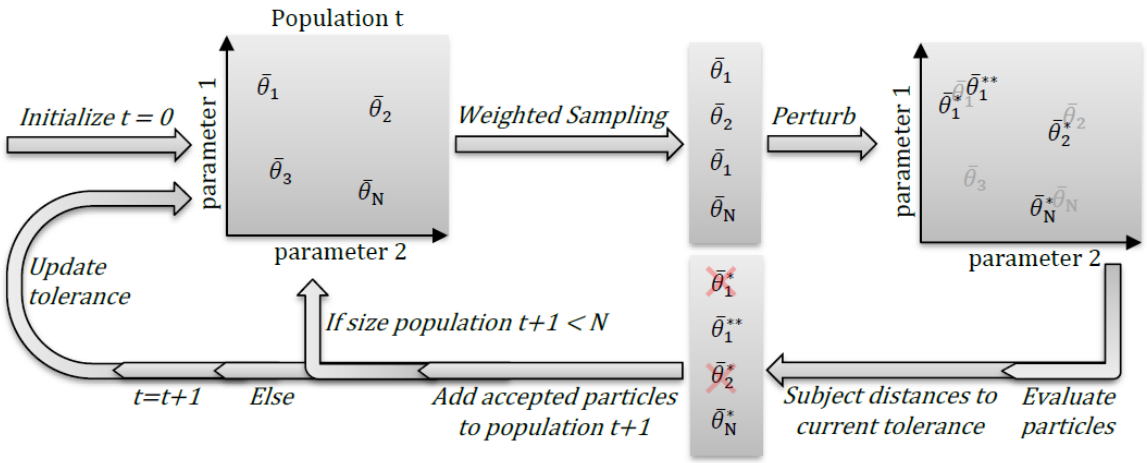


Figure 10: Graphical representation of the ABC-SMC algorithm

[8]Figure adapted from de Weerdt (2019)

In the ABC-SMC extension to the ABC variations described by Turner and van Zandt (2012), a tolerance level ε indicates the fit of the simulated outcomes to what is observed. This tolerance level ε is decreased sequentially at each iteration. Each subsequent iteration, therefore, leads to parameters that provide a better fit to the observed data. This leads to a sequence of intermediate distributions (Figure 21). These are used to move from the prior distribution to the posterior distribution which is achieved as follows. We start the algorithm at iteration zero and draw P (N in Figure 10) parameter sets, called particles, from the priors. This means that each particle is a vector containing one value for each calibration parameter. In each iteration, a new population, called a generation, is generated from the previous one using weighted sampling. Subsequently, a perturbation step is performed where the observations in the newly generated population are adjusted in such a way that the set of parameters for that generation is as close as possible, but not identical, to the previous generation. This encourages exploration of the parameter space, finding multiple other promising parameter sets, and ultimately diversifying the particle population. After the perturbation step, the particles are used for obtaining simulation outcomes. Their distance ρ is computed and compared with the tolerance level ε. An accepted particle will be included in the following child generation from which new particles are drawn, weighted, perturbed, and assessed until again P particles are accepted and form the intermediate distribution for that generation. The algorithm then terminates when the minimum ε or the maximum number of iterations has been reached. From the pseudo-code in Table 5 it is seen that at the end of each iteration, all weights are normalized such that they sum to 1.

In de Weerdt (2019), the authors set the particle population size at 40. Turner and van Zandt (2012) set this number at 500 particles per population, yet do not provide a motivation for this number. We observe that an increase in the particle population size leads to an increase in computational time for each iteration. Particularly, when the ESS is low, many parameter sets are evaluated within each iteration in order to end up with P particles. In order to reduce the computational demand of the ABC-SMC algorithm to a time similar to the NUTS algorithm, we set the particle population size at 25. In addition, we set the distance function ρ as the overall GOF that results from the individual targets' GOF. This ensures that simulation outcomes are scaled correctly with respect to the observed outcomes and directly provides us with GOF estimates for the algorithm. Hence, the role of the GOF is two-fold in ABC-SMC. Although setting the tolerance level ε to the desired level ensures the desired fit is reached, it is possible that this tolerance level is never reached until the maximum number of iterations has been reached. Moreover, according to Chong and Lam (2017), setting the maximum number of iterations equal for all algorithms allows for later comparison of convergence of the algorithms. As such, we decide to set ε arbitrarily low at a level of 1000 in order to ensure the maximum number of iterations is reached. Turner and van Zandt (2012) argue that

Table 5: ABC-SMC algorithm

1.	Set the number of generations (iterations) T and the number of particles (parameter samples) P
2.	Set $t = 1$ and the tolerances $\varepsilon_1 < \varepsilon_1 < \varepsilon_2 < \ldots < \varepsilon_T$
3.	Set the first particle indicator $i = 1$
4.	If $t = 1$: sample particle $\theta_i^{**\text{b}}$ from its prior $\pi(\theta)$ for $i = 1, \ldots, P$ If $t > 1$: sample particle $\theta_i^{*\text{a}}$ from generation $(\theta_{\mathbf{1}}, \ldots, \theta_{\mathbf{P}})_{t-1}$ for $i = 1, \ldots, P$, with weights $\{\mathbf{w}_{t-1}\}$ and perturb it to obtain $\theta_{\mathbf{i}}^{**} \sim K(\theta \mid \theta^*)$
5.	If $p(\theta^{**}) = 0$, go to step 4
6.	Generate $\hat{\mathbf{y}}^{**} \mid \theta^{**}$, calculate $\rho(y, y^{**})$ and compare with ε_t
7.	If rejected: return to step 4. If accepted: set $\{\theta_i\}_t = \theta_i^{**}$
8.	If $t = 1$: $\{w_i\}_t = \frac{\pi(\theta)}{\sum_{j=1}^{P} \{w_j\}_t K(\{\theta_i\}_t \mid \{\theta_j\}_{t-1})}$
9.	If $i < P$: $i = i + 1$
10.	Normalize weights such that $\sum_{i=1}^{P} \{w_i\}_t$
11.	If $t < T$: set $t = t + 1$ and return to step 3.

[a] θ^* intermediate particle candidate
[b] θ^{**} weighted and perturbed particle candidate

no proper guidelines are known for setting ε and the maximum number of iterations. As a result, the maximum number of iterations is set at a level such that both algorithms terminate within approximately one week. The exact number is discussed at the end of this chapter, after the discussion of the NUTS algorithm.

5.3.2 No-U-Turn sampler

In order to elaborate on how the NUTS sampler efficiently explores the parameter space, the HMC sampler needs to be covered. This should provide a clear basis for understanding the advantages the NUTS sampler provides. Betancourt (2017) argue that even in high-dimensional problems, such as our calibration problem, the HMC sampler is able to make larger jumps away from the starting point θ^*. This leads to more efficient exploration of the parameter space, again ultimately diversifying the parameter sets that constitute the posterior. How the HMC sampler does this is shown in the pseudo-code provided in Table 6.

An auxiliary parameter ρ is introduced whose prior distribution is user-specified and typically set as a normal distribution with mean 0 and covariance matrix Σ_ρ. Other specifications are possible as well, subject to the researcher and the problem at hand (Betancourt et al., 2014; Betancourt et al., 2017). Subsequently, ρ can be used to compute the posterior distribution of our parameters θ by $p(\rho, \theta|y) = p(\rho)p(\theta|y)$. Since $p(\rho)$ is user-specified, samples from θ can easily be obtained from constructing the joint posterior distribution $p(\rho, \theta|y)$. First, initial values for ρ and θ are drawn from their priors $p(\rho)$ and

$p(\theta)$. Then, for a specified number of steps L, we repeatedly calculate new values for ρ and θ that are based on the first-order gradients and the hyperparameter ε.

This hyperparameter should not be chosen too large but not too small either. This is because ε influences how accurate the parameters found by the Leapfrog estimator are. The Leapfrog is an estimator that approximates the differential equations seen in step three of Table 6. Because ε needs to be set optimally in HMC, the samples (ρ^*, θ^*) need to be corrected by an acceptance probability displayed in step four of the pseudo-code.

Table 6: HMC algorithm

1.	Propose an rv ρ with density $p(\rho) \sim N(0, \Sigma_\rho)$. Set Σ_ρ, ε and L and sample ρ.[a]
2.	Initialize θ from $\pi(\theta)$
3.	While $l < L$: repeat: 1. $p = p - \frac{\varepsilon}{2}\frac{\partial p(\theta\|y)}{\partial \theta}$ 2. $\theta = \theta + \varepsilon \cdot \Sigma_{rho}\rho$ 3.$p = p - \frac{\varepsilon}{2}\frac{\partial p(\theta\|y)}{\partial \theta}$
4.	Correct (ρ^*, θ^*) by $(p^{m+1}, \theta^{m+1}) = (-p^*, \theta^*)$ with $\alpha = min(\exp^H(\rho^m, \theta^m\|y) - H(\rho^*, \theta^*\|y), 1)$ $(p^{m+1}, \theta^{m+1}) = (p^m, \theta^m)$ else

[a] L represents the number of steps to repeat sampling, ε denotes the accuracy of the Leapfrog estimator, and Σ_ρ denotes the covariance matrix of ρ.

From the above explanation, it is clear that three hyperparameters need to be set by the user when employing HMC. Firstly, the covariance matrix of the prior $p(\rho)$ for ρ, secondly the number of steps L taken until we are satisfied with the parameter samples found and lastly the value of the step-size ε which determines how accurate the Leapfrog estimator can approximate the differential equations. When the number of steps is taken too small, the sampler may end up with samples not far from the starting point. However, when the number of steps is taken too large, the sampler may still end up with samples not far from the starting point. This is because, in n-dimensional spaces, the HMC sampler makes a U-turn and eventually ends up back at its starting point, when it takes too many steps. Accordingly, if the step-size ε is taken too small, it will take considerable time until the trajectory is explored. Conversely, if ε is taken too large, many proposals may be rejected. These downsides of poorly chosen L and ε illustrate why it is essential these are properly informed. Tuning these three parameters is a viable option, however, it may be very time-consuming and may negate the advantages HMC promises over the MH sampler when tuned poorly.

The bother of having to hand-tune the L and ε hyperparameters of the HMC sampler was eliminated by Hoffman and Gelman (2014) when they introduced the NUTS

sampler. NUTS recursively traces through a constructed set of parameter candidates and terminates when it starts to trace back to already visited solutions, hence when it starts to make a U-turn. The pseudo-code of the NUTS sampler is not intuitive and above all lengthy. As a result, we refer to Hoffman and Gelman (2014) for this. Hoffman and Gelman (2014) explain the way the NUTS sampler avoids making a U-turn is as follows. For each step, in each iteration, the Leapfrog estimator uses the differential equations to either move forward or backward from its starting point. It is crucial to realize moving forward and backward is not in terms of a two-dimensional space but in terms of an n-dimensional space. Each step is doubled from the steps taken before that. This means that we start out taking 1 step in one direction, then taking 2 steps in the subsequent direction, then taking 4 steps in the direction afterward, etc. This doubling of the trajectory at each step eventually constructs a binary tree, where the leaf nodes of the tree save the position of the sampler. Once a U-turn is made, the sampler terminates and the next iteration starts. Red paths in Figure 11 display such a U-turn, which happens when inner or outer trajectories alongside any balanced subtree of the binary tree start to double trace back, where it takes a step in one direction and subsequently, doubles that number of steps in the other direction. This ultimately leads the trajectory back to its starting point. Figure 11 visualizes such a binary tree for the trajectories the sampler may follow[9].

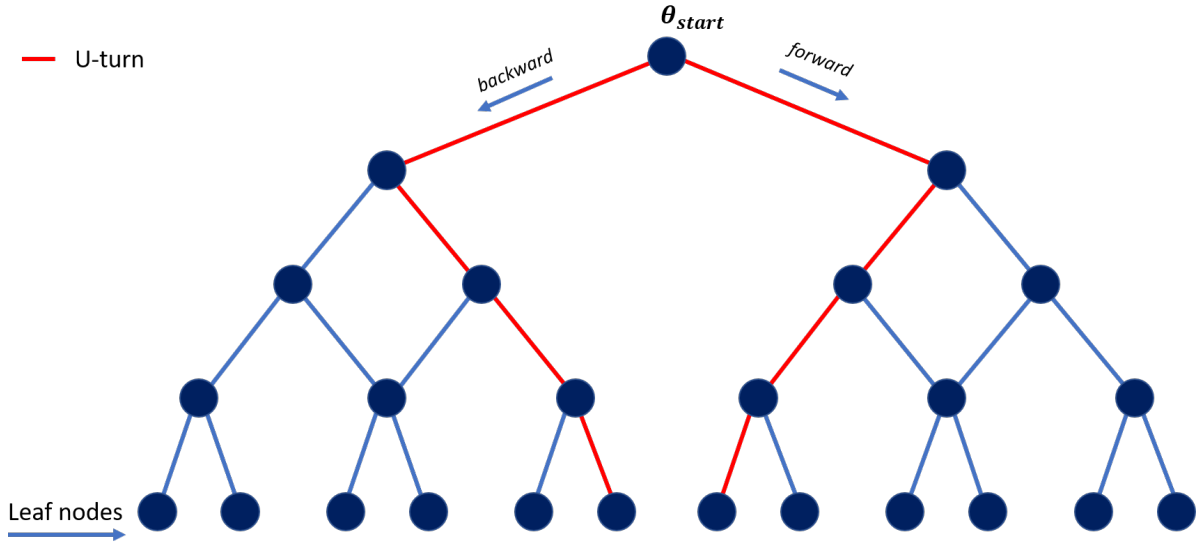


Figure 11: Binary tree indicating all possible forward and backward combinations (blue paths) and possible U-turns (red paths) for NUTS

Since the NUTS sampler removes any need for hand-tuning parameters, the only component that requires specification is the number of samples from which the posterior

[9]Blue paths indicate all feasible trajectories. Red paths indicate trajectories that terminate at leaf nodes due to a U-turn.

distribution is constructed. This number is set equal to the number of iterations set for the ABC-SMC algorithm in order to fairly compare the two algorithms. The last methodological decisions left are with which metrics convergence of the algorithms is determined, what possible stopping rules are set for each algorithm and how the performances of the algorithms on calibrating MISCAN-Colon's parameters are assessed and compared.

5.4 Convergence criteria, stopping rules and comparison measures

In Chapter 2, multiple metrics for determining the quality of convergence are proposed. For feasibility purposes, we decide convergence of the algorithms will be visually assessed by comparing plots of the posteriors with their priors. In addition, the acceptance probabilities and ESSs are calculated and reported for both algorithms in order to report on their mixing abilities. In order to be able to fairly compare the results of the algorithms, we decide the stopping rule for both will be setting the same maximum number of iterations. We argue that the minimum tolerance level ε of the ABC-SMC algorithm is therefore set arbitrarily low at a level of 1000 in order to ensure the algorithm does not prematurely terminate. From preliminary runs, we find that the final tolerance level is far from this level for the number of iterations that we set. As a result, we assume it to be an appropriate minimum tolerance level such that the maximum number of iterations is reached. The number of iterations is then chosen such that the computational time needed for running the algorithms is feasible and comparable across the two algorithms. As a result, we run both algorithms on both 15 as well as on 25 iterations. Lastly, we suggest that convergence criteria can simultaneously be used as a measure of performance for comparing the two algorithms. The algorithm that converges best under the fixed number of iterations is then a best-performing candidate. We say this because the best converging algorithm may not provide the best fit. Therefore, their performance is additionally evaluated based on their overall GOF. Nevertheless, there is a limit to how far we can base conclusions on the GOF values. Since they are simply a quantitative measure, their values can only be compared but do not indicate a proper fit. Therefore, the simulated outcomes are visually compared with the observed outcomes, including CIs in order to determine which algorithm yields parameters that lead to outcomes that resemble the observed national screening outcomes the best.

In the final analysis, we run both algorithms for both the standard sigmoid functional form for all test characteristics and the functional form where test sensitivity is modeled by a Richard curve and the other test characteristics by the standard sigmoid functional form. Each of those four scenarios is run for both 15 as well as 25 iterations. While for Bayesian algorithms these are low numbers of iterations, increasing the number of iterations is time infeasible. This is due to the high computational time for sequentially

running MISCAN-Colon for large population sizes. The population size is set at ten million individuals in order to minimize stochastic error as much as possible (Ozik et al., 2016).

6 Results

The results of running all scenarios for both algorithms are presented in two parts. First, their convergence based on the ESS, acceptance probabilities, and graphs of the priors versus the posteriors is examined. The test characteristics that result from the functional forms are visualized as well and show whether the calibrated coefficients lead to reasonable values. Based on the reported convergence metrics it is argued which functional form and which algorithm converges best. Then, the performance of all scenarios for both algorithms is examined based on the reported GOFs and graphs of the observed outcomes, CIs, and simulated outcomes. Lastly, we consider the outcomes on the convergence behavior and performance of the functional forms across the two algorithms and subsequently argue which functional form for which algorithm performs best amongst the 8 scenarios considered based on the reported findings.

6.1 Convergence

Convergence of ABC-SMC and NUTS is evaluated based on visual inspection of the priors versus the posteriors and the quantitative measures of the acceptance rates and ESSs. The latter are shown in Tables 7 and 8 for the ABC-SMC respectively NUTS algorithms.

For all scenarios of the ABC-SMC algorithm, initial acceptance probabilities start at or near the maximum reported rate and decrease as the algorithm moves from the prior to the posterior through its intermediate distributions. However, acceptance probabilities do not show a particular pattern for a higher number of iterations or between functional forms. Standard deviations, the minimum, and maximum values in Table 7 indicate a small range over which the acceptance probabilities vary and the maximum acceptance probabilities are low for all scenarios, never reaching over 0.57. Near the final iterations, the lower rates are reached, indicating that fewer particles are accepted. No better candidates are found, indicating that the algorithm is close to the posterior. The course in which the acceptance rates first decrease fast and then remain low suggests that the ABC-SMC algorithm initially mixes well but mixing slows down quickly.

The ESSs are small compared to the number of iterations ran for both sigmoid scenarios of the ABC-SMC algorithm. This indicates that there is still significant autocorrelation left in their samples which decreases only slowly. Reducing correlation between samples can be achieved by incorporating thinning. This is the practice of repeatedly throwing away samples after every cycle of k iterations. However, since we throw away samples

Table 7: Acceptance rates and ESS statistics on the convergence of the ABC-SMC algorithm by scenario

ABC-SMC

Functional form	*Iterations*		mean (SD)	min	max
Sigmoid	*15*	acceptance rate	0.25 (0.10)	0.10	0.48
		ESS	4.98 (6.00)	1.00	25.00
	25	acceptance rate	0.32 (0.15)	0.04	0.54
		ESS	4.22 (4.84)	1.20	25.00
Richard	*15*	acceptance rate	0.34 (0.08)	0.23	0.57
		ESS	17.37 (3.82)	10.46	25.00
	25	acceptance rate	0.33 (0.08)	0.33	0.51
		ESS	19.55 (3.34)	11.05	25.00

after each kth iteration, more iterations are needed to obtain convergence and for exploring the entire posterior. This study already highlights the computational burden and time constraints that lead us to restrict the number of iterations. Therefore, thinning, as of now, is not used in this study. In addition, this implies that the summary statistics calculated from the posterior are biased. Whereas the ESS statistic is low for the sigmoid functional forms, it is higher for the Richard functional form, which has smaller standard deviations as well. As a result, Richard appears to suffer less from autocorrelation between samples than the sigmoid functional form in the ABC-SMC algorithm leading to less biased estimates.

