Title

Nonhomogeneous Markov Chain for Estimating the Cumulative Risk of Multiple False Positive Screening Tests

Journal

Biometrics, 78(3)

ISSN

0006-341X

Authors

Golmakani, Marzieh K

Hubbard, Rebecca A

Miglioretti, Diana L

Publication Date

2022-09-01

DOI

10.1111/biom.13484

Peer reviewed

HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2022 October 04.

Published in final edited form as:

Biometrics. 2022 September ; 78(3): 1244–1256. doi:10.1111/biom.13484.

Non-Homogeneous Markov Chain for Estimating the Cumulative Risk of Multiple False Positive Screening Tests

Marzieh K Golmakani[1,*], **Rebecca A Hubbard**[2,**], **Diana L Miglioretti**[3,***]

[1]Pfizer Inc., San Diego, CA

[2]Department of Biostatistics, Epidemiology & Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA

[3]Department of Public Health Sciences, University of California at Davis, Davis, CA

Summary:

Screening tests are widely recommended for the early detection of disease among asymptomatic individuals. While detecting disease at an earlier stage has the potential to improve outcomes, screening also has negative consequences, including false positive results which may lead to anxiety, unnecessary diagnostic procedures and increased healthcare costs. In addition, multiple false positive results could discourage participating at subsequent screening rounds. Screening guidelines typically recommend repeated screening over a period of many years, but little prior research has investigated how often individuals receive multiple false positive test results. Estimating the cumulative risk of multiple false positive results over the course of multiple rounds of screening is challenging due to the presence of censoring and competing risks, which may depend on false positive risk, screening round and number of prior false positive results. To address the general challenge of estimating the cumulative risk of multiple false positive test results, we propose a non-homogeneous multi-state model to describe the screening process including competing events. We developed alternative approaches for estimating the cumulative risk of multiple false positive results using this multi-state model based on existing estimators for cumulative risk of a single false positive. We compared the performance of the newly proposed models through simulation studies and illustrate model performance using data on screening mammography from the Breast Cancer Surveillance Consortium. Across most simulation scenarios, the multi-state extension of a censoring bias model demonstrated lower bias compared to other approaches. In the context of screening mammography, we found that the cumulative risk of multiple false positive results is high. For instance, based on the censoring bias model, for a high risk individual, the cumulative probability of at least two false positive mammography results after 10 rounds of annual screening is 40.4

[*] kgolmakani@ucdavis.edu . [**] rhubb@pennmedicine.upenn.edu . [***] dmiglioretti@ucdavis.edu .

Supporting Information

Web Appendices, Tables, and Figures referenced in Sections 3 – 7 are available with this paper at the Biometrics website on Wiley Online Library. These include additional technical details, simulation results and data analyses. Code to reproduce all results is available online at https://github.com/kgolmakani/CumulativeRisk.git.

Keywords

BCSC; Censoring; Competing risk; Mammography; Multi-state model; Stochastic model

1 Introduction

The goal of a screening test is to identify the presence of undiagnosed disease among asymptomatic individuals and advance the time of diagnosis. While the benefit of early disease detection can be substantial if treatments are more effective for early compared to advanced stage disease , this benefit is offset, in part, by negative consequences of screening including adverse effects of the test itself, such as radiation exposure from imaging tests, overdiagnosis, overtreatment and false positive test results. An ideal screening test would result in a positive result if and only if the individual actually had the disease and a negative if and only if the individual did not have the disease; however, screening tests typically fall short of this ideal. For most screening tests, the most common harms are false positive results, which occur when the screening test erroneously indicates that the disease is present and may lead to anxiety, additional testing and associated medical costs, and invasive diagnostic procedures. When screening tests are used to diagnose a rare disease such as cancer, even a very specific test can result in many more false positive results than true positive results. For example, for screening mammography, the false positive rate is 110 per 1000 screens while the true positive rate is 5.1 per 1000 screens (Lehman et al. (2016)).

