Received: 15 March 2022 | Revised: 4 April 2023 | Accepted: 13 April 2023

DOI: 10.1002/sim.9756

RESEARCH ARTICLE

Statistics in Medicine WILEY

Approximate Bayesian computation for the natural history of breast cancer, with application to data from a Milan cohort study

Laura Bondi[1] | **Marco Bonetti**[2] | **Denitsa Grigorova**[3] | **Antonio Russo**[4]

[1]MRC Biostatistics Unit, Cambridge University, Cambridge, UK

[2]Department of Social and Political Sciences, Dondena Research Center, and Bocconi Institute for Data Science and Analytics, Bocconi University, Milan, Italy

[3]Big Data for Smart Society Institute, and Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski", Sofia, Bulgaria

[4]UOC Osservatorio Epidemiologico, ATS, Milan, Italy

Correspondence
Laura Bondi, MRC Biostatistics Unit, East Forvie Building, Cambridge Biomedical Campus, Cambridge CB2 0SR, UK.
Email: laura.bondi@mrc-bsu.cam.ac.uk

Summary

We explore models for the natural history of breast cancer, where the main events of interest are the start of asymptomatic detectability of the disease (through screening) and the time of symptomatic detection (through symptoms). We develop several parametric specifications based on a cure rate structure, and present the results of the analysis of data collected as part of a motivating study from Milan. Participants in the study were part of a regional breast cancer screening program, and their ten-year trajectories were obtained from administrative data available from the Italian national health care system. We first present a tractable model for which we develop the likelihood contributions of the observed trajectories and perform maximum likelihood inference on the latent process. Likelihood based inference is not feasible for more flexible models, and we implement approximate Bayesian computation (ABC) for inference. Issues that arise from the use of ABC for model choice and parameter estimation are discussed, including the problem of choosing appropriate summary statistics. The estimated parameters of the underlying disease process allow for the study of the effect of different examination schedules (age range and frequency of screening examinations) on a population of asymptomatic subjects.

KEYWORDS

approximate Bayesian computation, breast cancer, missing data, multi-state modeling, natural history, screening

1 | INTRODUCTION

Cancer screening is defined as the examination of asymptomatic subjects in order to classify them as likely or unlikely to be diseased.[1] Some recent reviews assess breast cancer screening.[2-5] The expected positive aspects of screening are the reduction in mortality and the avoidance of advanced morbidity. However, along with the benefits there may be negative effects of screening such as overdiagnosis, overtreatment, and false positive results that may lead to psychological distress.

Once a screening program is established in a country, it is difficult to conduct randomized trials to assess the effectiveness of screening. Sound, updated, and country-specific evidence is needed to decide whether to establish breast cancer screening programs and to identify the optimal screening policy with respect to the age range of the women invited, and the lag between successive examinations. As a consequence, there is a strong interest in learning about the natural history of the disease from observational data collected administratively.

Hu and Zelen[6] discuss a theoretical model for planning screening trials in order to compare mortality rates between a control group and a screened group. The authors model the natural history of the disease and how the disease could be detected by regular screening examinations. The work is used for planning the National Lung Screening Trial.

A commentary by Aalen[7] introduces a different class of models whose aim is to understand disease processes beyond the simple survival setting and integrating into the analysis all the information collected at each clinical examination.[8,9] In Sweeting et al,[8] the authors implement multi-state Markov models to analyze the longitudinal disease progression when transition times between disease states are interval censored, and taking into account different assumptions on the possibly non-ignorable missing data process occurring during follow-up. This setting reflects the screening context in which, even though examination times are scheduled, subjects can decide not to attend them and the decision to adhere to the scheduled examinations is possibly not independent of the underlying disease status or of the (perceived or real) risk of the subject. Similarly, Chen et al[9] is concerned with the analysis of incomplete longitudinal data, where the observation process may contain information about the life history of the disease. They consider progressive multi-state Markov response models where the parameter estimation is performed by maximizing the likelihood function.

An alternative to multi-state Markov models is the modeling of the underlying biological tumour growth as a continuous process. Recent work[10,11] proposes a continuous tumour growth model and derives theoretical results for jointly estimating tumour growth, time to symptomatic detection and mammography screening sensitivity as a function of mammographic density. These models evaluate mammography screening in terms of mortality, to estimate overdiagnosis, and to estimate the impact of lead-time bias when comparing survival times between screen-detected cancers and cancers found outside of the screening program. The models are implemented using likelihood-based estimation, with recent work exploring a likelihood-free approach consisting of calibrating the parameters via summary statistics at the population level.[12]

Another relevant work[13] is based on a likelihood-free estimation procedure designed to replicate standardized incidence rates of breast cancer. However, their focus is not on the natural history of the disease, which is only partly estimated from the data, but on quantifying the magnitude of overdiagnosis for invasive cancers and for carcinoma is situ cases.

The aim of this article is to explore several statistical models to describe the natural history of breast cancer, focusing on the insurgence of the disease, and on the detection of cases as it progresses from asymptomatic to symptomatic.

In Section 2, we describe the motivating observational study conducted in Milan. While observational studies do not typically provide trusted evidence to answer the same questions as randomized trials do, here the goal is to reconstruct the underlying latent process through the probabilistic description of the occurrence and development of breast cancer from a combination of data obtained from a screening program and from administrative health data streams.

All the models that we discuss can be seen as multi-state semi-Markov models, where the future evolution depends not only on the current state, but also on the entry time into that state. The estimation procedure that we employ depends on the complexity of the model. In principle, it is possible to compute the observed data likelihood[14] of each model, in order to find the maximum likelihood estimates for the parameters. However, the likelihood calculation and maximization can be numerically complicated or not feasible, unless the model has a simple structure.

Section 3 describes the modeling approach, and describes one such simple model for which likelihood inference is feasible. In Section 4, we move to the Bayesian inferential framework and develop a likelihood-free estimation procedure based on approximate Bayesian computation (ABC)[15] that allows one to implement a variety of models and to perform both model selection and parameter estimation on the motivating data. In Section 5, we discuss the use of ABC in this setting, and close with some final remarks.

2 | THE MOTIVATING DATA

The data that motivated this study concern a cohort of $n = 78051$ women, aged between 41 and 76 years, resident in the municipality of Milan, who were invited to participate to the mammographic screening program and in particular to a study with the acronym of FRiCaM (Risk Factors for Breast Cancer: Fattori di Rischio per il Carcinoma della Mammella), supported by a specific grant of the Italian League of Cancer Prevention. Italy does not have a universal screening program

for all regions in the country, but currently all Italian regions have implemented screening programs.[16] Screening examinations in Milan are normally offered to women 50–74 years old every two years (recently extended from the previous 50–69 policy), but under specific circumstances high-risk women can also be included in the program. All women had to be disease-free when they entered the study.

To collect data for the motivating study, a questionnaire was sent by mail or handed out to a total of 151,246 eligible women who had received no diagnosis of breast cancer at the time of entry, and about 60% of them completed it and returned it at their upcoming screening examination, or through postal delivery. The date when a woman filled out the questionnaire (which included the informed consent form) marked her date of entry into the study. Study entry dates range from January 1, 2003 to December 31, 2007.

The subjects' health trajectories were obtained from administrative data collected by the Italian National Health Service and from the Cancer Registry database. Follow-up ended when an invasive cancer diagnosis occurred or, for women without an observed diagnosis, when censoring occurred. The censoring date coincides with the earliest among date of administrative censoring (December 31, 2016), date of cancellation from the study, date of emigration, and date of death. The median follow-up was 12.29 years.

The available data also include the date of birth, the timing of the screening examinations (either mammograms or ultrasounds, which we treat equally) that were performed, and the dates of the outside-screening examinations and of the diagnoses (invasive tumors only). Due to lack of permission to obtain such information, the data did not include the individual examination results, and we had to infer whether each examination likely gave a positive or negative result based on some assumptions. Different assumptions may lead to different conclusions, and our analysis were therefore repeated under several scenarios. Even when changing the assumptions for the reconstruction of the examination outcomes and for the dates and kinds of detections, the results did not show considerable change.

Below we present the results obtained under what seemed to be the most plausible set of assumptions, also after discussion with an investigator who is familiar with the data. In the Supplemental Material, we include results produced under one different set of assumptions.

For women without an observed diagnosis of breast cancer, we assumed that all the examinations had given a negative result. For women having a breast cancer diagnosis recorded in the Cancer Registry, we had to determine whether the detection was symptomatic or asymptomatic, and to establish the date of the last negative examination before detection. A key piece of information was available from the variable which differentiated between in-screening and out-of-screening examinations. Indeed, out-of-screening examinations may be due to suspicious symptoms. We first checked if there were any screening examinations within the six months prior to the diagnosis. If yes, then the last one before the diagnosis was assumed to have yielded a positive result, and to have led to an asymptomatic detection. In this case the date of detection was defined as the date of that positive exam.

If, instead, there were no screening exams within 6 months of the date of diagnosis, we classified that detection as symptomatic, and we set the date of detection equal to the date of the most recent out-of-screening exam, if there were any within the 6 months prior to diagnosis. If no exams at all were recorded in the 6 months prior to diagnosis, then we set the date of the symptomatic detection back by a number of days equal to the average shift applied to the symptomatic detections which had that information (42.6 days).

Once the dates of detection were defined, we picked the last negative exam as the most recent exam performed at least 6 months before the detection. We decided to impose a distance of at least 6 months between the last negative exam and the detection because most diagnoses are preceded by a few examinations very close to each other, and those were likely performed to confirm the presence of the tumor.

These limitations of the available data are such that the results of our analyses should be taken with some caution (for example, no sensitivity/specificity of the examinations can be taken into account). However, also given the large sample size, we feel that they still provide useful information in particular on the effect of covariates, and most importantly these analyses let one explore the issues that one must address when developing and estimating disease history models from administrative data.

Out of the 78,051 women in the sample, 3034 (3.89%) were diagnosed with invasive breast cancer during the observation period and 75,017 (96.11%) were without diagnosis at the end of their follow-up. The total number of women who died after breast cancer diagnosis is 380 (12.5%) but here we only studied detection. We do not consider DCIS (ductal carcinoma in situ) cases, which were not included in the Cancer Registry database. Under the assumptions described above, the asymptomatic detections were 572 and the symptomatic ones 2462. The total number of exams was 396,183, performed on 74,345 women. The remaining 3706 women did not undergo any examination during the observation period. For additional descriptive statistics we refer to Table 1.

TABLE 1 Descriptive statistics of the data.

	Min	Median	Mean	Max
Age at questionnaire	41.30	60.91	60.82	76.85
Age at first exam after entry	41.37	61.02	60.80	84.64
Age at asymptomatic detection	45.05	64.93	64.18	76.23
Age at symptomatic detection	46.40	67.74	67.34	86.35
# screening examinations		**Mammographies**	**Ultrasounds**	**Total**
	0	12,215 (0.16)	75,694 (0.970)	12,213 (0.16)
	1	21,452 (0.28)	2219 (0.028)	21,412 (0.27)
	2	17,679 (0.23)	131 (0.002)	17,459 (0.22)
	3	12,612 (0.16)	7 (0.000)	12,303 (0.16)
	≥ 4	14,003 (0.18)	0	14,664 (0.19)
# outside-screening examinations		**Mammographies**	**Ultrasounds**	**Total**
	0	8812 (0.11)	62,609 (0.80)	8256 (0.11)
	1	10,903 (0.14)	7866 (0.10)	10,283 (0.13)
	2	26,900 (0.34)	2699 (0.03)	25,063 (0.32)
	3	18,451 (0.24)	1466 (0.02)	17,239 (0.22)
	≥ 4	12,985 (0.17)	3411 (0.04)	17,210 (0.22)
Breast cancer diagnoses	**Yes**	**No**		
	3034 (0.04)	75,017 (0.96)		
Observed follow-up	**Median**	**Mean**	**Min**	**Max**
	12.29	11.66	0	13.93
Status at end of follow-up (only for non-diseased subjects)	**Alive**	**Cancelled**	**Dead**	**Emigrated**
	65,494 (0.873)	232 (0.003)	7410 (0.099)	1881 (0.025)
	No	**Yes**	**Missing***	
At least one birth (X_1)	11,933	63,935	2183	
High level of education (X_2)	47,315	29,994	742	
Family history of cancer (X_3)	47,419	30,562	70	

Note: Time is measured from birth (in years).

*There were 2845 subjects with one or more of these covariates missing.

Additional variables, including level of education, comorbidities, family structure and family history of cancer, were collected by means of a questionnaire filled by the participants. We focused on three dichotomous covariates, which divide the women in eight groups as shown in Table 2: having had at least one birth X_1 (0 = no, 1 = yes); level of education X_2 (0 = low, 1 = high); and family history of cancer X_3 (0 = no, 1 = yes). These are indeed the three non-race main risk factors for breast cancer (among women with no previous history of the disease).[17] In addition to these, we have included education as a proxy for lifestyle-related factors. Genetic factors and breast density are also often discussed as relevant, but that information was not available to us. The effect of comorbidities is not as established, also due to the large number of comorbidities that occur in the population. Our choice of risk factors also allowed us to maintain the dimensionality of the model tractable, without creating groups that include too few subjects. Table 2 also shows the number of asymptomatic and symptomatic detections and the median age at detection, both in the total sample and within the eight covariate groups. Single imputation of missing values was performed on the three covariates by replacing them with draws from independent Bernoulli variables with parameters equal to the proportion of ones among the non-missing values for each variable.

TABLE 2 Observed outcomes in each covariate group and in the total sample.

Group	(x_1, x_2, x_3)	Size	Dx (%)	#Asymp Dx	#Symp Dx	% Symp Dx Among all Dx	Median age Asymp Dx	Median age Symp Dx
1	(0,0,0)	3377	142 (4.2%)	27	115	81%	65.49	69.10
2	(0,0,1)	2107	115 (5.5%)	31	84	73%	66.24	66.80
3	(0,1,0)	3430	154 (4.5%)	24	130	84%	59.36	64.99
4	(0,1,1)	3354	153 (4.6%)	23	130	85%	63.70	64.54
5	(1,0,0)	27,964	939 (3.4%)	183	756	81%	65.67	69.69
6	(1,0,1)	14,338	599 (4.2%)	109	490	82%	65.83	68.54
7	(1,1,0)	12,694	479 (3.8%)	98	381	80%	62.95	66.60
8	(1,1,1)	10,787	453 (4.2%)	77	376	83%	62.78	64.45
Total		78,051	3034 (3.9%)	572	2462	81%	64.93	67.74

Note: Ages are measured in years. X_1 = at least one birth (0: No, 1: Yes); X_2 = Education level (0: Low/Medium, 1: High); X_3 = Family history of cancer (0: No, 1: Yes).

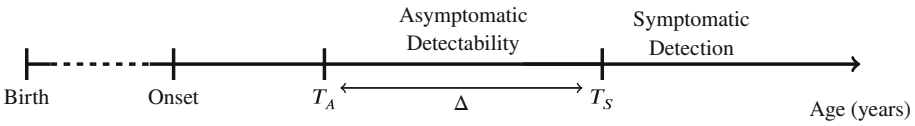


FIGURE 1 A graphical representation of the natural history from onset until detectability of the disease.

Note that the three covariates were assessed at the time of entry into the motivating study. However, given the rather advanced age at entry, we may consider the first two as being definitively measured at that time. On the other hand, family history is still potentially evolving (we will study that specific issue in a separate manuscript). In our models we treat these covariates as baseline covariates that summarize the life-long effect of parity, education and family history on breast cancer development and evolution.

We now turn to the description of a first, treatable model.

3 | A FIRST MODEL: CONSTRUCTION OF THE OBSERVED DATA LIKELIHOOD

All times are measured from birth of the woman. We assume that after the onset of the disease (which may or may not occur) there is a time interval in which not even a screening examination is able to detect the presence of the disease (see Figure 1). The two main quantities of interest are the time (from birth) to the start of asymptomatic detectability of the disease (which we denote by T_A) and the time to the symptomatic detection of the disease (denoted by T_S). At time T_A the disease becomes detectable through screening. Between time T_A and T_S the tumor can only be detected through screening (the "sojourn time," denoted by Δ), while at time T_S the disease becomes evident because of symptoms. In other words we have $T_S = T_A + \Delta$. Further, we assume that symptomatic detection occurs exactly when the first symptoms appear.

While studying the latent evolution of the disease, we are also interested in studying the probability of insurgence of the disease in a woman's lifetime. To allow for the direct estimation of such probability, we introduce a cure rate structure, that is, a proportion of women, which we call the "cured proportion", denoted by (1 − p) with $p \in (0, 1)$, who will never experience the event of developing breast cancer. This is equivalent to assigning positive probability $(1 - p)$ to the event $\{T_A = +\infty, T_S = +\infty\}$, where the probability p is therefore one of the parameters of the model. The standard terminology "cure" is however confusing in this context, so we will instead refer to the fraction p of women who will develop the disease as to the "susceptible" proportion, and to such women as "cases". Note that these should be considered to be latent cases and not observed cases.

We work under the stable disease population assumption, in which the rate of births and the distribution of ages at tumor onset are constant across calendar time.[11] We also assume stationarity of the joint distribution of (T_A, Δ) across birth cohorts.

Note that the goal is to draw conclusions about quantities that are mostly unobservable. Indeed, both T_A and T_S are never observed on any woman, and clearly the observed data would not be a good representation of the latent variables of interest. First of all, the time to the start of the asymptomatic detectability T_A is always interval censored. That is, even when we observe an asymptomatic detection, we never observe T_A precisely but we can only conclude that it happened before the observed age at detection. Second, there is a selection of women who enter the study (and the sample), since women who have already had a breast cancer diagnosis before entering the screening program are excluded from the sample. Third, once a woman has entered the study, she is not typically followed until her death, but follow-up lasts around 12 years when the trajectory is right censored. Therefore, we do not have any information about tumors with onset, or that will be detected, later on.

Note the relationship of these latent quantities with the observed data: the mean of the ages at *observed* symptomatic detection in the sample should be smaller than the expected value of T_S in the population, due to selection into the set of the observed T_S ages; indeed, subjects with larger sojourn time Δ (eg, T_S) are less likely to have their T_S value observed (since asymptomatic detection is more likely). Hence the distribution of the observed ages at detection, asymptomatic or symptomatic, would clearly not represent a good estimate of the underlying disease history, and the proportion of observed diagnoses out of the total may be very different from the probability of ever developing breast cancer.

When defining a model, there are basically three decisions to make that characterize its structure. The first one is the choice of the marginal distributions for T_A and Δ for the diseased subjects. Any distribution having support on the non-negative real line may work, but even distributions on the real line could be appropriate under some specific parameter combinations that make the negative tail negligible.

The second assumption concerns the dependence structure between T_A and Δ. While modeling them as independent random variables may facilitate the form of the likelihood function and the estimation of the model parameters, such assumption may be too simplistic and not reflect the link between these two quantities that has been documented in the literature.[18]

Lastly, one should decide on how to include covariates (and which ones) in the model, both as modifiers of the joint distribution of (T_A, Δ) and of the probability p.

In principle, it is possible to compute the observed data likelihood, and obtain the maximum likelihood estimates for the parameters. However, the calculation and maximization of the observed data likelihood can be complicated or not feasible, especially when the number of parameters grows. Indeed, such observed data likelihood involves many (bivariate) integrals which may not be solvable in closed form, but may need to be approximated numerically–thus introducing numerical difficulties in the estimation process.

Indeed, as we have seen above, each screening examination provides some information about the value of T_A, which is necessarily interval censored. On the other hand, T_S is either observed precisely in the case of symptomatic (outside-screening) detections, or we only have partial information on it. Integrating the joint probability density of (T_A, T_S), denoted by $f_{(T_A,T_S)}(t_a, t_s)$, on an appropriate subset of the domain as determined by the observed events, provides the observed data likelihood contribution, which we denote by L_i, for a generic i-th subject.