Table 8: Acceptance rates and ESS statistics on the convergence of the NUTS algorithm by scenario

NUTS

Functional form	*Iterations*		mean (SD)	min	max
Sigmoid	*15*	acceptance rate	0.94 (0.02)	0.89	0.96
		ESS	25.51 (83.63)	-204.88	273.47
	25	acceptance rate	0.92 (0.01)	0.90	0.95
		ESS	7.74 (92.54)	-358.78	62.76
Richard	*15*	acceptance rate	0.81 (0.14)	0.39	0.91
		ESS	33.83 (44.65)	8.95	225.18
	25	acceptance rate	0.90 (0.03)	0.86	0.98
		ESS	29.95 (10.04)	16.92	49.13

[a] Negative ESS indicate negative autocorrelation between odd lags in $ESS = \frac{n}{1+2\sum_{k=1}^{\infty}\rho_k}$. This is embedded within the Monte Carlo integration scheme in order to reduce variance without having to increase the sample size.

In contrast to the ABC-SMC algorithm, NUTS does not show a particular pattern for its acceptance probabilities. For all four scenarios, acceptance probabilities fluctuate between 0.88 and 0.96 around 0.90. This suggests that more candidates are accepted gradually. In addition, since many candidates are accepted with a probability higher than

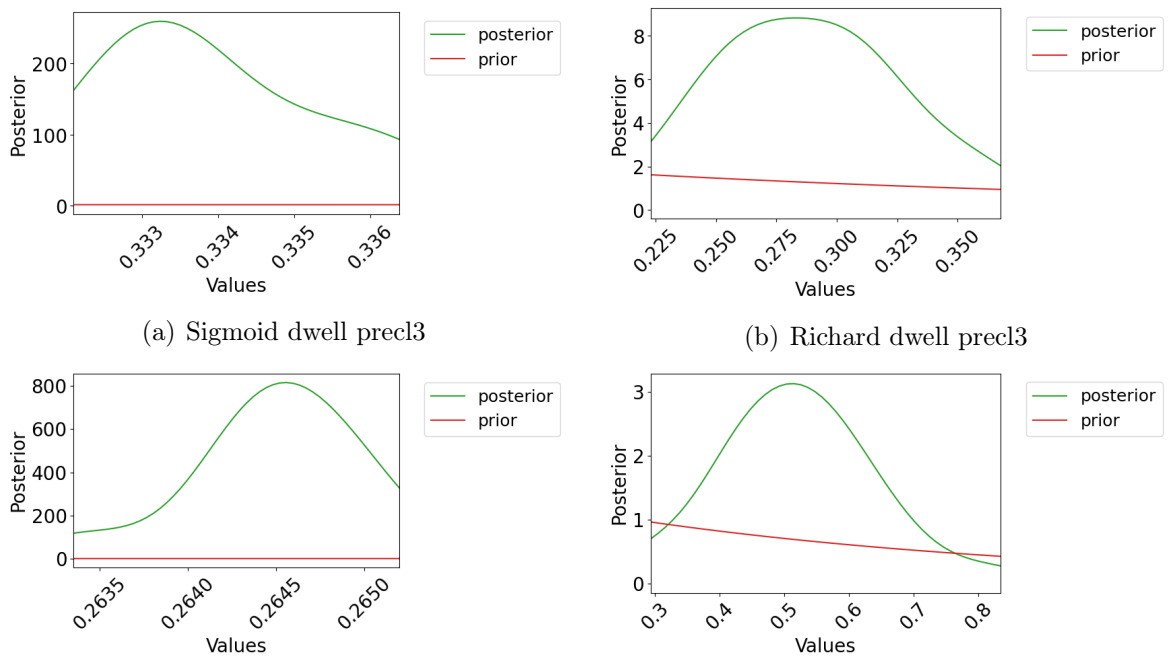
the required 80% level, the algorithm seems to be stable and converge properly. However, another possibility is that for these numbers of iterations not very diverse candidate sets are explored yet, which explains why their acceptance probabilities are similar. The acceptance probabilities indicate no distinct differences between the two functional forms.

In contrast to the ESSs reported for the ABC-SMC algorithm, the mean ESSs of the NUTS scenarios are higher, even for the sigmoid scenario with 25 iterations. Furthermore, much more extreme values are reported that regularly surpass the number of iterations ran. This happens when the correlation between lags of k is negative. Because of the truncation of even and odd lags in the denominator of the ESS formula[10], this results in the values shown in Table 8. It is argued on Carpenter (2018) that this indicates that such chains mix quickly. Yet, no real conclusion can be based on these values since they are found for such small numbers of iterations.

Although the ESS in the NUTS algorithm is sometimes larger, in absolute terms, than the number of iterations run, in theory, it still holds that larger ESSs indicate that there is less autocorrelation found for the NUTS algorithm than for the ABC-SMC algorithm. However, we expect less influence of negative autocorrelation when running NUTS for larger numbers of iterations because the introduction of negative autocorrelation is a way of reducing uncertainty for smaller numbers of iterations. Negative autocorrelation may also be an indicator that the sampler alternates between samples and hasn't fully left its warm-up phase. In addition, from the absence of a particular pattern for the acceptance probabilities of the NUTS algorithm compared to their decrease for the ABC-SMC algorithm, it is unclear whether the NUTS algorithm has truly converged. As a result, it is unclear from the ESS and acceptance probabilities alone whether NUTS indeed is converging better than ABC-SMC. Therefore, we discuss the shift behavior of the posteriors from the prior distributions and the test characteristics that follow from the calibrated parameters in order to further assess whether the algorithms have converged, how sensible the calibrated parameters are, and finally how informative the reported ESS and acceptance probabilities are.

In order to support the quantitative measures shown in Tables 7 and 8, it is assessed if and how far the posteriors move from their prior. Priors and posteriors do not overlap when the posterior has moved away from its prior. We say that the prior was uninformative and that the actual posterior is thus different from the original parameters. Consequently, posteriors that do show overlap with their prior have not moved away and are largely informed by their prior. For the ABC-SMC algorithm, some parameters have shifted from their priors completely, while others are strongly centered towards it. For both the sigmoid and Richard functional forms, all preclinical cancer dwelling times remain at their prior (Figures 12(a)-12(d)). In the figures, priors are denoted in red and posteriors in green.

[10] $ESS = \frac{n}{1+2\sum_{k=1}^{\infty}\rho_k}$



(a) Sigmoid dwell precl3

(b) Richard dwell precl3

(c) Sigmoid dwell precl4

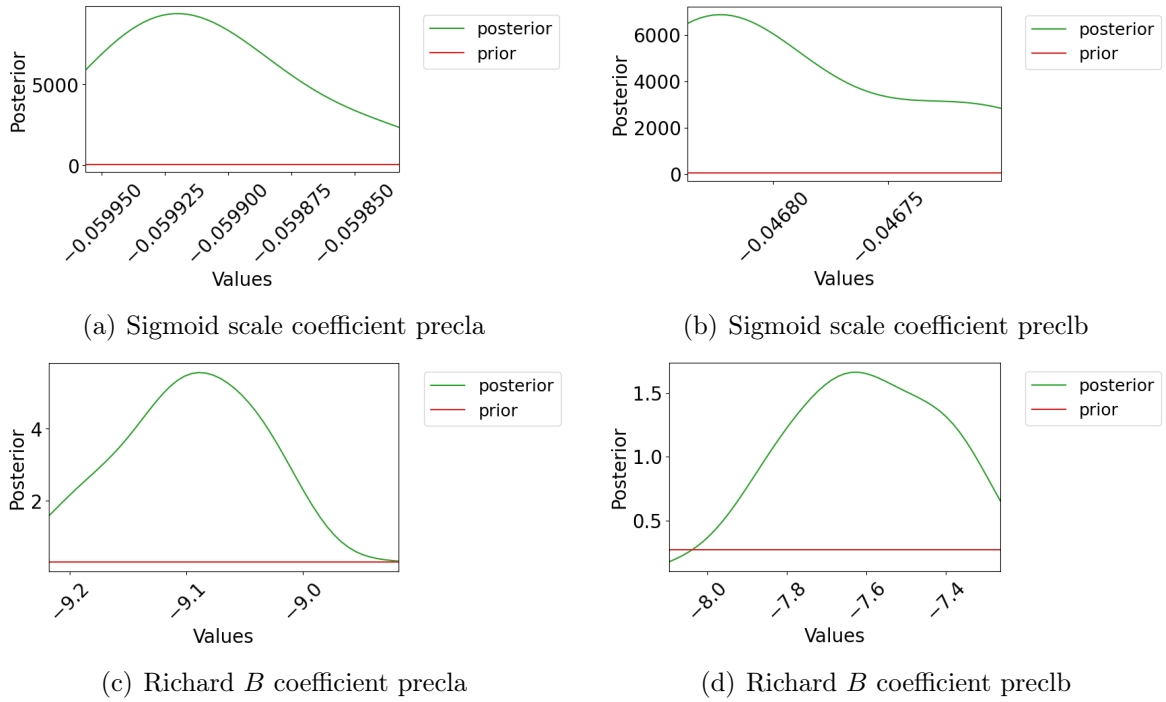
(d) Richard dwell precl4

Figure 12: Priors (red) and posteriors (green) for sigmoid (a) and Richard (b) functional form calibrated preclinical cancer stage III dwelling times and sigmoid (c) and Richard (d) functional form calibrated preclinical cancer stage IV dwelling times for 25 iterations of the ABC-SMC algorithm

Figures for preclinical cancer stage I and II dwelling times are not displayed but follow the same behavior as preclinical cancer stage III dwelling times (Figures 12(a) and 12(b)). Although preclinical cancer stage IV dwelling times remain centered around their uniform prior in both functional form specifications, their focus lies towards opposite extremes. This indicates that more iterations are needed to obtain a stable posterior. The posteriors' overlap with their prior indicates that the ABC-SMC algorithm finds preclinical dwelling times similar to the original dwelling times. However, since more iterations are needed to obtain stable results, it is unclear whether indeed preclinical cancers dwelling times are initially correctly calibrated.

For test characteristics, shift behavior largely coincides. For both functional forms, the lack of specificity coefficients and systematic lack of sensitivity coefficients for medium and large adenomas have shifted from their prior. Estimated functional forms for lack of specificity and systematic lack of test sensitivity for medium and large adenomas indeed coincide for both functional forms (Figures 22(g), 22(e) and 22(f) and Figures 23(c), 23(a) and 23(b) in Appendix B.5. Lack of specificity (Figures 22(g) and 23(c)) is defined on the same small scale as the original parameters. In contrast, systematic lack of specificity for both size groups is estimated near the upper bound and is, therefore, more extreme than those specified in Figures 18(e) and 18(f). These estimated systematic lack of sensitivities would suggest that we miss all non-bleeding adenomas during CRC screening. Though

not impossible, further research should be done with respect to this observation in order to determine whether indeed such functional forms are reasonable.



(a) Sigmoid scale coefficient precla

(b) Sigmoid scale coefficient preclb

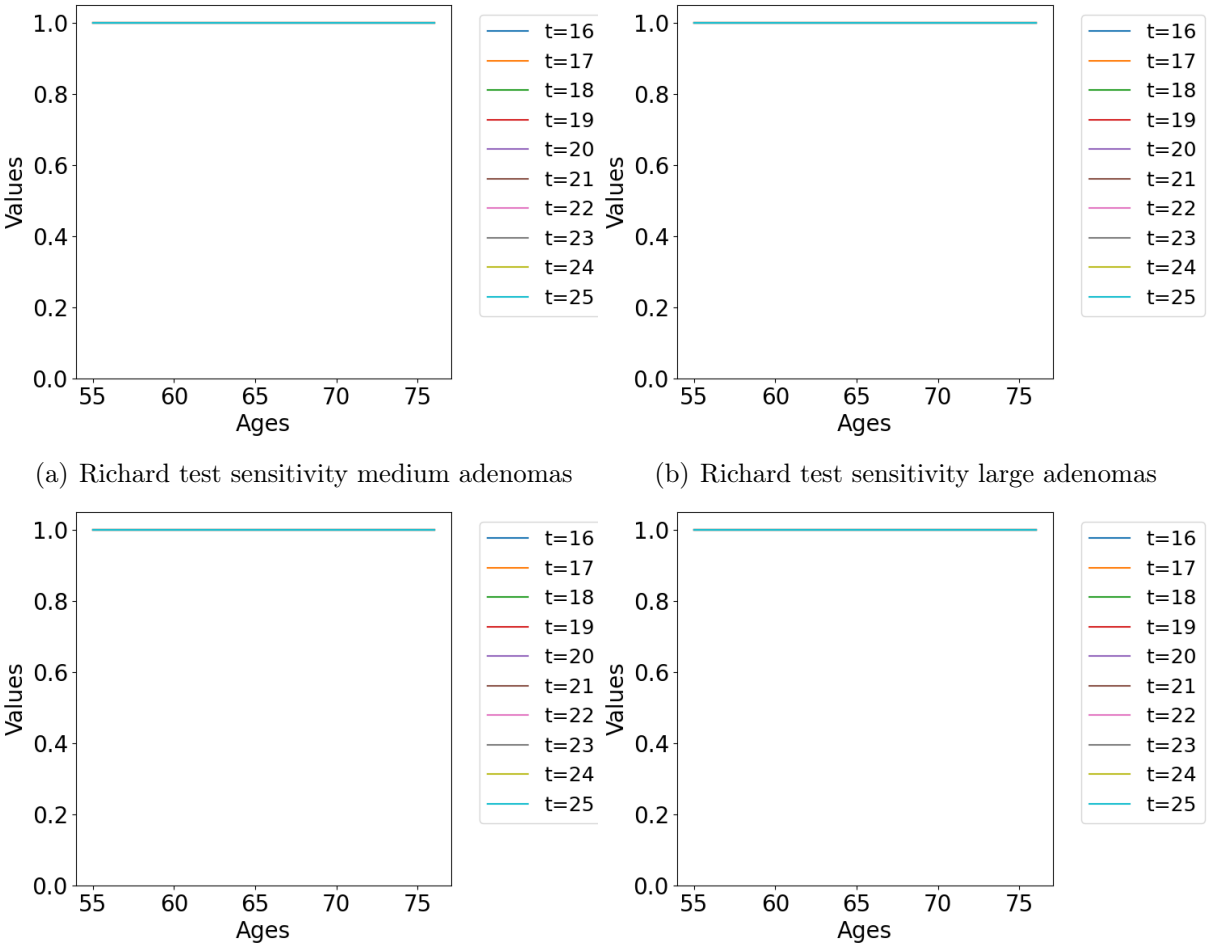
(c) Richard B coefficient precla

(d) Richard B coefficient preclb

Figure 13: Priors (red), intermediate distributions (blue, orange) and posteriors (green) for sigmoid functional form calibrated FIT sensitivity scale coefficients for preclinical cancer a (a) and preclinical cancer b (b) and Richard functional form calibrated FIT sensitivity B coefficients for preclinical cancer a (c) and b (d) for 25 iterations of the ABC-SMC algorithm

For both functional forms of the ABC-SMC algorithm, all coefficients for FIT sensitivity have shifted from their prior with the exception of sigmoid functional form scale coefficients and Richard functional form B coefficients for preclinical cancer a and b (Figure 13). For the Richard functional forms, this leads to test sensitivities for medium and large adenomas and preclinical cancer a and b that lie on or near the upper bound for all ages (Figure 14). Although their parameter sets differ across the iterations, the functional forms estimated for both adenoma sizes and preclinical cancers overlap for all iterations $t = 16$ to 25. Graphs of the other test characteristics estimated for the sigmoid and the Richard settings are found in Appendix B.5. Despite the posterior shifting behavior being similar for the two functional forms, the sigmoid test characteristics show less extreme behavior than the Richard functional form for test sensitivities as well as remain closer to the originally specified forms found in Appendix B.2.

For the NUTS algorithm, all preclinical cancer dwelling times are informed by their prior as well (Figures 24 and 25 in Appendix B.6). For the test characteristics, this is not the case for all coefficients. Most coefficients have partly or completely moved away from their prior. However, the systematic lack of sensitivity scale and shift coefficients for



(a) Richard test sensitivity medium adenomas

(b) Richard test sensitivity large adenomas

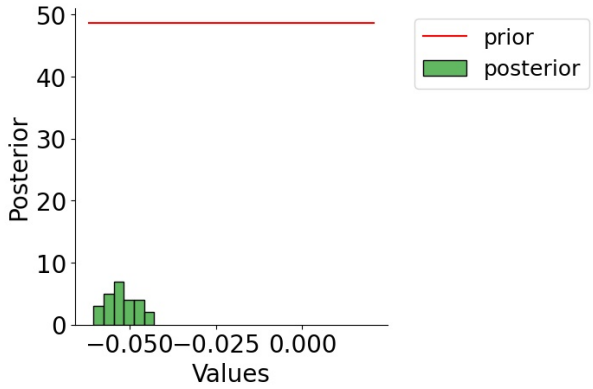
(c) Richard test sensitivity preclinical cancer a

(d) Richard test sensitivity preclinical cancer b

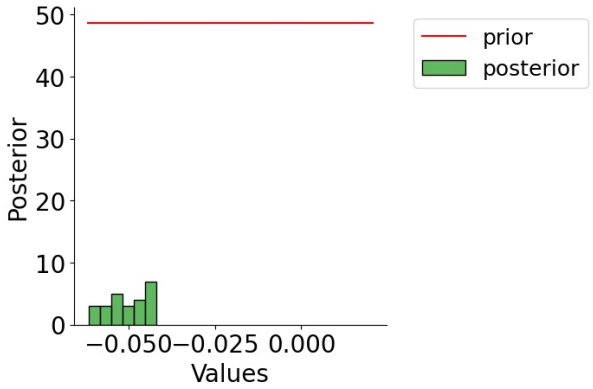
Figure 14: Calibrated medium (a) and large (b) adenoma and preclinical cancer a (c) and b (d) test sensitivity for the Richard functional form for iterations $t = 16, \ldots, 25$ of the ABC-SMC algorithm

medium and large adenomas show the same shifting behavior for both functional forms. Both their scale and shift coefficients are primarily uninformed by their prior and the posteriors' domains large overlap for all. This is reflected by the estimated FIT systematic lack of sensitivities of both adenoma size groups seen in Figures 26(e), 26(f), 27(e) and 27(f) in Appendix B.7. Both functional forms push the systematic lack of sensitivity for both size groups towards the upper bound for all ages. The lack of specificity coefficients' posteriors show the same overlapping behavior across functional forms as the systematic lack of sensitivity coefficients. Figures 26(d) and 27(d) in Appendix B.7 reflect this as well. In addition, both lack of specificity functional forms remain close to the functional form for the original parameters (Figure 18(d) in Appendix B.2).

