Supplement 1: Technical Appendix

Estimation of Breast Cancer Overdiagnosis in a US Breast Screening Cohort

Marc D. Ryser*, Jane Lange, Lurdes Inoue, Ellen O'Meara, Charlotte Gard, Diana L. Miglioretti, Jean-Luc Bulliard, Andrew F. Brouwer, E. Shelley Hwang, Ruth B. Etzioni

Contents

*Contact: marc.ryser@duke.edu

1 Mathematical Model of Breast Cancer Natural History

We modeled the natural history of disease using a progressive-indolent mixture model (see Figure 1E in main text) [1, 2, 3]. Starting in a healthy state, women can develop preclinical breast cancers that are asymptomatic but screen detectable. A fraction f_i of preclinical cancers remain in the preclinical state indefinitely, the remaining $1 - f_i$ progress to symptomatic breast cancer after a finite sojourn time. We defined the sensitivity of a screening test as the probability that, given a woman has preclinical cancer, the testing episode (including follow-up imaging and diagnostic work-up) results in a breast cancer diagnosis. The mixture model was parameterized as a stochastic Markov model with piecewise constant transition rates as detailed below.

Onset of preclinical disease is modeled using a piecewise exponential distribution with changepoints $s = (s_0, s_1, s_2)$ and age-dependent hazard function $h(t)$ as follows:

$$h(t) = \begin{cases} 0, & t \in [0, s_0) \\ h_0, & t \in [s_0, s_1) \\ h_1, & t \in [s_1, s_2) \\ h_2, & t \in [s_2, +\infty). \end{cases} \tag{1}$$

We chose a piecewise constant hazard function because, in contrast to smooth alternatives such as the Weibull or gamma distributions, it allows for non-monotone dependencies on age, and is amenable to analytic computation of the likelihood function (see Section 2.1). As explained in detail in Section 2.3, the changepoints are not considered model parameters and are instead chosen based on a combination of etiological considerations and formal model selection. We further denote by $\Lambda(t)$ the cumulative hazard function

$$\Lambda(t) = \exp\left(-\int_0^t h(t)dt\right), \tag{2}$$

and by $f(t)$ and $F(t)$ the probability density and distribution functions of the preclinical onset time.

Preclinical cancers are assumed to be progressive with probability ψ, and indolent or non-progressive with probability $1-\psi$. Progressive tumors are assumed to progress to clinical disease after a finite time, whereas indolent cancers remain in the preclinical state indefinitely.

Among progressive cancers, the sojourn time (that is the time between entering the preclinical

state and progressing to the clinical state) is modeled as an exponentially distributed random variable with rate λ, that is with a mean sojourn time of $1/\lambda$.

Both indolent and progressive pre-clinical cancers are detectable by screening mammography. The screening sensitivity β, representing the conditional probability of detecting a cancer given that it is present at the time of screening, is assumed to be the same for both lesion types.

Finally, we denote by θ the parameter vector for the mixture model as

$$\theta = (h_0, h_1, h_2, \psi, \lambda, \beta). \tag{3}$$

2 Bayesian Inference with Stan

The natural history model was fit to individual-level screening mammography and cancer diagnosis data using Bayesian estimation methods which are particularly useful for quantifying uncertainty in parameter estimates and accounting for results from previous studies (prior information). Specifically, we used the Hamiltonian Monte Carlo sampler Stan [4], as implemented in the software R (version 4.0.3), using the packages rstan (version 2.21.2) and loo (version 2.4.1).

This Bayesian algorithm samples from the joint posterior parameter distribution which naturally provides both parameter estimates and uncertainty intervals, referred to as credible intervals (analogous to the classical confidence intervals). Most importantly, being able to sample parameter values from a credible range based on the fitted model permits then simulating disease natural history and incidence for a hypothetical cohort accounting for uncertainty in the natural history parameters.

Each model run with Stan consisted of two independent chains of length 2,000 were, including 200 warm-up steps that were discarded prior to analysis. Chain convergence was assessed through visual inspection, and convergence measures including the estimated effective sample size and the $\hat{R}$ metric.