Importantly, we condition on the observed mammography/ultrasound exams. Depending on the presence of a positive or negative exam, diagnosis and/or right censoring, one can observe different types of data configurations: cases with a symptomatic detection, cases with an asymptomatic detection and cases without an observed diagnosis. These three kinds of configurations contribute to the observed data likelihood in different ways (Recall that we are assuming perfect sensitivity and specificity of the examinations.).

For a subject with an observed symptomatic detection, T_S is fixed at the observed value t_s and one should integrate the joint density function over all possible values of T_A. The lower bound of the integral (l) is the last negative examination if there is one, or the lower bound of the support otherwise. Note that, clearly, $T_A < T_S$ with probability one (since $\Delta > 0$). Thus, the contribution of such configuration to the observed data likelihood is

$$L_i = p \cdot \int_l^\infty f_{T_A,T_S|\text{case}}(u, t_s)du.$$

For a subject with an observed asymptomatic detection, T_S is greater than the last observed exam (denoted by d since it coincides with the date of detection) and T_A is necessarily between l, the last negative exam if there is one, and the detection time d. This defines the integration region for this kind of trajectories:

$$L_i = p \cdot \int_l^d \int_d^\infty f_{T_A,T_S|\text{case}}(u,v)dvdu.$$

Lastly, a subject who has not developed the disease (yet) may experience breast cancer after the last negative exam or the end of follow-up (with probability p), or never experience it (with probability $(1-p)$). In the first case, the likelihood contribution L_i is obtained by integrating the joint density of $(T_A, T_S)|$case over all values of T_A greater than the last negative exam l and over values of T_S greater than the age at the end of follow-up c. In the second case, the result of the analogous integration is 1 since the conditional distribution of $(T_A, T_S)|$non-case is concentrated on $\{T_A = +\infty, T_S = +\infty\}$. Since these two events are disjoint, the total contribution to the likelihood is the sum of their probabilities:

$$L_i = (1-p) + p \cdot \int_l^\infty \int_c^\infty f_{T_A,T_S|\text{case}}(u,v)dvdu.$$

Lastly, all likelihood contributions should take into account the fact that only asymptomatic women can enter the study, i.e.the distributions of the quantities of interest should all be conditional on the event $\{T_S > \text{Age at entry}\}$: each likelihood contribution L_i should be divided by the probability of the conditioning event

$$\text{c}_i = P(T_S > \text{Age at entry}|\text{Age at entry}) = (1-p) + p\ P(T_S > \text{Age at entry}|\text{Case, Age at entry}).$$

Note how this expression is also based on the assumptions that once T_S is reached, a symptomatic detection (and diagnosis) is immediately observed. While this is exactly not the case, we believe that it is most consistent with the study entry requirement. Notably, the condition does *not* require that $T_A >$ Age at entry.

The observed data likelihood is then given by the product of all the (independent) subjects' contributions: $L = \prod_{i=1}^n \frac{L_i}{\text{c}_i}$.

Even if not explicitly indicated in the notation above, L is clearly a function of all the model parameters. For numerical maximization (and for estimate ion of the variance-covariance matrix of the MLEs), it is more convenient to work with the log-likelihood $l = \sum_{i=1}^n log(L_i) - \sum_{i=1}^n log(\text{c}_i)$.

The degree of difficulty of calculating the two-dimensional integrals which form the observed data likelihood function varies greatly according to the specific distributional assumptions. Only specific model formulations lead to analytical or partially analytical solutions. In the following section we describe one such simple model.

3.1 | Model specification and results

We consider a simple model that assumes independence between T_A and Δ and does not include any covariates. In particular, for the cases, we assume $T_A \sim N(\mu, \sigma^2)$ and $\Delta \sim Exp(\lambda)$, with Δ independent of T_A, where $\mu \in \mathbb{R}$, $\sigma > 0$ and $\lambda > 0$. Easily, $T_S = T_A + \Delta$ has density $f_{T_S}(t) = \lambda\, e^{\frac{\lambda^2\sigma^2}{2} + \lambda(\mu - t)}\, \Phi(t, \mu + \lambda\sigma^2, \sigma^2)$, where $\Phi(\cdot, \mu + \lambda\sigma^2, \sigma^2)$ is the cdf of a $N(\mu + \lambda\sigma^2, \sigma^2)$. Also, the conditional density of $T_S|T_A$ is $f_{T_S|T_A}(v|u) = \lambda e^{-\lambda(v-u)} I_{(u,\infty)}(v)$. Note that, marginally, T_S follows an exponentially modified Gaussian (emg) distribution with parameters (μ, σ, λ). The contributions to the observed data likelihood are as follows (for the derivation please refer to the Supplementary Material).

Using the notation introduced earlier, we have: (i) for a subject with an observed symptomatic detection

$$L_i = p \cdot f_{T_S}(t_S) \cdot \left(1 - \frac{\Phi(l, \mu + \lambda\sigma^2, \sigma^2)}{\Phi(t_s, \mu + \lambda\sigma^2, \sigma^2)}\right);$$

(ii) for a subject with an observed asymptomatic detection

$$L_i = p \cdot e^{\frac{\lambda^2\sigma^2}{2} + \lambda(\mu - d)} \cdot \left(\Phi(d, \mu + \lambda\sigma^2, \sigma^2) - \Phi(l, \mu + \lambda\sigma^2, \sigma^2)\right);$$

TABLE 3 MLEs and 95% confidence intervals for the model parameters.

T_A		Δ	Cure rate
μ	σ	λ	p
64.9 (64.5, 65.3)	22.3 (17.6, 27.0)	1.62 (1.51, 1.73)	0.179 (0.172, 0.186)

Note: Time is measured in years.

and (iii) for a subject without observed diagnosis

$$L_i = (1-p) + p \cdot \left(\frac{f_{T_S}(c) - e^{\lambda(l-c)} f_{T_S}(l)}{\lambda} + 1 - \Phi(c, \mu, \sigma^2) \right).$$

The probability of the conditioning event $\{T_S >$ Age at entry|Age at entry$\}$ is equal to

$$c_i = (1-p) + p \cdot \left(1 - \Phi(f, \mu, \sigma^2) + \frac{f_{T_S}(f)}{\lambda} \right).$$

Table 3 shows the estimates obtained from the maximization of the observed data likelihood with respect to the four model parameters $(\mu, \sigma, \lambda, p)$. The likelihood maximization is performed using the `R` function `maxLik`.[19] For the maximization, we reparameterized all models in such a way that the resulting parameter space becomes the whole $\mathbb{R}^4$, that is, with no constraints. In particular, we applied a logarithmic transformation to all parameters with a positivity constraint, while for the parameter p, constrained to take values in the interval $[0, 1]$, we used a logistic reparametrization. Relying on invariance of maximum likelihood estimators one then obtains the estimates for the original parameters. Application of the delta method (details not shown) allows one to then compute their standard errors. As noted in the Discussion Section, estimation of the standard errors for such models (applied to very partially observed data) requires care due to instability in the estimation of the Hessian matrix. We decided against adding further details on this specific implementation of the delta method, as not to give too much importance to this first, simple model.

The estimated latent proportion of women experiencing the disease in their lifetime is around 18% (recall that our model does not impose any constraint on the upper bound of the subjects' lifespan). One may compare such rate to the estimated lifetime risk of breast cancer, that has been estimated as being one out of eight, or 12.5%.[20] As expected, although the observed proportion of diagnoses in the sample was around 4%, the model reconstructs the frequency of many more lifetime diagnoses than those observed during the limited follow-up of the subjects in the study.

The start of the asymptomatic detectability is on average close to the age of 65 years, ranging between 20 and 110 with 95% estimated frequency. The numbers 20 and 110 are the values taken by the two (consistent) estimators for the percentiles 2.5% and 95% of the normal distribution of T_A, where consistency clearly follows from the continuous mapping theorem applied to the MLEs. Note that this is a wide interval; in Section 4.3 we will see that including some covariates in the model will have the effect of reducing such marginal variability of T_A.

Somewhat surprisingly, the model suggests that the sojourn time Δ is quite short, lasting on average 7–8 months, with an exponential tail. This result is different from current estimates from previous studies, which suggest a mean sojourn time between 2 and 7 years.[21] The exclusion of DCIS cases from our analysis is very likely a factor that contributes to obtain shorter estimates for Δ.[21] However, we believe that the main reason for such small estimate for the sojourn time is possibly the lack of detailed information on the examination results and on the kind of detection from our data (see our comment on this in Section 2). Indeed, we should also recall that T_A has been defined here starting from the dates of the observed diagnoses, and that it is defined as the time when detectability starts. Thus a shorter sojourn time may be compatible with an over-estimation of T_A.

We now move to more flexible and informative models, which will require a different likelihood-free inferential procedure.

4 | MORE FLEXIBLE MODELS

4.1 | Approximate Bayesian computation

As we have pointed out, the calculation of the observed data likelihood for latent processes with large amounts of missing data can be challenging even for relatively simple models. In general, every small change to the model requires the observed likelihood function to be constructed and implemented. For example, the inclusion of a dependence structure between T_A and Δ requires solving complicated integrals through numerical approximations that determine loss of accuracy, as well as a significant increase in the difficulty by the optimization algorithms in identifying the maximum likelihood estimates.

An estimation procedure that allowed one to quickly implement several different models would greatly increase the flexibility in modeling. This is possible by implementing a likelihood-free approach, where the observed likelihood function does not need to be calculated explicitly, nor maximized. A likelihood-free approach that seems particularly promising for disease history models is approximate Bayesian computation (ABC).[15]

The first step of ABC consists of setting prior distributions for the model parameters. One then samples a parameter vector from their prior distribution, and generates a dataset from the corresponding model. In the basic version of ABC, if the simulated data are "close enough" to the real data, that parameter combination is retained and included in the sample of parameter values that approximates the posterior distribution of the parameters given the data. Indeed, implementing this procedure a very large number of times (here 200, 000) and selecting only a very small proportion (called tolerance or retention rate) of samples, then allows one to approximate the parameters' posterior distribution.

It is also common to post-process the ABC output to improve the selected posterior sample by applying a so-called "regression adjustment." The idea is to regress each parameter (or to perform a multivariate regression with all the parameters as response vector) on the set of summary statistics and to apply a correction based on the difference between observed and simulated summaries.[22,23]

Measuring the distance between two datasets (observed vs. model-generated) is not trivial: one should use informative summary statistics of the data, which reduce the dimensionality of the data but still retain the information needed to perform accurate inference on the parameters. Indeed, only by using *sufficient* statistics and by conditioning on the event that their values are identical (and not just close) in the observed data and in the model generated data, one would ensure that the sample of retained parameter values represents a sample from their exact posterior distribution.[23] More recently, various approaches for efficiently comparing observed and generated data without defining summary statistics have been explored.[24]

There is a vast literature on the choice of summary statistics in ABC, and a variety of approaches have been proposed.[25] Most of the methods, however, do not propose any constructive procedure, but only suggest techniques to select a subset of summary statistics among a bigger set of proposals (subset selection methods) or to combine them to reduce the dimensionality (projection methods).

In our models we include the three binary covariates described in Section 2, which partition the subjects into eight groups, and consider a set of the same summary statistics computed on each of the eight groups. In particular, we build "Metric 1" to measure the dissimilarity between the observed and a model-generated dataset, based on a total of 32 summary statistics (4 for each of the eight groups of women): proportion of observed detections, proportion of observed symptomatic detections among the total number of observed detections, median age at observed asymptomatic detection, and median age at observed symptomatic detection. The distance between the two datasets is then defined as the L^2-distance between the standardized summary statistics of the two datasets. The standardization is performed by dividing each summary statistic by a robust estimate of its standard deviation (the median absolute deviation).

"Metric 2" refines "Metric 1" by also considering the entire distribution of the observed age at detection. This metric makes use of the classical test statistic for the comparison of two proportions and of the Kolmogorov-Smirnov test statistic to assess if two observed samples can be considered to be generated by the same underlying distribution. We perform the first 16 tests to compare the proportions of observed detections and of observed symptomatic detections in each of the eight covariate groups. Then, we perform 16 additional tests to compare the distributions of the age at asymptomatic and symptomatic detection, again for each group. We believe that the test statistics, or the corresponding p-values, could provide a good measure of the distance between two objects (two proportions or two distributions, depending on the test). There are many ways to combine the test outputs (test statistics or p-values) into a distance function between the two datasets. In the Supplemental Material we briefly explore the relative performance of "Metric 1" and a version of "Metric 2" on two simulated datasets, and "Metric 1" seems to produce estimates of the parameter values which are closer to the true values.

Hence, in Section 4.3 we present the results obtained by using "Metric 1", while fitting different models to the motivating data. The retention (tolerance) rate is chosen through a leave-one-out cross-validation procedure, which is implemented and available in the `R` package `abc`.[26] We make use of local linear regression to correct the posterior samples by regression adjustment.

In the simulation process, we generate the screening examinations with the same planned schedule of the real screening program, and assuming a constant adherence rate of 0.6 to the prescribed examinations.[27] Hence, the screening parameters are fixed, and not object of inference. For the subjects belonging to the susceptible proportion, the disease history is then superimposed to the performed examinations to produce the observed age at detection (if it happens inside the interval of follow-up), the detection mode (symptomatic or asymptomatic), and the age at last negative examination. For the non-cases, we identify the age at their last negative examination, if there is one, before the end of the follow-up. We thus obtain a dataset containing information that has structure similar to that of the observed data.

To make the simulated data as comparable as possible to the observed ones, we keep approximately the same distribution for the covariates. The approximation comes from the fact that one needs to generate a slightly larger sample of women because some of them will experience a symptomatic detection of the disease before the age at entry in the study, and therefore will be excluded from the effective sample. Through some simulations, we estimated this proportion to be roughly 4% of women, so for each simulated sample we generated $78{,}051/0.96 \approx 81{,}305$ women. We assign the 78,051 observed covariate vectors to the first 78,051 women in the simulated sample, and take a random sample of the covariate vectors for the remaining $81{,}305 - 78{,}051 = 3254$ women.

Note that the ABC procedure described above, known as ABC-rejection algorithm, is very computationally demanding since only a small fraction of the generated samples are retained and contribute to the posterior distribution approximation. There exist many refinements of the ABC algorithm, aimed at reducing the inefficiency due to sampling from very uninformative prior distributions by exploiting the information of already accepted parameter values.[28] These refined algorithms could bring a substantial computational gain, but have the main drawback of not being easy parallelizable on multiple cores. Having the possibility to work on a server with many processors, we decided to implement the simpler ABC-rejection procedure (For the implementation of all models we used the software `R`[29] on a server with 176 cores.).

4.2 | Models

Recall the three binary covariates described in Section 2: X_1 ="at least one birth," X_2 ="high level of education," and X_3 ="family history of cancer," all coded as 0 = no and 1 = yes. We posit models such that the susceptible proportion depends on the observed covariates $x = (x_1, x_2, x_3)$ through the logit link:

$$p(x) = \frac{e^{p_0+p_1x_1+p_2x_2+p_3x_3}}{1 + e^{p_0+p_1x_1+p_2x_2+p_3x_3}}.$$

For the cases (those who will eventually develop the disease) subjects, the evolution of the disease is described by the time to its asymptomatic detectability T_A and by its sojourn time Δ. We let the mean of T_A depend on the covariates linearly, while the variance of T_A is assumed constant across covariate groups.

The distribution of Δ is then defined conditionally on the observed value of T_A, and it may reflect the effect of the covariates but only indirectly (see below). Note that any form of dependence between T_A and Δ is easily manageable through ABC, since the simulated value of T_A is already available when one generates the value of Δ from the distribution of $\Delta|T_A$.

We have exploited the flexibility of ABC by exploring several different models. We do not report all details, such as the prior distributions, for all of them here. Parameters associated with covariates had uninformative prior distributions centered at zero. The prior distribution for the mean of T_A in the baseline group, denoted with β_0, was chosen to be N(65, 10): indeed, from the literature and from the simple model in Section 3.1 (see MLEs in Table 3), we expect a mean of 65 to be reasonable[30] but we still keep a variance large enough to let the data bring in relevant information on β_0. Similarly, p_0 represents the proportion of women who will develop the disease in the baseline group and we assign to it a rather informative prior: $p_0 \sim \text{logit}\,(\text{Beta}(3,\ 21))$ around the lifetime risk of 1 in eight that has been repeatedly suggested as a possible consensus value in the literature.[31] Indeed, the prior distribution corresponds to a woman in the baseline group has on average a probability of $3/(3 + 21) = 0.125$ of belonging to the diseased (non-cured) group.

Here below is the list of ten models. The number of parameters to be estimated, indicated below between square brackets, is always equal to 4, for the cured proportion regression, plus 7 or 8 for the disease history.

1. **Normal + Exponential** [4 + 7 = 11 parameters]

$$T_A \mid \beta_0, ..., \beta_3, \sigma \sim \mathrm{N}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,\ \sigma^2);$$
$$\Delta \mid \{T_A = t_A\}, \gamma_0, \gamma_1 \sim \mathrm{Exp}(e^{\gamma_0 + \gamma_1 t_A}).$$

2. **Normal + Exponential (log-scale)** [4 + 7 = 11 parameters]

$$T_A \mid \beta_0, ..., \beta_3, \sigma \sim logN\left(\mu = m\left(\frac{s^2}{m^2} + 1\right)^{-\frac{1}{2}},\ \sigma^2 = \log\left(\frac{s^2}{m^2} + 1\right)\right);$$
$$\tilde{\Delta} = log(T_S) - log(T_A) \mid \{T_A = t_A\}, \gamma_0, \gamma_1 \sim \mathrm{Exp}(e^{\gamma_0 + \gamma_1 t_A}),$$

where $m = E(T_A) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ and $s^2 = Var(T_A)$. This parameterization is used to let the variance of T_A (in the original scale) be independent of covariates, that is, the same across groups.

3. **Bivariate normal** [4 + 8 = 12 parameters]

$$(T_A, \Delta) \mid \beta_0, ..., \beta_3, \mu_\Delta, \sigma_1, \sigma_2, \rho \sim N_2(\mu, \Sigma),$$

where $\mu = (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,\ \mu_\Delta)$ and $\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$.

4. **Bivariate normal (log-scale)** [4 + 8 = 12 parameters]

Let $\tilde{\Delta} = \log(T_S) - \log(T_A)$. We assume

$$(\log(T_A), \tilde{\Delta}) \mid \beta_0, ..., \beta_3, \mu_\Delta, \sigma_1, \sigma_2, \rho \sim N_2(\mu, \Sigma),$$

where $\mu = \left(m\left(\frac{s^2}{m^2} + 1\right)^{-\frac{1}{2}},\ \mu_\Delta\right)$ and $\Sigma = \begin{bmatrix} \log\left(\frac{s^2}{m^2} + 1\right) & \rho \log\left(\frac{s^2}{m^2} + 1\right)^{\frac{1}{2}} \sigma_2 \\ \rho \log\left(\frac{s^2}{m^2} + 1\right)^{\frac{1}{2}} \sigma_2 & \sigma_2^2 \end{bmatrix}$.

As in all the previous models, again here $m = E(T_A) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ depends on the covariates, while $s^2 = Var(T_A)$ does not.

5. **Gamma + Weibull** [4 + 8 = 12 parameters]

$$T_A \mid \beta_0, ..., \beta_3, \sigma \sim \mathrm{Gamma}\left(\frac{(\mu(x))^2}{\sigma^2},\ \frac{\mu(x)}{\sigma^2}\right);$$
$$\Delta \mid \{T_A = t_A\}, \gamma_0, \gamma_1, k \sim \mathrm{Weibull}\left(\lambda(t_A), k\right),$$

where $E(T_A) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ and $\lambda(t_A) = e^{\gamma_0 + \gamma_1 t_A}$ and k has a prior distribution that includes one (corresponding to the exponential case).