Screening tests are typically repeated many times over the course of an individual's lifetime. For example, the American Cancer Society recommends that women between 45 and 54 years old with an average risk of breast cancer undergo annual screening mammography and women 55 years and older either transition to biennial screening or continue screening annually, depending on preferences, until a life expectancy of less than 10 years (Oeffinger et al. (2015)). In addition, they recommend that women age 40 - 44 years should have the choice to start annual screening based on their personal preferences. A woman who complies with these recommendations and is breast cancer free through age 79 would undergo up to 40 screening mammograms. Given false positive results are common for screening mammography, women may experience multiple false positive results over the course of repeated screens which could discourage continued screening (Klompenhouwer et al. (2014). Multiple false positive screening mammograms also increase unnecessary radiation exposure from diagnostic imaging and risk of radiation-induced cancers (Miglioretti et al. (2016)). Examining the impact of repeat screening is thus important in order to quantify the cumulative benefits and burdens of different recommended screening regimens based on screening interval, starting age, and stopping age. Estimation of the cumulative risk of multiple false positive results using observational data is complicated because most individuals will be censored before completing all recommended screening rounds and due to possible dependence of false positive risk and competing risks, such as diagnosis of the disease of interest and death, on censoring time and the number of prior false positive results.

Existing estimators of cumulative risk of a single false positive result include a discrete survival model (Gelfand and Wang (2000)), a population average model (Xu et al. (2004) and Hubbard et al. (2010)) and a censoring bias model (Hubbard and Miglioretti (2013)). The discrete survival model is limited by an assumption of independence of event and censoring time, which is often violated in the case of medical screening test for several reasons. For example, individuals receiving a false positive may be more or less likely to return for additional screening (Burman et al. (1999)). In the context of screening mammography, women at higher breast cancer risk might be more likely to return for additional screening and might have a different probability of a false positive result than a lower risk women. Also, individuals who are screened more frequently will have more observed screening exams during the study period by definition, and for screening mammography, the false positive probability decreases as the screening interval decreases (Hubbard et al. (2010)). Proposed population average approaches relax the assumption of independent censoring but are limited by parametric assumptions about variation in risk across screening rounds (Hubbard et al. (2010) and Xu et al. (2004)). Population average models estimate the total false positive risk associated with the screening program if all eligible individuals were to participate in all recommended rounds of screening rather than assuming uncensored individuals are representative of censored individuals. The model proposed by Xu et al. (2004) assumes constant risk across screening rounds after censoring which is unrealistic in most screening contexts. For example, false positive risk is substantially higher at the first screening mammogram compared to subsequent mammograms because comparison images are not available for assessing change (Hubbard and Miglioretti (2013)). Hubbard et al. (2010) relaxed the assumption of constant risk by modeling risk as a function of screening round, but this assumption is unverifiable for unobserved screening rounds. The censoring bias model is a more flexible semi-parametric approach that relaxes the strong assumptions required by other models. It allows for dependent censoring without imposing a fixed functional form for variation in risk across screening rounds. However, all existing models can only estimate the cumulative risk of a single false positive screening test. Although the stochastic model developed by Miglioretti et al. (2016) estimated the probability of false positive screening results followed by additional imaging or biopsy, it did not account for dependent censoring and the number of prior false positive results. Thus, appropriate methods are needed to estimate the cumulative probability of experiencing multiple false positives.

In this paper, we propose a non-homogeneous Markov chain for modeling repeat screening test results and use this model to develop estimators of the cumulative risk of receiving multiple false positive results across a recommended program of repeat screening that addresses dependent censoring and competing risks. We compare the performance of these estimators under eight scenarios for variation in risk as a function of screening round, censoring time and number of prior false positive results. We also illustrate the performance of these models for estimating the cumulative risk of multiple false positive results using data collected by six breast imaging registries in the U.S. Breast cancer Surveillance Consortium (BCSC). We conclude with a summary and discussion of our approach and findings.