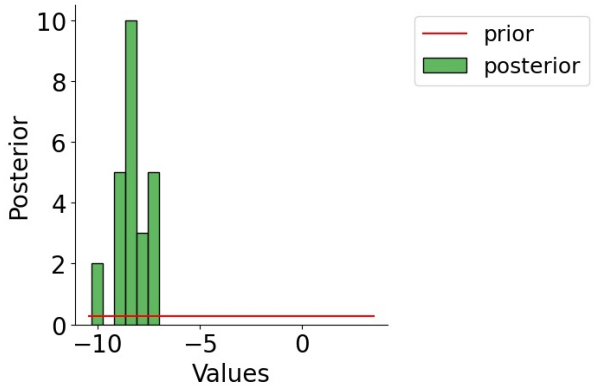
Similarl to the ABC-SMC algorithm, all coefficients for FIT sensitivity have shifted from their prior with the exception of sigmoid functional form scale coefficients and Richard functional form B coefficients for preclinical cancer a and b (Figure 15). However, their shifting behavior leads to estimated test sensitivities that are less extreme than those



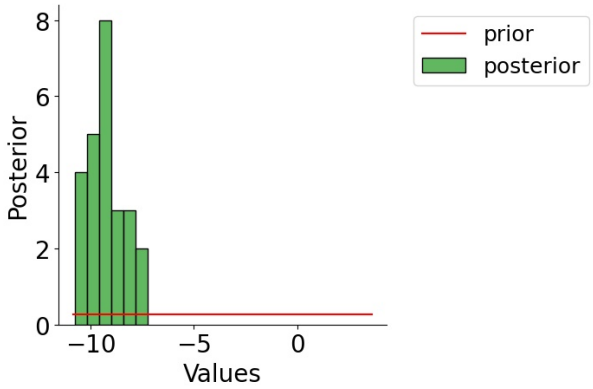
(a) Sigmoid scale coefficient precla



(b) Sigmoid scale coefficient preclb



(c) Richard B coefficient precla

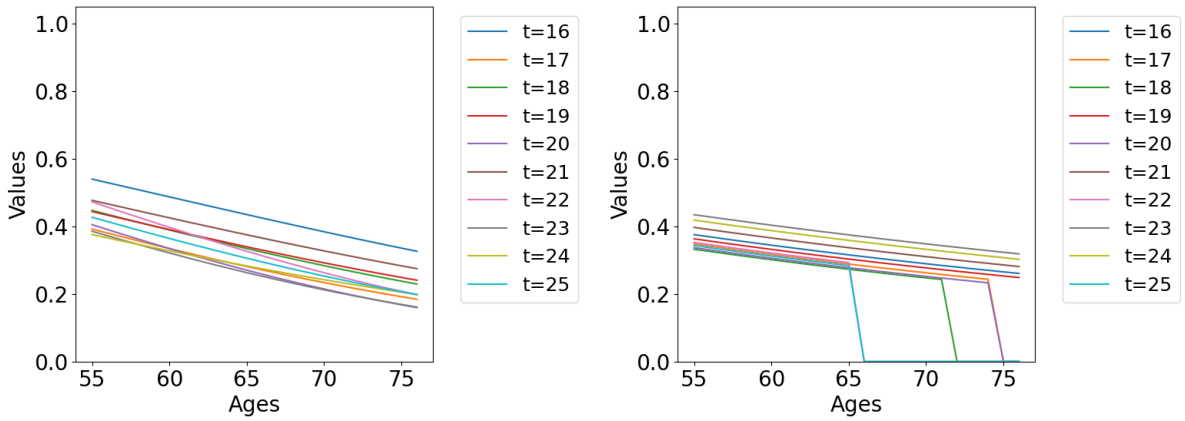


(d) Richard B coefficient preclb

Figure 15: Priors (red) and posteriors (green) for sigmoid functional form calibrated scale coefficients for preclinical cancer a (a) and b (b) and Richard functional form calibrated B coefficients for preclinical cancer a (c) and b (d) for 25 iterations of the NUTS algorithm

found for the ABC-SMC algorithm with the Richard functional form and more similar to those found for the ABC-SMC algorithm with the sigmoid functional form (Figures 26(a)-26(c), 27(a)-27(c), and 16(a)-16(b)). Despite their less extreme shape, NUTS preclinical cancer b test sensitivity estimated for the Richard functional form shows a sudden break for iterations 18, 22, and 25. This is due to the combination of the Q, B, and ν coefficients found for these iterations. This confirms that the algorithms are unable to take into account the simultaneous role these three coefficients have in determining an appropriate level for test sensitivities. As a result, we question whether modeling test characteristics as a function of screening age is desired.

In addition to the jumping behavior in Figure 16(a), extreme functional forms for ABC-SMC test sensitivities found for the Richard curve are implausible, since this would indicate that the FIT would correctly detect adenomas and preclinical cancers in all cases. Considering interval cancers have been reported from 2014 to 2017, test sensitivities cannot follow this estimated shape in reality. In addition, this is very different from the previously calibrated parameters from Figure 17 in Appendix B.1. Consequently, various factors may contribute to this effect. Firstly, since so few iterations are run for both



(a) Sigmoid test sensitivity preclinical cancer b (b) Richard test sensitivity preclinical cancer b

Figure 16: Calibrated preclinical cancer b test sensitivity for the sigmoid (a) and Richard (b) functional forms for iterations $t = 16, \ldots, 25$ of the NUTS algorithm

algorithms, the entire posterior has not been explored yet but only a subset of it. This is justified by the negative autocorrelation found between samples. This subset of the posterior may lead to test characteristics that are on or near the upper bound. Secondly, poorly set priors for the Richard functional form in combination with poorly explored posteriors may further impede the ability of the sampler to find combinations of the Q, B, and ν coefficients that lead to reasonable test characteristics. Lastly, modeling the test characteristics through another function may complicate finding reasonable test characteristics. Regardless of the factors leading up to this behavior, it is not acceptable and more research should be done in order to determine which combination of the proposed factors lead up to it and how to resolve it. Conversely to the extreme behavior found for the ABC-SMC Richard functional form test sensitivities, test sensitivity for medium and large adenomas and preclinical cancer a estimated for the Richard functional form for the NUTS algorithm show less extreme behavior, which is more in line with our expectations. However, the visualized test characteristics show that a different functional form is estimated for each iteration for both algorithms across functional forms. This indicates that more iterations are necessary in order for the algorithms to properly converge. This is supported by Figures 28 and 29 in Appendix B.7, which display the estimated test characteristics for the sigmoid and Richard functional forms for the last ten iterations of the NUTS algorithm when it is run for 100 iterations. Indeed, for the last ten iterations, the sigmoid estimated test characteristics overlap. For the Richard functional form, this effect is weaker though more prominent than for the 25 iterations shown here. The ABC-SMC algorithm takes significantly longer to run for 100 iterations and these results are therefore infeasible to obtain.

6.2 Performance

Differences between the convergence of the two algorithms follow from the reported statistics. However, the visualized distributions and estimated test characteristics show multiple similarities between the functional forms as well. In order to further assess which functional form for which algorithm prevails, GOF statistics are examined and visualized simulations and observed outcomes are compared. Some summarizing statistics on the GOFs are given for both the ABC-SMC and NUTS algorithms in Tables 9 and 10.

The GOF indeed decreases with the number of iterations for all scenarios of the ABC-SMC algorithm, where the maximum is found for the first iteration and the minimum is found for the last (Table 14). Remarkably, the mean GOF and mean tolerance levels for 25 iterations of the Sigmoid functional form are higher than those for 15 iterations. This can be attributed to the 25 iterations scenario starting with a higher initial GOF than the 15 iterations scenario, seen from the reported maxima. It illustrates the uncertainty of the algorithm and emphasizes the need for letting the algorithm run long enough such that we can approximate the posterior better. For the NUTS algorithm, the GOF increases with the number of iterations (Table 15). Additionally, mean GOFs reported for all NUTS scenarios are higher than all ABC-SMC scenarios. The increasing GOFs for the NUTS algorithm clearly display the difference between the two algorithms. ABC-SMC is required to accept particle candidates that lead to a distance that is accepted based on the current tolerance level. Indeed, these tolerances are higher than the distances as shown in Table 9. As a result of the adaptive lowering of the tolerance level, the distances are required to decrease alongside it. However, the NUTS algorithm solely accepts candidates conditional on the data. It, therefore, does not require a bound for the distances. Because of this incremental distinction between the workings of the algorithms, simulated and observed outcomes are compared in Figures 31 to 34 in Appendix B.11. In accordance with the discussion in Section 5.2, the reported CIs for interval cancers are calculated from the Poisson distribution and for screen-detected adenomas and cancers from the binomial distribution. All confidence levels are set at 95%.

Figure 31 in Appendix B.11 shows outcomes for the ABC-SMC algorithm with the sigmoid functional form on the left, and the outcomes for the NUTS algorithm with the sigmoid functional form on the right for the first screening round of the Dutch national CRC screening program over 2014 to 2017. Overall, few simulated outcomes lie within the CIs of the observed outcomes. However, screen-detected (non-) advanced adenomas for the first two age groups and cancer stages II and IV seem to be estimated best. Here, the ABC-SMC algorithm leads to screen-detected non-advanced adenoma rates (Figure 31(a) in Appendix B.11) that fit the first two age groups[11] better than the rates by the NUTS algorithm do (Figure 31(b) in Appendix B.11). However, with the exception of

[11] [55, 60), [60, 65), [65, 70), [70+)

Table 9: Summarizing statistics on the GOFs and tolerance levels of the ABC-SMC algorithm by scenario

Functional form	*Iterations*		mean (SD)	CV	min	max
Sigmoid	*15*	GOF	64,306.93 (14,054.77)	21.86	57,366.00	104,732.00
		ε	68,104.70 (22,416.38)	32.91	57,396.36	132,817.63
	25	GOF	76,847.32 (11,368.88)	14.79	71,978.00	126,280.00
		ε	81,112.01 (23,460.07)	28.92	72,022.36	178,668.12
Richard	*15*	GOF	68,958.87 (13,022.24)	18.88	57,276.00	104,703.00
		ε	73,956.11 (18,853.53)	25.49	58,487.13	121,277.43
	25	GOF	59,780.92 (14,580.58)	24.39	49,147.00	115,009.00
		ε	64,303.44 (23,589.10)	36.68	49,642.19	156,213.33

Table 10: Summarizing statistics on the GOFs of the NUTS algorithm by scenario

Functional form	*Iterations*		mean (SD)	CV	min	max
Sigmoid	*15*	GOF	265,531.47 (12,9047.97)	48.60	34,168.00	480,985.00
	25	GOF	409,236.40 (232,143.79)	56.73	36,722.00	781,918.00
Richard	*15*	GOF	243,211.07 (130,049.39)	53.47	33,014.00	437,076.00
	25	GOF	325,812.68 (184,331.40)	56.58	23,030.00	619,332.00

screen-detected CRC, both algorithms overestimate all screening program outcomes for the highest ages for the first screening round. This effect is most prominent for reported interval cancers. In general, both algorithms show very similar fits for both adenoma size groups and interval cancers. Screen-detected CRC fit for stage I-IV is similar as well, although the ABC-SMC algorithm estimates screen-detected CRC stage II better, and the NUTS algorithm estimates screen-detected CRC stage IV cancers better. For these two observations, the fit for all age groups lies within or near the reported CIs. Screen-detected CRCs for stages I and III seem equally underestimated. This may be due to the distribution of observations over them. Since outcomes are identified according to age groups [55, 60), [60, 65), [65, 70) and [70+) it is possible that few observations in any of them compared to the number of individuals in that category lead to the simulations shown here. This illustrates the importance of gathering more data on the outcomes of the Dutch national CRC screening program up until 2021. Additionally, in order to mitigate some of these effects, groups in which few observations were found may be merged in future calibrations. However, besides the distribution of the observations over age groups, the underestimated screen-detected CRC may alternatively be an indicator that the calibrated sigmoid scale and shift coefficients for both algorithms lead to estimated test sensitivities that are too low for preclinical cancers and too high for adenomas. Too low estimated test sensitivities for preclinical cancers may further explain the overestimation of interval cancers since if we miss those cancers, they surely will have to be clinically detected in combination when combined with short preclinical cancer dwelling times.

Outcomes for the second screening round under the same conditions (Figure 32 in Appendix B.11) largely show the same behavior. However, screen-detected advanced adenomas are now underestimated instead of overestimated. Additionally, instead of ABC-SMC prevailing, non-advanced adenoma fit is better and more similar for both algorithms, especially for the first three age groups. Interval cancers remain largely overestimated for the second screening round as well, and screen-detected cancers remain underestimated as well for both algorithms. However, screen-detected CRC stages II and IV still show the best fit for both algorithms, which is now slightly better indicated by more simulations falling within the CIs. The same explanation about the distribution of observations over age groups and the estimated test characteristics may hold for the underestimation and overestimation seen for the second screening round outcomes.

Figures 33 and 34 in Appendix B.11 show the simulated and observed outcomes for the first and second screening rounds of the Dutch national screening program over 2014 to 2017 estimated with the ABC-SMC respectively NUTS algorithm using the Richard functional form for test sensitivity. Fit for (non-) advanced adenomas is similar to that of the sigmoid functional form for the first screening round for both algorithms (Figures 31(a) and 31(b) in Appendix B.11). Although both adenoma size groups are still overestimated, more observations fall within the CIs for the ABC-SMC algorithm as compared to the sigmoid functional form. Similar patterns as for screen-detected CRCs for the sigmoid functional form for the first screening round 1 are found where the ABC-SMC algorithm estimates stage II screen-detected CRC near perfectly, and stage IV screen-detected CRC is fit best by the NUTS algorithm, though this fit is similar for the ABC-SMC algorithm. Interval cancers show a similar fit to those estimated with the sigmoid functional form as well and are once again largely overestimated. This effect is again largest for the highest age groups and found in both adenoma size groups as well.

For the second screening round with the Richard functional form outcomes, screen-detected (non-) advanced adenomas show the same behavior as for the second screening round with the sigmoid functional form. Regardless, the fit is slightly better for the first three age groups for the NUTS algorithm this time, despite overestimating non-advanced adenomas for the last age group more gravely than the ABC-SMC algorithm again. Despite screen-detected CRCs showing the best fit for stages II and IV once more, their fit is reversed. Specifically, stage II screen-detected CRC is now simulated best for the NUTS algorithm, where three out of four age groups fall within the CIs. Conversely, stage IV screen-detected CRC is now fit best for the ABC-SMC algorithm, where all simulations lie within the CIs. However, three out of four age groups fall within the CIs for NUTS stage IV screen-detected CRC as well. Lastly, interval cancers show no difference with the other four scenarios and remain largely overestimated, particularly for the highest age group. The scale of overestimation remains the same and is comparable to the sigmoid functional form.

Figures 31 to 34 in Appendix B.11 show similar behavior for the different functional forms and across the two algorithms. Particularly, for both adenoma size groups and interval cancers, the last age group is systematically overestimated. Additionally, for all scenarios with 25 iterations, interval cancers are more heavily overestimated than they are for the original parameters. The scale of overestimation is equal for both functional forms. Overestimation of the last age groups mirrors stringent fit for most outcomes of the first age group. This group contains few observations due to the phased roll-out of the Dutch national CRC screening program. This may subsequently lead to improper fit for the other age groups, further explaining underestimation of screen-detected CRCs and overestimation of interval cancers. A possible way to resolve this issue in future works may be by gathering more data on first and second screening round outcomes as well as merging age groups with few observations. Regardless, of the under- and overestimated outcomes, most simulations that lie within the CIs are found for the ABC-SMC algorithm with the Richard functional form. This scenario does show a lower mean GOF than the sigmoid functional form with 25 iterations and both functional forms for the NUTS algorithm with 25 iterations. However, the Richard functional form estimates all test sensitivity shapes near the upper bound for all ages for the ABC-SMC algorithm and leads to extreme jumping behavior for NUTS estimated preclinical cancer b test sensitivity. As a result, even though the Richard test sensitivities lead to the best fit with respect to the reported statistics and displayed visualizations, their superiority should be accepted with care. The sigmoid functional form shows more stable behavior and may therefore be preferred. Alternatively, it may be evaluated whether the modeling test characteristics as a continuous function of age should be considered at all.

7 Conclusion

In this work, we compared the ABC-SMC and NUTS algorithms in order to determine which methodology is most effective for calibrating MISCAN-colon's parameters. Various convergence statistics were reported to determine convergence performance. In addition, GOF statistics, as well as visualization measures, were used to determine the algorithms' fit to the observed data. We additionally compared the sigmoid and Richard functional forms for estimating the test characteristics as continuous functions of age for both algorithms in order to determine how modeling test characteristics as a continuous function of screening age affects the performance of the two algorithms.

Claims on the convergence and performance of both algorithms across their functional forms were supported by reported acceptance rates, ESS statistics, and GOF. For the ABC-SMC algorithm, the acceptance rates indicated that the chain initially mixed well but that mixing rapidly slowed down. For the NUTS algorithm, acceptance rates lied closest to the required acceptance rate level of 80%. It was unclear from the constant level

of the acceptance rates and the ESSs whether the algorithm had converged yet. This is plausible since MCMC algorithms usually require more iterations to converge. Although the convergence statistics were ambiguous in identifying which of the two algorithms performed better, the ABC-SMC algorithm did show smaller GOF as compared to the NUTS algorithm. In addition, although the fit was similar across the functional forms and targets, the ABC-SMC algorithm with the Richard functional form led to the best fit, where the most simulated outcomes lied within the CIs. This suggests that the ABC-SMC algorithm with the Richard functional form is the best methodology for calibrating MISCAN-Colon's parameters.

The Richard functional form led to better performance than the sigmoid functional form, despite multiple test characteristics showing similar estimated shapes across the two. However, regardless of these best-fit findings, the Richard functional form led to questionable shapes for ABC-SMC test sensitivities. In addition, modeling the test characteristics through a function resulted in greater overestimation of outcomes than the original parameters' fit. It was argued that these test sensitivities cannot be trusted and that multiple factors may contribute to this finding. In combination with having only explored a subset of the posterior for only 25 iterations, poorly set priors may impede the algorithms' ability to find proper combinations of the Richard Coefficients that lead to reasonable test characteristics. On top of that, both functional forms differed across the 25 iterations, implying that neither of the two algorithms had converged yet for either test characteristics functional forms for 15 and 25 iterations. The comparison of posteriors with priors further emphasized that more iterations were needed to obtain a proper posterior distribution. Another indicator that the algorithm's number of iterations and the particle population size were set too small was the overfitting to small rates. Despite this, the number of observations available for each age group played a role there as well.