6. **Gamma + piecewise exponential** [4 + 8 = 12 parameters]

$$T_A \mid \beta_0, ..., \beta_3, \sigma \sim \mathrm{Gamma}\left(\frac{(\mu(x))^2}{\sigma^2},\ \frac{\mu(x)}{\sigma^2}\right);$$
$$\Delta \mid \{T_A = t_A\}, \lambda_1, \lambda_2, \lambda_3 \sim \mathrm{Exp}(\lambda_1 \cdot 1(t_A \le 55) + \lambda_2 \cdot 1(55 < t_A \le 65) + \lambda_3 \cdot 1(t_A > 65)),$$

where $E(T_A) = \mu(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$.

7. **Rescaled Beta + Exp** [4 + 7 = 11 parameters]

$$T_A \mid \beta_0, ..., \beta_3, \sigma \sim 100 \cdot \mathrm{Beta}(\alpha,\ \beta);$$
$$\Delta \mid \{T_A = t_A\}, \gamma_0, \gamma_1 \sim \mathrm{Exp}(e^{\gamma_0 + \gamma_1 t_A}),$$

where $E(T_A) = 100 \cdot \frac{\alpha}{\alpha+\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ and $\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

8. **Rescaled Beta + Weibull** [4 + 8 = 12 parameters]

$$T_A \mid \beta_0, ..., \beta_3, \sigma \sim 100 \cdot \text{Beta}(\alpha, \beta);$$
$$\Delta \mid \{T_A = t_A\}, \gamma_0, \gamma_1, k \sim \text{Weibull}\,(\lambda(t_A), k),$$

where $E(T_A) = 100 \cdot \frac{\alpha}{\alpha+\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, $\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$, and $\lambda(t_A) = e^{\gamma_0+\gamma_1 t_A}$. Here, too, k has a prior distributions that includes one.

9. **Rescaled Beta + piecewise exponential** [4 + 8 = 12 parameters]

$$T_A \mid \beta_0, ..., \beta_3, \sigma \sim 100 \cdot \text{Beta}(\alpha, \beta);$$
$$\Delta \mid \{T_A = t_A\}, \lambda_1, \lambda_2, \lambda_3 \sim \text{Exp}(\lambda_1 \cdot 1(t_A \le 55) + \lambda_2 \cdot 1(55 < t_A \le 65) + \lambda_3 \cdot 1(t_A > 65)),$$

where $E(T_A) = 100 \cdot \frac{\alpha}{\alpha+\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, $\sigma^2 = 100^2 \cdot \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

10. **Normal + piecewise exponential** [4 + 8 = 12 parameters]

$$T_A \mid \beta_0, ..., \beta_3, \sigma \sim \text{N}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3, \sigma^2);$$
$$\Delta \mid \{T_A = t_A\}, \lambda_1, \lambda_2, \lambda_3 \sim \text{Exp}(\lambda_1 \cdot 1(t_A \le 55) + \lambda_2 \cdot 1(55 < t_A \le 65) + \lambda_3 \cdot 1(t_A > 65)).$$

Note that both the normal and the gamma distributions have decreasing densities for older ages (with the gamma density decreasing more slowly, in addition to not imposing symmetry and not allowing for negative values). Note also that very limited data are available for older ages, due to right censoring which also includes death. One may expect the three models based on the rescaled beta density to provide a more realistic shape for the right tail of T_A.

In the next section we discuss the results of the ABC-based model selection procedure to choose among these models.

4.3 | Model selection and results

To select the best model among the ones described above, we simulate 200,000 samples from each model.[15,32] The metric used to quantify the distance between each simulated sample and the observed one is "Metric 1", based, for each covariate-defined stratum, on the proportion of observed detections and of observed symptomatic detections, and on the median age at observed asymptomatic and symptomatic detection (see also Section 4.1). Then, from the pooled set of samples produced by all the models, we select the samples that have the smallest distance from the observed data, keeping track of which model generated each sample. The resulting sample of parameter values and model index can be regarded as a sample from the approximate joint posterior distribution of the parameter and the model index. The number of retained samples generated by a specific model, divided by the total number of retained samples, thus represents an approximation of the posterior probability of that model. For a more detailed description of this procedure.[32]

Since the initial number of samples (200,000) was the same for each model, we are assuming a uniform prior distribution over the ten models. Table 4 contains the numerical values of the approximate posterior probabilities. Model 10, Model 6 and Model 9 clearly show the highest (by far) posterior probabilities (0.225, 0.214, and 0.202).

The ABC model choice procedure described above presents some potential pitfalls.[33] Indeed, as it has been highlighted by Marin et al,[32] in many cases it may even fail to converge to a Dirac distribution on the true model as the size of the observed dataset grows to infinity. In other words, the so-called "curse of insufficiency"[32] is likely to occur, thus leading to arbitrariness in the construction of the Bayes factor (and thus of the posterior probabilities of the models).

TABLE 4 Posterior probabilities of the ten models (global retention rate = 0.005).

Model	1	2	3	4	5	6	7	8	9	10
Posterior probability	0.073	0.076	0.024	0.009	0.052	0.214	0.064	0.061	0.202	0.225

TABLE 5 Counts of votes for the ten models out of a total of 1000 trees composing the random forest.

Model	1	2	3	4	5	6	7	8	9	10
Votes	64	92	8	8	52	182	70	84	252	188

Given these concerns, some alternative techniques to conduct model choice in the context of ABC have been proposed, and we also implement an alternative approach based on random forests.[34] For an introduction to random forests, which are a machine learning tool consisting of the aggregation of simple classifiers (called trees) that can be used both for classification and regression purposes, we refer to chapter 15 of the book by Hastie et al.[35] Model selection through ABC is reformulated as a classification problem, and it is split into two steps.[34] The first step trains a random forest that predicts, for each possible value of the summary statistics, the model that best fits the data. In other words, the random forest is a classifier that associates to each vector of summary statistics a predicted model among the ten proposed. The training set is represented by the pooled set of simulations performed for the ten models. Once the classifier is trained, the predicted model for the set of observed summary statistics represents the selected model, that is, the model that obtained the majority of votes among the classification trees of the random forest. Table 5 shows that, given a trained random forest made of 1000 trees, Model 9 obtained the majority of votes (252) and it is, therefore, the model selected for having the best fit to the observed data.

In the second step[34], the posterior probability of the selected model is computed through a secondary random forest. The binary model prediction errors (Model 9 vs. all the other models) are computed for each observation using the out-of-bag classifiers (see here Reference 35 for the description of out-of-bag classifier in a random forest). This secondary random forest, which is again trained on the pooled set of simulations performed for the ten models, performs a regression of the prediction error on the summary statistics. Lastly, the posterior probability of the selected model is computed as the random forest regression estimate associated to the vector of observed summary statistics. In our case, this procedure resulted in a posterior probability for Model 9 equal to 0.247.

The results from this alternative procedure for model selection disagree slightly with those from the simpler algorithm described at the beginning of this section. However, the two approaches agree on the best three models being Model 9, Model 10, and Model 6.

An assessment on the overall performance of model selection in ABC is difficult since, among other issues: (i) it is based on one very specific model; (ii) it depends on the specific summary measure that one implements; (iii) it depends on the collection of alternative models that one considers. Given the motivation provided in the literature to consider the approach based on random forests more reliable,[32] and the additional simulation study that we performed to assess its ability to discriminate among our proposed models (see Section S3 of the Supplementary Material), we now focus on the results of the ABC estimation procedure for Model 9, the "Rescaled Beta + piecewise Exponential" model.

We assumed the following independent prior distributions for the model parameters: $\beta_0 \sim \text{N}(0.65, 0.05)$, $\beta_i \sim \text{N}(0, 0.25)$, for $i = 1, 2, 3$, $\sigma \sim \text{Unif}(0.02, 0.25)$, $\lambda_i \sim \text{Unif}\,(0.1, 4)$, for $i = 1, 2, 3$, $p_0 \sim \text{logit}\,(\text{Beta}(3, 21))$, $p_i \sim \text{Unif}(-2, 2)$, for $i = 1, 2, 3$.

A retention (or tolerance) rate of 0.02 was chosen via a leave-one-out cross-validation procedure, by comparing the quality of several posterior estimates obtained using different tolerance rates. The posterior distributions shown in Figure 2 are thus based on a sample of $200,\ 000 \times 0.02 = 4000$ selected parameter values. Following a comment by a reviewer, the results reported in the rest of this Section have been slightly refined by applying the post-processing regression adjustment to the transformed parameters to ensure that all posterior values fall within the support of the corresponding prior distributions.

We note that the posterior distributions of the model parameters are much more concentrated than the prior distributions. The only exception is parameter λ_1, whose posterior distribution is still quite flat. This lack of posterior information is probably due to the small number of cases observed among women younger than 55 years old. Table 6 shows the posterior modes and the 95% intervals corresponding to the regions of the approximate posterior distributions that have the highest density (HPD intervals).

Some interesting observations on the effect of the covariates arise from the estimated posterior distributions: (i) women with at least one child tend to have a lower probability of ever experiencing breast cancer, and a later T_A if they do (posterior distributions for p_1 and β_1); (ii) having a family history of cancer has the opposite effects, according to the posterior distributions for p_3 and β_3; (iii) women with a high level of education experience breast cancer earlier than women with a

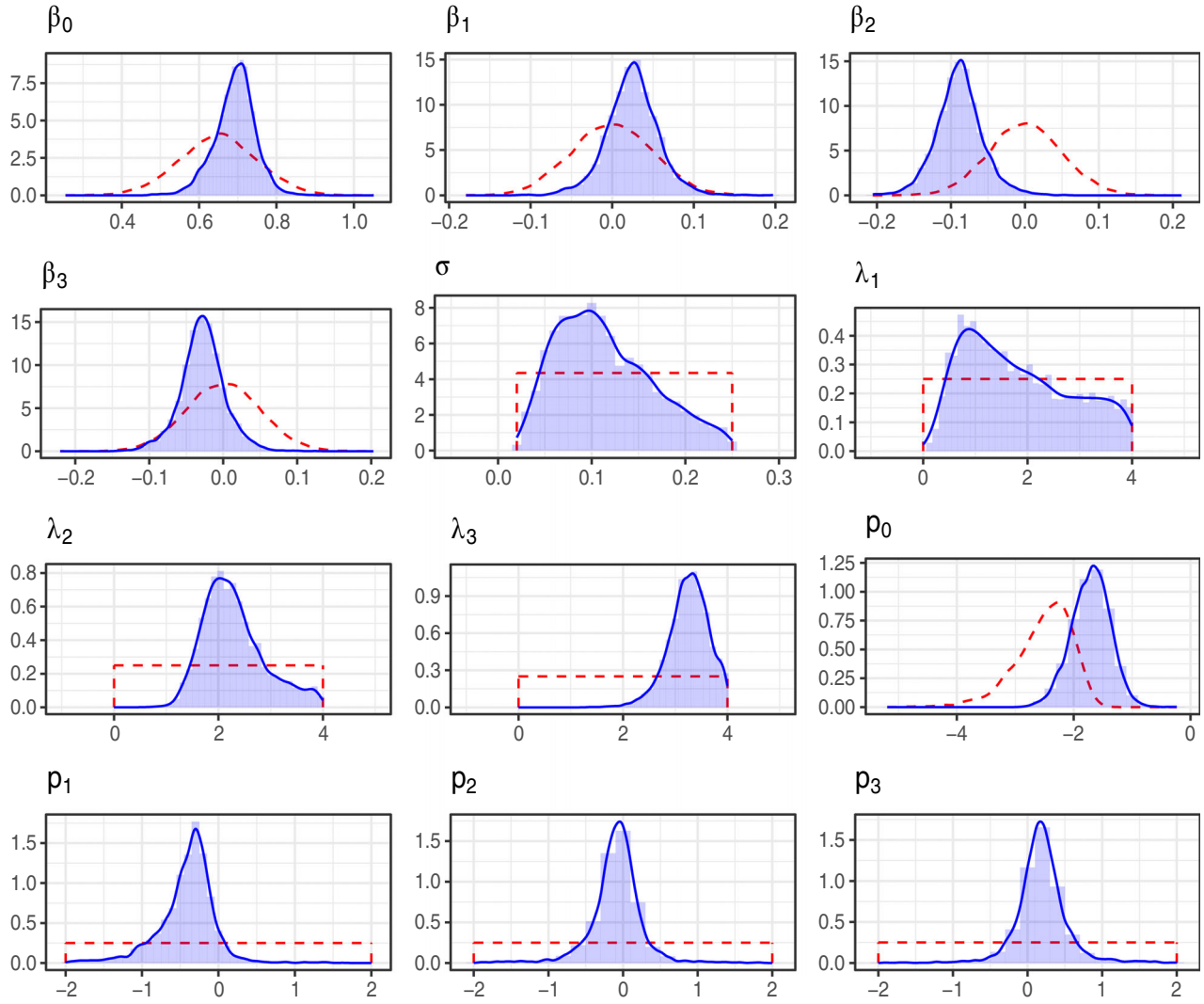


FIGURE 2 Prior (red dashed line) and local linear regression adjusted approximate posterior (blue histogram and solid line) densities for each parameter of the "Rescaled Beta + piecewise exponential" model.

TABLE 6 Posterior modes and the 95% highest posterior density (HPD) intervals.

Parameter	β_0	β_1	β_2	β_3	σ	λ_1
Mode	0.708	0.027	−0.086	−0.028	0.097	0.877
HPDI	(0.586, 0.7490)	(−0.039, 0.089)	(−0.153, −0.024)	(−0.088, 0.036)	(0.030, 0.219)	(0.291, 3.783)
Parameter	λ_2	λ_3	p_0	p_1	p_2	p_3
Mode	2.023	3.334	−1.661	−0.298	−0.044	0.172
HPDI	(1.240, 3.578)	(2.583, 3.998)	(−2.339, −1.064)	(−1.327, 0.173)	(−0.798, 0.556)	(−0.507, 0.851)

lower education level, but this variable is probably not very relevant in modifying the susceptible proportion p (posterior for p_2 almost symmetric around 0).

To gain a clearer idea on how covariates influence the mean of T_A, which is defined as $\mu(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, we may combine the posterior distributions of $\beta_0, \beta_1, \beta_2$ and β_3 according to the covariate combination of each group (see Table 7). The resulting boxplots are shown in the left panel of Figure 3. We can see that covariates do indeed play an important role in determining $E(T_A)$, whose estimated posterior median ranges from a minimum of 58 to a maximum of 72 years old.

TABLE 7 X_1= At least one birth (0:No, 1:Yes); X_2= Education level (0:Low/Medium, 1:High); X_3= Family history of cancer (0:No, 1:Yes).

Group	X_1	X_2	X_3
1	0	0	0
2	0	0	1
3	0	1	0
4	0	1	1
5	1	0	0
6	1	0	1
7	1	1	0
8	1	1	1

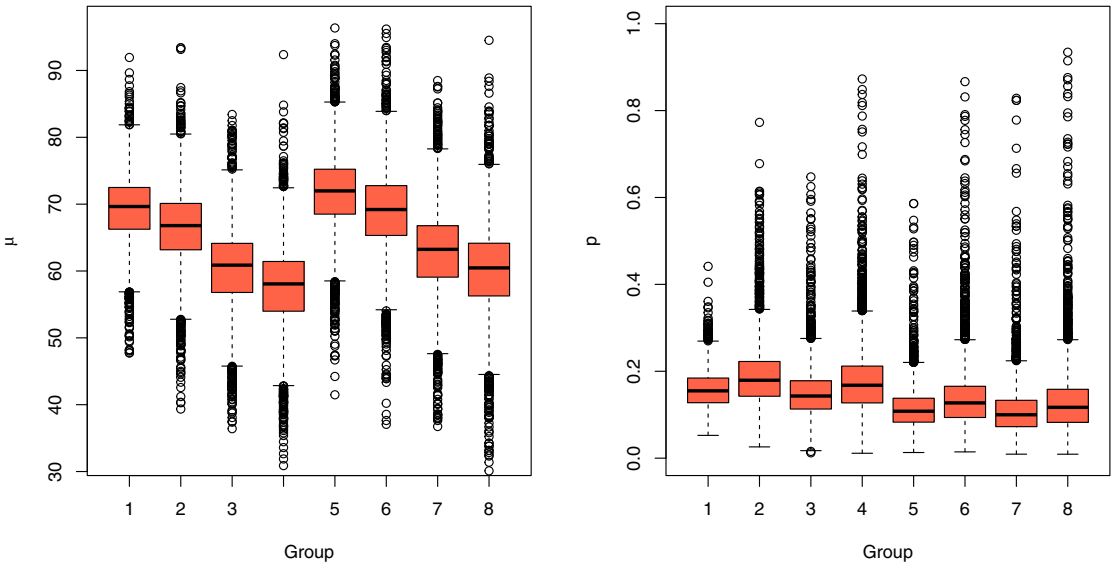


FIGURE 3 Approximate posterior distribution of the mean age at asymptomatic detectability $\mu(x)$ and of the susceptible proportion p(x) across covariate groups.

Similarly, combining the posterior distributions of p_0, p_1, p_2 and p_3, we can compute the posterior distribution of the susceptible proportion p(x) in the eight covariate groups. As we can see in right panel of Figure 3, the probability for a woman of developing breast cancer varies across groups. In particular, its median ranges from a minimum of about 10%–11% for women in groups 5 and 7 (having at least one birth and with no family history of cancer) to a maximum of about 17%–18% for women in groups 2 and 4 (without any birth and with family history of cancer).

Once an approximation of the posterior distribution of the parameters is available, it is also possible to compute approximate predictive distributions for T_A in each covariate group, as well as for Δ given the observed value of T_A. Given a specific covariate configuration, we have a joint posterior sample for the mean and for the standard deviation of T_A, $\{(\mu_i, \sigma_i),\ i = 1, \ldots, 4000\}$. For each couple (μ_i, σ_i), we then draw a value of t_A^i from the model, that is, we generate

$$t_A^i | \mu_i, \sigma_i \overset{ind}{\sim} 100 \cdot \text{Beta}(\mu_i,\ \sigma_i), \quad \text{for } i = 1, \ldots, 4000,$$

where Beta(μ_i, σ_i) denotes a Beta random variable having mean μ_i and variance σ_i^2. The set of generated values $\{t_A^i,\ i = 1, \ldots, 4000\}$ then represents a sample from the ABC approximation of the predictive distribution of T_A in that group.[36]

We can repeat this procedure for each covariate group, obtaining the eight distributions shown by the boxplots in the left-hand side of Figure 4.

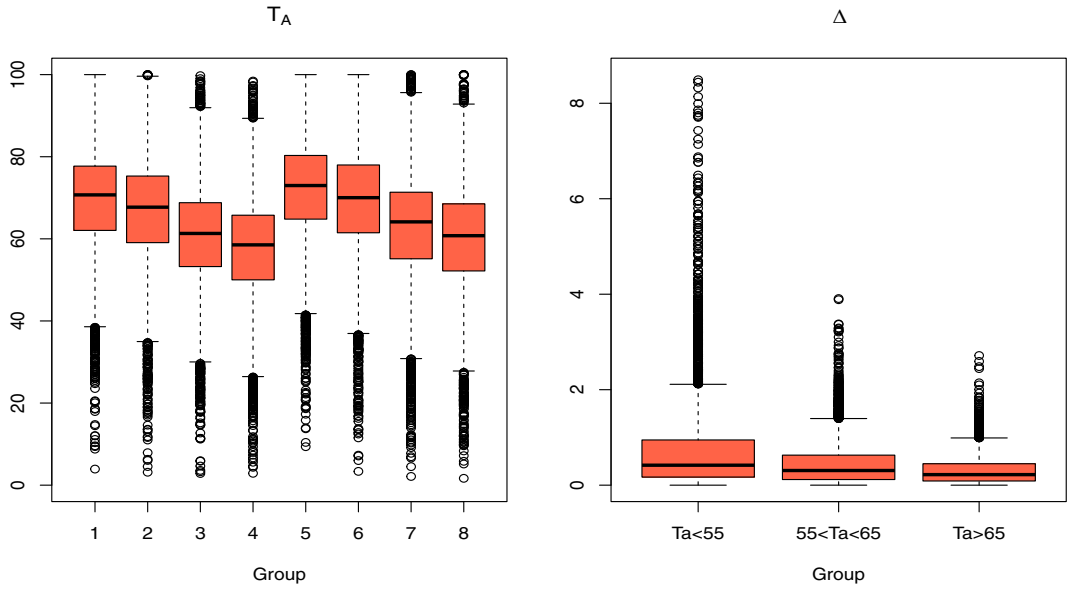


FIGURE 4 Predictive distributions for T_A in each covariate group and for Δ given the observed value of T_A.