2 Definitions and notation

For ease of notation, we suppress the index for the subject throughout. Let S represent the censoring time, defined as the total number of screening rounds observed for each subject, and W_ℓ represent the event time, defined as the screening round at which each subject receives the ℓ^{th} false positive result. Let Y_j represent the outcome of the j^{th} screen for each subject, taking values 0 if a true negative result (TN), 1 if a false positive result (FP), and 2 if the disease of interest is diagnosed (DD). Let $\mathbf{Y}_j = (Y_1, \ldots, Y_j)$ represent the vector of all screening outcomes up to round j. We assume subjects are observed for a maximum of M rounds. In addition, we assume that subjects who do not receive ℓ false positive results by round M would receive them at arbitrary rounds after M. The cumulative risk of receiving ℓ false positive results over the course of k rounds of screening is defined as $P(W_\ell \leqslant k)$.

3 A multi-state model for repeat screening

In this section, we propose a non-homogeneous Markov chain to describe transition across states defined by the results of a repeated screening test. Let Y_j, the possible outcomes of a screening test at round j (0 = TN, 1 = FP, 2 = DD), represent the states of a Markov chain. Disease diagnosis (DD) is treated as a competing event for false positive results and can be considered an absorbing state. We define transition across these states over multiple rounds of screening using a first order Markov chain.

First, we define the baseline model which describes the probability of receiving a false positive, a true negative or disease diagnosis at the first screening round. This baseline model depends on whether or not an individual is censored at the first round.

Let p_{y0} be the probability of receiving an outcome y_0 at the first screening round, $j = 1$, for an individual censored at round $S \geqslant 2$. Since follow-up ends once an individual is diagnosed with the disease, we need to define p_{y0} for individuals with $S \geqslant 2$ only for $y_0 =$ 0 or 1; then we have $p_1 = \text{expit}(\beta_{0(1)} + \beta_{1(1)}S)$ and $p_0 = 1 - p_1$, where $\text{expit}(x) = \frac{\exp(x)}{1+\exp(x)}$. We also define the probability of receiving a false positive ($y_0 = 1$) or disease diagnosis ($y_0 = 2$) at the first screening round for an individual censored at round $S = 1$ as $p'_{y0} = \frac{\exp(\beta_{0(y_0)})}{1 + \sum_{y_0=1}^{2} \exp(\beta_{0(y_0)})}$; Then, the probability of receiving a true negative ($y_0 = 0$) is written as $p'_0 = \frac{1}{1 + \sum_{y_0=1}^{2} \exp(\beta_{0(y_0)})}$. In order to provide an estimation of personalized risks, subject characteristics can also be incorporated into the baseline probability models.

Next, we describe the transition probabilities for round $j = 2\ldots, S$ conditional on the state of the prior round. Let $p_{y_1y_2} = P(Y_j = y_2 | Y_{j-1} = y_1)$ where $\sum_{y_2=0}^{2} p_{y_1y_2} = 1$ for $y_1 = 0, 1, 2$ and first order Markov property holds; that is,

$$P(Y_j = y_j \mid Y_1 = y_1, \ldots Y_{j-1} = y_{j-1}) = P(Y_j = y_j \mid Y_{j-1} = y_{j-1}).$$

Since disease diagnosis is considered as an absorbing state, we have $p_{20} = p_{21} = 0$ and $p_{22} = 1$. In order to accommodate variation in transition probabilities across screening rounds and allow for different transition probabilities for patients with varying censoring times, we introduce two transition probability matrices with elements dependent on screening round and censoring time, one for all transitions prior to the final observed transition and one for the final transition. Because disease diagnosis is an absorbing state, it is known a priori that the probability of transition to this state prior to the last observed screening round is 0. Thus, for transition probabilities at round j for an individual censored at round S where $2 \leqslant j < S$, we define the transition probability matrix $P = \begin{pmatrix} 1-p_{01} & p_{01} & 0 \\ 1-p_{11} & p_{11} & 0 \\ 0 & 0 & 1 \end{pmatrix}$. The transition probabilities in this matrix are defined as functions of screening round and censoring time, such that for $y_1 = 0, 1$ and $y_2 = 1$ we have $p_{y_1y_2} = \text{expit}(\beta_{0(y_1y_2)} + \beta_{1(y_1y_2)}S + \beta_{2(y_1y_2)}j + \beta_{3(y_1y_2)}Sj)$.