In summary, the presented findings emphasize the importance of running such Bayesian algorithms for higher numbers of iterations as well as the possible impediment of calibration by modeling test characteristics through a function. Ideally, more than 100 iterations are run for a population size of at least ten million individuals to facilitate convergence of the algorithms and minimize stochastic error. However, computational time is high and since we ran eight scenarios, computational time mattered greatly in this work. As a result, running 100 iterations for both algorithms would lead to a computational time of over two weeks. The 15 and 25 iterations ran here led to a total computational time of 8 days, which was acceptable in our case. The reason computational time is so high for these algorithms is on one hand because both algorithms sample multiple parameter sets and need to evaluate all of those samples. When the population sample size of our microsimulation model is high, once such model evaluation takes already up to 450 seconds. As a result, evaluating the model sequentially for many samples will add to the computational time needed. In addition, even though the proposed methods were partic-

ularly chosen bearing large problems in mind, the problem at hand became increasingly complex by including many parameters. This further increased computational time since large vectors of parameters were evaluated. Due to the high computational demand for Bayesian algorithms, we, therefore, considered a maximum of 25 iterations in this work. Because it is uncertain whether the low number of iterations lead to the unacceptable Richard functional forms, further research should aim to clarify whether increasing the number of iterations will mitigate this effect, whether we can improve the functional form estimation or whether the reduction in calibration parameters does not outweigh the disadvantages of finding non-reasonable test characteristics. Such research may perhaps indicate if modeling test characteristics as a function of age should not be done at all and instead, we may revert back to the original approach.

Besides highlighting the shortcoming of this work, we suggest various other starting points for further research. Throwing out some samples repeatedly after a fixed number of iterations, known as thinning, may reduce autocorrelation between samples. However, this does require a significant increase in the number of iterations run, in addition to the advised increase in the number of iterations. Additionally, although we calibrated preclinical dwelling times and test characteristics together because we expect them to simultaneously affect our outcomes, it would be of interest to reduce the number of parameters to be calibrated and re-establish the performance and computational time needed for the algorithms on this reduced set. Furthermore, it was argued it may be interesting to merge age groups for outcomes of the second screening round. Since we haven't observed many individuals in the lower age groups, the algorithm appears to overfit on these data points, leading to an improper fit on the remaining age groups. The GOF might possibly decrease as a result of the merge as well. Lastly, a re-calibration after data on the years 2018, 2019 and 2020 becomes available would be interesting to examine if the algorithms will perform better once more information on the first and second screening round for all screening ages becomes available. Another side note to this issue is, however, that the COVID-19 related screening delay will have to be modeled into MISCAN-Colon as well to replicate the 2020, and 2021 situations as close as possible.

A Tables

Table 11: Sex and age distributions of round 1 and round 2 screening participants

	Round 1 (n=2,598,654)	Round 2 (n=867,792)
Sex (%)		
Male	1,256,451 (48%)	416,294 (48%)
Female	1,342,203 (52%)	451,498 (52%)
Age groups (%)		
55-59	435,855 (17%)	1 (0%)[a]
60-64	859,980 (33%)	104,857 (12%)
65-69	712,546 (27%)	636,994 (74%)
70+	590,273 (23%)	125,940 (15%)

[a] 0.000115%

Table 12: Overview of the selected calibration parameters ordered by cancer dwelling times and test characteristics

Parameters
Description
1. Model specific parameters
Dwelling times
Preclinical cancer I
Preclinical cancer II
Preclinical cancer III
Preclinical cancer IV
Lack of specificity
Sensitivity
medium adenomas
large adenomas
short before clinical cancers
long before clinical cancers
Systematic lack of sensitivity
medium adenomas
large adenomas
2. Sigmoid function specific parameters

Dwelling times
Preclinical cancer I
Preclinical cancer II
Preclinical cancer III
Preclinical cancer IV
Lack of specificity
Scale coefficient
Shift coefficient
Sensitivity
Medium adenomas scale coefficient
Medium adenomas shift coefficient
Large adenomas scale coefficient
Large adenomas shift coefficient
Short before clinical cancers scale coefficient
Short before clinical cancers shift coefficient
Long before clinical cancers scale coefficient
Long before clinical cancers shift coefficient
Systematic lack of sensitivity
Medium adenomas scale coefficient
Medium adenomas shift coefficient
Large adenomas scale coefficient
Large adenomas shift coefficient
3. Richard function specific parameters
Dwelling times
Preclinical cancer I
Preclinical cancer II
Preclinical cancer III
Preclinical cancer IV
Lack of specificity
Scale coefficient
Shift coefficient
Sensitivity
Medium adenomas Q coefficient
Medium adenomas B coefficient
Medium adenomas ν coefficient
Large adenomas Q coefficient
Large adenomas B coefficient
Large adenomas ν coefficient

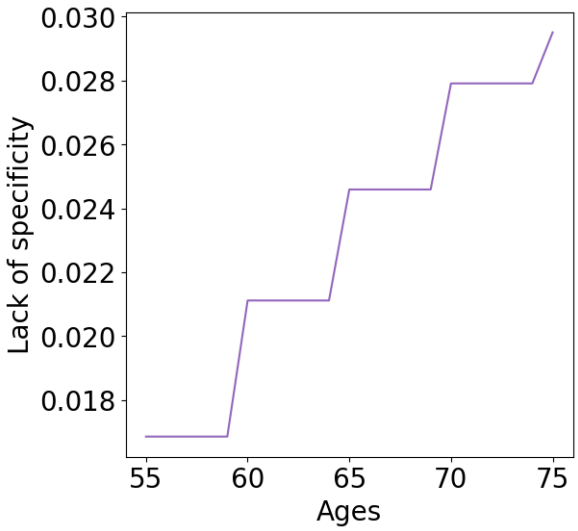
Short before clinical cancers Q coefficient
Short before clinical cancers B coefficient
Short before clinical cancers ν coefficient
Long before clinical cancers Q coefficient
Long before clinical cancers B coefficient
Long before clinical cancers ν coefficient
Systematic lack of sensitivity
Medium adenomas scale coefficient
Medium adenomas shift coefficient
Large adenomas scale coefficient
Large adenomas shift coefficient

Table 13: Overview of the selected calibration targets ordered by adenomas by size group and screen-detected cancers by stage and interval cancers

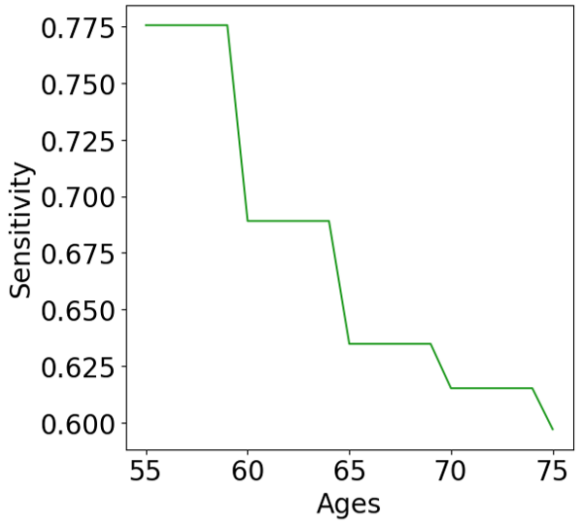
Targets
Description
Adenomas
screen-detected non-advanced adenomas
screen-detected advanced adenomas
screen-detected cancers
screen-detected CRC stage I
screen-detected CRC stage II
screen-detected CRC stage III
screen-detected CRC stage IV
Clinical cancers
Interval cancer

B Figures

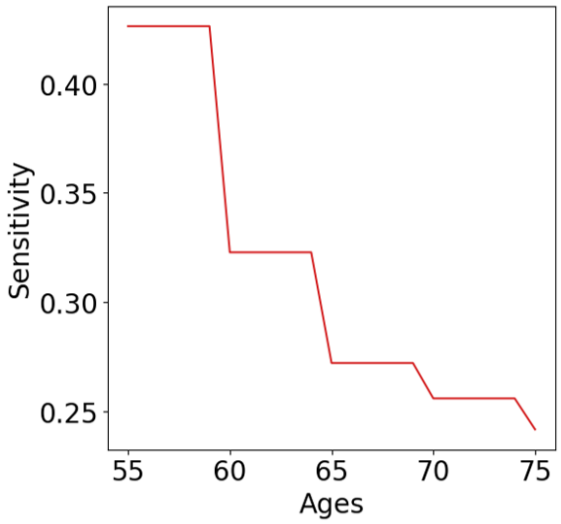
B.1 Original parameters



(a) Lack of specificity



(b) preclinical a



(c) preclinical b

Figure 17: Original lack of specificity (a) shown on its own scale by age group, preclinical cancer a sensitivity by age group (b) and preclinical cancer b by age group (c)

B.2 Sigmoid functional form for test characteristics

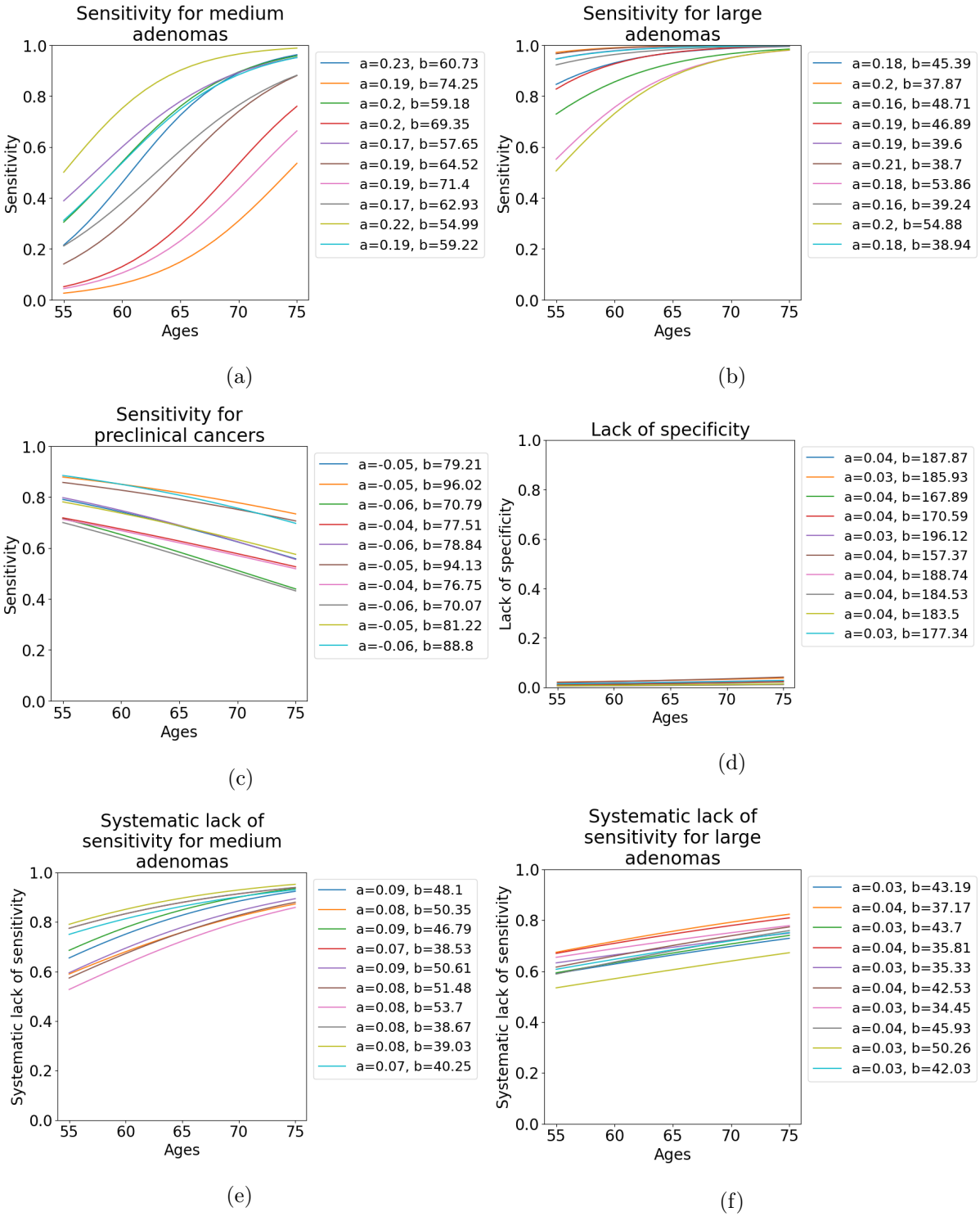
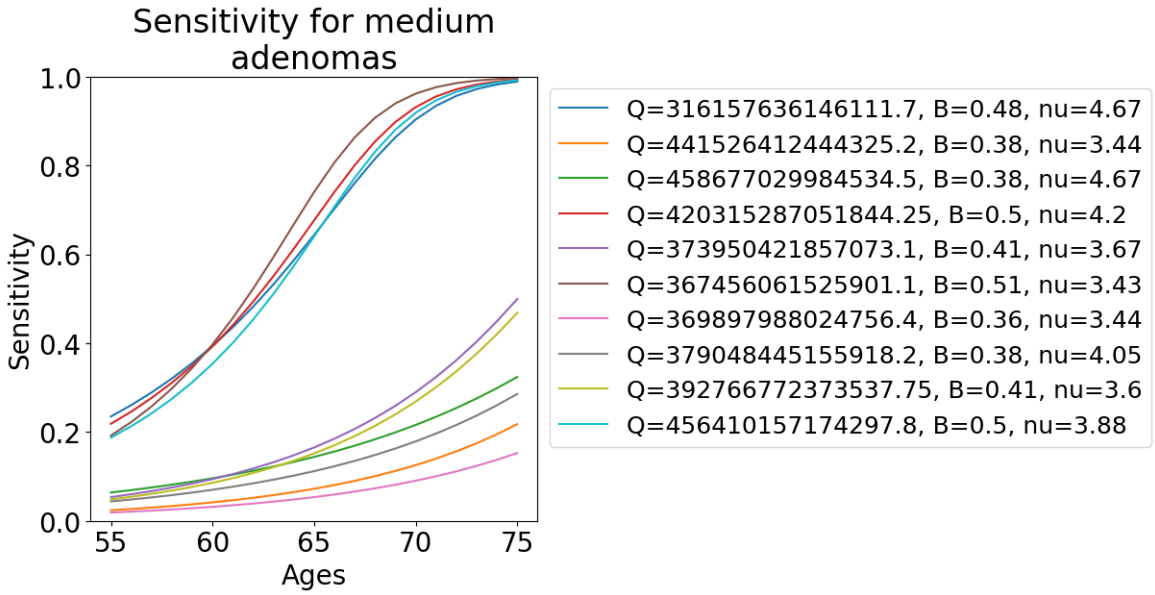
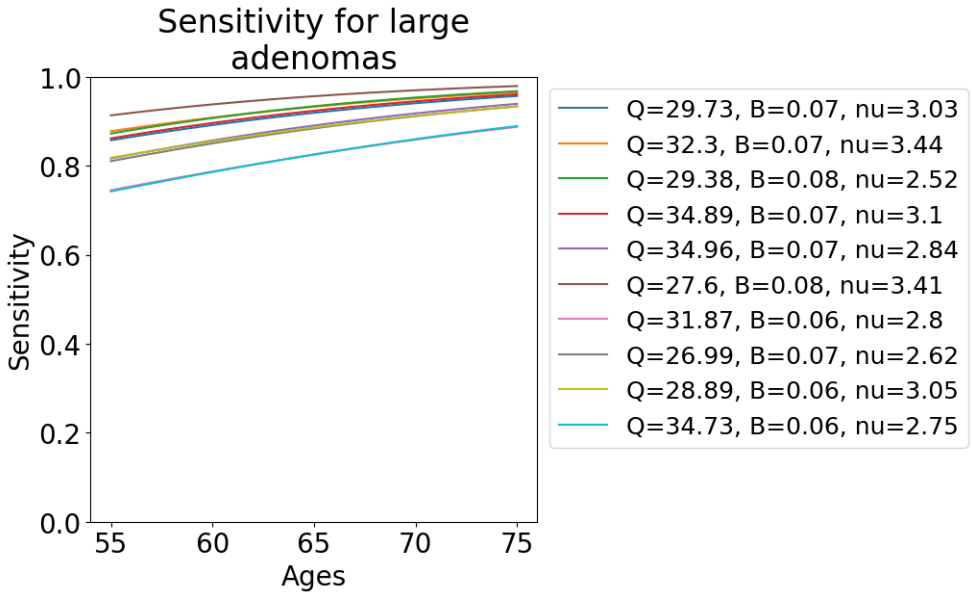


Figure 18: Sigmoid functional form for medium (a) and large (b) adenoma and preclinical cancer a and b (c) sensitivity, lack of specificity (d) and medium (e) and large (f) adenoma systematic lack of sensitivity

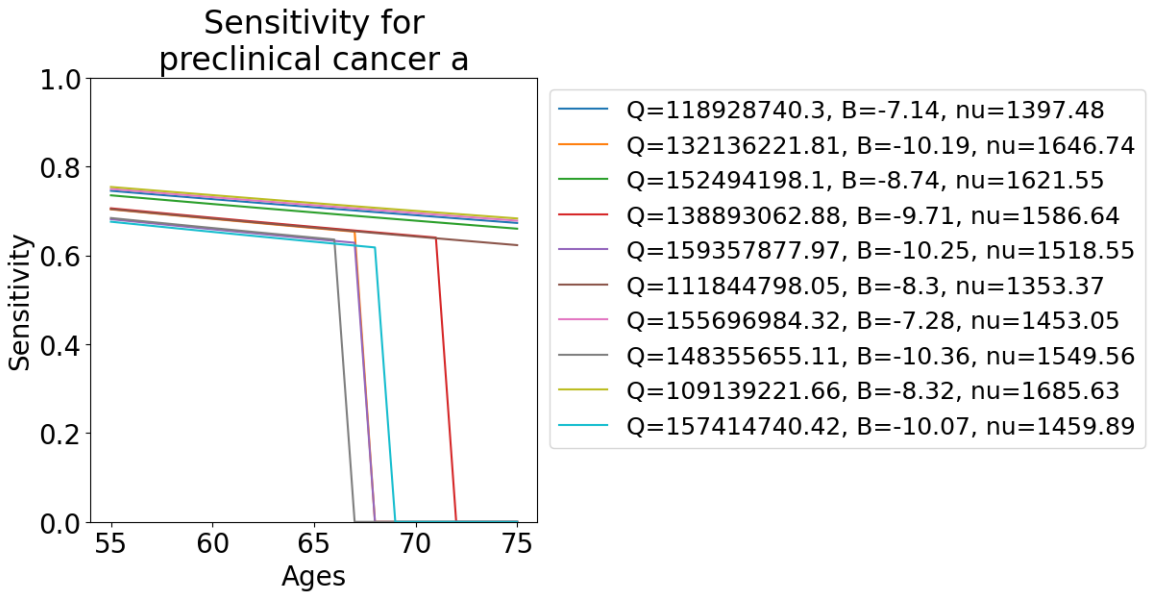
B.3 Richard functional form for test sensitivity



