A model for the natural history of breast cancer: application to a Norwegian screening dataset

Laura Bondi[a], Marco Bonetti[b], and Solveig Hofvind[c]

[a]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK, and Dondena Research Center, Bocconi University, Milan, Italy; laura.bondi@mrc-bsu.cam.ac.uk
[b]Department of Social and Political Sciences, Dondena Research Center, and Bocconi Institute for Data Science and Analytics, Bocconi University, Milan, Italy; marco.bonetti@unibocconi.it
[c]Section of Breast Cancer Screening, Cancer Registry of Norway, Oslo and Department of Health and Care Sciences, UiT The Arctic University of Norway, Tromsø, Norway; sshh@kreftregisteret.no

Abstract

In this work we present some preliminary results on the analysis of data collected by BreastScreen Norway to estimate the natural history of the disease. Learning about the disease occurrence and evolution is crucial to identify the optimal screening schedule, with respect to the age range of the invited women and the lag between successive examinations. The model is a multi-state semi-Markov model with a cure rate structure, where the main quantities of interest are the probability of developing the disease, the age at start of asymptomatic detectability of the disease and the sojourn time, i.e. the time interval during which the disease is screen-detectable but not yet symptomatic.

Keywords: approximate Bayesian computation (ABC), breast cancer, cure rate model, disease history, multi-state model

1. Introduction

Cancer screening is defined as the examination of asymptomatic subjects in order to detect tumours before they become evident because of symptoms (1). In the past decades screening programs have been implemented in many countries, therefore making randomized trials difficult to perform and those performed difficult to evaluate due to changes in screening techniques and diagnostic tools. As a consequence, researchers can only rely on observational data collected administratively to learn about the natural history of the disease and to identify the optimal screening policy (in terms of the age range of the women invited, and lag between consecutive examinations), and risk-based tailoring of their schedules.

In this work, we apply a multi-state semi-Markov model proposed in (2) to the breast cancer screening data from Norway (BreastScreen Norway). This model aims at reconstructing the latent process of occurrence and development of breast cancer.

All times are measured from birth of the woman. For those women who do experience the disease, we assume that after the onset of the disease 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 to the start of asymptomatic detectability (through screening) of the disease (which we denote by T_A) and the time to the symptomatic detection of the disease (denoted by T_S). Between time T_A and T_S the tumour 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.

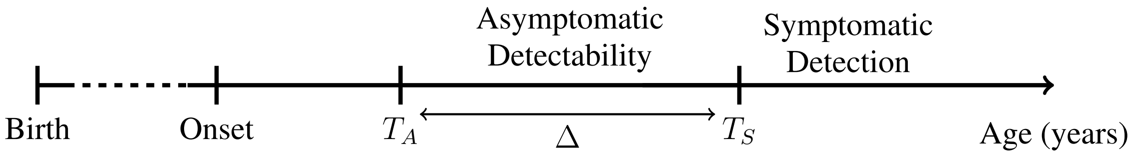

Figure 1: A graphical representation of the natural history from onset until detectability of the disease.

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. For this purpose, the model presents a cure rate structure, i.e. only a proportion of women, which we call the "susceptible proportion", denoted by p with $p \in (0, 1)$ will experience the event of developing breast cancer. Note that these should be considered to be latent cases and not necessarily observed cases.

Given by the limited follow-up, the lag between screening examinations and the impossibility to observe T_S when the tumour is screen-detected, this problem presents a complex missing data structure, which makes the observed data likelihood (3) of the model complicated if not intractable.

2. The Norwegian breast cancer screening data

Breast cancer screening in Norway was first introduced in 1995 (and became nationwide in 2005) and targets women aged 50-69 for biennial mammographic screening. We focus on a cohort of women born between 1948 and 1952 without any breast cancer diagnoses before entering the screening program at age 50 (with a small variability due to the invitations rounds). For these women we have information on their screening invitations, attendance and result for each examination, possible date of breast cancer diagnosis (DCIS or invasive), type of detection (in-screening or symptomatic), some additional tumour characteristics, as well as on a number of covariates collected through a questionnaire.