Finally, if $S \geqslant 2$ and $j = S$ we define the following transition matrix

$$P' = \begin{pmatrix} 1-(p'_{01}+p'_{02}) & p'_{01} & p'_{02} \\ 1-(p'_{11}+p'_{12}) & p'_{11} & p'_{12} \\ 0 & 0 & 1 \end{pmatrix}.$$

For $y_1 = 0, 1$ and $y_2 = 1, 2$ the transition probability in the matrix above is defined as $p'_{y_1y_2} = \frac{\exp(\beta_{y_1y_2}^T U)}{1 + \sum_{y_2=1}^{2} \exp(\beta_{y_1y_2}^T U)}$, where $U^T = (1, S, j, Sj)$ and $\beta_{y_1y_2}^T = (\beta_{0(y_1y_2)}, \beta_{1(y_1y_2)}, \beta_{2(y_1y_2)}, \beta_{3(y_1y_2)})$.

Let $p_{y_1y_2}^{(h,j)}$ represent the probability that state y_1 is followed by y_2 in round j when the total number of screening rounds attended is $S = h$. $p_{y_1y_2}^{(h,j)}$ constitutes the elements of matrices P and P'. We define

$$p_{y_1y_2}^{(i,j)} = \begin{cases} p_{y0} & 1 = j < h \\ p'_{y0} & j = h = 1 \\ p_{y_1y_2} & 2 \leqslant j < h \\ p'_{y_1y_2} & 2 \leqslant j = h \end{cases} \tag{1}$$

Further, assume that $n_{y_1y_2}^{(h,j)}$ is the number of observations when state y_1 is followed by y_2 in round j when the censoring time is $S = h$. Then, the likelihood function can be written as

$$L = \prod_{h=1}^{M} \prod_{y_1=0}^{2} \prod_{y_2=0}^{2} \{ P(S=h)^{\sum_{j=1}^{h} n_{y_1y_2}^{(h,j)}} \prod_{j=1}^{M} p_{y_1y_2}^{(h,j)^{n_{y_1y_2}^{(h,j)}}} \}. \tag{2}$$

The estimates of $P(S = h)$ and $p_{y_1y_2}^{(h,j)}$ can be calculated by maximizing L subject to the constraint $\sum_{h=1}^{M} p(S = h) = 1$ and $\sum_{y_2=0}^{2} p_{y_1y_2}^{(h,j)} = 1$; thus, we have $\hat{P}(S = h) = \frac{\sum_{y_1=0}^{2}\sum_{y_2=0}^{2}\sum_{j=1}^{h} n_{y_1y_2}^{(h,j)}}{\sum_{h=1}^{M}\sum_{y_1=0}^{2}\sum_{y_2=0}^{2}\sum_{j=1}^{h} n_{y_1y_2}^{(h,j)}}$ and $\hat{p}_{y_1y_2}^{(h,j)} = \frac{n_{y_1y_2}^{(h,j)}}{\sum_{y_2=0}^{2} n_{y_1y_2}^{(h,j)}}$. When $j \leqslant h$, this estimate can be directly calculated from the data.

For More detailed estimation procedure please see Web Appendix A.

4 Using the Markov model to estimate cumulative risk of multiple FPs

Several methods have been used to estimate the cumulative false positive risk associated with repeat screening tests. The cumulative risk of a single false positive result can be considered the cumulative incidence function in a discrete survival model. The event time is the screening round at which the first false positive test result occurs and the censoring time is the number of screening rounds observed for a participant. In this section, we use the proposed stochastic process formulation to develop alternative estimators for the cumulative risk of multiple false positive test results, following the approaches to handling dependence of false positive risk on screening round and censoring time previously proposed for the case of a single false positive test result. The cumulative risk of experiencing at least ℓ false positive results in k screens ($\ell \leqslant k$) can be written as