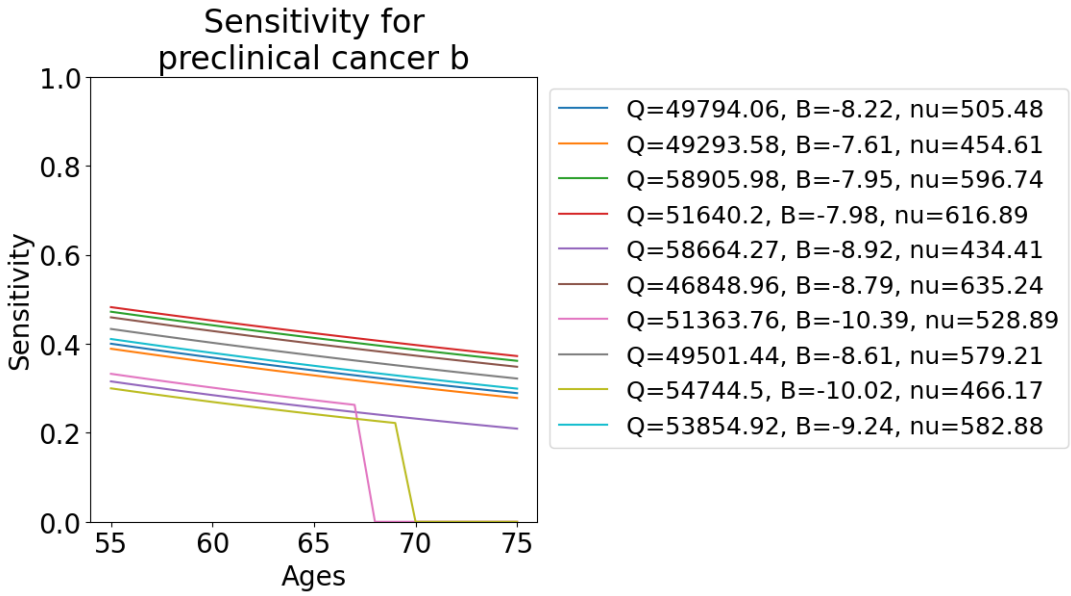
(a)



(b)



(c)



(d)

Figure 19: Richard functional form for medium (a) and large (b) adenoma and preclinical cancer a (c) and b (d) sensitivity

B.4 Miscellaneous figures

$$y = sens = \sigma(a, b, age) = \frac{1}{1 + e^{-a(age-b)}}$$

$$y = \ lack\ of\ spec = \frac{1}{1 + e^{-a(age-b)}}$$

$$y = sys\ lack\ of\ sens = \frac{1}{1 + e^{-a(age-b)}}$$

(a) Sigmoid functional form for the test characteristics

$$y = sens = \frac{1}{(1 + Qe^{-B \cdot age})^{\frac{1}{\nu}}}$$

(b) Richard functional form for test sensitivity

Figure 20: Sigmoid (a) and Richard (b) functional forms of the test characteristics as function of age, scale a and shift b coefficients respectively age, Q, B and ν coefficients

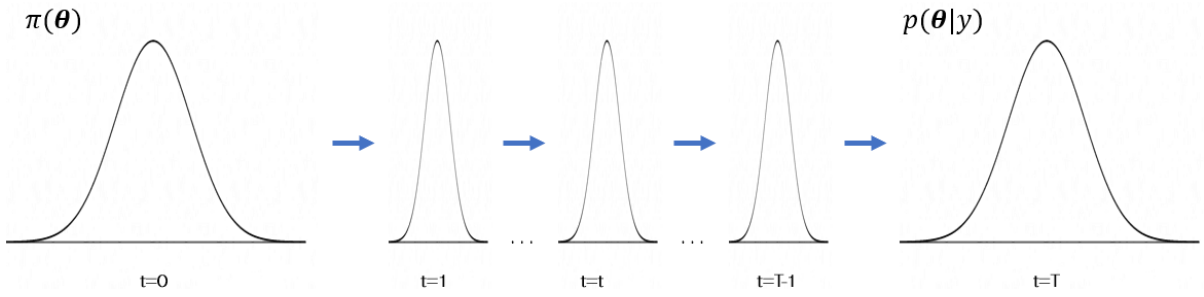
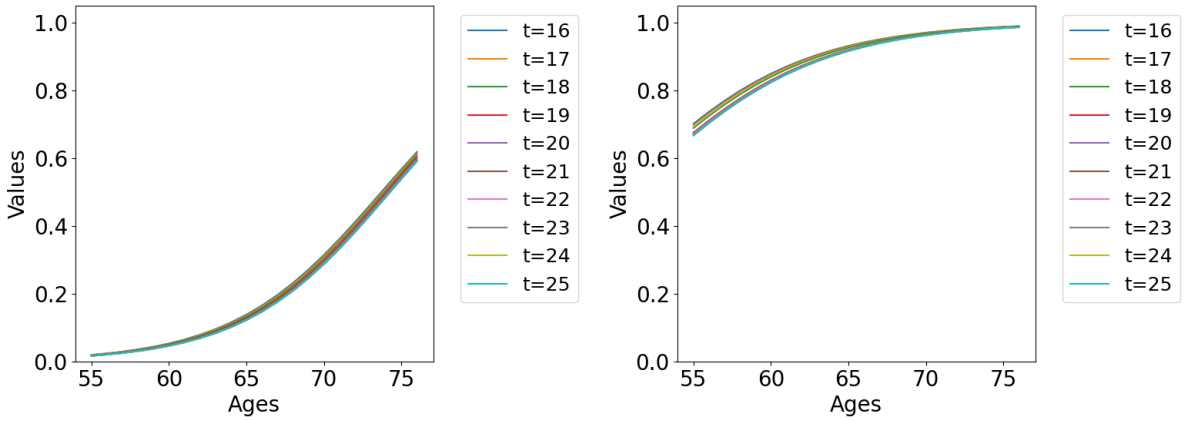


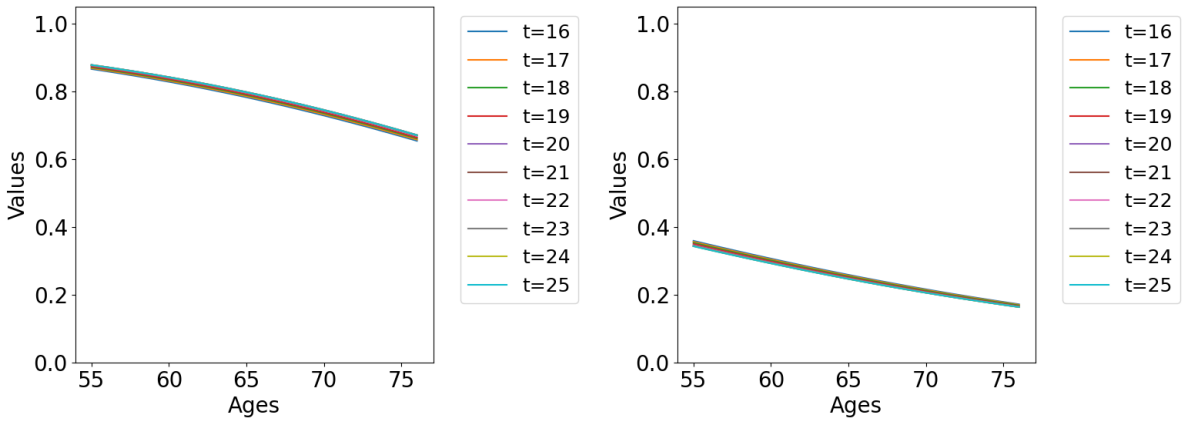
Figure 21: Graphical representation of moving from the prior to the posterior distribution in ABC-SMC

B.5 Estimated test characteristics for the ABC-SMC algorithm



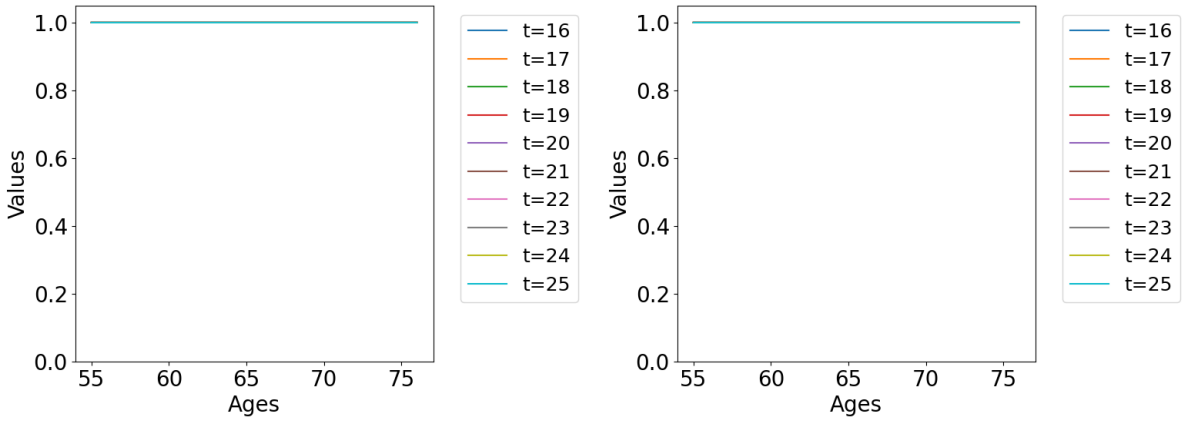
(a) Medium adenoma sensitivity

(b) Large adenoma sensitivity



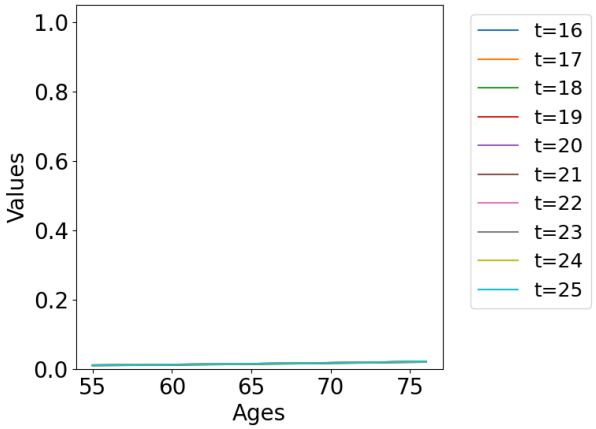
(c) Precla sensitivity

(d) Preclb sensitivity



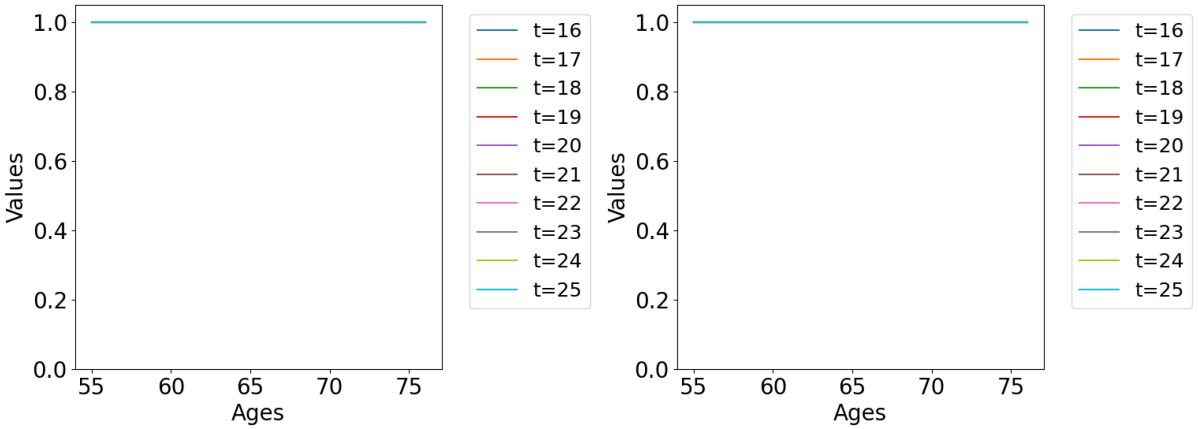
(e) Systematic lack of medium adenoma sensitivity

(f) Systematic lack of large adenoma sensitivity

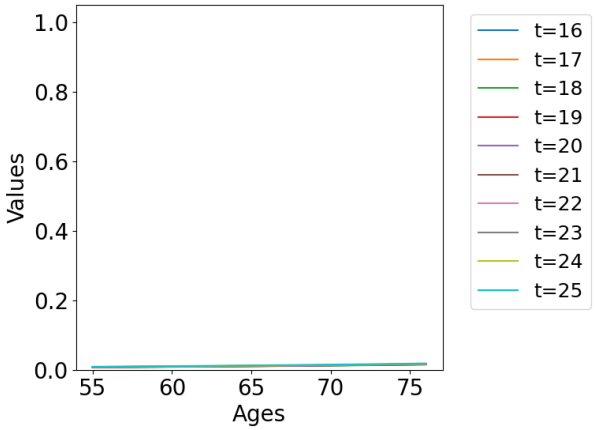


(g) Lack of specificity

Figure 22: Calibrated test sensitivity for medium (a) and large (b) adenomas and pre-clinical cancer a (c) and b (d), systematic lack of sensitivity for medium (e) and large (f) adenomas and lack of specificity (g) for the sigmoid scenario of the ABC-SMC algorithm displayed for particles $t = 16, \ldots, 25$ of the last generation



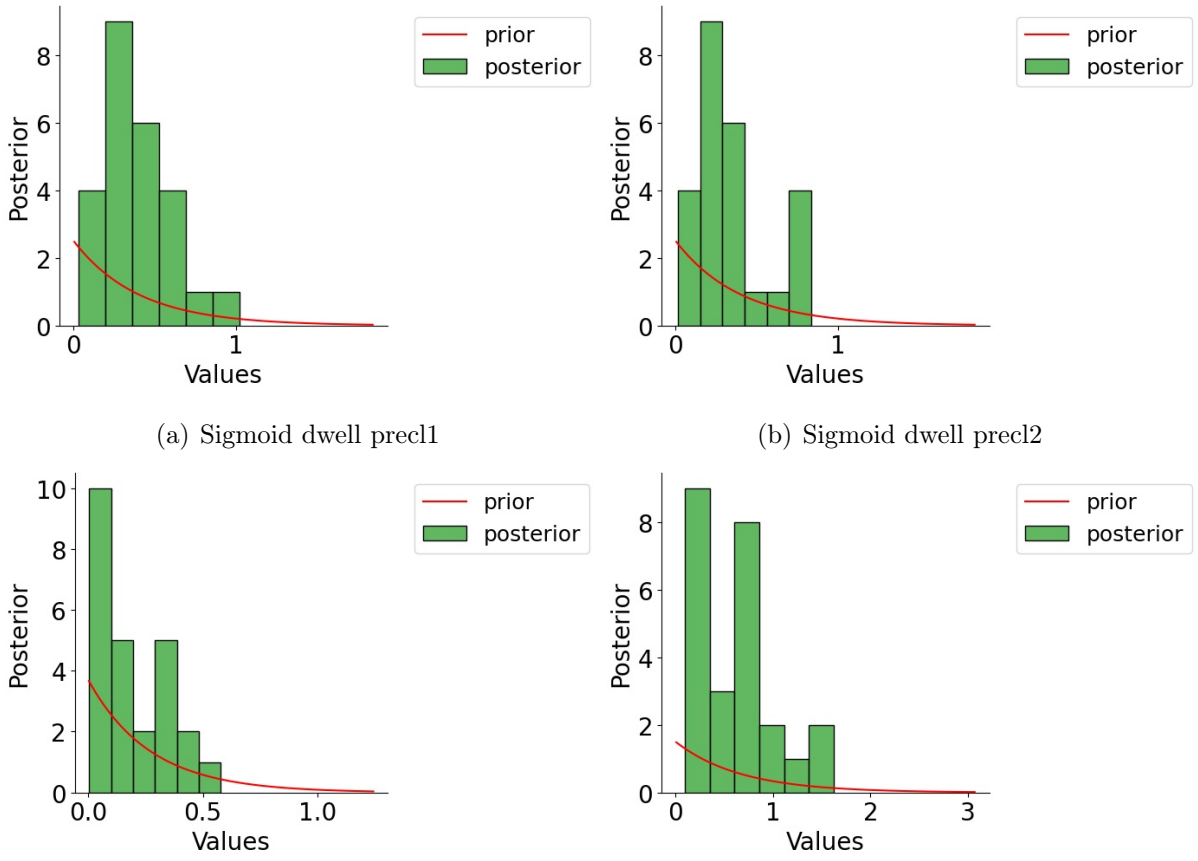
(a) Systematic lack of medium adenoma sensitivity (b) Systematic lack of large adenoma sensitivity



(c) Lack of specificity

Figure 23: Calibrated systematic lack of sensitivity for medium (a) and large (b) adenomas and lack of specificity (c) for the Richard scenario of the ABC-SMC algorithm displayed for particles $t = 16, \ldots, 25$ of the last generation

B.6 Calibrated preclinical cancer dwelling times for the NUTS algorithm



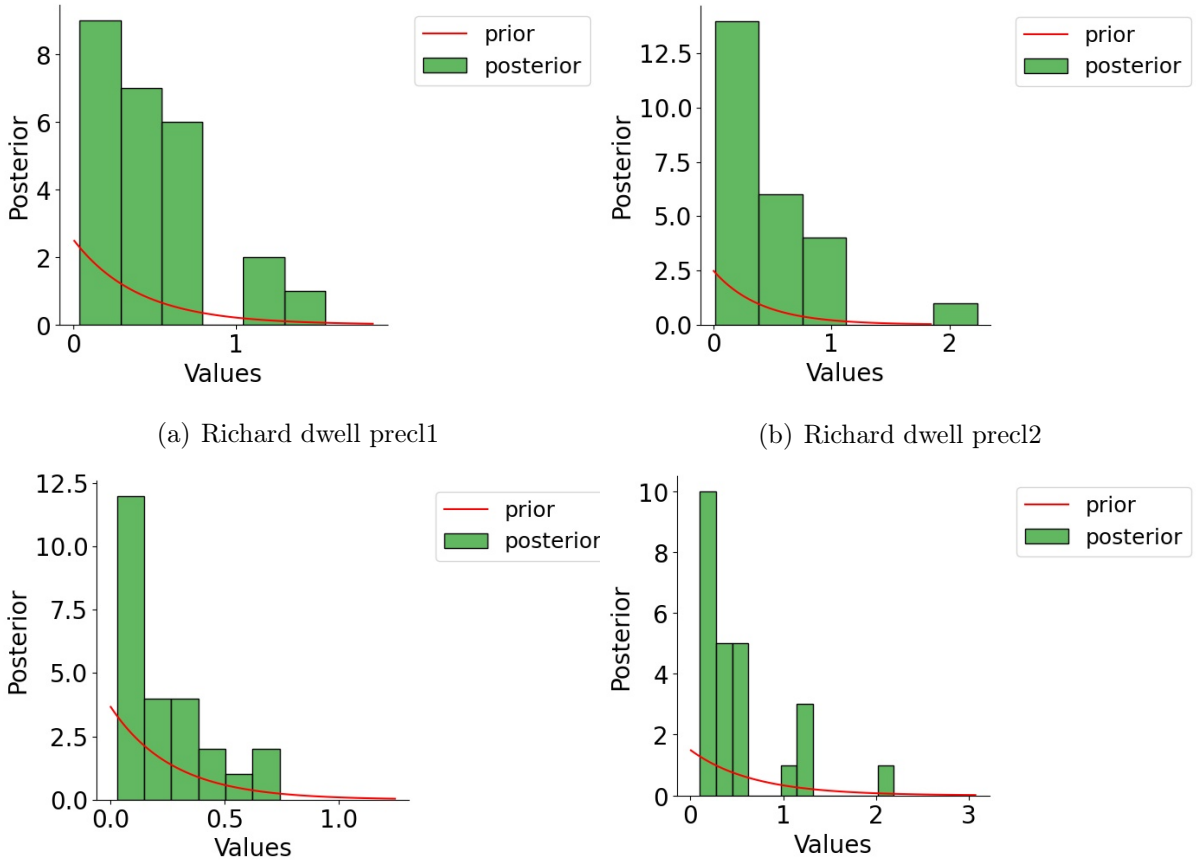
(a) Sigmoid dwell precl1

(b) Sigmoid dwell precl2

(c) Sigmoid dwell precl3

(d) Sigmoid dwell precl4

Figure 24: Priors (red) and posteriors (green) for sigmoid functional form calibrated dwelling times for preclinical cancer stage I (a), II (b), III (c) and IV (d) for 25 iterations of the NUTS algorithm