Similarly, the posterior sample of size 4000 for λ_1, λ_2, and λ_3 can be used to generate a sample from the approximate predictive distribution of Δ given T_A (see the right-hand side of Figure 4), by using:

$$\delta_1^i | \{T_A \leq 55\}, \lambda_1^i \overset{ind}{\sim} \text{Exp}(\lambda_1^i), \quad \text{for } i = 1, \ldots, 4000;$$
$$\delta_2^i | \{55 < T_A \leq 65\}, \lambda_2^i \overset{ind}{\sim} \text{Exp}(\lambda_2^i), \quad \text{for } i = 1, \ldots, 4000;$$
$$\delta_3^i | \{T_A > 65\}, \lambda_3^i \overset{ind}{\sim} \text{Exp}(\lambda_3^i), \quad \text{for } i = 1, \ldots, 4000.$$

Note that these results suggest that Δ (slightly) decreases when T_A increases, which seems to be in contrast with the medical literature.[18]

Clearly, the predictive distributions of T_A and Δ cannot be directly compared to the observed data. In the Supplementary Material (Section S9) we provide an example where, under simplified assumptions, one can compute the distributions of the observed age at asymptomatic and symptomatic detection analytically. One way to explore the goodness of fit of these models would be to generate data from them and to compare such data to the observed data through some summaries. However, this is exactly how ABC has produced the estimated model parameters, so that the algorithm is indeed already based on a goodness-of-fit maximizing procedure (see also Section 4.4). As an additional validation of the estimated model, we have performed goodness-of-fit for Model 9 in a cross-validation fashion. We partitioned the data into five folds and each time used four of them to estimate the posterior distributions which are then used to generate a sample with the same covariate distribution as in the left-out fold. The resulting summaries, reported in Table 8, show rather small discrepancies between the observed and the simulated data. Thus, the estimated model provides a reasonable fit to the data in terms of the observed summaries.

Relatedly, in the next section we analyze the effect of different screening policies (in terms of observed detections) given the estimated latent disease process.

4.4 | Comparing alternative screening strategies

After estimating the parameters of the models, one can use this information to compare different screening strategies to help identify an optimal screening strategy.

In particular, we now compare several screening strategies, which differ with respect to the gap between consecutive examinations, the proportion of attended examinations out of the total number of invitations (adherence), and the screening age range. In particular, we start from the screening strategy offered in Lombardy (denoted by "Screening strategy 1")

TABLE 8 Summaries obtained from the five-fold cross-validation in the observed and generated data.

% Dx	% Symp Dx	Median age Asymp Dx	Median age Symp Dx
Observed in the left-out fold			
0.0381	0.7882	64.8638	68.7885
0.0400	0.8173	66.8350	68.1780
0.0420	0.7939	64.6242	67.9890
0.0359	0.8128	63.2334	66.0643
0.0384	0.8464	65.0609	67.1622
0.0389	0.8117	64.9235	67.6364
Simulated from posterior model			
0.0300	0.8522	64.1530	68.8799
0.0349	0.8108	64.9741	68.6950
0.0318	0.8245	65.2285	67.4301
0.0329	0.8538	62.8602	67.6026
0.0309	0.8375	63.3131	67.8522
0.0321	0.8358	64.1058	68.0920

Note: The bottom line of each table shows the average of the results from the five folds.

TABLE 9 Observed summary statistics on a sample of size 100,000 generated from the estimated "Rescaled beta + piecewise exponential" model under several different screening strategies.

Screening Strategy	% Dx	% Asymp Dx	Median age Asymp Dx	Median age Symp Dx	Median Lead time
(50–69, 2 years, 60%)	5.45%	15.4%	59.99	62.65	0.370
(50–69, 2 years, 80%)	5.61%	19.2%	59.92	62.39	0.362
(50–74, 2 years, 60%)	7.25%	14.0%	62.33	65.38	0.342
(50–74, 2 years, 80%)	7.37%	18.0%	62.18	65.25	0.326
(50–74, 1 years, 60%)	7.36%	24.7%	62.27	65.50	0.321
(50–74, 1 years, 80%)	7.42%	30.9%	62.62	65.62	0.328

Note: The screening strategies are defined by the screening age range, the gap between subsequent exams, and the overall adherence proportion.

and we measure the effect of varying some of its features on the total number of observed detections during the screening age interval, the percentage of asymptomatic detections, and the median age at observed asymptomatic and symptomatic detection. The underlying assumption (as supported by many studies[3]), is that the moment when a tumor is detected could make a difference on the outcome of the disease. Indeed, detecting the disease earlier rather than when symptoms would have emerged, that is, at a less advanced stage, should allow one to treat it with more success.

The six screening strategies that we have considered are shown in Table 9. All the screening strategies are implemented on a sample of size 100,000 generated from the estimated predictive distributions for the "Rescaled Beta + piecewise Exponential" model. In the simulated samples we assume an administrative follow-up interval that coincides with the screening interval (ie, 50–69 or 50–74 depending on the policy), except for a small proportion of about 5% of the subjects, for whom censoring for other causes occurs earlier.

As expected, reducing the gap between consecutive screening examinations from two years to one year results in an increase in the percentage of asymptomatic detections out of all detections by 72%–76% (from 14.0% to 24.7% or from 18.0% to 30.9%), depending on adherence. Clearly, such an increase would come with a substantial increase in the cost of the program.

Another possibility to increase the percentage of tumors diagnosed before becoming symptomatic would be to increase the adherence to the screening program. From our results we estimate that increasing it from the current level of about 60% to an adherence of 80% would make the proportion of asymptomatic detections increase by 25%–29%. Thus, even without modifying the screening strategy, it seems crucial to find ways to raise the awareness on the importance of breast cancer screening. As adherence likely depends on subjects' covariates and is not constant over time, campaigns to encourage women to attend the screening examinations regularly should target categories of women who tend to adhere less.[37]

Interestingly, intensifying the screening examinations (either by reducing the gap or by increasing the adherence) does not seem to imply a relevant difference on the age at observed asymptomatic and symptomatic detections, but only on the total number of observed diagnoses.

Another observation concerns the effect of extending the end of the screening interval from the age of 69 to the age of 74 years old (this change has been recently implemented in the Lombardy screening program). The total number of tumors detected during the screening period (which is longer) increases by 30%. However, the proportion of asymptomatic detections slightly decreases by 4%–6%. We can explain this result by recalling that tumors at older ages are (slightly) faster in becoming symptomatic according to our model, so screening in the age range 69–74 is less "efficient" (produces slightly fewer asymptomatic detections) than screening at younger ages.

We should also point out that, despite the small values of the (latent) quantity Δ predicted by our model, the difference between the median age at observed asymptomatic and symptomatic detections is around 3 years, similar to the gap observed in the motivating data. Such observed difference seems to be due to the fact that women over 69 (or 74 with the new screening policy) are not screened, and therefore detections that occur after that age can only be symptomatic, making the median age at observed symptomatic detection increase.

This also shows, once again, that the data filtered by the partial observation mechanism do not give a clear picture of the underlying latent disease process in absence of a proper inferential model. Indeed, for more details on the results see Table 9.

5 | DISCUSSION

We have analysed several parametric models to describe the natural history of breast cancer, where the main events of interest are the start of asymptomatic detectability of the disease and the time of symptomatic detection (T_A and T_S). The models differ in their parametric assumptions, but they all share a cure rate structure that takes into account that a fraction of the women will never experience the disease. Estimating how long tumors stay in the latent phase between time T_A and time T_S (ie, estimating the sojourn time Δ) is of great importance for planning an efficient screening policy.

We have obtained the distribution of these random quantities by estimating the model parameters from data collected as part of the motivating study. While the results seem to provide useful information, they should be handled with some care given the described lack of some information (and thus their reconstruction) in the available data. At the same time, the exclusion of DCIS cases (not available in the data) from our analysis makes the comparison with other previous studies which include them not immediate.

Depending on the complexity of each model, we have employed a likelihood-based or a likelihood-free estimation procedure. Given the complex missing data structure, it has shown to be very challenging and in most cases not feasible to obtain maximum likelihood estimates for the model parameters. The calculation and the maximization of the observed data likelihood rely on numerical algorithms, and even for relatively simple models they have been found to be computationally unstable. The numerical approximation of the Hessian matrix used to obtain standard errors for the parameters has also been found to be difficult to compute.

On the other hand, approximate Bayesian computation (ABC) allowed us to perform both model selection and parameter estimation without the need to maximize nor calculate explicitly the observed data likelihood function. However, we recall that inference based on ABC is subject to several levels of approximation: (i) the metric chosen to assess the dissimilarity between generated and observed data; (ii) the tolerance for acceptance of a generated parameter value; (iii) the use of Monte Carlo to estimate the posterior distributions; and (iv) the use of post-processing adjustments.[15]

We experimented with two different metrics to evaluate the distance between simulated and observed data and, based on some simulations, we chose one of them. One could try to refine the way of calculating the distance between the two datasets by using different statistics to measure the difference between the distributions of the ages at observed diagnosis. An alternative approach to quantify the distance between the datasets may be to consider the accuracy of a classification method implemented to distinguish between observed and simulated data.[38] Lastly, while this is not a major

concern in our application, the standard regression adjustment may produce samples from the approximate posterior distribution of the parameters also beyond the support of their prior distributions. Extensions of the regression adjustment approach through reparametrization may be explored to provide a more refined output.[15] Similarly, modified kernel density estimates can also be used to bound the density estimates used for visualization.

The results from the model in Section 3 and the model selected in Section 4 are not directly comparable, since the MLEs obtained in Section 3 refer to a model that does not include covariates. However, Table 3 shows that the MLEs reflect an average across groups of the estimates found from the model with covariates, and a general agreement between the two models can be appreciated. In the Supplemental Material we further compare the results from the two models.

Also, note that the time when asymptomatic detectability starts (T_A) depends on the accuracy of the technology used to perform the examination that, therefore, should be the same for all the visits included in the estimation procedure. An improvement in the examination technique could make T_A move backwards, and the length of the asymptomatic detectability interval increase.

The theoretical distribution of the observed age at asymptomatic and symptomatic detection can be computed analytically from a theoretical model, after superimposing the screening examinations. In the Supplementary Material we obtain the analytical form of the distributions of the observed age at detection, both symptomatic and asymptomatic, for one such simple model. The resulting expressions are rather complicated, and in most cases simulations are probably a more suitable tool to study the effect of the selection process on the observed detections under complex models and screening strategies.

Summing up, in this work we have highlighted that latent (realistic) models for disease histories are challenging to develop and implement, but that ABC is a very flexible and conceptually simple tool, that looks especially suitable in this setting where it is relatively easy to generate data even from rather complex models and filter them through non-trivial observation processes.

We should point out that goodness-of-fit of the models here is evaluated conditionally on the choice of the prior distributions for the parameters of each model. Therefore, it is possible that a model is penalized by a poor choice of the prior or, on the contrary, that a model performs well thanks to a good prior choice. In particular, a change in the prior distributions may lead to a different result in ABC model choice. Note that model selection between two non-nested parametric models could also be performed by using Vuong's test.[39] However, Vuong's test is based on the ratio of the likelihood functions under the two models, and as a consequence also requires that one be able to compute them.

Our models have assumed perfect screening sensitivity and specificity. However, they can be extended to estimate them from the data, and to take into account the dependence between the subject-specific adherence pattern and the latent disease process. These extensions could not be implemented fully on the motivating data, given that detailed information about screening invitations and examinations results was not available to us. However, we have conducted a small experiment in this direction. We have extended the selected model by introducing an additional parameter for the sensitivity of the screening examinations. The ABC estimation procedure did not work too well: despite using quite an informative prior for the new parameter (Beta with mean equal to 5/6), stability issues in the estimation of the susceptible proportions emerged, and this could be due to the choice of the distance (the summary statistics) or to the lack of sufficient information in the data. Indeed, while in general sensitivity may be identifiable, this experiment suggests that the choice of the metric to be used in ABC may make the identifiability of some parameters more difficult. As an additional sensitivity analysis, we have inserted in the data generation process a hard coded value of sensitivity of 0.9, and that did not seem to change the posterior distributions of the other parameters of the model.

Access to data with longer follow-up could allow one to study the effect of screening and treatments on survival. In general, it will be of great interest to apply the models that we have developed to other, similar datasets to confirm the information on the latent process that emerged here.

Clearly, changing the way in which event times are observed, for example by changing the screening schedule, cannot impact the latent process. One way to check whether these models describe the latent process well would therefore be to also use data collected under different screening policies. For example, we know that the Covid-19 pandemic is causing a drop in screening adherence. Therefore, it will be important to apply these models to data collected by screening programs during, and after this period.

In this work we did not mention overdiagnosis due to mammography screening, that is the detection of a breast cancer that would not be detected during the woman's lifetime in the absence of screening. In other words, an overdiagnosed cancer would have never become symptomatic, because of its very slow evolution, and would have never led to death. Many authors discussed this issue and proposed several methods to quantify the risk of overdiagnosis.[40-42] However, given that we have data on invasive cases only, overdiagnosis is probably less of a concern in our data.[13] A possibility to extend

our models to address such question could be to implement a cure rate structure on Δ for the in-screening detected cases or, equivalently, to assign a positive probability to the event $\{T_A < +\infty, T_S = +\infty\}$. Identifiability and estimability for such extended models, both in general and for even large sample sizes, are open questions that will need to be addressed.

ACKNOWLEDGMENTS

The author would like to thank Keith Humphreys, Antonietta Mira and the anonymous referees for their constructive comments on the manuscript.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are not publicly available due to privacy or ethical restrictions. Upon request, the Milan Health Care Agency will consider the possibility of releasing the data. A mock dataset and the code analyzing it are available at https://laurabondi.github.io/research/.

ORCID


Laura Bondi https://orcid.org/0000-0002-7034-9406

REFERENCES

1. Van Oortmarssen G, Boer R, Habbema J. Modelling issues in cancer screening. *Stat Methods Med Res*. 1995;4(1):33-54.
2. Marmot M, Altman D, Cameron D, et al. The benefits and harms of breast cancer screening: an independent review. *Lancet*. 2012;380(9855):1778-1786.
3. Nelson H, Fu R, Cantor A, et al. Effectiveness of breast cancer screening: systematic review and meta-analysis to update the 2009 U.S. preventive services task force recommendation. *Ann Intern Med*. 2016;164(4):244-255.
4. Gøtzsche P, Jørgensen K. Screening for breast cancer with mammography. *Cochrane Database Syst Rev*. 2013;(6):CD001877.
5. Hutchison G, Shapiro S. Lead time gained by diagnostic screening for breast cancer. *J Natl Cancer Inst*. 1968;41(3):665-681.
6. Hu P, Zelen M. Planning clinical trials to evaluate early detection programmes. *Biometrika*. 1997;84(4):817-829.
7. Aalen O. Understanding disease processes. *Stat Med*. 2010;29(11):1159-1160.
8. Sweetin M, Farewell V, De Angelis D. Multi-state Markov models for disease progression in the presence of informative examination times: an application to hepatitis C. *Stat Med*. 2010;29(11):1161-1174.
9. Chen B, Yi G, Cook R. Analysis of interval-censored disease progression data via multi-state models under a nonignorable inspection process. *Stat Med*. 2010;29(11):1175-1189.
10. Abrahamsson L, Humphreys K. A statistical model of breast cancer tumour growth with estimation of screening sensitivity as a function of mammographic density. *Stat Methods Med Res*. 2013;25(4):1620-1637.
11. Isheden G, Humphreys K. Modelling breast cancer tumour growth for a stable disease population. *Stat Methods Med Res*. 2019;28(3):681-702.
12. Bergqvist O. Calibration of Breast Cancer Natural History Models Using Approximate Bayesian Computation. 2020 http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-273605
13. Seigneurin A, François O, Labarère J, Oudeville P, Monlong J, Colonna M. Overdiagnosis from non-progressive cancer detected by screening mammography: stochastic simulation study with calibration to population based registry data. *BMJ*. 2011;343:d7017.
14. Little R, Rubin D. *Statistical Analysis with Missing Data*. Second ed. New York: Wiley; 2002.
15. Sisson S, Fan Y, Beaumont M. *Handbook of Approximate Bayesian Computation*. New York: Chapman and Hall/CRC; 2018.
16. Ventura L, Giorgi D, Giordano L, et al. Mammographic breast cancer screening in Italy: 2011–2012 survey. *Epidemiol Prev*. 2015;39(3):21-29.
17. Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst*. 1989;81(24):1879-1886.
18. Weedon-Fekjær H, Lindqvist B, Vatten L, et al. Breast cancer tumor growth estimated through mammography screening data. *Breast Cancer Res*. 2008;10(3):1-13.
19. Henningsen A, Toomet O. Maxlik: a package for maximum likelihood estimation in R. *Comput Stat*. 2011;26(3):443-458.
20. Waks A, Winer E. Breast cancer treatment: a review. *JAMA*. 2019;321(3):288-300.
21. Weedon-Fekjær H, Vatten L, Aalen OO, et al. Estimating mean sojourn time and screening test sensitivity in breast cancer mammography screening: new results. *J Med Screen*. 2005;12(4):172-178.
22. Beaumont M, Zhang W, Balding D. Approximate Bayesian computation in population genetics. *Genetics*. 2002;162(4):2025-2035.
23. Lintusaari J, Gutmann M, Dutta R, et al. Fundamentals and recent developments in approximate Bayesian computation. *Syst Biol*. 2017;66(1):e66-e82.
24. Drovandi C. Frazier, DT a comparison of likelihood-free methods with and without summary statistics. *Stat Comput*. 2022;32:42.
25. Blum M, Nunes M, Prangle D, et al. A comparative review of dimension reduction methods in approximate Bayesian computation. *Stat Sci*. 2013;28(2):189-208.
26. Csilléry K, François O, Blum M. Abc: an R package for approximate Bayesian computation (ABC). *Methods Ecol Evol*. 2012;3(3):475-479.