The screening invitations stop at age 69 but the breast cancer diagnoses are updated from the Cancer Registry until May 2022. Therefore the lentgh of follow-up is between 20 and 24 years. We focus only on invasive cancers and on three binary covariates, which divide the women in eight groups as shown in Table 1: 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 breast 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) (4).

Table 1 also shows the sample size, the proportion of observed detections, the proportion of symptomatic detections and the mean age at asymptomatic/symptomatic detections for each covariate group.

Group	(x_1, x_2, x_3)	Size	% Dx	% Symp Dx among all Dx	Mean age Asymp Dx	Mean age Symp Dx
1	(0,0,0)	1240	5.2%	27%	63.3	65.6
2	(0,0,1)	351	7.4%	35%	63.1	63.2
3	(0,1,0)	3731	4.9%	39%	62.9	65.3
4	(0,1,1)	1329	9.0%	38%	63.1	64.0
5	(1,0,0)	16241	3.8%	28%	63.8	66.0
6	(1,0,1)	4627	5.7%	36%	64.1	66.2
7	(1,1,0)	39330	4.2%	33%	63.7	65.1
8	(1,1,1)	13252	5.7%	35%	64.0	65.0
Total		80101				

Table 1: Observed outcomes in each covariate group. Ages are measured in years. X_1= at least one child birth (0:No, 1:Yes); X_2=Education level (0:Low, 1:Medium/High); X_3=Family history of breast cancer (0:No, 1:Yes).

3. Model and preliminary results

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 breast cancer," all coded as $0 =$ No and $1 =$ Yes. The susceptible proportion is modelled as function of the observed covariates $\boldsymbol{x} = (x_1, x_2, x_3)$ through the logit link:

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

For the subjects who will eventually develop the disease, the disease evolution is described through the joint distribution of the couple (T_A, Δ). The model assumes that the mean of T_A depends on the covariates linearly, and that the variance of T_A is constant across covariate groups. The distribution of Δ is defined conditionally on the observed value of T_A, and it may depend on the covariates but only indirectly (see below). The definition of the model for the disease process is as follows:

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

where $E(T_A) = \mu(\boldsymbol{x}) = 100 \cdot \frac{\alpha}{\alpha+\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, $\sigma^2 = 100^2 \cdot \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. This model, called "Rescaled beta + piecewise exponential" model, has been introduced in (2), where it was selected among ten competing models to describe breast cancer natural history. We refer to this reference for additional details on the model, such as the prior distribution for the 12 parameters $(\beta_0, \beta_1, \beta_2, \beta_3, \sigma, \lambda_1, \lambda_2, \lambda_3, p_o, p_1, p_2, p_3)$.

Approximate posterior distributions for the model parameters are obtained via likelihood-free inference. More specifically, relying on approximate Bayesian computation (ABC) allows us to avoid deriving the complex observed data likelihood of the model. Details on the metric function employed to measure the distance between the real and generated data are not shown here, but a complete description, together with a discussion of the advantages and drawbacks of ABC in this setting, can be found in (2). The results presented here are based on 200,000 generated datasets.

Figure 2 shows the boxplots of the posterior distributions for the mean of T_A, $\mu(\boldsymbol{x})$, and for the susceptible proportion $p(\boldsymbol{x})$ across the eight covariate groups. The left panel of Figure 2 highlights that women with at least one child tend to experience breast cancer later than those without children. From the right panel, instead, clearly emerges that having a family history of breast cancer is associated with an increased risk of belonging to the group of the susceptible subjects. Note that the posterior distributions of the p's are shifted toward slightly higher values than expected from previous studies (see e.g. (5)). This is likely due to the difficulties in estimating a weakly identified cure rate model with limited follow-up, i.e. little data at higher ages.