2.1 Model Likelihood

For each cohort participant, we have the following longitudinal data available: age at first screen, age at subsequent screens (if any), age and type (screen-detected vs clinical) of cancer diagnosis

(if applicable) and age at censoring (if no cancer diagnosis). To derive the likelihood we cluster participants into three mutually exclusive groups: (i) participants who had one or more screens and no cancer diagnosis; (ii) participants who had one or more screens and a screen-detected cancer diagnosis at the last recorded screen; and (iii) participants who had one or more screens and a clinical cancer diagnosis after the last recorded screen. The following derivations follow closely our previous work [1].

2.1.1 Group 1: No Cancer

We denote by $\tau_1^j, \tau_2^j, \ldots, \tau_n^j$ the ages when participant j undergoes screening. Omitting the superscript j for the sake of simplicity, we assign to each participant a vector of length $l = n + 2$

$$t^j \equiv t = (t_k) = (s_0, \tau_1, \ldots, \tau_n, \tau_c), \tag{4}$$

where s_0 is the age at which the onset hazard (1) becomes positive and τ_c is the time of censoring, set at 18 months after the last negative screen at time τ_n (see main text for details on censoring).

There are three contributions to this likelihood, accounting for the possibilities that (i) there is preclinical onset of an indolent cancer prior to censoring, but the lesion is not detected on remaining screens until censoring; (ii) there is preclinical onset of a progressive cancer prior to censoring, but the lesion is neither detected on remaining screens nor does it progress to clinical disease until censoring; and (iii) there is no preclinical disease onset prior to the censoring date. Assembling the three contributions yields, for participant j,

$$\begin{aligned} L_j(\theta) = \psi &\sum_{k=2}^{l} (1-\beta)^{l-k} \left[F(t_k) - F(t_{k-1})\right] \\ &+ (1-\psi) \sum_{k=2}^{l} (1-\beta)^{l-k} \Omega(t_{k-1}, t_k, t_l, s, \theta) \\ &+ 1 - F(t_l), \end{aligned} \tag{5}$$

where

$$\begin{aligned}\Omega(t_a, t_b, t_c, s, \theta) = &\mathbb{1}_{t_a<s_1, t_b>s_0}\ R\left(\max(s_0, t_a), \min(s_1, t_b), h_0, s_0, \lambda, t_c\right) \\ &+ \mathbb{1}_{t_a<s_2, t_b>s_1}\ e^{-h_0(s_1-s_0)}\ R\left(\max(s_1, t_a), \min(s_2, t_b), h_1, s_1, \lambda, t_c\right) \\ &+ \mathbb{1}_{t_b>s_2}\ e^{-h_0(s_1-s_0)-h_1(s_2-s_1)}\ R\left(\max(s_2, t_a), t_b, h_2, s_2, \lambda, t_c\right),\end{aligned} \tag{6}$$

and

$$R(L, U, h, s, \lambda, t) = \frac{h}{\lambda - h}\left(e^{U(\lambda-h)} - e^{L(\lambda-h)}\right) e^{hs-\lambda t}. \tag{7}$$

2.1.2 Group 2: Screen-detected Cancers

We assign each participant j with a screen-detected cancer a vector of length $l = n + 1$

$$t^j \equiv t = (s_0, \tau_1, \ldots, \tau_n), \tag{8}$$

where τ_n is the age at the screen which leads to the cancer diagnosis.

There are two contributions to this likelihood, accounting for the possibilities that (i) there is preclinical onset of an indolent cancer that is detected on a subsequent screen; and (ii) there is preclinical onset of a progressive cancer that is detected on a subsequent screen, prior to clinical progression. Taken together, we obtain the following likelihood contribution for participant j

$$\begin{aligned}D_j(\theta) = &\beta\psi \sum_{k=2}^{l} (1-\beta)^{l-k} \left[F(t_k) - F(t_{k-1})\right] \\ &+ (1-\psi)\beta \sum_{k=2}^{l} (1-\beta)^{l-k} \Omega(t_{k-1}, t_k, t_l, s, \theta),\end{aligned} \tag{9}$$

where $\Omega(\ldots)$ is defined in (6).