27. Struttura Promozione della Salute e Screening. Gli screening oncologici in Lombardia. Report dati 2015 (survey 2016) e dati 2016 (survey 2017 prima parte). Technical report. Regione Lombardia 2017 https://www.regione.lombardia.it/wps/wcm/connect/5cfe19ff-e3c7-4b1e-a336-f1e1b9670c42/Report_screening_2016_Luglio2017.pdf?MOD=AJPERES&CACHEID=ROOTWORKSPACE-5cfe19ff-e3c7-4b1e-a336-f1e1b9670c42-n0c4p8H
28. Marjoram P. Approximation Bayesian computation. *OA Genetics*. 2013;1(1):1-5.
29. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2013 http://www.R-project.org/
30. Bidoli E, Virdone S, Hamdi-Cherif M, et al. Worldwide age at onset of female breast cancer: a 25-year population-based cancer registry study. *Sci Rep*. 2019;9(1):1-8.
31. Howlader N, Noone AM, Krapcho M, et al. SEER explorer. *Breast Cancer-Stage Distribution of SEER Incidence Cases, 2007-2016 by Sex*. Bethesda: National Cancer Institute; 2019.
32. Marin J, Pudlo P, Estoup A, et al. Likelihood-free Model Choice. 2015 1–21; Arxiv:1503.07689.
33. Robert CP, Cornuet JM, Marin JM, Pillai NS. Lack of confidence in approximate Bayesian computation model choice. *Proc Natl Acad Sci*. 2011;108(37):15112-15117.
34. Pudlo P, Marin J-M, Estoup A, Cornuet J-M, Gautier M, Robert CP. Reliable ABC model choice via random forests. *Bioinformatics*. 2015;32(6):859-866.
35. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. New York, NY: Springer New York Inc; 2001.
36. Bernardo J, Smith A. *Bayesian Theory*. New York, NY: Wiley; 2008.
37. Bhargava S, Tsuruda K, Moen K, et al. Lower attendance rates in immigrant versus non-immigrant women in the Norwegian breast cancer screening programme. *J Med Screen*. 2018;25(3):155-161.
38. Gutmann M, Dutta R, Kaski S, et al. Likelihood-free inference via classification. *Stat Comput*. 2018;28(2):411-425.
39. Vuong Q. Likelihood ratio tests for model selection and non-nested hypotheses. *Econom; J Econom Soc*. 1989;57(2):307-333.
40. Beckmann K, Duffy SW, Lynch J, et al. Estimates of over-diagnosis of breast cancer due to population-based mammography screening in South Australia after adjustment for lead time effects. *J Med Screen*. 2015;22(3):127-135.
41. Pathirana T, Hayen A, Doust J, et al. Lifetime risk of prostate cancer overdiagnosis in Australia: quantifying the risk of over diagnosis associated with prostate cancer screening in Australia using a novel lifetime risk approach. *BMJ Open*. 2019;9(3):1-7.
42. Ahn H, Kim H, Welch H. Korea's thyroid-cancer "epidemic"–screening and over diagnosis. *N Engl J Med*. 2014;371(19):1765-1767.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Bondi L, Bonetti M, Grigorova D, Russo A. Approximate Bayesian computation for the natural history of breast cancer, with application to data from a Milan cohort study. *Statistics in Medicine*. 2023;1-21. doi: 10.1002/sim.9756

SUPPLEMENTARY MATERIAL FOR

"APPROXIMATE BAYESIAN COMPUTATION (ABC) FOR THE NATURAL HISTORY OF BREAST CANCER, WITH APPLICATION TO DATA FROM A MILAN COHORT STUDY"

S1 | OBSERVED DATA LIKELIHOOD FOR THE MODEL IN SECTION 3

In this Section, we present the derivation of the observed data likelihood function of the model studied in Section 3 of the main article. As already mentioned, we have three kinds of data configurations:

(i) for a subject with an observed symptomatic detection at age t_s,

$$\begin{aligned} L_i &= p \int_l^{t_s} f_{T_A,T_S}(x,t_s)dx = p \int_l^{t_s} f_{T_A}(x) f_{T_S|T_A}(t_S|x)dx = p \int_l^{t_s} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \lambda e^{-\lambda(t_S-x)}dx \\ &= p\ \lambda\ e^{\frac{\lambda^2\sigma^2}{2}+\lambda(\mu-t_S)} \Big(\Phi(t_S,\mu+\lambda\sigma^2,\sigma^2) - \Phi(l,\mu+\lambda\sigma^2,\sigma^2)\Big) \\ &= p\ f_{T_S}(t_S)\Big(1 - \frac{\Phi(l,\mu+\lambda\sigma^2,\sigma^2)}{\Phi(t_S,\mu+\lambda\sigma^2,\sigma^2)}\Big) = p\ \Big(f_{T_S}(t_S) - e^{\lambda(l-t_S)} f_{T_S}(l)\Big), \end{aligned}$$

where l is the smallest possible value for the asymptomatic detectability, which can be the age at the last negative examination, if there is one, or the lower bound of the support of T_A;

(ii) for a subject with an observed asymptomatic detection at age d,

$$\begin{aligned} L_i &= p \int_l^d \int_d^\infty f_{T_A,T_S}(x,y)dydx = p \int_l^d \int_d^\infty f_{T_A}(x) f_{T_S|T_A}(y|x)dydx \\ &= p \int_l^d \int_d^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \lambda e^{-\lambda(y-x)}dydx = p \int_l^d \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} e^{\lambda x} \int_d^\infty \lambda e^{-\lambda y}dydx \\ &= p \int_l^d \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} e^{\lambda x} e^{-\lambda d}dx = p \int_l^d \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\lambda d + \frac{\lambda(\sigma^2\lambda+2\mu)}{2}} e^{-\frac{(x-(\mu+\sigma^2\lambda))^2}{2\sigma^2}}dx \\ &= p\ e^{\frac{\lambda^2\sigma^2}{2}+\lambda(\mu-d)} \Big(\Phi(d,\mu+\lambda\sigma^2,\sigma^2) - \Phi(l,\mu+\lambda\sigma^2,\sigma^2)\Big) = p\ \frac{f_{T_S}(d) - e^{\lambda(l-d)} f_{T_S}(l)}{\lambda}, \end{aligned}$$

where l is the smallest possible value for the asymptomatic detectability, which can be the age at the last negative examination, if there is one, or the lower bound of the support of T_A;

(iii) for a subject without an observed diagnosis at the censoring time c,

$$L_i = (1-p) + p\Big(\int_l^c \int_c^\infty f_{T_A,T_S}(x,y)dydx + \int_c^\infty \int_x^\infty f_{T_A,T_S}(x,y)dydx\Big),$$

where

$$\begin{aligned} \int_l^c \int_c^\infty f_{T_A,T_S}(x,y)dydx &= e^{\frac{\lambda^2\sigma^2}{2}+\lambda(\mu-c)}\Big(\Phi(c,\mu+\lambda\sigma^2,\sigma^2) - \Phi(l,\mu+\lambda\sigma^2,\sigma^2)\Big) \\ &= \frac{f_{T_S}(c) - e^{\lambda(l-c)} f_{T_S}(l)}{\lambda}, \end{aligned}$$

and

$$\int_c^\infty \int_x^\infty f_{T_A,T_S}(x,y)dydx = \int_c^\infty \int_x^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \lambda e^{-\lambda(y-x)} dydx$$
$$= \int_c^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \int_x^\infty \lambda e^{-\lambda(y-x)} dydx = \int_c^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 1 - \Phi(c,\mu,\sigma^2).$$

Note that, this contribution is equal to $L_i = (1-p) + p \int_l^\infty \int_c^\infty f_{T_A,T_S}(x,y)dydx$. However, from the model assumptions, $T_A < T_S$ with probability one, and the integral $\int_c^\infty \int_c^x f_{T_A,T_S}(x,y)dydx$ takes the value zero.

Lastly, we compute the probability of the conditioning event (entry requirement for the motivating study):

$$c_i = P(T_S > \text{Age at entry} | \text{Age at entry}) = (1-p) + p\ P(T_S > \text{Age at entry} | \text{Diseased, Age at entry}).$$

Denoting by f the age at entry in the program, we have that

$$P(T_S > \text{Age at entry} | \text{Diseased, Age at entry}) = 1 - \int_{-\infty}^f \int_x^f f_{T_A,T_S}(x,y)dydx$$
$$= 1 - \int_{-\infty}^f f_{T_A}(x) \int_x^f \lambda e^{-\lambda(y-x)} dydx = 1 - \int_{-\infty}^f \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \left(1 - e^{\lambda(x-f)}\right)$$
$$= 1 - \Phi(f,\mu,\sigma^2) + e^{\frac{\lambda^2\sigma^2}{2} + \lambda(\mu-f)} \Phi(f, \mu+\lambda\sigma^2, \sigma^2) = 1 - \Phi(f,\mu,\sigma^2) + \frac{f_{T_S}(f)}{\lambda}.$$

The observed data likelihood is the product of the subjects' contributions, divided by the probability of the conditioning event: $L = \prod_{i=1}^n \frac{L_i}{c_i}$.

S2 | ABC VS MLE

We compare the ABC approximate posterior distributions to the MLEs obtained for the model described in Section 3.1 of the article, the "Normal + Exponential" model without covariates.

Figure S1 shows that the ABC approximate posterior distributions are concentrated on regions of the parameters' supports which are not far from the confidence intervals around the maximum likelihood estimates. Only the approximate posterior distribution for λ seems to overestimate its magnitude, while those for μ, σ and p show a substantial agreement with the MLEs.

Figure S2 shows that using the post-processing regression adjustment does not modify the ABC results significantly. Importantly, although confidence intervals and posterior density intervals are clearly different in their very meaning, the results from ABC suggest less precise inference compared to the MLEs. This seems to be a potential drawback in the use of the more flexible ABC approach, unless one is willing to impose more concentrated (and thus potentially misleading) prior distributions for the parameters.

S3 | ASSESSMENT OF THE PERFORMANCE OF ABC MODEL SELECTION

A conclusive assessment of the performance of the ABC model selection procedures is still lacking. In general, such an assessment is difficult since, among other issues: (i) it is based on one very specific model; (ii) it depends on the specific summary measure that one implements; (iii) it depends on the collection of alternative models that one considers.

Specifically, we have experimented with model selection within our context, generating 50 datasets from the prior distributions of each model. As shown by Table S1, the random forest-based algorithm does quite a satisfactory job at discriminating between the different models, recovering the true data generating model in most cases for nine of the ten considered models. We again need to stress the highly specific nature of our context, being careful not to overstate the generality of these results.

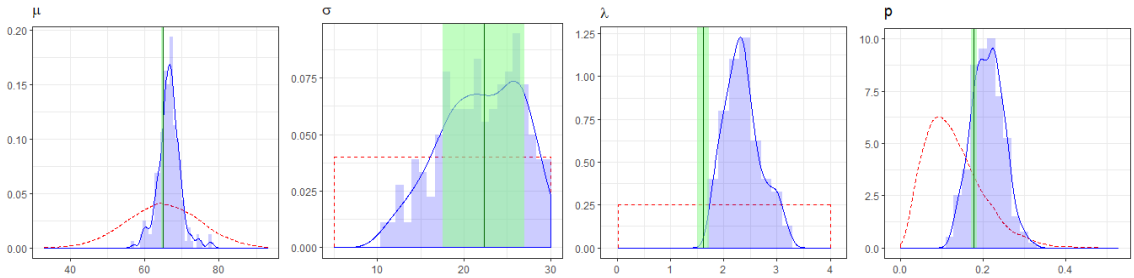


FIGURE S1 Prior (red dashed line) and unadjusted posterior (blue histogram and solid line) densities for each parameter of the "Normal + Exponential" model without covariates. The green vertical lines represent the MLEs, together with their 95% confidence intervals (green areas).

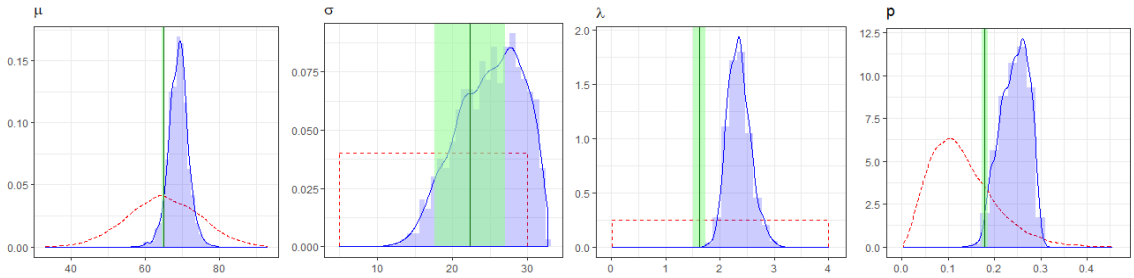


FIGURE S2 Prior (red dashed line) and local-linear-regression adjusted posterior (blue histogram and solid line) densities for each parameter of the "Normal + Exponential" model without covariates. The green vertical lines represent the MLEs, together with their 95% confidence intervals (green areas).

7	5	5	5	5	5	12	2	2	2	50
4	17	1	6	2	10	4	0	4	2	50
0	4	37	4	0	1	1	0	3	0	50
0	1	2	43	0	0	0	1	1	2	50
5	7	4	2	13	5	2	4	8	0	50
0	0	3	3	0	29	0	0	9	6	50
4	2	10	4	4	0	11	7	7	1	50
5	5	8	7	1	3	4	14	3	0	50
0	1	1	2	0	10	0	0	27	9	50
0	1	0	2	0	10	0	0	17	20	50
25	43	71	78	25	73	34	28	81	42	

TABLE S1 Misclassification table from the RF-based algorithm using 50 prior samples from each model. Rows correspond to the data generating models (from 1 to 10) and columns to the selected models.

S4 | ABC ESTIMATION FOR THE "GAMMA + PIECEWISE EXPONENTIAL" MODEL

We present the results of the ABC estimation procedure applied to Model 6, the "Gamma + piecewise Exponential" model, in order to compare them to the results of the "Rescaled Beta + piecewise Exponential" model. We present the local-linear-regression adjusted results obtained by using Metric 1. We assumed the following independent prior distributions for the model parameters: $\beta_0 \sim \text{N}(65, 100)$, $\beta_i \sim \text{N}(0, 25)$, for $i = 1, 2, 3$, $\sigma \sim \text{Unif}(0, 25)$, $\lambda_i \sim \text{Unif}\big(0.1, 4\big)$, for $i = 1, 2, 3$, $p_0 \sim \text{logit}\big(\text{Beta}(3, 21)\big)$, $p_i \sim \text{Unif}(-2, 2)$, for $i = 1, 2, 3$. A retention (or tolerance) rate of 0.01 is chosen via a leave-one-out cross-validation procedure, as performed by the `R` function `cv4abc`. This means that the posterior distributions are estimated from a sample of $200{,}000 \times 0.01 = 2{,}000$ retained parameter values.

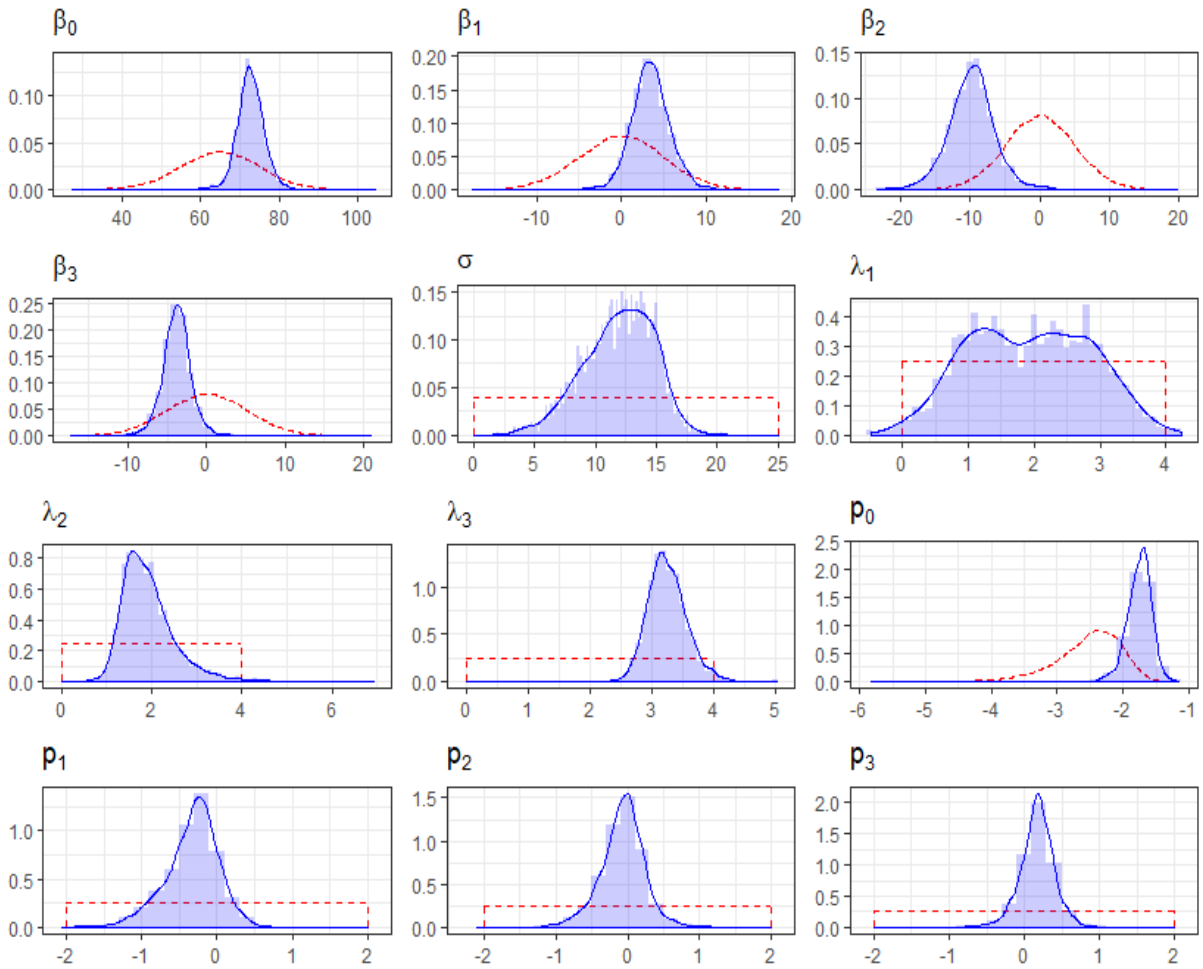


FIGURE S3 Prior (red dashed line) and local linear regression adjusted approximate posterior (blue histogram and solid line) densities for each parameter of the "Gamma + piecewise Exponential" model.

The posterior distributions of the model parameters are much more concentrated than the prior distributions (see Figure S3). The main observations about the effect of each covariate that arise from the estimated posteriors are very similar to those highlighted in Section 4.3 for the "Rescaled Beta + piecewise Exponential" model: (i) women with at least one child tend to have a lower probability to experience breast cancer and a later T_A when they do (posterior distributions of p_1 and β_1); (ii) having a family history of cancer has the opposite effect, according to the posterior distributions of p_3 and β_3; (iii) women with a high level of education experience breast cancer around 10 years earlier than women with a lower level, but they have a lower probability of getting diseased, according to the posterior distributions for β_2 and p_2.

To gain a clear idea of how covariates influence the mean of T_A, which is given by $E(T_A) = \mu(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, we again combine the posterior distributions of $\beta_0, \beta_1, \beta_2$ and β_3 according to the covariate values combination of each group. The resulting boxplots are shown in the left panel of Figure S4. We can see how covariates play an important role in determining $E(T_A)$, whose median ranges from a minimum of 59 years old to a maximum of 76 years. The similarity with the left panel of Figure 3 is evident, showing an agreement of the two models on the estimate of $\mu(x)$. The right panel of Figure S4 is the analogous of the right panel of Figure 3 and shows how the probability for a woman of developing breast cancer varies across groups. Its posterior median ranges from a minimum of about 10% for women in groups 5 and 7 (having at least one birth and with no family history of cancer) to a maximum of about 20% for women in groups 2 and 4 (without any birth and with family history of cancer).