(a) Richard dwell precl1

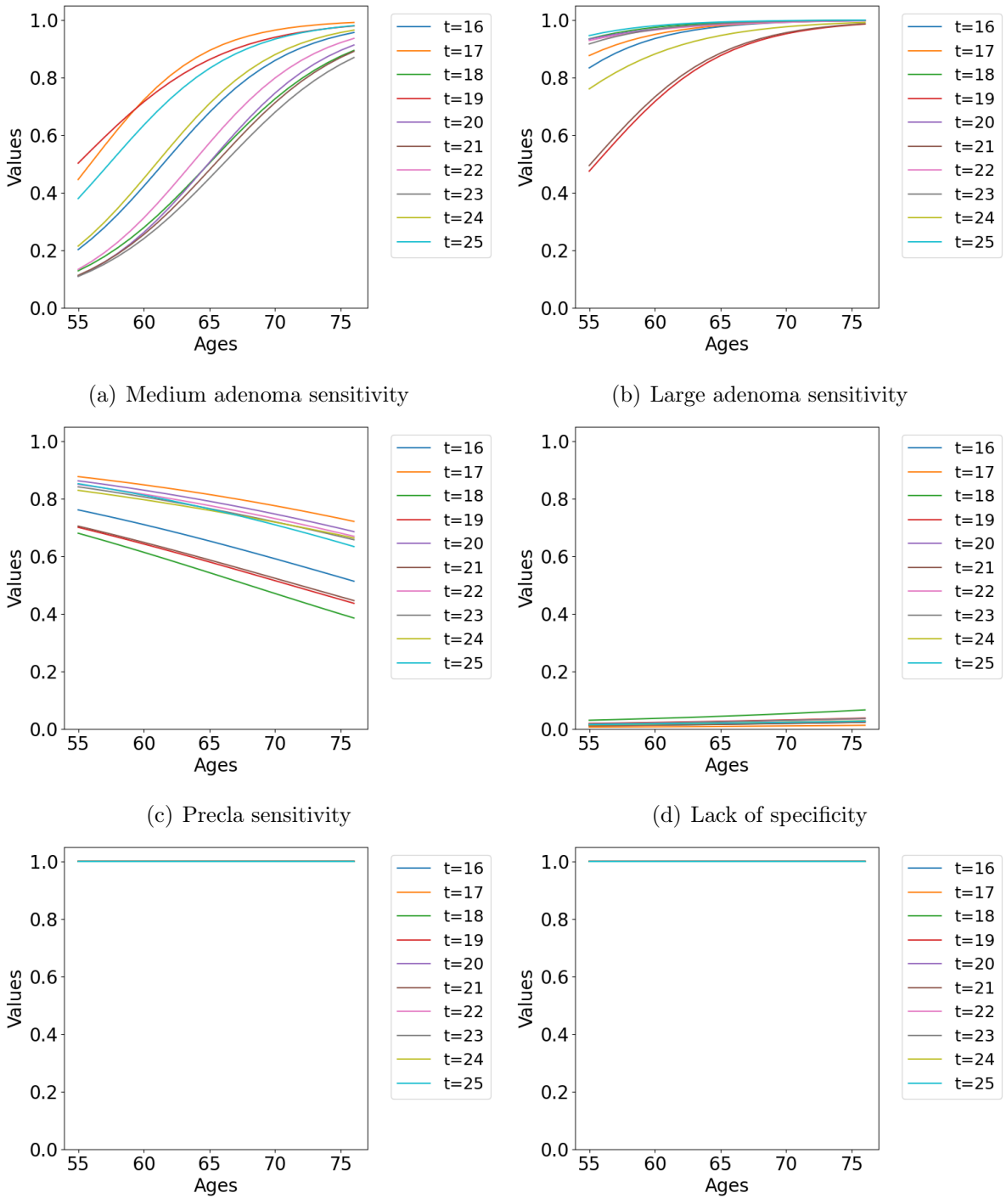
(b) Richard dwell precl2

(c) Richard dwell precl3

(d) Richard dwell precl4

Figure 25: Priors (red) and posteriors (green) for Richard functional form calibrated dwelling times for preclinical cancer stage I (a), II (b), III (c) and IV (d) for 25 iterations of the NUTS algorithm

B.7 Estimated test characteristics for the NUTS algorithm



(a) Medium adenoma sensitivity

(b) Large adenoma sensitivity

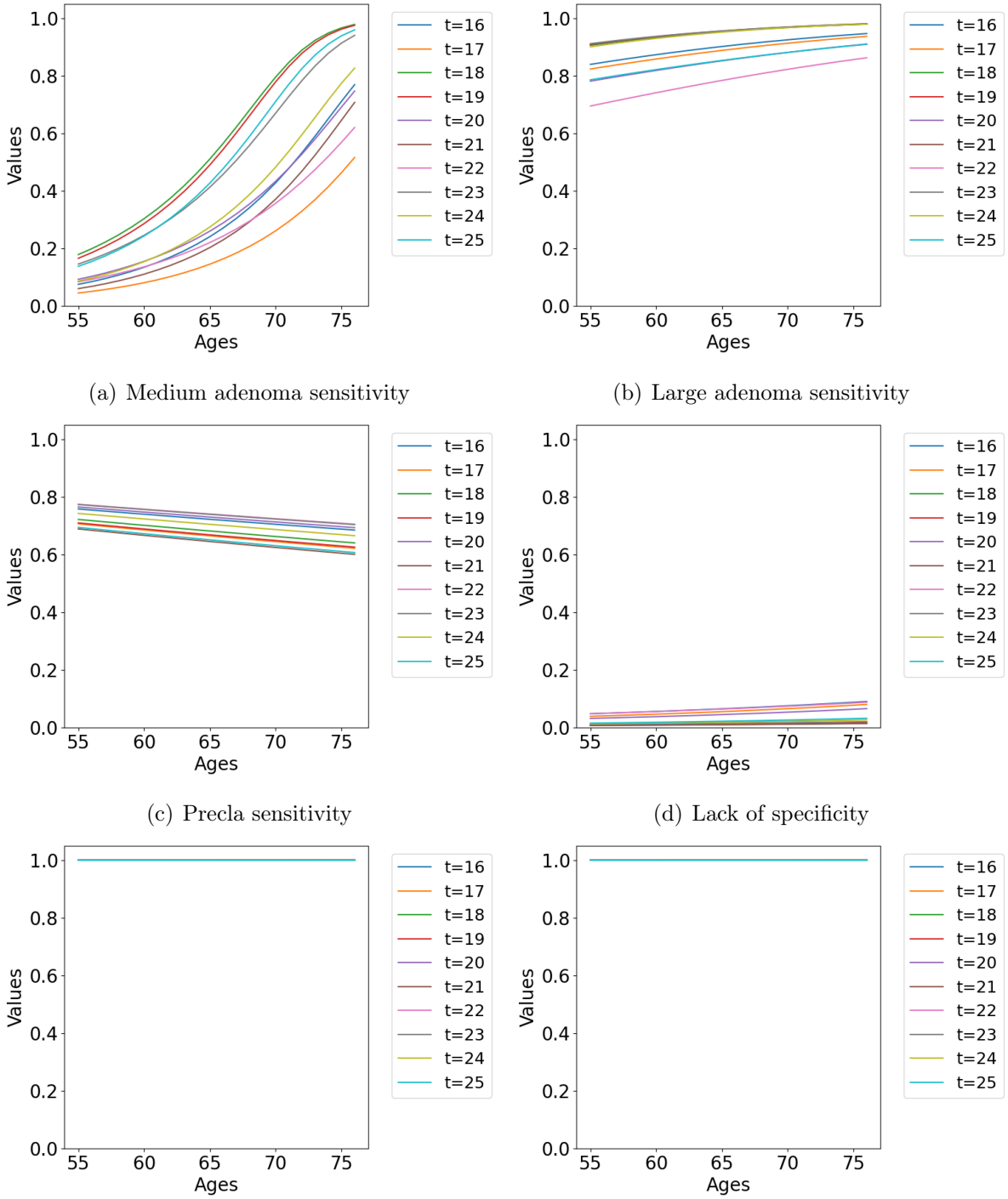
(c) Precla sensitivity

(d) Lack of specificity

(e) Systematic lack of medium adenoma sensitivity

(f) Systematic lack of large adenoma sensitivity

Figure 26: Calibrated test sensitivity for medium (a) and large (b) adenomas and preclinical cancer a (c), lack of specificity (d) and systematic lack of sensitivity for medium (e) and large (f) adenomas for the Sigmoid scenario of the NUTS algorithm displayed for iterations $t = 16, \ldots, 25$



(a) Medium adenoma sensitivity

(b) Large adenoma sensitivity

(c) Precla sensitivity

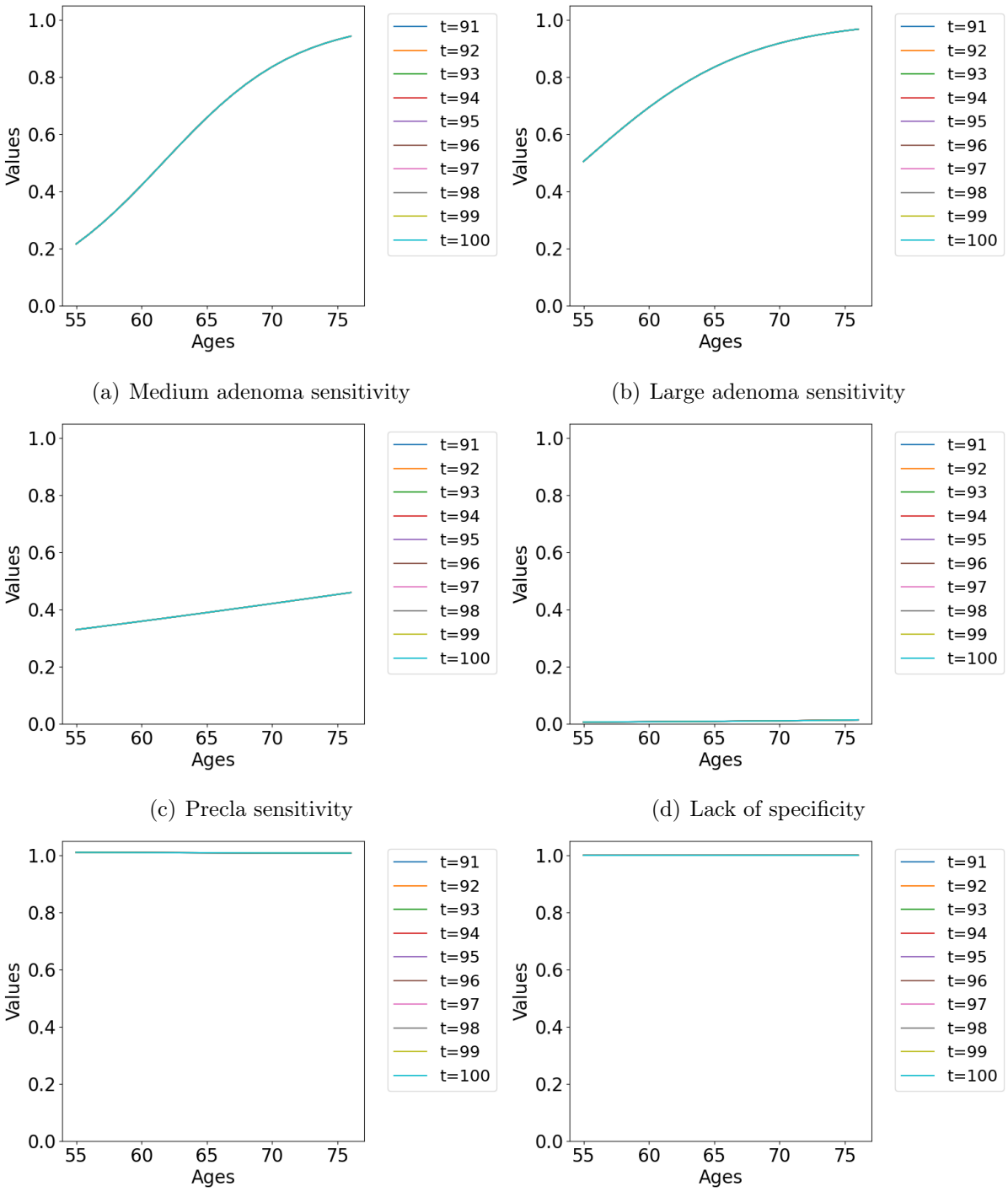
(d) Lack of specificity

(e) Systematic lack of medium adenoma sensitivity

(f) Systematic lack of large adenoma sensitivity

Figure 27: Calibrated test sensitivity for medium (a) and large (b) adenomas and pre-clinical cancer a (c), lack of specificity (d) and systematic lack of sensitivity for medium (e) and large (f) adenomas for the Richard scenario of the NUTS algorithm displayed for iterations $t = 16, \ldots, 25$

B.8 Estimated test characteristics for the NUTS algorithm with 100 iterations



(a) Medium adenoma sensitivity

(b) Large adenoma sensitivity

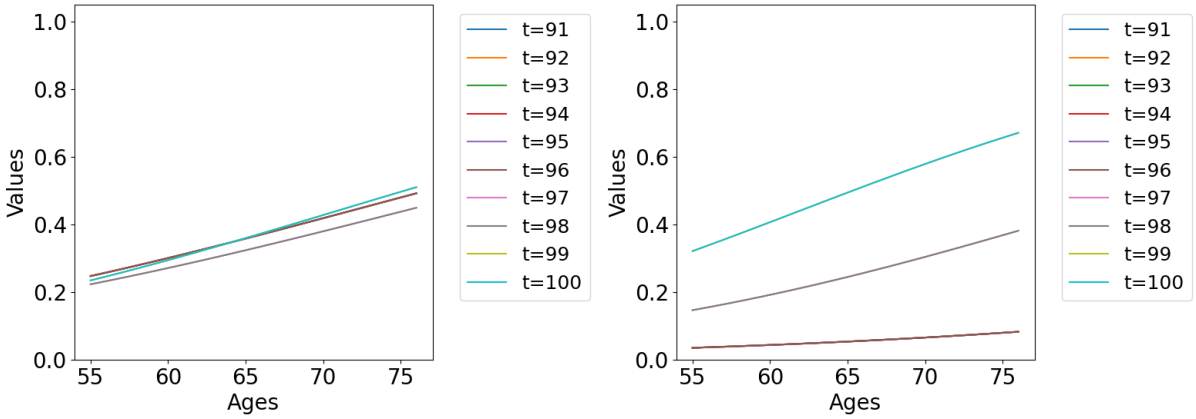
(c) Precla sensitivity

(d) Lack of specificity

(e) Systematic lack of medium adenoma sensitivity

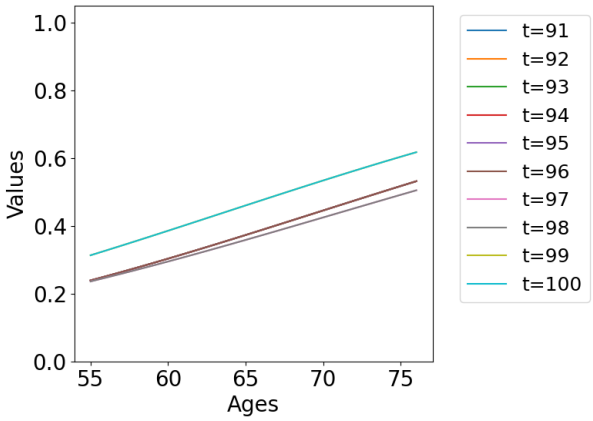
(f) Systematic lack of large adenoma sensitivity

Figure 28: Calibrated test sensitivity for medium (a) and large (b) adenomas and preclinical cancer a (c), lack of specificity (d) and systematic lack of sensitivity for medium (e) and large (f) adenomas for the Sigmoid scenario of the NUTS algorithm displayed for iterations $t = 91, \ldots, 100$

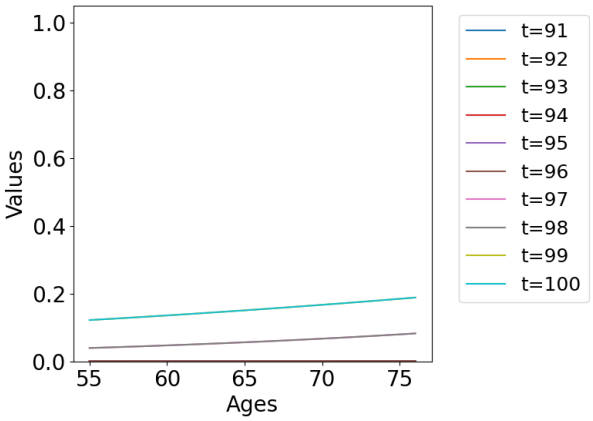


(a) Medium adenoma sensitivity

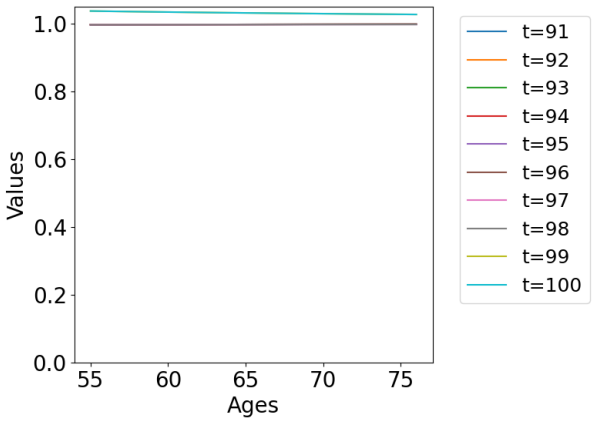
(b) Large adenoma sensitivity



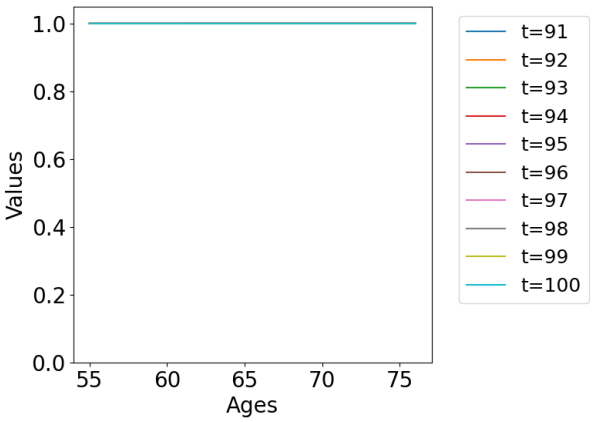
(c) Precla sensitivity



(d) Lack of specificity



(e) Systematic lack of medium adenoma sensitivity



(f) Systematic lack of large adenoma sensitivity

Figure 29: Calibrated test sensitivity for medium (a) and large (b) adenomas and preclinical cancer a (c), lack of specificity (d) and systematic lack of sensitivity for medium (e) and large (f) adenomas for the Richard scenario of the NUTS algorithm displayed for iterations $t = 91, \ldots, 100$