2.1.3 Group 3: Clinical Cancers

To each participant with a clinical cancer diagnosis, we assign a vector of length $l = n + 2$

$$t^j \equiv t = (s_0, \tau_1, \ldots, \tau_n, \tau_c), \tag{10}$$

where τ_n is the time of the last screen before clinical diagnosis, and τ_c is the time of cancer diagnosis.

There is a single contribution to this likelihood, accounting for preclinical onset and clinical progression prior to screen detection. Differentiating between the two scenarios of preclinical onset taking place before or after the last screen prior to diagnosis, we obtain the following likelihood contribution for participant j

$$\begin{aligned} I_j(\theta) = (1-\psi)\sum_{k=2}^{l-1}(1-\beta)^{l-k}\left[\Omega(t_{k-1},t_k,t_{l-1},s,\theta)-\Omega(t_{k-1},t_k,t_l,s,\theta)\right] \\ +(1-\psi)\left[F(t_l)-F(t_{l-1})-\Omega(t_{l-1},t_l,t_l,s,\theta)\right]. \end{aligned} \tag{11}$$

2.1.4 Full Likelihood

Because only women without a history of clinical cancer were included in the cohort, we conditioned on not having had a clinical cancer diagnosis prior to the first screen in the likelihood using the normalizing factor (which is the same for participants of all groups)

$$z_j(\theta) = \psi + (1-\psi)\left[1-F(\tau_2)+\Omega(\tau_1,\tau_1,\tau_2,s,\theta)\right]. \tag{12}$$

Finally, denoting by $\mathcal{I}_L$, $\mathcal{I}_D$ and $\mathcal{I}_I$ the indices of participants in the non-cancer, screen-detected cancer and clinical cancer groups, respectively, we obtain the following expression for the full likelihood

$$LKH(\theta) \sim \sum_{j\in\mathcal{I}_L}\frac{L_j(\theta)}{z_j(\theta)} + \sum_{j\in\mathcal{I}_D}\frac{D_j(\theta)}{z_j(\theta)} + \sum_{j\in\mathcal{I}_I}\frac{I_j(\theta)}{z_j(\theta)}. \tag{13}$$

2.2 Prior Distributions

With the exception of the screening sensitivity we chose non-informative prior distributions for the model parameters. More specifically, for the preclinical onset rates h_i $(i = 1, 2, 3)$ and the progression rate λ from preclinical to clinical disease we used

$$h_i, \lambda \ \sim \text{Exp}(0.01), \tag{14}$$

and for the fraction of indolent cancers

$$\psi \sim \text{Beta}(1, 1). \tag{15}$$

The screening sensitivity of mammography has been studied extensively and is generally estimated to be around 70%-90%. Because using an informative prior for the sensitivity β not only reflects this prior knowledge but also enhances the practical identifiability for the other model parameters, we chose as its prior distribution

$$\beta \sim \text{Beta}(38.5, 5.8), \tag{16}$$

The mean of this prior distribution is 87%, reflective of an empirical sensitivity estimate derived from the Breast Cancer Screening Consortium of 86.9% (95% CI: 86.3-87.6%) [5]. Given that the modeled screening test sensitivity is a composite accuracy of not only the imaging modality but also that of the subsequent diagnostic work-up and the biopsy referral practice, we kept the prior moderately vague (95% CI: 75.7-95.0%).

2.3 Changepoint Selection for Preclinical Onset Rate

The change points of the piece-wise constant preclinical onset rate were selected from a set of biologically plausible scenarios by comparing the resulting model fits using Bayesian leave-one-out cross-validation (LOO-CV), see [6] and chapter 7 in [7].