Similarly to what we have done in the previous section, we estimate the predictive distributions for T_A in each covariate group, as well as for Δ given the observed value of T_A. Note how the distributions shown in Figure S5 are again consistent with the

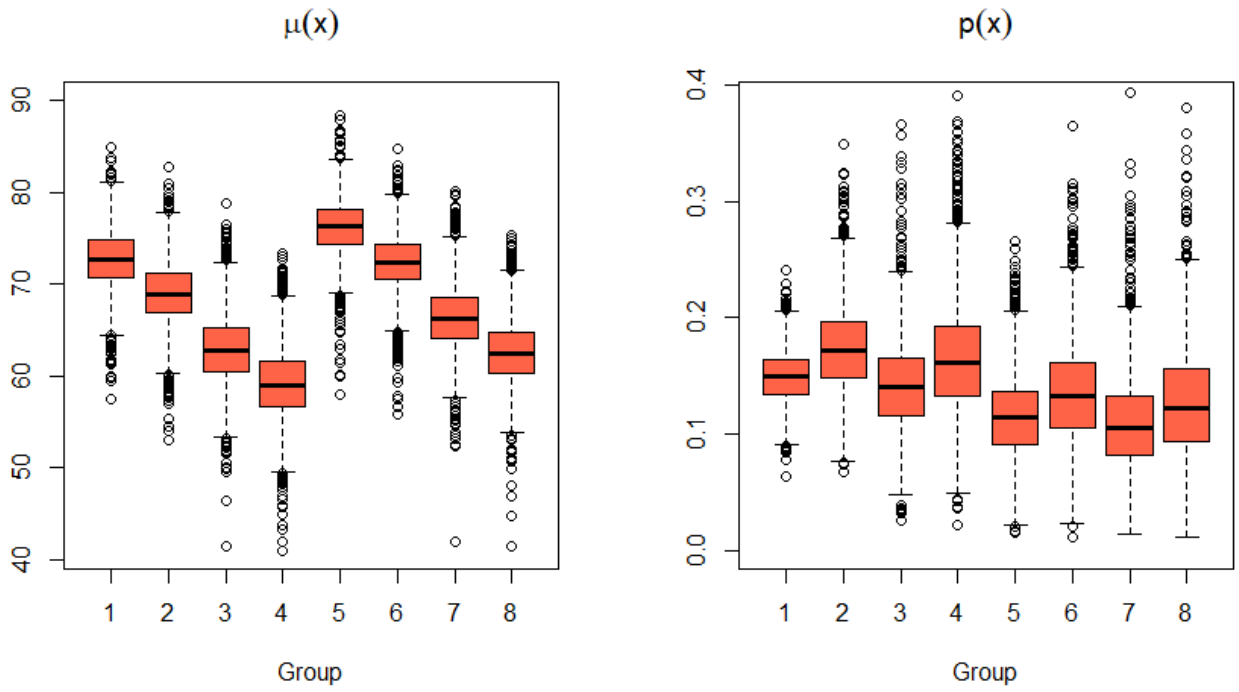


FIGURE S4 Approximate posterior distribution of the mean age at asymptomatic detectability $\mu(x)$ and of the susceptible proportion p(x), across the eight covariate groups, for the "Gamma + piecewise Exponential" model.

results obtained from the "Rescaled Beta + piecewise Exponential" model. Indeed, the set of boxplots on the left-hand side of the figure (predictive distributions for T_A) looks very similar to that shown in the left panel of Figure 4.

Similarly, the posterior sample of size 2000 for λ_1, λ_2 and λ_3 can be used to generate a sample from the approximate predictive distribution of Δ given T_A (see the right-hand side of Figure S5). From these results the two models substantially agree in concluding that tumors with a later T_A seem to evolve faster, and therefore to have a shorter Δ, than tumors with earlier T_A.

S5 | RESULTS UNDER DIFFERENT DATA ASSUMPTIONS

In this Section we show the results obtained under an alternative set of assumptions on the data used to define symptomatic and asymptomatic detections.

Compared to the first set of assumptions, that we presented in the main body of the article, here we obtain a larger proportion of asymptomatic detections. Indeed, we checked if there was at least one screening examination within one year (previously we set this interval to be less than six months) prior to the diagnosis. If yes, then the last one before the diagnosis was assumed to have given a positive result and to have led to an asymptomatic detection. In this case the date of detection was defined as the date of that positive exam.

If, instead, there were no screening exams within one year of the date of diagnosis, we classified that detection as symptomatic, and we set the date of detection equal to the date of the most recent out-of-screening exam, if there were any within the six months prior to diagnosis. If no exams at all were recorded in the six months prior to diagnosis, then we set the date of the symptomatic detection back by a number of days equal to the average shift applied to the symptomatic detections which had that information (72 days).

Once the dates of detection were defined, we picked the last negative exam as the most recent exam performed at least one year before the detection. In other words, we imposed a larger minimum distance (one year instead of six months) between the last negative exam and the detection, as compared to the first set of assumptions.

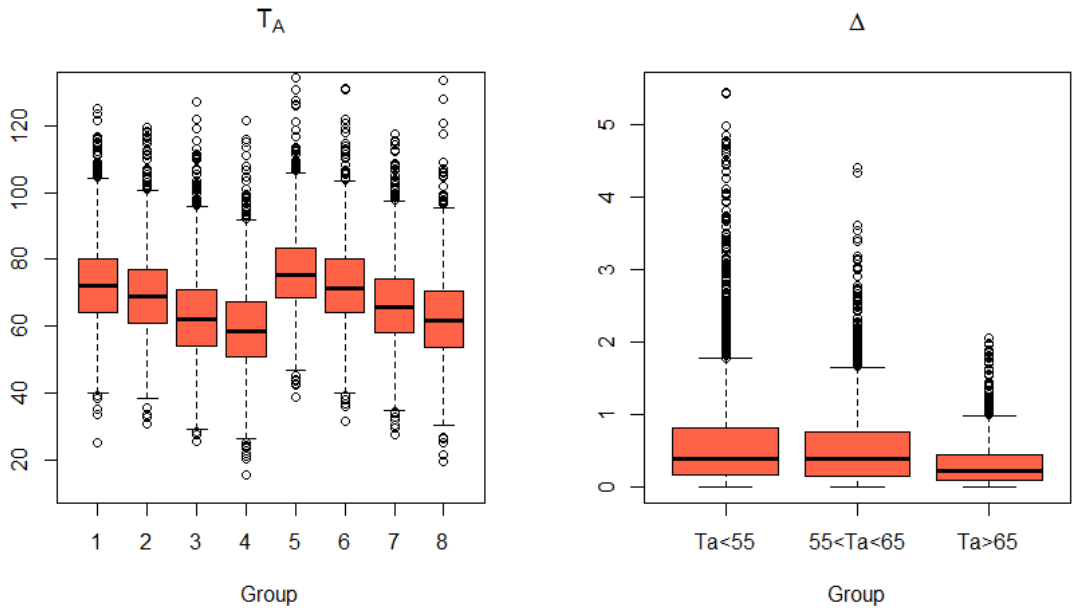


FIGURE S5 Predictive distributions for T_A in each covariate group and for Δ given the observed value of T_A, for the "Gamma + piecewise Exponential" model.

By following these rules, we obtained 728 asymptomatic and 2306 symptomatic detections. The total number of diagnoses is of course unchanged (3034).

We performed model selection using the observed summaries from this second set of assumptions and the results are very similar to those obtained from the first set of assumptions. Model 10 (the new one) is selected by the naive procedure (posterior probability of 0.23), but Model 9 still wins when using the random forest-based algorithm (posterior probability of 0.28)

We then performed ABC parameter estimation for Model 9 (the "Beta + piecewise Exponential" model), using Metric 1 and applying the local-linear-regression adjustment to the posterior sample. Figure S6 shows the posterior distributions obtained retaining 0.5% (chosen via cross-validation) of the proposed parameter values. If one compares these distributions with those shown in Figure 2 of the article, he can notice that the differences are negligible for almost all parameters. The only small changes regard the posterior distributions of λ_2 and λ_3, which are now slightly shifted towards smaller values (and thus towards larger values of Δ).

The resulting predictive distributions for T_A are nearly identical to those presented in the article, while the predicted values for Δ are slightly larger than before (see Figure S7). The predicted mean sojourn times are now 12 months, 8 months and 5 months, for $T_A \leq 55$, $55 < T_A \leq 65$ and $T_A > 65$ respectively.

S6 | SENSITIVITY ANALYSIS FOR THE IMPUTATION OF MISSING DATA

To test the sensitivity of our results to the missing data imputation method, we considered four extreme cases. In this analysis we excluded subjects with a missing value for family history (since they are less than $1/1000$) and imputed the missing values for X_1 and X_2 in the following four different ways: all (0,0), all (0,1), all (1,0), all (1,1). The observed summaries (data not shown) were very similar between the four resulting datasets and the original dataset.

The model selection procedure selected the same model under all four scenarios (Model 10 from the naïve model selection and Model 9 from the procedure based on random forests). The posterior probabilities of Model 9 from the random forest approach were equal to 0.28, 0.27, 0.28 and 0.26, respectively for the four cases.

The posterior distributions for the parameters of the chosen model are substantially overlapping for all the cases, as shown in Figure S8, where the vertical red segments represent the 95% HPD intervals for the model parameters under the original data (obtained from marginal imputation of the three covariates) and the four extreme imputation scenarios considered here.

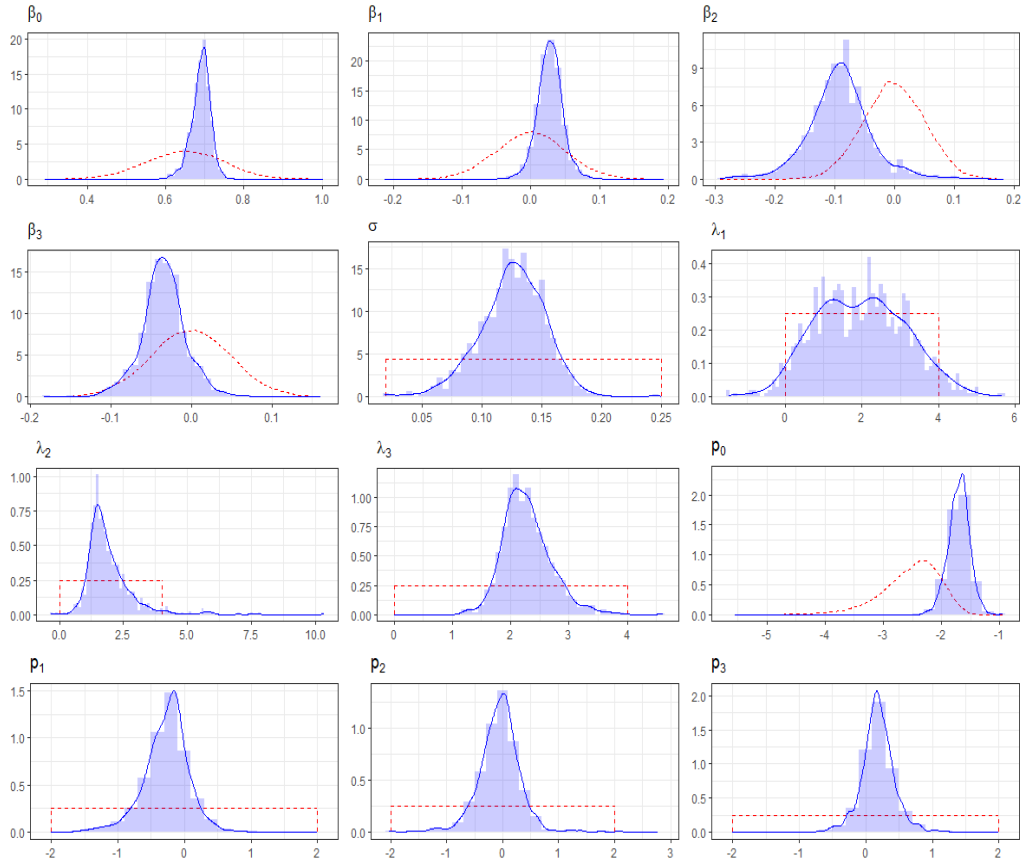


FIGURE S6 Prior (red dashed line) and local linear regression adjusted posterior (blue histogram and solid line) densities for each parameter of the "Beta + piecewise Exponential" model.

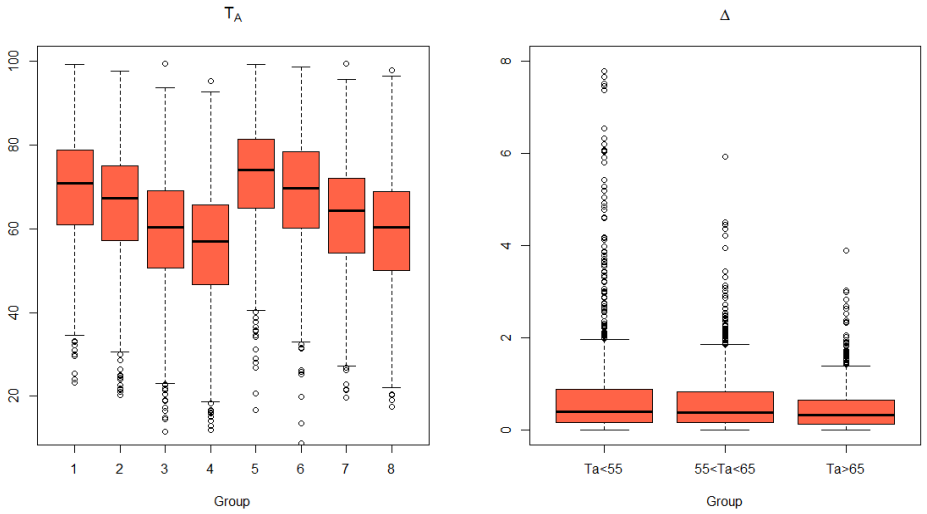


FIGURE S7 Predictive distributions for T_A in each covariate group and for Δ given the observed value of T_A.

S7 | SENSITIVITY ANALYSIS FOR THE ABC METRIC

We compare the results when using Metric 1 and Metric 2 on a "target" sample dataset simulated from the "Gamma + piecewise Exponential" model under some plausible parameter values. Metric 2 was built by considering the L^2-norm of the vector of

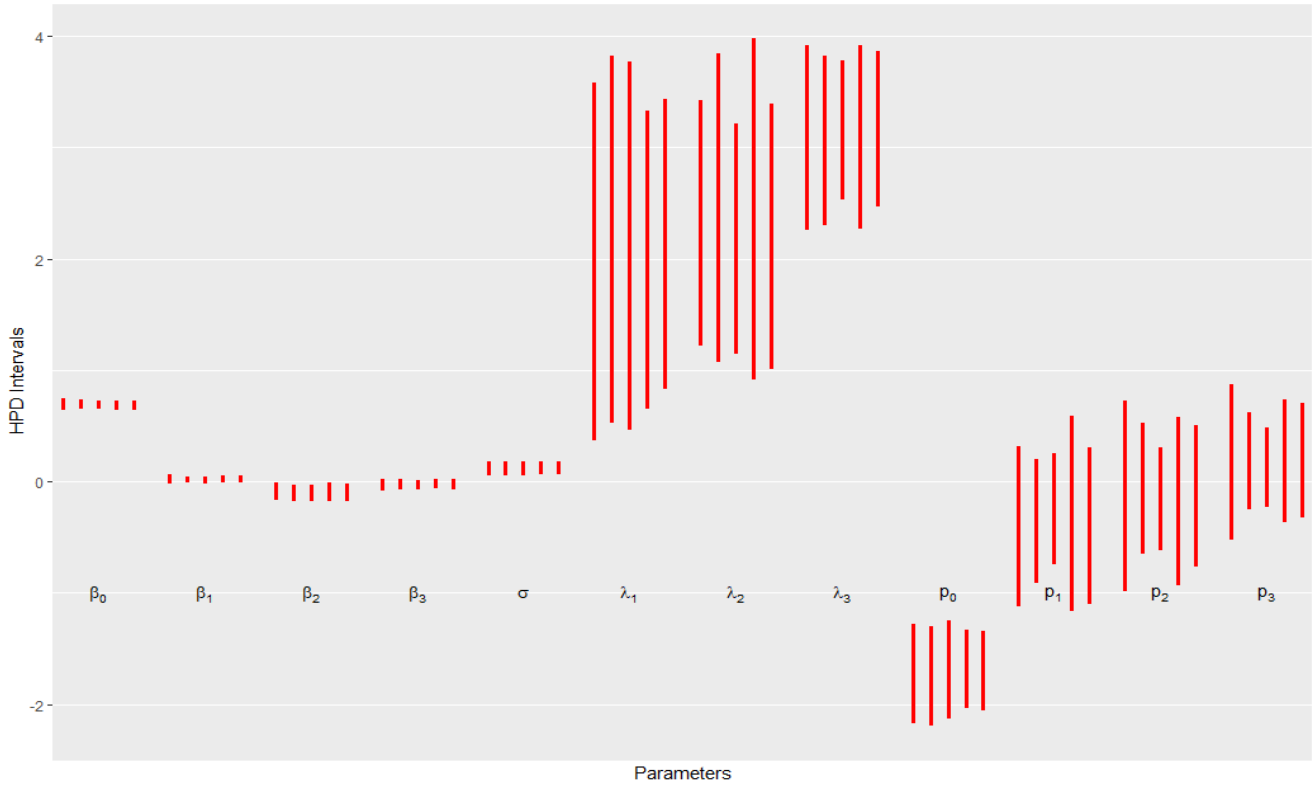


FIGURE S8 95% HPD intervals for the model parameters under the original data and the four scenarios with extreme imputation.

the 32 test statistics, as described in the main body of the article, i.e. selecting those parameter combinations that lead to the minimum L^2-distance between the test statistics and the point $(0, \dots, 0)$.

For each metric, we report the results obtained with and without the local-linear-regression adjustment. In each of the four procedures, the retention rate was chosen via a leave-one-out cross-validation procedure. Table S2 shows the posterior modes together with the 95% highest posterior density intervals (HPDI). The regression adjustment improves the performance of Metric 1 drastically, while it does not have a clear effect for Metric 2, which shows in both cases a poor fit to the (known) parameter values used to generate the data.

The regression-adjusted version of Metric 1 is strongly preferable over the other three options, showing a great ability to recover all the parameter values with precision. Figure S9 shows a graphical comparison between the resulting posterior distributions and the true parameter values. The plots show that the posterior distributions are located around the true value of each parameter, and that in most cases have a much smaller variance than the corresponding prior distribution.

This example confirmed our preference of using Metric 1 with regression adjustment in the analysis of the motivating data.

S8 | INCIDENCE

Given the estimated latent disease process, one can study what the resulting annual incidence for the eight covariate groups (as defined in Table 2).

Recall that the incidence is defined as the probability of being diagnosed with the disease over a specified interval of time, given that the subject was disease free until the start of that interval.

Here we are interested in comparing the shapes of the annual incidence curves when one considers the entire population (see Figure S10b) or when one restricts the calculation only to the susceptible proportion of subjects, i.e. those who will experience the disease in their lifetime (see Figure S10c). While the incidence conditional on being in the susceptible proportion of subjects is, as expected, monotone increasing, the *observed* incidence in the population is not. Figure S10b allows for a more direct

	True	Metric 1		Metric 2	
		Unadjusted	Adjusted	Unadjusted	Adjusted
β_0	67	65.25 (49.06,79.55)	66.15 (57.53,74.75)	64.72 (52.15,75.53)	69.78 (67.89,71.58)
β_1	-7	-0.27 (-9.76,8.16)	-6.59 (-11.74,-0.21)	-1.67 (-9.30 ,7.57)	-13.73 (-16.08,-11.13)
β_2	2	0.28 (-8.83,8.34)	1.24 (-3.22,5.92)	0.61 (-9.32,8.48)	5.87 (2.12,8.93)
β_3	4	0.46 (-8.94,8.89)	4.26 (-2.86,10.22)	0.98 (-8.39,8.65)	5.49 (-0.48,10.75)
σ	13	19.65 (5.82,24.83)	13.95 (-1.87,21.69)	21.83 (10.46,24.98)	16.35 (8.94,19.96)
λ_1	0.5	0.33 (0.10,3.75)	0.81 (-0.20,4.19)	0.66 (0.14,3.79)	0.69 (0.00,2.87)
λ_2	0.3	0.16 (0.10,2.15)	0.29 (-0.01,1.17)	0.60 (0.11 ,2.25)	-0.34 (-0.37,-0.29)
λ_3	0.2	0.14 (0.10,1.60)	0.17 (-0.01,0.79)	0.47 (0.13 ,1.14)	0.06 (0.04,0.10)
p_0	-1.9	-1.96 (-3.13,-1.04)	-2.23 (-3.20,-1.29)	-1.8 (-2.61,-0.98)	-1.87 (-2.23,-1.43)
p_1	-0.3	-0.29 (-1.96,1.16)	-0.29 (-0.99,0.10)	-0.31 (-1.75 ,0.97)	-0.25 (-0.84,0.68)
p_2	-0.03	-0.13 (-1.93,1.49)	0.04 (-0.84,0.70)	0.11 (-1.49 ,1.40)	0.16 (-2.09,2.38)
p_3	0.2	0.12 (-1.67,1.77)	0.27 (-0.97,1.24)	0.21 (-1.19 ,1.71)	1.19 (0.86,1.66)

TABLE S2 Posterior modes and 95% highest posterior density intervals (HPDI).