Given the posterior distributions, one can then compute the predictive distributions for the quantities of interest, T_A and Δ (see Figure 3). The median predicted values for T_A vary between 69.2 (in group 4) and 72.5 (in group 6). The standard deviation of T_A is estimated to be around 7.5 years. The predictive distributions for Δ in the three groups defined by the value of T_A have medians 2.4, 4.5 and 0.9 years, respectively. While the first two appear to be in line with what is known by previous studies, the predictive distribution for Δ in the third group does not seem very reliable. Similarly to what was highlighted when commenting the posterior distributions for the p's, this can be probably attributed to the lack of information about T_A for tumours occurring after the age of 65, given that screening examinations stop at age 69 .

4. Conclusions

We have estimated a parametric model to describe the insurgence and the evolution of breast cancer, where the main quantities of interest are the probability of developing the disease (p), the start of asymptomatic detectability of the disease and the time of symptomatic detection (T_A and T_S). Given the estimated latent disease process, it is possible to simulate the effect of different screening schedules, possibly tailored to the individual risk factors, on the number and kind of diagnoses in the populations.

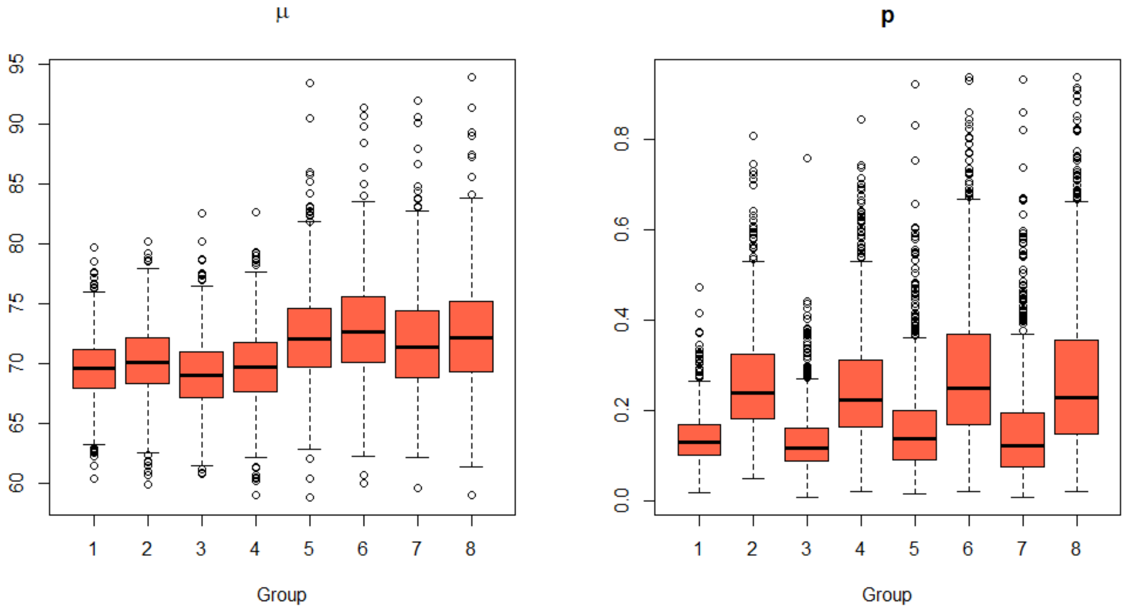

Figure 2: Approximate posterior distributions of the mean age at asymptomatic detectability $\mu(\boldsymbol{x})$ and of the susceptible proportion $\mathrm{p}(\boldsymbol{x})$ across covariate groups.

This work might be extended by including in the model a parameter for the sensitivity of the screening examinations. Indeed, the assumption of a 100% sensitivity, as we made here, corresponds to setting the probability of false negative results equal to zero, which is not realistic (6). Moreover, additional covariates, such as breast density, could be included in the model.

Other interesting analyses could be based on different cohorts of women included in the BreastScreen Norway database, which could confirm the estimated disease process or highlight differences in the occurrence between different generations.