For the selection of the changepoints $s = (s_0, s_1, s_2)$ in the piecewise constant formulation of the onset hazard in (1) we first identified a series of etiologically plausible scenarios:

- We assumed the onset hazard to be negligible prior to some age $s_0 > 0$. A natural lower bound for s_0 is the age at thelarche. However, because we modeled average-risk women only, we assumed the earliest age of onset to be between 35 to 45 years. We note that this does not preclude the accumulation of premalignant changes prior to this age; it merely implies that fully malignant and screen-detectable neoplasms do not arise prior to this age.
- The next change-point $s_1 > s_0$ was chosen to reflect the onset of menopause because of its implications for the biology and etiology of breast cancer [8]. Given a median age of menopause onset of 51 years [9] we considered 50 and 55 years as candidate ages for s_1.
- The final changepoint $s_2 > s_1$ was chosen such that the end of menopause would be completed in the vast majority of women by that age [8]. Further requiring an interval of at least 10 years between s_1 and s_2, we varied the latter changepoint between 60 and 65 years.

Based on these considerations, we considered the following etiologically plausible changepoint scenarios: $G1 = (35, 50, 65)$, $G2 = (40, 50, 65)$, $G3 = (35, 50, 60)$, $G4 = (40, 50, 60)$, $G5 = (45, 55, 65)$, $G6 = (40, 55, 65)$.

We then used Bayesian model comparison to select the overall best fitting scenario. More precisely, comparison was achieved through Bayesian leave-one-out cross-validation (LOO-CV) [10, 6] as implemented in the R package *loo* (version 2.4.1). Bayesian LOO-CV provides an estimate of the expected log predictive density (ELPD), which quantifies the out-of-sample predictive accuracy of the fitted model. Intuitively, the ELPD is a Bayesian analog of the commonly used frequentist predictive accuracy measures AIC and BIC. Higher ELPD values indicate a better fit to the data. Pairwise model comparisons were performed based on the ELPD differences and their standard errors. The results of the model comparison are shown in Supplementary Table 1 in Supplement 2; the best fitting changepoint model was $G6 = (40, 55, 65)$.

The key model parameters, including mean sojourn time and fraction indolent preclinical cancers, were largely insensitive to the specific choice of changepoints, see Supplementary Figure S1 in Supplement 2.

2.4 Parameter Identifiability and Model Checking

Parameter identifiability means that the combination of model structure and available data permits unique estimation of the model parameters [11, 12]. Following our prior work, for each model fit, we assessed local identifiability by checking that the marginal posterior parameter distributions (Supplementary Figure S2 in Supplement 2) were unimodal and had acceptably narrow 95% credible intervals [1]. We also examined the pairwise correlations of the model parameters to ensure that there were no strong correlations which could be indicative of identifiability issues.

Once identifiability was verified, we performed posterior predictive checks, a common procedure in Bayesian modeling that ascertains how well a model fits the data (29). In brief, we repeatedly sampled parameters from the full posterior distribution, used them to simulate natural histories for cohorts of women undergoing screening, and recorded the number of screen-detected and clinical cancers in each cohort. We then compared the resulting screen-detected and clinical cancer incidence in each screening round against the observed data and found that the model fit the data reasonable well, see Supplementary Figure S3 in Supplement 2.

3 Overdiagnosis Prediction

We accounted for two overdiagnosis contributions: one due to the detection of non-progressive (or indolent) cancers, and the other due to detection of progressive cancers that do not progress to

clinical disease before breast cancer unrelated death of the patient.

3.1 Indolent Cancer Contribution

We denote by $\mathcal{I}$ an indicator for the type of screen-detected cancer: $\mathcal{I} = 1$ for indolent cancers and $\mathcal{I} = 0$ for progressive cancers. We further denote by SD_n the event of a screen-detected cancer at screen n. With this notation, and applying Bayes' rule, we obtain the probability that a screen-detected cancer is indolent as

$$\mathbb{P}(\mathcal{I} = 1|SD_n) = \frac{\mathbb{P}\left(SD_n|\mathcal{I} = 1\right)\mathbb{P}\left(\mathcal{I} = 1\right)}{\mathbb{P}(SD_n)}, \tag{17}$$

where

$$\mathbb{P}(SD_n) = \mathbb{P}\left(SD_n|\mathcal{I} = 1\right)\mathbb{P}\left(\mathcal{I} = 1\right) + \mathbb{P}\left(SD_n|{=}0\right)\mathbb{P}\left(\mathcal{I} = 0\right). \tag{18}$$