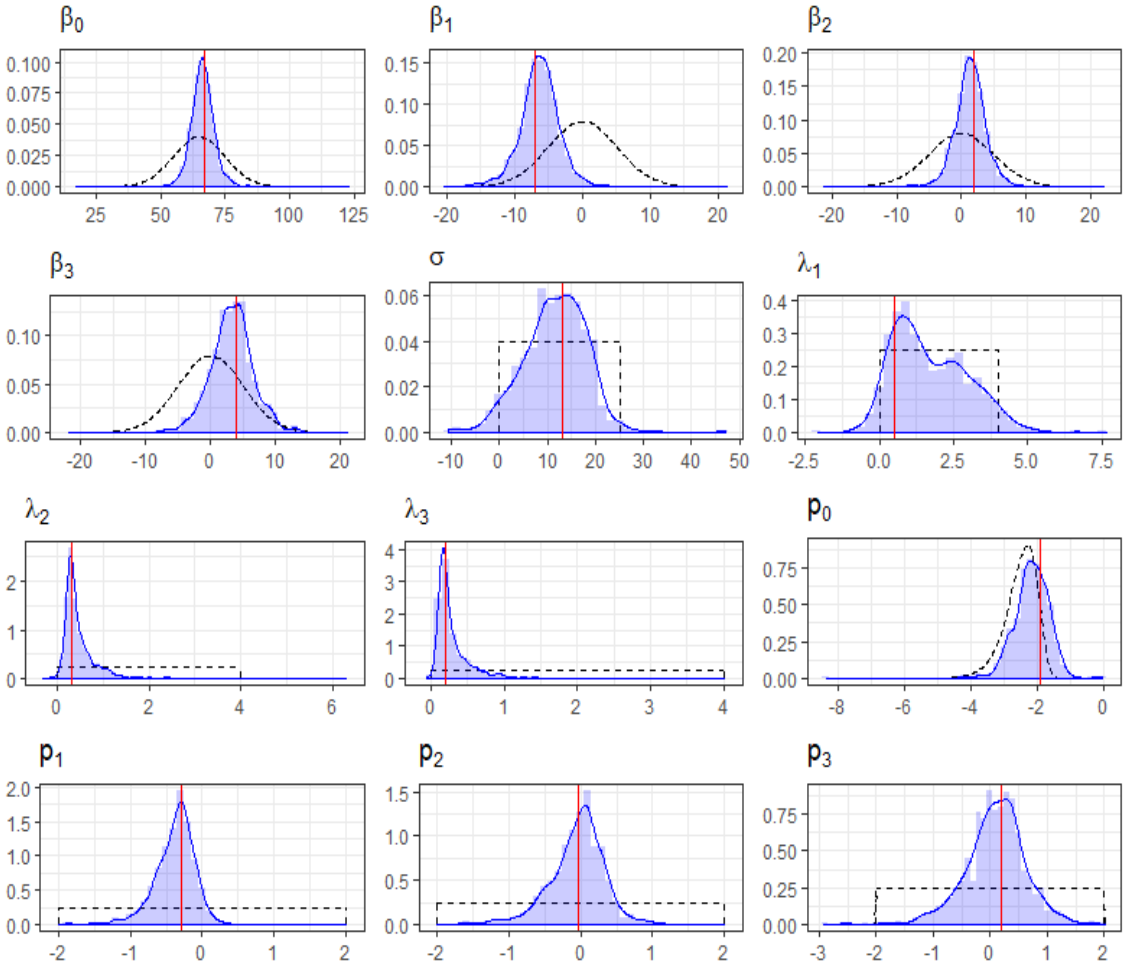
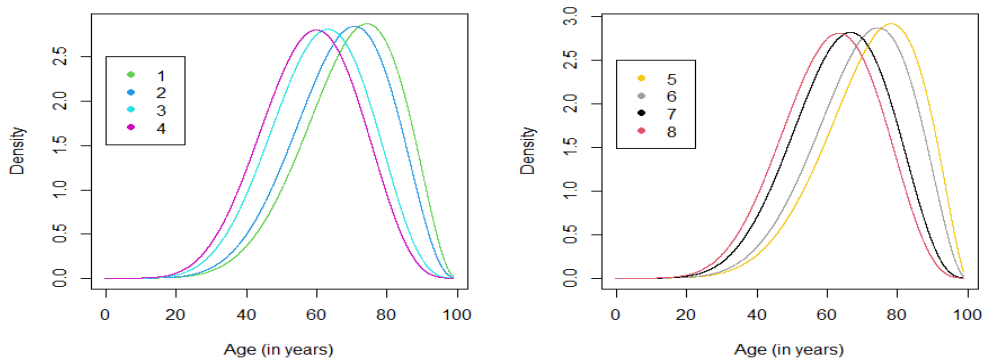
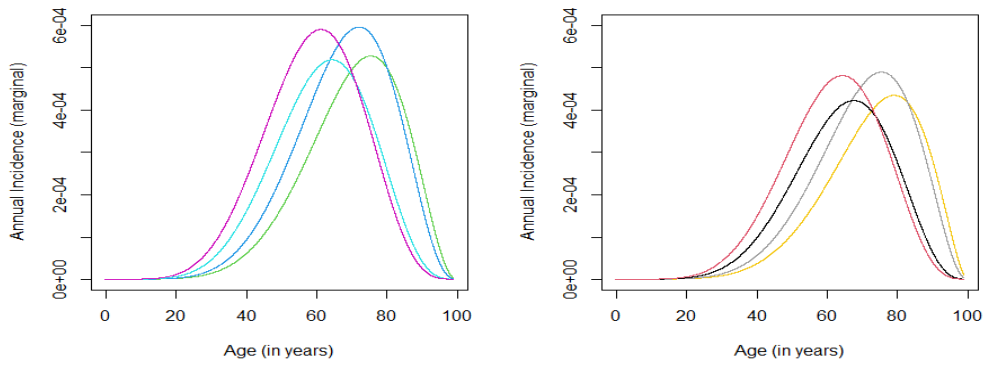


FIGURE S9 True values (red vertical lines), prior (black dashed lines), and local linear regression adjusted posterior (blue histograms and blue solid lines) densities obtained using Metric 1 and the "Gamma + piecewise Exponential" model.

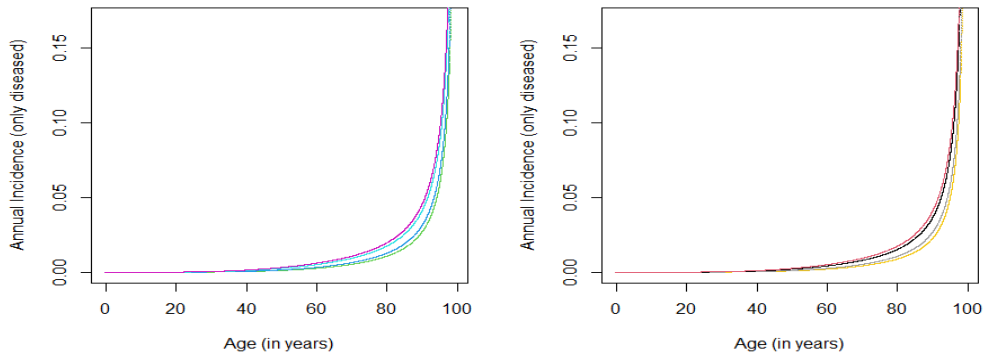
comparison with the official published statistics.[43] The similarity between observed and predicted annual incidence in the population supports the ability of our model to capture the underlying disease history.



(a) Estimated density for T_A using the posterior mode as a point estimate for the parameters of the rescaled Beta distribution.



(b) Annual incidence in the population.



(c) Annual incidence conditional on being diseased.

FIGURE S10

S9 | DISTRIBUTION OF THE AGE AT OBSERVED AGE AT DETECTION WITHIN A SCREENING PROGRAM: AN ANALYTICAL EXAMPLE

In this section, we show an analytical example of how the exact distributions of the observed age at asymptomatic and symptomatic detections can be computed.

Let B indicate the calendar time of birth. We assume a homogeneous Poisson process for the births, so that conditionally on the number of events the times are distributed uniformly over a time interval, which we take as being [Bmin, 0), i.e. $B \sim U(Bmin, 0)$. Recall the usual definitions of the potential values (T_A, T_S) associated with each individual in the population, where T_A is the age at which asymptomatic detectability starts and T_S is the age at which symptomatic detectability starts (and symptomatic detection occurs if disease is not detected asymptomatically prior to T_S). In addition, T_D indicates the age at death in the absence

of breast cancer. We assume stationarity with respect to birth time, and in particular:

$$\begin{aligned} T_A &= +\infty \; w.p.\; (1-p_A) \text{ and } T_A \sim N(\mu_A, \sigma_A^2) \; w.p.\; p_A \text{ (call the latter density } f_A^*); \\ T_S &= T_A + \Delta, \text{ such that } \Delta \sim Exp(\lambda) \text{ and } \Delta \perp\!\!\!\perp T_A; \\ T_D &\sim N(\mu_D, \sigma_D^2) \text{ and } T_D \perp\!\!\!\perp (T_A, T_S). \end{aligned}$$

We describe the effect of the selection process that leads to the observation of either $\widetilde{T}_A$ or of T_S, where $\widetilde{T}_A$ is the age at asymptomatic detection as determined from a screening examination, as defined below.

Bias in the observed data arise from three sources: (i) Selection into the study (in particular, $B + T_S > 0$); (ii) Varying screening frequency and/or compliance as a function of age; and (iii) Right censoring due to ending the study. Here we focus on the first kind.

The sampling design is as follows: at calendar time zero subjects enter the study. The criteria for entry are being alive and not having shown symptoms of the disease yet. In other words, $\{\{B+T_D > 0\} \cap \{B+T_S > 0\}\}$. So subjects enter at the same time but with different ages, as described by the random variable $-B$.

Once in the study, subjects are monitored by screening examination every gap years, exactly. We let the screening examinations continue until death. Below we assume perfect compliance to the screening visits. Each screening examination clearly occurs only if the subject is alive at that time, i.e. if $B + T_D$ is greater than the calendar time of that screening examination. At each examination the subject can be found to be negative (indicating that T_A has not occurred yet) or positive (indicating that T_A is prior to the examinations schedule. In the latter case the new variable $\widetilde{T}_A$ is set equal to the time of the examination.

Notably, one reaches the examination time only if T_S has not occurred yet. Indeed, if T_S occurs prior to the examination time, then one would not observe $\widetilde{T}_A$ but rather T_S itself.

For each observed $\widetilde{T}_A$ we expect the value to always be larger than the corresponding T_A for that subject, due to the non-continuous monitoring performed by the screening. We also expect an indirect effect of the selection into the study, since conditioning on $T_S > 0$ carries with it a rather high (but not equal to one) probability that $T_A > 0$ as well. The screening process produces the observation of some $\widetilde{T}_A$ instead of the T_S that would have been observed had screening not been performed. Overall, the overall effect of these phenomena is not clear-cut.

Our goal is to derive, under the assumptions above, the exact distribution of the two observed variables $\widetilde{T}_A$ and T_S.

S9.1 | Distribution of the observed values of the variable T_S.

We derive the distribution of the random variable $T_S|T_S$ is observed. Let $B = b$. For any $t > -b$, The event $\{T_S \in [t, t+dt), T_S \text{ obs}\}$ is equivalent to the event

$$A_t = \left\{ T_S \in [t, t+dt), b+T_D > 0, b+T_S > 0, b+T_D > b+T_S, b+T_A \in \left(\left\lfloor \frac{b+t}{gap} \right\rfloor gap, b+t \right) \right\}, \tag{1}$$

which describes the fact that for T_S to be observed and to be equal to t, the subject should have entered the study, be alive at calendar time $b + T_S$, and be such that the start of asymptomatic detectability of the disease should fall between the last screening examination before calendar time $b+t$ and time $b+t$. Indeed, screening examinations occur at calendar times $(j \cdot gap)$ for $j = 0, 1, 2, \ldots$, and $\left\lfloor \frac{u}{gap} \right\rfloor gap = j \cdot gap$ for $u \in [j \cdot gap, (j+1) \cdot gap)$. For ease of notation we call $l(u) = \left\lfloor \frac{u}{gap} \right\rfloor gap$.
The event $\{b+T_D > b+T_S\} \subseteq \{b+T_D > 0\}$, so that only the former event needs to be included. Note that the event $\{b+T_S > 0\}$ is such that the event in (1) is empty for $t < -b$, so that we only need consider its probability for $t > -b$.
Given $B = b$ (with B independent of all other random variables) we have

$$\begin{aligned} P(A_t|B=b) &= P(T_D > t) P\left(T_S \in [t, t+dt), b+T_A \in (l(b+t), b+t)\right) \cdot \mathbb{1}\left(t > -b\right) \\ &= S_{T_D}(t) P\left(T_S \in [t, t+dt), T_A \in (l(b+t)-b, t)\right) \cdot \mathbb{1}\left(t > -b\right) \end{aligned} \tag{2}$$

The term $P(T_D > t) = S_{T_D}(t)$ corresponds to the first event in (1), and the fact that we consider all of these expressions as dt tends to zero, so that $f_{T_S, T_S \text{obs}}(t)$ can be obtained from $(dt) f_{T_S, T_S \text{obs}}(t) \approx P(A_t)$.

Now, let us focus on the last term of (2). This can be written as a one-dimensional integral

$$\begin{aligned}
P(T_S \in [t, t+dt), T_A \in (l(b+t)-b, t)) &\approx (dt) \int_{l(b+t)-b}^{t} f_{T_A,T_S}(u,t)du \\
&= (dt) \int_{l(b+t)-b}^{t} f_{T_S|T_A}(t|u)\, p_A f^*_{T_A}(u)du = (dt) \int_{l(b+t)-b}^{t} f_{T_S|T_A}(t|u)\, p_A f^*_{T_A}(u)du \\
&= (dt) \int_{l(b+t)-b}^{t} f_{\Delta|T_A}(t-u|u)\, p_A f^*_{T_A}(u)du \overset{\perp\!\!\!\perp}{=} (dt) \int_{l(b+t)-b}^{t} f_{\Delta}(t-u)\, p_A f^*_{T_A}(u)du \\
&= (dt)\, p_A \int_{l(b+t)-b}^{t} \mathrm{e}^{-\lambda(t-u)} \frac{1}{\sqrt{2\pi}\sigma_A} \mathrm{e}^{-\frac{1}{2\sigma_A^2}(u-\mu_A)^2} du
\end{aligned} \tag{3}$$

Expanding the square in the exponent and collecting the terms allows one to recognize the kernel of the $N(\mu^*, \sigma_A^2)$ density, with $\mu^* = \mu_A + \lambda\sigma_A^2$. Call the associated cumulative distribution function Φ^*. After some algebra we obtain

$$P(A_t|B=b) \approx (dt) S_{T_D}(t)\, p_A \left[\lambda \mathrm{e}^{-\lambda t - \frac{1}{2\sigma_A^2}(\mu_A^2 - (\mu^*)^2)}\right] \left[\Phi^*(t) - \Phi^*(l(b+t)-b)\right] \cdot \mathbb{1}\big(t > -b\big). \tag{4}$$

We now integrate with respect to the (uniform) birth time B:

$$\begin{aligned}
P(T_S \in [t, t+dt), T_S \text{ obs}) &= \int_{Bmin}^{0} P(T_S \in [t, t+dt), T_S \text{ obs}|B=b) f_B(b)db \\
&= \int_{Bmin}^{0} P(A_t|B=b) f_B(b)db \\
&= (dt) \int_{Bmin}^{0} S_{T_D}(t)\, p_A \left[\lambda \mathrm{e}^{-\lambda t - \frac{1}{2\sigma_A^2}(\mu_A^2 - (\mu^*)^2)}\right] \left[\Phi^*(t) - \Phi^*(l(b+t)-b)\right] \frac{1}{-Bmin} \cdot \mathbb{1}\big(b > -t\big)db \\
&= (dt) \frac{p_A \left[1 - \Phi_{T_D}(t)\right]}{-Bmin} \left[\lambda \mathrm{e}^{-\lambda t - \frac{1}{2\sigma_A^2}(\mu_A^2 - (\mu^*)^2)}\right] \int_{\max(Bmin, -t)}^{0} \left[\Phi^*(t) - \Phi^*(l(b+t)-b)\right] db,
\end{aligned}$$

so that taking limits as $dt \to 0$ yields the final form

$$f_{T_S}(t; T_S \text{ obs}) = \frac{p_A \left[1 - \Phi_{T_D}(t)\right]}{-Bmin} \left[\lambda \mathrm{e}^{-\lambda t - \frac{1}{2\sigma_A^2}(\mu_A^2 - (\mu^*)^2)}\right] \int_{\max(Bmin, -t)}^{0} \left[\Phi^*(t) - \Phi^*(l(b+t)-b)\right] db \tag{5}$$

which requires numerical integration of the normal cumulative distribution function Φ^*.

Finally, note from (1) that the distribution in (5) is actually the distribution of T_S and T_S observed. The desired distribution of T_S conditionally on T_S being observed is obtained by taking the ratio of (5) and the normalizing constant, which can also be obtained numerically.

Figure S11 shows a sample output of the simulations performed in R. In particular, the density function in (5) is superimposed on the histogram of the observed values of T_S from a simulate sample from the model.

S9.2 | Distribution of the observed variable $\widetilde{T}_A$.

We define the new random variable $\widetilde{T}_A$ which indicates the time from birth until an observed asymptomatic diagnosis, i.e. a detection that has occurred at one of the planned screening examinations. Receiving such a diagnosis occurs when T_A is prior to the screening examination, while T_S is after it. In addition, the subject should be in the study, and not have died prior to the screening examination. If these events do not happen, then $\widetilde{T}_A$ is not observed (N/A).