B.9 GOF development along the number of iterations

Table 14: GOF development along the number of iterations for the ABC-SMC algorithm

GOF				
Iterations	Sigmoid		Richard	
1	104732.00	126280.00	104703.00	115009.00
2	90270.00	95607.00	85041.00	84224.00
3	67667.00	80876.00	81184.00	75576.00
4	64276.00	78417.00	76226.00	71283.00
5	59482.00	76043.00	71233.00	66659.00
6	58589.00	75480.00	67445.00	64597.00
7	58273.00	75371.00	65007.00	62436.00
8	58112.00	74902.00	63364.00	59535.00
9	58007.00	74112.00	62510.00	58549.00
10	57815.00	73633.00	61739.00	57374.00
11	57657.00	73468.00	61123.00	56575.00
12	57558.00	73266.00	60205.00	55438.00
13	57415.00	73268.00	59353.00	54282.00
14	57385.00	73100.00	57974.00	53451.00
15	57366.00	72984.00	57276.00	52838.00
16	-	72922.00	-	52518.00
17	-	72869.00	-	51906.00
18	-	72805.00	-	51418.00
19	-	72587.00	-	50997.00
20	-	72418.00	-	50679.00
21	-	72322.00	-	50377.00
22	-	72245.00	-	50145.00
23	-	72156.00	-	50046.00
24	-	72074.00	-	49464.00
25	-	71978.00	-	49147.00

Table 15: GOF development along the number of iterations for the NUTS algorithm

GOF				
Iterations	Sigmoid		Richard	
1	34168.00	36722.00	33014.00	23030.00
2	84831.00	66091.00	69847.00	49934.00
3	135061.00	81645.00	87238.00	72193.00
4	158884.00	145228.00	113977.00	108781.00
5	191404.00	165567.00	167007.00	127227.00
6	228526.00	187889.00	183962.00	144077.00
7	245214.00	213154.00	223505.00	172481.00
8	273641.00	246953.00	266886.00	199437.00
9	294118.00	285891.00	287575.00	229186.00
10	315317.00	311281.00	304797.00	265091.00
11	342949.00	338686.00	318905.00	281333.00
12	372069.00	357413.00	358372.00	297474.00
13	393793.00	393113.00	375759.00	318488.00
14	432012.00	420436.00	420246.00	331553.00
15	480985.00	482324.00	437076.00	374031.00
16	-	505289.00	-	407851.00
17	-	547801.00	-	436132.00
18	-	582112.00	-	452279.00
19	-	621476.00	-	471946.00
20	-	643563.00	-	507414.00
21	-	665136.00	-	523569.00
22	-	689407.00	-	548734.00
23	-	716606.00	-	578825.00
24	-	745209.00	-	604919.00
25	-	781918.00	-	619332.00

B.10 Validation of the original parameters

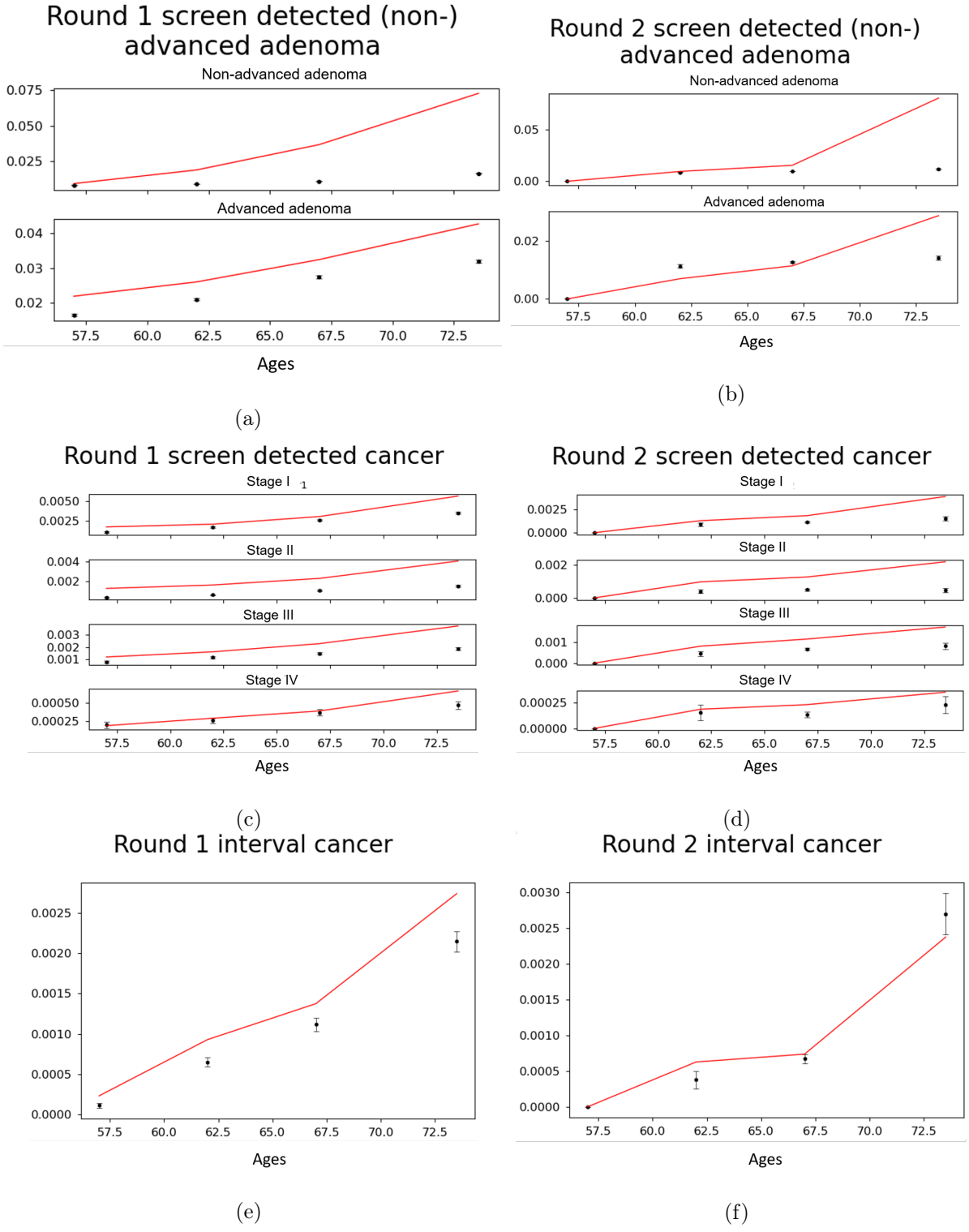
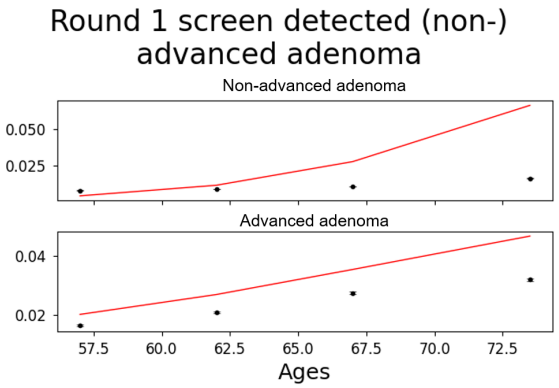
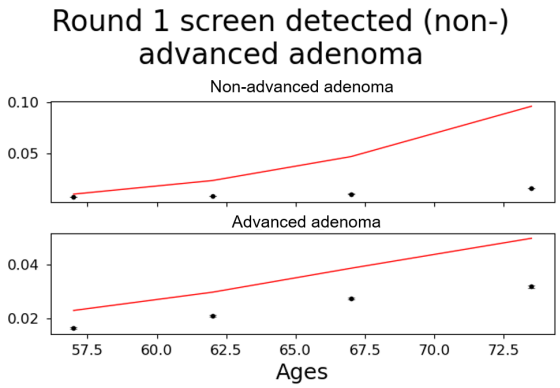


Figure 30: MISCAN-Colon original parameters estimated (red) versus observed (black) screen-detected (non-) advanced adenomas for the first (a) and second (b) screening round, screen-detected CRC for the first (c) and second (d) screening round and interval cancers for the first (e) and second (f) screening round with CIs resulting from the Dutch national CRC screening program in the Netherlands over 2014 to 2017

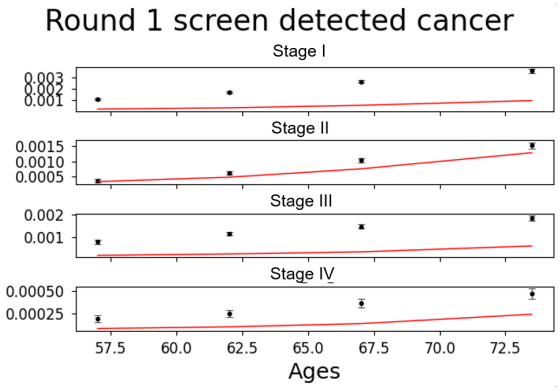
B.11 Estimated and observed outcomes



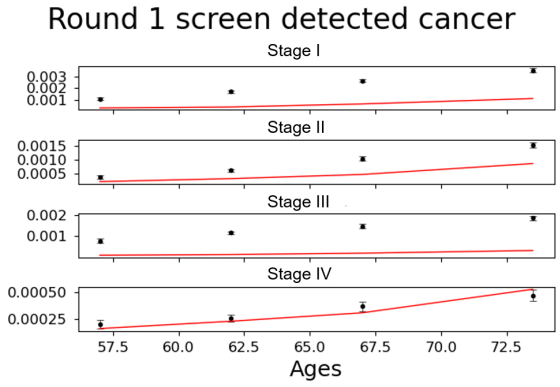
(a) ABC sigmoid functional form



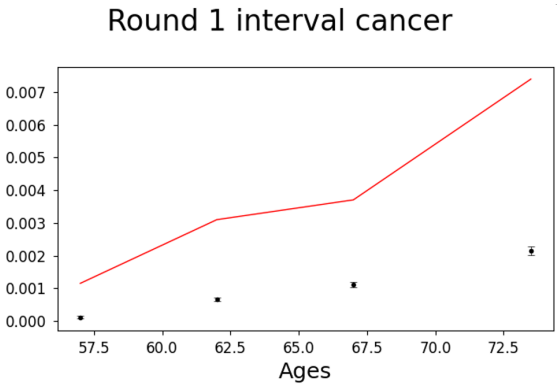
(b) NUTS sigmoid functional form



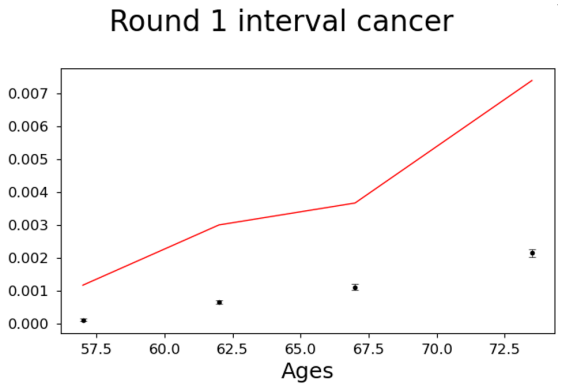
(c) ABC sigmoid functional form



(d) NUTS sigmoid functional form

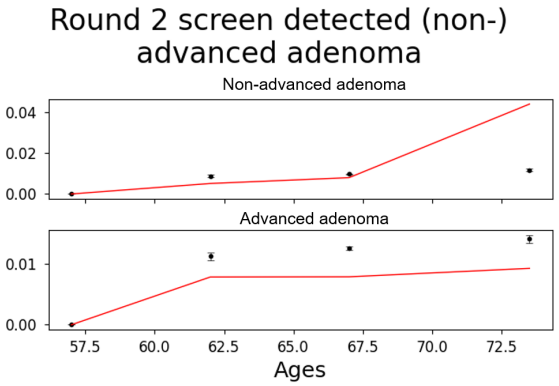


(e) ABC sigmoid functional form

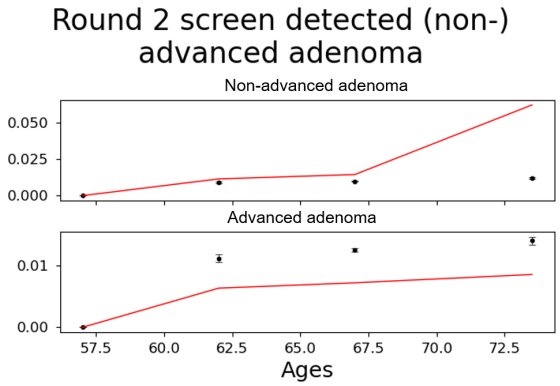


(f) NUTS sigmoid functional form

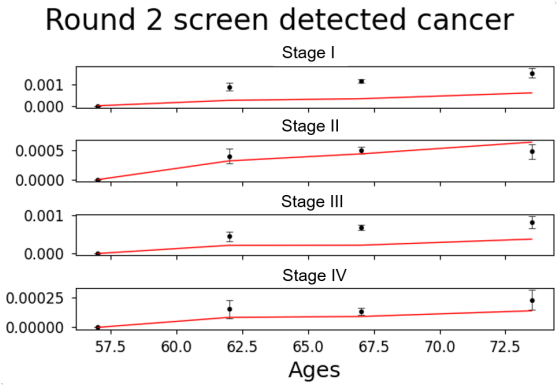
Figure 31: MISCAN-Colon estimated (red) versus observed (black) screen-detected (non-) advanced adenomas for the ABC (a) and NUTS (b) algorithms, screen-detected CRC for the ABC (c) and NUTS (d) algorithms and interval cancers for the ABC (e) and NUTS (f) algorithms displayed for the first screening round of the Dutch national CRC screening program over 2014 to 2017 found for 25 iterations with the sigmoid functional form



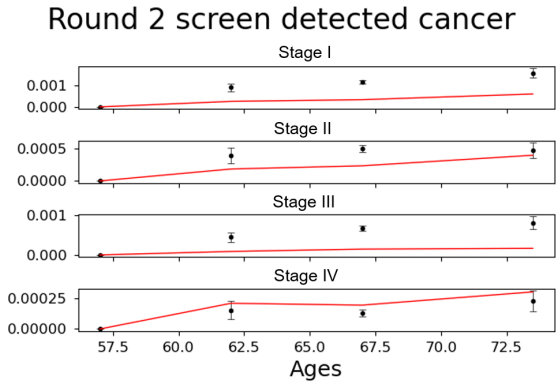
(a) ABC sigmoid functional form



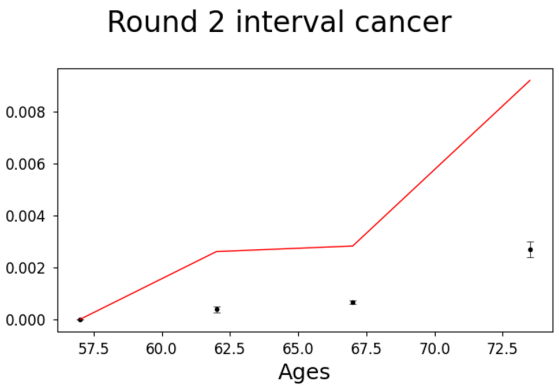
(b) NUTS sigmoid functional form



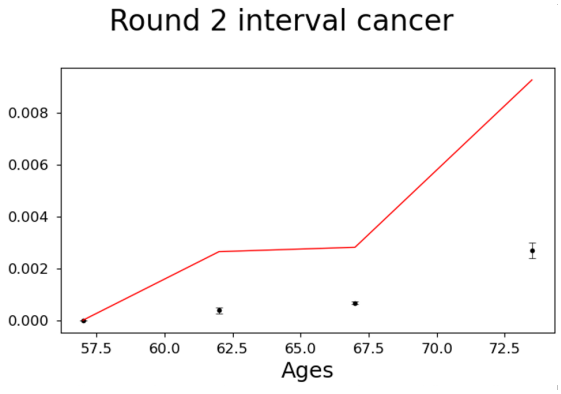
(c) ABC sigmoid functional form



(d) NUTS sigmoid functional form

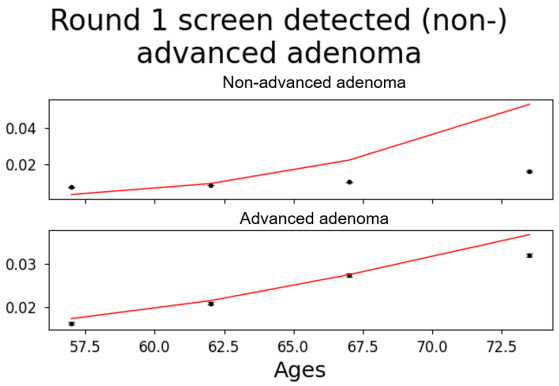


(e) ABC sigmoid functional form

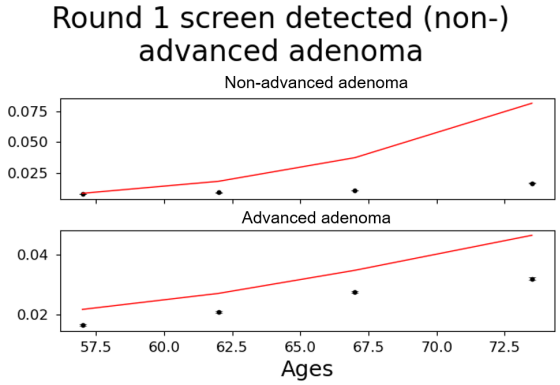


(f) NUTS sigmoid functional form

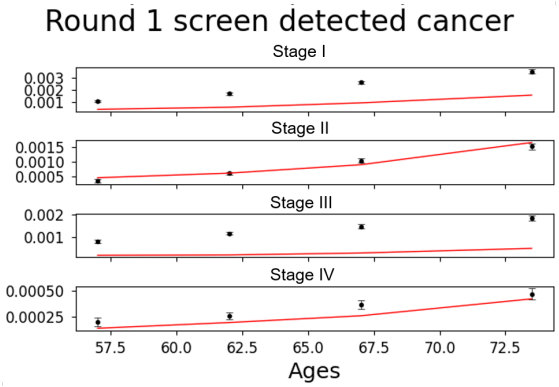
Figure 32: MISCAN-Colon estimated (red) versus observed (black) screen-detected (non-) advanced adenomas for the ABC (a) and NUTS (b) algorithms, screen-detected CRC for the ABC (c) and NUTS (d) algorithms and interval cancers for the ABC (e) and NUTS (f) algorithms displayed for the second screening round of the Dutch national CRC screening program over 2014 to 2017 found for 25 iterations with the sigmoid functional form