In (17), $\mathbb{P}(\mathcal{I} = 1) = \psi$ and $\mathbb{P}(\mathcal{I} = 0) = 1 - \psi$ and the conditional probabilities of screen-detection given $\mathcal{I}$ correspond to the two terms in the likelihood expression for screen-detected cancers (9). More precisely, introducing the vector $t = (s_0, \tau_1, \ldots, \tau_n)$ of $l = n + 1$ we obtain

$$\mathbb{P}\left(SD_n|\mathcal{I} = 1\right) = \beta \sum_{k=2}^{l} (1 - \beta)^{l-k} \left[F(t_k) - F(t_{k-1})\right] \tag{19}$$

and

$$\mathbb{P}\left(SD_n|\mathcal{I} = 0\right) = \beta \sum_{i=2}^{l} (1 - \beta)^{l-k} \Omega(t_{k-1}, t_k, t_l, s, \theta). \tag{20}$$

Inserting equations (18) - (20) into (17) we obtain the overdiagnosis contribution due to the detection of an indolent cancer at screen n.

3.2 Progressive Cancer Contribution

Denote by U the sojourn time from preclinical onset to clinical progression and by T_n the time between the n^{th} screen and death from a breast cancer unrelated cause. The overdiagnosis contribution at screen n due to competing mortality is then expressed as the conditional probability of developing a progressive preclinical cancer prior to the time of the screen (τ_n) and to die from a

cause other than breast cancer before clinical progression

$$\begin{aligned}\mathbb{P}(\mathcal{I}=0, T_n < U | SD_n) &= \mathbb{P}(\mathcal{I}=0|SD_n)\ \mathbb{P}(T_n < U) \\ &= \frac{\mathbb{P}(SD_n|\mathcal{I}=0)\mathbb{P}(\mathcal{I}=0)}{\mathbb{P}(SD)}\ \mathbb{P}(T_n < U),\end{aligned} \tag{21}$$

where

$$\begin{aligned}\mathbb{P}(T_n < U) &= 1 - \mathbb{P}(T_n > U) \\ &= 1 - \int_0^\infty \mathbb{P}(T_n > t)\mathbb{P}(U = t)dt \\ &= 1 - \int_0^\infty \exp\left(-\int_{\tau_n}^{\tau_n+t} \xi_0(s)ds\right)\lambda e^{-\lambda t}dt,\end{aligned} \tag{22}$$

where $\xi_0(t)$ is the other cause mortality hazard function. Inserting equation (22), together with (18) - (20), into (21) yields the overdiagnosis contribution due competing mortality at screen n. Regarding implementation of the integral in (22), we used for $\xi_0(t)$ an age-cohort adjusted estimate of the risk of other cause death from a published study [13]. We used 1971 as the birth year of the simulated cohort for overdiagnosis estimation. In other words, in our annual and biennial scenarios the first screen (age 50) took place in 2021.

3.3 Programmatic Overdiagnosis

A screening program is characterized by the its start age a_i, its end age a_f and its screening frequency, which, assuming regular screening intervals, is determined by the total number of screens M between a_i and a_f. We define the overall overdiagnosis as the fraction of overdiagnosed among all screen-detected cancers. Importantly, we account for the event that a woman will die before reaching the last screen of the program. Denoting by SD the event of a screen-detected cancer during the screening period, and by ODX the event of an overdiagnosed screen-detected cancer during the screening period, the overall overdiagnosis of the program is

$$\mathbb{P}(ODX|SD) = \frac{\mathbb{P}(ODX)}{\mathbb{P}(SD)}. \tag{23}$$

To compute the two event probabilities, we define T as the age at death and find that

$$\mathbb{P}(SD) = \sum_{k=1}^{M} \mathbb{P}(SD_k)\mathbb{P}(T > \tau_k | T > a_i), \tag{24}$$

where $\mathbb{P}(SD_k)$ is found in (18) and

$$\mathbb{P}(T > \tau_k | T > a_i) = \exp\left(-\int_{a_i}^{\tau_k} \xi_0(s)ds\right). \tag{25}$$