Also, assuming a stable disease population (in which the rate of births and the distribution of ages at tumour onset are constant across calendar time (7)), one could analyse a larger cohort of women at once to exploit the information provided by women at different ages and therefore to obtain a more precise inference. This would however come with a higher computational cost and a longer time needed to perform the ABC simulations.

Future developments of this work will also involve the numerical computation of the observed data likelihood function to perform exact Bayesian inference through MCMC.

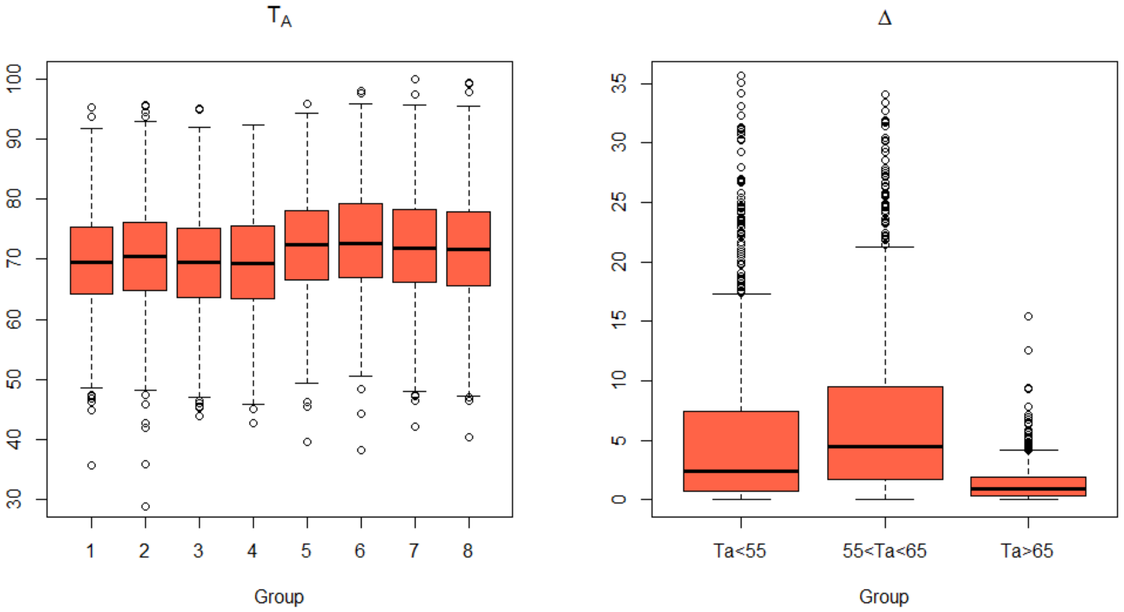

Figure 3: Approximate predictive distributions of the age at asymptomatic detectability T_A and of the sojourn time Δ across covariate groups.

References

[1] Van Oortmarssen G, Boer R and Habbema J. Modelling issues in cancer screening. *Statistical Methods in Medical Research.* 1995; 4(1): 33–54.

[2] Bondi L, Bonetti M, Grigorova D and Russo A. Approximate Bayesian Computation (ABC) for the natural history of breast cancer, with application to data from a Milan cohort study. *Statistics in Medicine.* 2023; In press.

[3] Little R and Rubin D. *Statistical analysis with missing data.* Second ed. New York: Wiley, 2002.

[4] Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute.* 1989; 81(24):1879-86.

[5] Howlader N, Noone AM, Krapcho M, et al. SEER Explorer. Breast Cancer-Stage Distribution of SEER Incidence Cases, 2007-2016 by Sex. *National Cancer Institute, Bethesda*. 2019.

[6] Abrahamsson L and Humphreys K. A statistical model of breast cancer tumour growth with estimation of screening sensitivity as a function of mammographic density. *Statistical Methods in Medical Research.* 2013; 25(4): 1620–1637.

[7] Isheden G and Humphreys K. Modelling breast cancer tumour growth for a stable disease population. *Statistical Methods in Medical Research.* 2019; 28(3): 681–702.