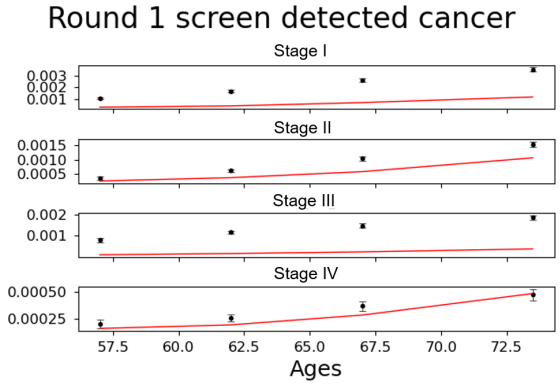
(a) ABC Richard functional form



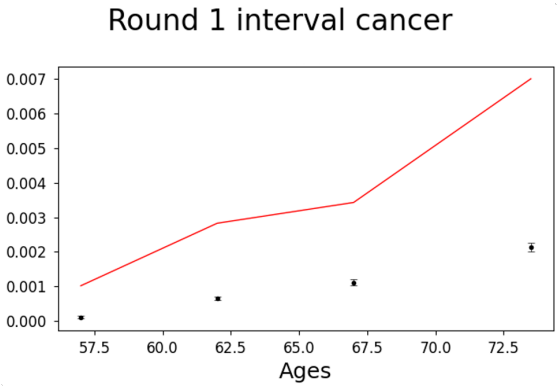
(b) NUTS Richard functional form



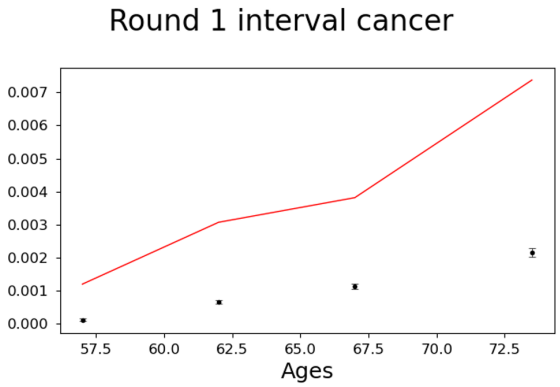
(c) ABC Richard functional form



(d) NUTS Richard functional form

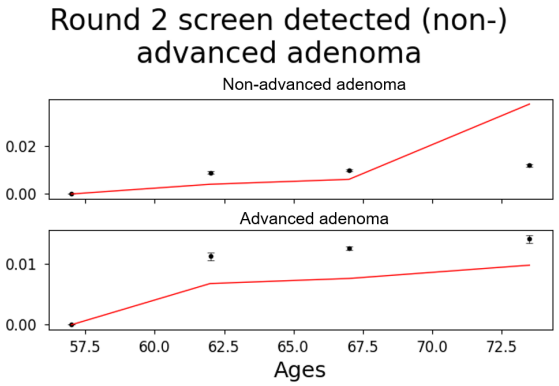


(e) ABC Richard functional form

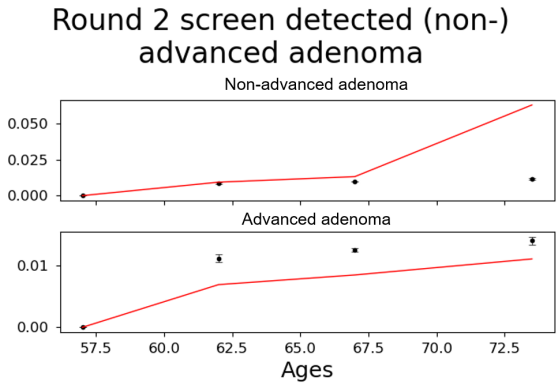


(f) NUTS Richard functional form

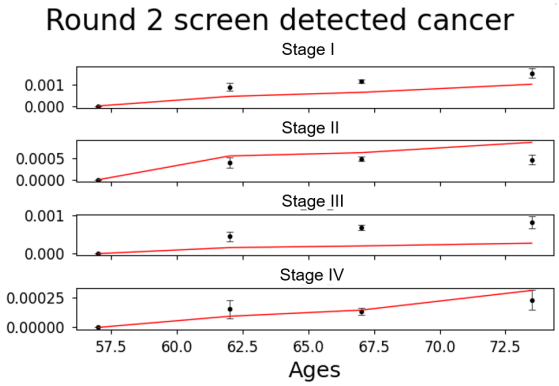
Figure 33: MISCAN-Colon estimated (red) versus observed (black) screen-detected (non-) advanced adenomas for the ABC (a) and NUTS (b) algorithms, screen-detected CRC for the ABC (c) and NUTS (d) algorithms and interval cancers for the ABC (e) and NUTS (f) algorithms displayed for the first screening round of the Dutch national CRC screening program over 2014 to 2017 found for 25 iterations with the Richard functional form



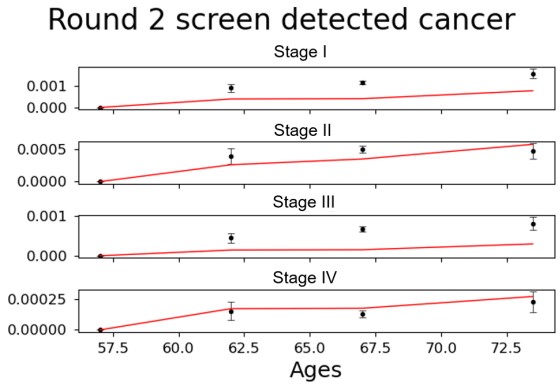
(a) ABC Richard functional form



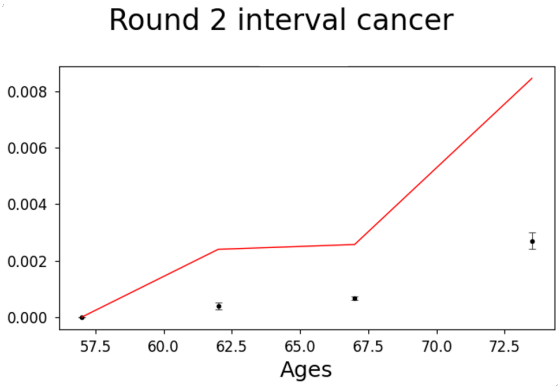
(b) NUTS Richard functional form



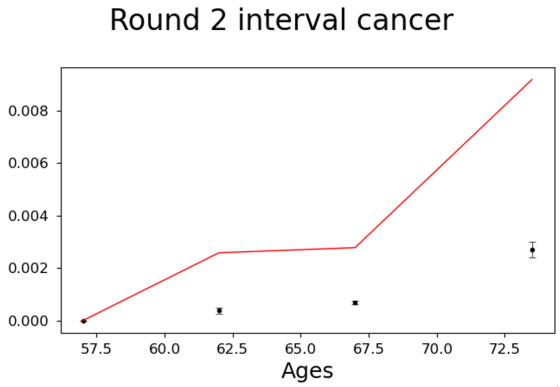
(c) ABC Richard functional form



(d) NUTS Richard functional form



(e) ABC Richard functional form



(f) NUTS Richard functional form

Figure 34: MISCAN-Colon estimated (red) versus observed (black) screen-detected (non-) advanced adenomas for the ABC (a) and NUTS (b) algorithms, screen-detected CRC for the ABC (c) and NUTS (d) algorithms and interval cancers for the ABC (e) and NUTS (f) algorithms displayed for the second screening round of the Dutch national CRC screening program over 2014 to 2017 found for 25 iterations with the Richard functional form

References

Alarid-Escudero, F., Knudsen, A. B., Ozik, J., Collier, N., and Kuntz, K. M. (2019). Characterization and valuation of uncertainty of calibrated parameters in stochastic decision models. *arXiv e-prints*, page arXiv:1906.04668.

American Cancer Society (2018). Colorectal cancer stages. https://www.cancer.org/treatment/understanding-your-diagnosis/staging.html. Accessed: 2020-08-17.

American Cancer Society (2020). Cancer staging. https://www.cancer.org/treatment/understanding-your-diagnosis/staging.html. Accessed: 2020-08-17.

Bergqvist, O. (2020). Calibration of breast cancer natural history models using approximate bayesian computation. Master's thesis, KTH Royal Institute of Technology, School of Engineering Sciences, Stockholm, Sweden.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv e-prints*, page arXiv:1701.02434.

Betancourt, M., Byrne, S., and Girolami, M. (2014). Optimizing the integrator step size for Hamiltonian Monte Carlo. *arXiv e-prints*, page arXiv:1411.6669.

Betancourt, M., Byrne, S., Livingstone, S., and Girolami, M. (2017). The feometric foundations of Hamiltonian Monte Carlo. *Bernoulli*, 23(4A):2257–2298.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6):394–424.

Cancer Intervention and Surveillance Modeling Network (2015). Colorectal cancer model profile. https://cisnet.cancer.gov/resources/profiles. Accessed: 2020-04-21.

Carpenter, B. (2018). We were measuring the speed of stan incorrectly—it's faster than we thought in some cases due to antithetical sampling. https://statmodeling.stat.columbia.edu/2018/01/18/measuring-speed-stan-incorrectly-faster-thought-cases-due-antithetical-sampling/. Accessed: 2021-03-20.

Chong, A. and Lam, K. (2017). A comparison of MCMC algorithms for the Bayesian calibration of building energy models. In *Proceedings of the 15th IBPSA Conference*, pages 1319–1328, San Francisco, CA, USA. International Building Performance Simulation Organization.

Corley, D. A., Jensen, C. D., Marks, A. R., Zhao, W. K., de Boer, J., Levin, T. R., Doubeni, C., Fireman, B. H., and Quesenberry, C. P. (2013). Variation of adenoma prevalence by age, sex, race, and colon location in a large population: implications for screening and quality programs. *Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association*, 11(2):172–180.

de Jonge, L. (2019). Calibrating parameters in the MISCAN model using a genetic algorithm. Master's thesis, Erasmus University Rotterdam, Rotterdam, ZH.

de Weerdt, A. (2019). Comparative simulation study on calibrating MISCAN-colon using ABC-SMC with adaptive multi-dimensional tolerance updating. Master's thesis, Erasmus University Rotterdam, Rotterdam, ZH.

Ferlay, J., Steliarova-Foucher, E., Lortet-Tieulent, J., Rosso, S., Coebergh, J., Comber, H., Forman, D., and Bray, F. (2013). Cancer incidence and mortality patterns in Europe: Estimates for 40 countries in 2012. *European Journal of Cancer*, 49(6):1374–1403.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC, 3rd edition.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

Geweke, J. F. (1991). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Staff Report 148, Federal Reserve Bank of Minneapolis.

Greenberg, E. (2014). *Introduction to Bayesian Econometrics*. Cambridge University Press, 2nd edition.

Habbema, J., van Oortmarssen, G., Lubbe, J., and van der Maas, P. (1985). The MISCAN simulation program for the evaluation of screening for disease. *Computer Methods and Programs in Biomedicine*, 20(1):79 – 93.

Hazelbag, C. M., Dushoff, J., Dominic, E. M., Mthombothi, Z. E., and Delva, W. (2020). Calibration of individual-based models to epidemiological data: A systematic review. *PLOS Computational Biology*, 16(5):1–17.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research.*, 15(1):1593–1623.

Karnon, J. and Vanni, T. (2011). Calibrating models in economic evaluation, a comparison of alternative measures of goodness of fit, parameter search strategies and convergence criteria. *Pharmaeconomics*, 29(1):51–62.

Kong, C. Y., McMahon, P. M., and Gazelle, G. S. (2009). Calibration of disease simulation model using an engineering approach. *Value in Health*, 12(4):521 – 529.

Loeve, F., Brown, M. L., Boer, R., van Ballegooijen, M., van Oortmarssen, G. J., and Habbema, J. D. F. (2000). Endoscopic Colorectal Cancer Screening: a Cost-Saving Analysis. *JNCI: Journal of the National Cancer Institute*, 92(7):557–563.

Meester, R. (2017). *Optimizing Outcomes for Colorectal Cancer Screening*. PhD thesis, Erasmus University Rotterdam, Rotterdam, ZH.

Minter, A. and Retkute, R. (2019). Approximate Bayesian computation for infectious disease modelling. *Epidemics*, 29:100368.

Morson, B. (1974). The polyp-cancer sequence in the large bowel. *Proceedings of the Royal Society of Medicine*, 67(6):451–457.

Naber, S. (2017). *Reducing Harms and Increasing Benefits of Screening for Cervical Cancer and Colorectal Cancer – A Model-based Approach*. PhD thesis, Erasmus University Rotterdam, Rotterdam, ZH.

Ozik, J., Collier, N., Wozniak, J., and Rutter, C. (2016). High performance calibration of a colorectal cancer natural history model with Incremental Mixture Importance Sampling. In *Computational Approaches for Cancer Workshop*, Salt Lake City, UT, USA. The International Conference on High Performance Computing, Networking, Storage and Analysis.

Paap, R. (2019). Bayesian econometrics (in finance). Erasmus University Rotterdam. Accessed: 2021–02-18.

Parkin, D. M. (2001). Global cancer statistics in the year 2000. *The Lancet. Oncology*, 2(9):533–543.

RIVM (2011). Colorectal cancer screening programme. https://www.rivm.nl/en/colorectal-cancer-screening-programme. Accessed: 2020-04-19.

RIVM (2013). Wat is darmkanker? https://www.rivm.nl/bevolkingsonderzoek-darmkanker/over-darmkanker. Accessed: 2020-04-19.

Rutter, C., Miglioretti, D., and Savarino, J. (2009). Bayesian calibration of microsimulation models. *Journal of the American Statistical Association*, 104:1338–1350.

Sai, A., Vivas-Valencia, C., Imperiale, T. F., and Kong, N. (2019). Multiobjective calibration of disease simulation models using Gaussian processes. *Medical Decision Making*, 39(5):540–552. PMID: 31375053.

Stout, N. K., Knudsen, A. B., Kong, C. Y., Mcmahon, P. M., and Gazelle, G. S. (2009). Calibration methods used in cancer simulation models and suggested reporting guidelines. *PharmacoEconomics*, 27(7):533–545.

Toes-Zoutendijk, E., Kooyker, A., Dekker, E., Spaander, M., Opstal van Winden, A., Ramakers, C., Buskermolen, M., van Vuuren, A., Kuipers, E., van Kemenade, F., Van Velthuysen, M.-L., Thomeer, M., Veldhuizen, H., Ballegooijen, M., Nagtegaal, I., Koning, H., Leerdam, M., and Lansdorp-Vogelaar, I. (2019). Incidence of interval colorectal cancer after negative results from first-round fecal immunochemical screening tests, by cutoff value and participant sex and age. *Clinical Gastroenterology and Hepatology*, 18:1493–1500.

Toes-Zoutendijk, E., Kooyker, A. I., Elferink, M. A., Spaander, M. C. W., Dekker, E., Koning, H. J. d., Lemmens, V. E., van Leerdam, M. E., and Lansdorp-Vogelaar, I. (2018). Stage distribution of screen-detected colorectal cancers in the Netherlands. *Gut*, 67(9):1745–1746.

Toes-Zoutendijk, E., van Leerdam, M. E., Dekker, E., van Hees, F., Penning, C., Nagtegaal, I., van der Meulen, M. P., van Vuuren, A. J., Kuipers, E. J., Bonfrer, J. M., Biermann, K., Thomeer, M. G., van Veldhuizen, H., Kroep, S., van Ballegooijen, M., Meijer, G. A., de Koning, H. J., Spaander, M. C., Lansdorp-Vogelaar, I., Schipper, D., Masclee, A., Wiersma, T., Otte, J., van der Beek, A., van Kemenade, F., Stoker, J., den Heeten, G., de Graaf, E., van Grevenstein, W., Kluiters, Y., and Blankenstein, M. (2017). Real-time monitoring of results during first year of Dutch colorectal cancer screening program and optimization by altering fecal immunochemical test cut-off levels. *Gastroenterology*, 152(4):767 – 775.e2.

Turner, B. M. and van Zandt, T. (2012). A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology*, 56:69–85.

Van der Meulen, M. P., Lansdorp-Vogelaar, I., van Heijningen, E.-M. B., Kuipers, E. J., and van Ballegooijen, M. (2016). Nonbleeding adenomas: Evidence of systematic false-negative fecal immunochemical test results and their implications for screening effectiveness—a modeling study. *Cancer*, 122(11):1680–1688.

van der Steen, A., van Rosmalen, J., Kroep, S., van Hees, F., Steyerberg, E. W., de Koning, H. J., van Ballegooijen, M., and Lansdorp-Vogelaar, I. (2016). Calibrating parameters for microsimulation disease models: A review and comparison of different goodness-of-fit criteria. *Medical Decision Making*, 36(5):652–665. PMID: 26957567.

van Hees, F., Habbema, J. D. F., Meester, R. G., Lansdorp-Vogelaar, I., van Ballegooijen, M., and Zauber, A. G. (2014). Should Colorectal Cancer Screening Be Considered in

Elderly Persons Without Previous Screening?: A Cost-Effectiveness Analysis. *Annals of Internal Medicine*, 160(11):750–759.

van Hees, F., Zauber, A. G., van Veldhuizen, H., Heijnen, M.-L. A., Penning, C., de Koning, H. J., van Ballegooijen, M., and Lansdorp-Vogelaar, I. (2015). The value of models in informing resource allocation in colorectal cancer screening: The case of the Netherlands. *Gut*, 64(12):1985–1997.

Vanni, T., Karnon, J., Madan, J., White, R. G., Edmunds, W. J., Foss, A. M., and Legood, R. (2011). Calibrating models in economic evaluation: A seven-step approach. *PharmacoEconomics*, 29(1):35–49.

Wilschut, J., Hol, L., Dekker, E., Jansen, J., Leerdam, M., Lansdorp-Vogelaar, I., Kuipers, E., Habbema, J., and Ballegooijen, M. (2011). Cost-effectiveness analysis of a quantitative immunochemical test for colorectal cancer screening. *Gastroenterology*, 141:1648–55.

Yildirim, I. (2012). Bayesian inference: Metropolis-Hastings sampling. http://www.mit.edu/~ilkery/papers/MetropolisHastingsSampling.pdf. Accessed: 2021-02-18.