Finally,

$$\begin{aligned}\mathbb{P}(ODX) = &\sum_{k=1}^{M} \mathbb{P}(\mathcal{I}=1, SD_k)\mathbb{P}(T > \tau_k | T > a_i) \\ &+ \mathbb{P}(\mathcal{I}=0, SD_k)\mathbb{P}(T \in [\tau_k, \tau_k + U] | T > a_i).\end{aligned} \tag{26}$$

4 Definition of Screen-Detected vs Interval Cancers

To classify diagnoses as either screen-detected or interval cancers, we used the `mode_of_detection` variable as coded by the BCSC, see page 9 of BCSC Standard Definitions, Version 3 [14]. This definition counts the last screen within 12 months as positive if it had a BI-RADS 3, 4 or 5. The BCSC considers 12 months to be most meaningful because most cancers diagnosed within one year after a BI-RADS 3 assessment were detected 6-8 months after the screening exam, consistent with detection via short-interval follow-up exam and not through symptomatic presentation [15]. As shown in the table below, only 9 (1.4%) of screen detected cancers are diagnosed between 9 and 12 months of the last mammogram.

	0-3m	3-6m	6-9m	9-12m
Screen-detected cancers, N (%)	601 (93.2)	8 (1.2)	27 (4.2)	9 (1.4)

Table 1: Time from last mammogram to screen-detected cancer (months)

References

[1] Marc D Ryser, Roman Gulati, Marisa C Eisenberg, Yu Shen, E Shelley Hwang, and Ruth B Etzioni. Identification of the fraction of indolent tumors and associated overdiagnosis in breast cancer screening trials. *American Journal of Epidemiology*, 188(1):197–205, 2019.

[2] HH Chen, SW Duffy, and L Tabar. A mover-stayer mixture of markov chain models for the assessment of dedifferentiation and tumour progression in breast cancer. *Journal of Applied Statistics*, 24(3):265–278, 1997.

[3] Anne Helene Olsen, Olorunsola F Agbaje, Jonathan P Myles, Elsebeth Lynge, and Stephen W Duffy. Overdiagnosis, sojourn time, and sensitivity in the copenhagen mammography screening program. *The Breast Journal*, 12(4):338–342, 2006.

[4] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1):1–32, 2017.

[5] Constance D Lehman, Robert F Arao, Brian L Sprague, Janie M Lee, Diana SM Buist, Karla Kerlikowske, Louise M Henderson, Tracy Onega, Anna NA Tosteson, Garth H Rauscher, et al. National performance benchmarks for modern screening digital mammography: update from the breast cancer surveillance consortium. *Radiology*, 283(1):49–58, 2017.

[6] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5):1413–1432, 2017.

[7] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

[8] Collaborative Group on Hormonal Factors in Breast Cancer et al. Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies. *The Lancet Oncology*, 13(11):1141–1151, 2012.

[9] Deborah Grady. Management of menopausal symptoms. *New England Journal of Medicine*, 355(22):2338–2347, 2006.

[10] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

[11] Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula Klingmüller, and Jens Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 2009.

[12] Andrew F Brouwer, Rafael Meza, and Marisa C Eisenberg. Parameter estimation for multistage clonal expansion models from cancer incidence data: A practical identifiability analysis. *PLoS computational biology*, 13(3):e1005431, 2017.

[13] Ronald E Gangnon, Natasha K Stout, Oguzhan Alagoz, John M Hampton, Brian L Sprague, and Amy Trentham-Dietz. Contribution of breast cancer to overall mortality for us women. *Medical Decision Making*, 38(1_suppl):24S–31S, 2018.

[14] BCSC Standard Definitions, Version 3. https://www.bcsc-research.org/application/files/6916/1429/6964/BCSC_Data_Definitions_v3__2020.09.23.pdf. Accessed: 2021-12-23.

[15] Brian L Sprague, Diana L Miglioretti, Christoph I Lee, Hannah Perry, Anna AN Tosteson, and Karla Kerlikowske. New mammography screening performance metrics based on the entire screening episode. *Cancer*, 126(14):3289–3296, 2020.