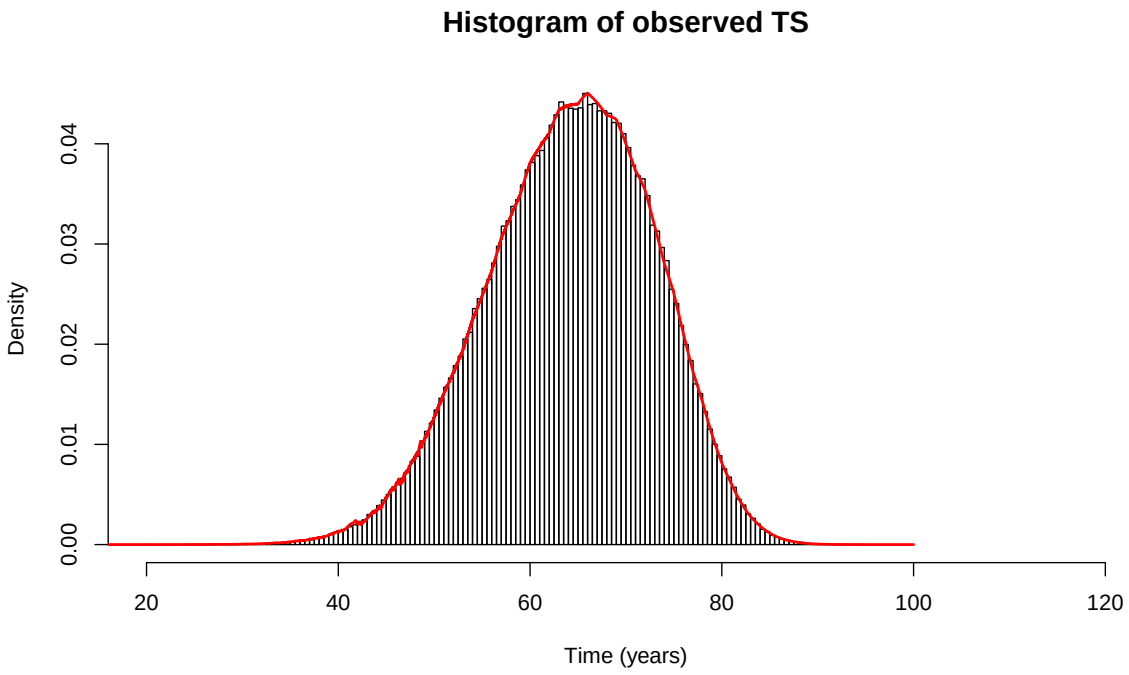


FIGURE S11 Sample of observed values T_S in a simulated sample from 10 million initial subjects. Parameter values were $p_A = 0.15$, $\mu_A = 65$, $\sigma_A^2 = 100$, $\lambda = 1/3$, $\mu_D = 80$, $\sigma_D^2 = 25$, $gap = 3$, and $Bmin = -50$. The red curve represents the density function in (5)

The distribution of the observed $\widetilde{T}_A$ can be obtained in closed form, and computed without numerical integration except for the (readily available) cumulative distribution function of a normal random variable, and for the normalizing constant.
Note that $\widetilde{T}_A$ can only be such that $B + \widetilde{T}_A$ is equal to one of the calendar times at which screening examinations are offered, i.e. calendar times $j \cdot gap$ for $j = 0, 1, 2, \ldots$. Hence

$$P(\widetilde{T}_A \in [t, t+dt), \widetilde{T}_A \text{ obs}) = \sum_{j=0}^{+\infty} P(\widetilde{T}_A \in [t, t+dt), B + \widetilde{T}_A = j \cdot gap, \widetilde{T}_A \text{ obs}). \tag{6}$$

Let us first focus on $j \geq 1$. For a fixed j the following events are identical:

$$\{B + \widetilde{T}_A = j \cdot gap\} = \left\{ \left\lfloor \frac{B + T_A}{gap} \right\rfloor = j - 1 \right\} \cap \left\{ B + T_S > j \cdot gap \right\}.$$

Since $B + t = j \cdot gap$, $j < \lfloor \frac{t}{gap} \rfloor$ must hold, so that the sum in (6) only needs to run until $\lfloor \frac{t}{gap} \rfloor$.

$$\left\{ B + \widetilde{T}_A \in [t, t+dt), \widetilde{T}_A \text{ obs} \right\} =$$
$$\left\{ B + T_S > 0, B + T_D > 0, \left\lfloor \frac{B + T_A}{gap} \right\rfloor = j - 1, B + T_S > j \cdot gap, B + T_D > j \cdot gap, B + t = j \cdot gap \right\} \tag{7}$$

where the first two events on the right hand side can be dropped as they are captured in their intersection with later events. Now, let $B = b$. Recall that B is independent of all other variables. Then, we can compute the conditional probability

$$\begin{aligned}
&P\left(\frac{B+T_S}{gap} > j, \frac{B+T_D}{gap} > j, \left\lfloor \frac{B+T_A}{gap} \right\rfloor = j-1 \middle| B = b\right) \\
&= P\left(\frac{b+T_A+\Delta}{gap} > j, \left\lfloor \frac{b+T_A}{gap} \right\rfloor = j-1, b+T_D > j\cdot gap, b = j\cdot gap - t\right) \\
&= P\left(\frac{b+T_A}{gap} > j - \frac{\Delta}{gap}, j-1 \le \frac{b+T_A}{gap} < j, b+T_D > j\cdot gap, b = j\cdot gap - t\right) \\
&= P\left(\max\left(j-1, j-\frac{\Delta}{gap}\right) \le \frac{b+T_A}{gap} < j, b+T_D > j\cdot gap, b = j\cdot gap - t\right) \\
&= P\left(\max\left(j-1, j-\frac{\Delta}{gap}\right) \le \frac{b+T_A}{gap} < j\right) P(T_D > -b + j\cdot gap)\cdot \mathbb{1}\left(b = j\cdot gap - t\right).
\end{aligned} \tag{8}$$

Let $Y = (b+T_A)/gap$ and $T = \Delta/gap$. From our assumptions it follows that $Y \perp\!\!\!\perp T$, and that

$$Y \sim N\left(\frac{\mu_A + b}{gap}, \frac{\sigma_A^2}{gap^2}\right) \text{ w.p. } p_A, \text{ and } +\infty \text{ w.p. } (1-p_A); \quad T \sim Exp(\lambda\cdot gap).$$

The expression in (8) is therefore equal to

$$\begin{aligned}
&S_{T_D}(-b + j\cdot gap)\, P(\max(j-1, j-T) \le Y < j)\cdot \mathbb{1}\left(b = j\cdot gap - t\right) \\
&= S_{T_D}(-b + j\cdot gap)\cdot \mathbb{1}\left(b = j\cdot gap - t\right)\, p_A \int_{j-1}^{j} \left[\int_{j-y}^{+\infty} f_{T|Y}(t|y)dt\right] f_Y(y)dy \\
&= S_{T_D}(t)\cdot \mathbb{1}\left(b = j\cdot gap - t\right)\, p_A \int_{j-1}^{j} S_T(j-y) f_Y(y)dy \\
&= S_{T_D}(t)\cdot \mathbb{1}\left(b = j\cdot gap - t\right)\, p_A \int_{j-1}^{j} \mathrm{e}^{-\lambda gap(j-y)} \frac{gap}{\sqrt{2\pi}\sigma_A} e^{-\frac{gap^2}{2\sigma_A^2}\left(y - \frac{\mu_A+b}{gap}\right)^2} dy \\
&= S_{T_D}(t)\cdot \mathbb{1}\left(b = j\cdot gap - t\right)\, p_A\, \mathrm{e}^{-\lambda\cdot j\cdot gap} \exp\left[-\frac{gap^2}{2\sigma_A^2}\left(\left(\frac{\mu_A+b}{gap}\right)^2 - \widetilde{\mu}_b^2\right)\right]\left[\widetilde{\Phi}_b(j) - \widetilde{\Phi}_b(j-1)\right],
\end{aligned} \tag{9}$$

where $\widetilde{\Phi}_b$ is the cdf of the $N\left(\widetilde{\mu}_b, \frac{\sigma_A^2}{gap^2}\right)$ distribution, with $\widetilde{\mu}_b = \frac{\sigma_A^2}{gap}\lambda + \frac{\mu_A+b}{gap}$.

Hence we finally obtain

$$\begin{aligned}
&P(\widetilde{T}_A \in [t, t+dt), B + \widetilde{T}_A = j\cdot gap, \widetilde{T}_A \text{ obs}) \\
&= P\left(\frac{B+T_S}{gap} > j, \frac{B+T_D}{gap} > j, \left\lfloor \frac{B+T_A}{gap} \right\rfloor = j-1 \middle| B = b\right)\cdot f_B(b) \\
&= \frac{S_{T_D}(t)}{(-\mathrm{Bmin})}\, p_A\, \mathrm{e}^{-\lambda\cdot j\cdot gap} \exp\left[-\frac{gap^2}{2\sigma_A^2}\left(\left(\frac{\mu_A + j\cdot gap - t}{gap}\right)^2 - \widetilde{\mu}^2_{(j\cdot gap - t)}\right)\right] \\
&\quad \cdot \left(\widetilde{\Phi}_{(j\cdot gap - t)}(j) - \widetilde{\Phi}_{(j\cdot gap - t)}(j-1)\right)\cdot \mathbb{1}\left(j\cdot gap < t < j\cdot gap - \mathrm{Bmin}\right).
\end{aligned} \tag{10}$$

Let us now turn to the case in which the disease is detected during the first screening examination after entry into the study, i.e. the case of $B + \widetilde{T}_A = 0$. This corresponds to the first element ($j = 0$) of the series in (6). Recall that for a subject to find herself in this situation she must have $B + T_A < 0$ and $B + T_S > 0$ (and be alive at calendar time zero). The probability of such

event is therefore equal to

$$
\begin{aligned}
&P(\widetilde{T}_A \in [t, t+dt), B + \widetilde{T}_A = 0, \widetilde{T}_A \text{ obs}) \\
&\quad = P(\widetilde{T}_A \in [t, t+dt), B + \widetilde{T}_A = 0, \widetilde{T}_A \text{ obs} | B = b) \cdot f_B(b) \\
&\quad \approx P(b + \widetilde{T}_A = 0, b + T_A \le 0, b + T_S > 0, b + T_D > 0) \cdot \mathbb{1}\big(b = -t\big) \frac{\mathbb{1}\big(b \in (\text{Bmin}, 0)\big)}{|\text{Bmin}|} \\
&\quad = \frac{S_{T_D}(t)}{|\text{Bmin}|} P(T_A \in (-\infty, -b], \Delta > -b - T_A) \cdot \mathbb{1}\big(t \in (0, |\text{Bmin}|)\big) \\
&\quad \overset{\Delta \perp\!\!\!\perp T_A}{=} p_A \frac{S_{T_D}(t)}{|\text{Bmin}|} \cdot \mathbb{1}\big(t \in (0, |\text{Bmin}|)\big) \int_{-\infty}^{t} \left[\int_{t-u}^{+\infty} f_\Delta(\delta) d\delta \right] f_{T_A}(u) du \\
&\quad = p_A \frac{S_{T_D}(t)}{|\text{Bmin}|} \cdot \mathbb{1}\big(t \in (0, |\text{Bmin}|)\big) \int_{-\infty}^{t} S_\Delta(t-u) f_{T_A}(u) du,
\end{aligned}
\tag{11}
$$

Therefore, with a bit of rearranging,

$$
\begin{aligned}
&P(\widetilde{T}_A \in [t, t+dt), B + \widetilde{T}_A = 0, \widetilde{T}_A \text{ obs}) \\
&\quad = p_A \frac{S_{T_D}(t)}{|\text{Bmin}|} \cdot \mathbb{1}\big(t \in (0, |\text{Bmin}|)\big) \int_{-\infty}^{t} \mathrm{e}^{-\lambda(t-u)} \frac{1}{\sqrt{2\pi}\sigma_A} \exp\left[-\frac{1}{2\sigma_A^2} \left(u - \mu_A\right)^2 \right] du \\
&\quad = p_A \frac{S_{T_D}(t)}{|\text{Bmin}|} \cdot \mathbb{1}\big(t \in (0, |\text{Bmin}|)\big) \mathrm{e}^{-\lambda t} \Phi^*(t) \exp\left[-\frac{1}{2\sigma_A^2} \left(\mu_A^2 - (\mu^*)^2\right) \right] \\
&\quad = p_A \frac{S_{T_D}(t)}{|\text{Bmin}|} \cdot \mathbb{1}\big(t \in (0, |\text{Bmin}|)\big) \mathrm{e}^{-\lambda t} \Phi^*(t) \exp\left(\frac{\lambda^2 \sigma_A^2}{2} + \lambda \mu_A \right),
\end{aligned}
\tag{12}
$$

where Φ^* is the cdf of the $N(\mu^*, \sigma_A^2)$ random variable, with $\mu^* = \mu_A + \lambda \sigma_A^2$. We now put all the terms together to obtain

$$
\begin{aligned}
P(\widetilde{T}_A \in [t, t+dt), \widetilde{T}_A \text{ obs}) \approx (dt)\, p_A \frac{S_{T_D}(t)}{|\text{Bmin}|} \Bigg\{ & \mathrm{e}^{-\lambda t} \Phi^*(t) \exp\left(\frac{\lambda^2 \sigma_A^2}{2} + \lambda \mu_A \right) \cdot \mathbb{1}\big(t \in (0, |\text{Bmin}|)\big) \\
& + \sum_{j=1}^{\lfloor \frac{t}{gap} \rfloor} \left[\mathrm{e}^{-\lambda \cdot j \cdot gap} \exp\left[-\frac{gap^2}{2\sigma_A^2} \left(\left(\frac{\mu_A + j \cdot gap - t}{gap} \right)^2 - \widetilde{\mu}^2_{(j \cdot gap - t)} \right) \right] \cdot \right. \\
& \left. \cdot \left[\widetilde{\Phi}_{(j \cdot gap - t)}(j) - \widetilde{\Phi}_{(j \cdot gap - t)}(j-1) \right] \cdot \mathbb{1}\big(t \in (j \cdot gap, j \cdot gap - \text{Bmin})\big) \right] \Bigg\}.
\end{aligned}
\tag{13}
$$

Figure S12 shows the sample output of simulations performed in R. In particular, the density function in (13) is superimposed on the histogram of the observed values of $\widetilde{T}_A$ from one simulated sample from the model.

Note that, conditionally on $B = b$, the distribution of $\widetilde{T}_A$ is discrete. In particular, if $\widetilde{T}_A$ is observed, then it takes the countable number of values $-b + j \cdot gap$ for $j = 0, 1, 2, \dots$, plus the additional value N/A (which we may also set as $+\infty$. Conditionally on $B = b$, these values are taken with the following probabilities:

$$
\widetilde{T}_A = \begin{cases} -b & P(b + T_A < 0, b + T_S > 0, b + T_D > 0) \\ -b + j \cdot gap & P(\lfloor \frac{b+T_A}{gap} \rfloor = j-1, \frac{b+T_S}{gap} > j, \frac{b+T_S}{gap} > j) \text{ for j} = 1, 2, \dots \\ N/A & k. \end{cases}
$$

Hence the probabilities of the non-N/A values are obtained as described above from the distributions of (T_a, δ) and T_D.

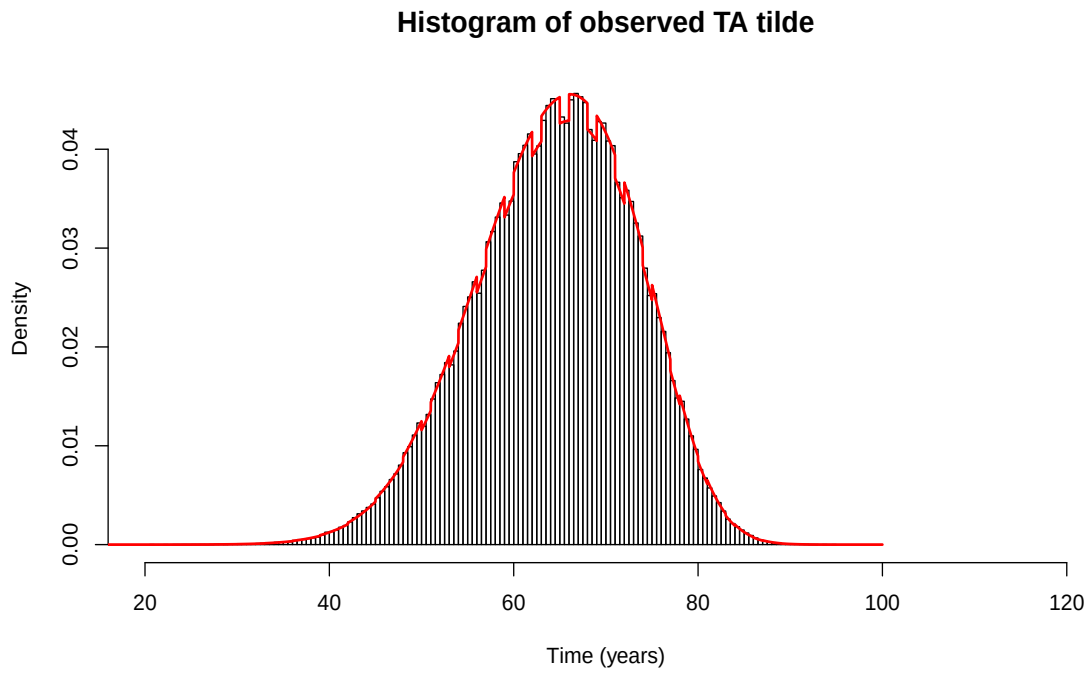


FIGURE S12 Sample of observed values $\widetilde{T}_A$ in a simulated sample from 10 million initial subjects. Parameter values were $p_A = 0.15$, $\mu_A = 65$, $\sigma_A^2 = 100$, $\lambda = 1/3$, $\mu_D = 80$, $\sigma_D^2 = 25$, $gap = 3$ and $Bmin = -50$. The red curve represents the density function in (13).

The conditional probability $k = P(\widetilde{T}_A = N/A|B = b)$ can be obtained as one minus the series of the probabilities of the other values taken by the random variable. The marginal probability $P(\widetilde{T}_A = N/A)$ is then

$$P(\widetilde{T}_A = N/A) = \int_{Bmin}^{0} P(\widetilde{T}_A = N/A) f_B(b) db = \int_{Bmin}^{0} P(\widetilde{T}_A = N/A|B = b) f_B(b) db,$$

that, as we have mentioned, can be obtained numerically.

S9.3 | Marginal probability of being in study

Lastly, we obtain the marginal probability that a randomly selected member of the population is included into the sample. Indeed,

$$\begin{aligned} P(B + T_A + \Delta > 0, B + T_D > 0) &= \int_{Bmin}^{0} P(B + T_A + \Delta > 0, B + T_D > 0|B = b) f_B(b) db \\ &= \int_{Bmin}^{0} P(b + T_A + \Delta > 0, b + T_D > 0) f_B(b) db \\ &= \int_{Bmin}^{0} P(b + T_A + \Delta > 0) P(b + T_D > 0) f_B(b) db, \end{aligned}$$

where

$$P(b+T_A+\Delta>0) = 1 - P(\Delta < -b - T_A) = 1 - p_A \int_{-\infty}^{-b}\int_{0}^{-b-t} f_\Delta(\delta)d\delta f_{T_A}(t)dt_A$$

$$= 1 - p_A \int_{-\infty}^{-b} \left[1 - e^{-\lambda(-b-t)}\right] f_{T_A}(t)dt$$

$$= 1 - p_A\, \Phi_{T_A}(-b) + p_A\, e^{\lambda b} \int_{-\infty}^{-b} e^{\lambda t} f_{T_A}(t)dt$$

$$= 1 - p_A\, \Phi_{T_A}(-b) + p_A\, e^{\lambda b} \int_{-\infty}^{-b} e^{\lambda t} \frac{1}{\sqrt{2\pi}\sigma_A} e^{-\frac{1}{2\sigma_A^2}(t-\mu_A)^2} dt$$

$$= 1 - p_A\, \Phi_{T_A}(-b) + p_A\, e^{\lambda b} \exp\left(\frac{\sigma_A^2\lambda^2}{2} + \mu_A\lambda\right) \int_{-\infty}^{-b} \frac{1}{\sqrt{2\pi}\sigma_A} e^{-\frac{1}{2\sigma_A^2}(t-(\mu_A+\sigma_A^2\lambda))^2} dt$$

$$= 1 - p_A\, \Phi_{T_A}(-b) + p_A\, e^{\lambda b} \exp\left(\frac{\sigma_A^2\lambda^2}{2} + \mu_A\lambda\right) \Phi_Z\left(\frac{-b-(\mu_A+\sigma_A^2\lambda)}{\sigma_A}\right)$$

$$= 1 - p_A\, \Phi_Z\left(\frac{-b-\mu_A}{\sigma_A}\right) + p_A\, \exp\left(\frac{\lambda b + \sigma_A^2\lambda^2}{2} + \mu_A\lambda\right) \Phi_Z\left(\frac{-b-(\mu_A+\sigma_A^2\lambda)}{\sigma_A}\right),$$

and therefore

$$P(B+T_A+\Delta>0, B+T_D>0) = \int_{Bmin}^{0} P(b+T_A+\Delta>0)P(b+T_D>0)\, f_B(b)\, db$$

$$= \int_{Bmin}^{0} P(b+T_A+\Delta>0)P(T_D>-b)\, f_B(b)\, db$$

$$= \int_{Bmin}^{0} P(b+T_A+\Delta>0)\left(1-\Phi_{T_D}(-b)\right) f_B(b) db$$

$$= \frac{-1}{Bmin} \int_{Bmin}^{0} P(b+T_A+\Delta>0)\left[1-\Phi_Z\left(\frac{-b-\mu_D}{\sigma_D}\right)\right] db,$$

which can be easily calculated numerically in R.

□