$$\begin{aligned} P(W_\ell \leqslant k) &= \sum_{h=1}^{M}\sum_{j=\ell}^{k} P(W_\ell = j \mid S = h)P(S = h) \\ &= \sum_{h=1}^{M}\sum_{j=\ell}^{k}\sum_{C_{\ell j}} p_{i_1}^{(h,1)} \prod_{m=2}^{j} p_{i_{(m-1)}i_m}^{(h,m)} P(S = h) \end{aligned} \tag{3}$$

where $C_{\ell j}$ represents the set of all possible combinations of states $i_1,\ldots,i_j$ such that $i_j = 1$ (means a false positive) and exactly ℓ of them are equal to 1. For example, if $h = 1$, $\ell = 2$ and $j = 3$ then, $C_{\ell j}$ has two tuples ($i_1 = 1$, $i_2 = 0$, $i_3 = 1$) and ($i_1 = 0$, $i_2 = 1$, $i_3 = 1$). Thus, the summand only includes these cases for the states.

Multi-state extensions of “discrete time survival model” and “population average model” are described in Web Appendix B.

4.1 Censoring bias model

The semi-parametric censoring bias model was first proposed for discrete time-to-event data under dependent censoring by Scharfstein et al. (2001) and was later adapted to the context of modeling risk of a single false positive screening test result by Hubbard and Miglioretti (2013), who showed that this model has greater flexibility compared to the other existing models as it allows for both dependent censoring and changes in false positive risk across screening rounds without requiring parametric assumptions about variation in risk across screening rounds following censoring. In this model, false positive risk is assumed to vary across screening rounds in the same way for subjects with $S = h$ rounds as for subjects with $S > h$ rounds of screening. However, variation in risk is not constrained to follow a specific

functional form (Hubbard and Miglioretti (2013)). Under this model, the cumulative risk of at least ℓ false positive test results can be written as

$$P(W_\ell \leqslant k) = \sum_{h=1}^{M} \sum_{j=\ell}^{k} P^*(W_\ell = j \mid S = h) P(S = h)\,. \tag{4}$$

To fully identify this model, we assume that

if $j \leqslant h$, then $P^*(W_\ell = j \mid S = h) = \sum_{C_{\ell j}} p_{i_1}^{(h,1)} \prod_{m=2}^{j} p_{i_{(m-1)} i_m}^{(h,m)}$;

if $j > h$, then $P^*(W_\ell = j \mid S = h) = \sum_{r=0}^{min\{\ell-1,\, h\}} P(W_\ell = j \mid S = h, n_h = r) P(n_h = r \mid S = h)$.

Where n_h is the number of false positives at or before round h. In order to identify the cumulative probability of W_ℓ this model assumes a relationship between the non-identified and identifiable components of equation (4). In other words, for $j > h$, this model identifies $P(W_\ell = j \mid S = h, n_h = r)$ by using information from all subjects with more than h screening rounds and the same number of false positives at or before round h. For $h = 1, \cdots, M-1$ and $j > h$ we define

$$P(W_\ell = j \mid S = h, n_h = r) = \frac{P(W_\ell = j \mid S > h, n_h = r) \exp(g_{hj}(\alpha_{\ell - r}))}{\sum_{i=\ell}^{M+1} \exp(g_{hi}(\alpha_{\ell - r})) P(W_\ell = i \mid S > h, n_h = r)} \tag{5}$$

where $P(W_\ell = M+1 \mid S = h, n_h = r)$ is defined to be $P(W_\ell > M \mid S = h, n_h = r)$. $\alpha_{\ell - r}$ is called the censoring bias parameter and $g_{hi}(\alpha_{\ell - r})$ is called the censoring bias function which can be any positive valued function. The censoring bias function specifies the relationship between the false positive risk among subjects with $S = h$ and those with $S > h$. Scharfstein et al. (2001) showed that the followings hold for the censoring bias function, $g_{hi}(\alpha_{\ell - r})$:

1. Specification of $g_{hi}(\alpha_{\ell - r})$ identifies the distribution of W_ℓ
2. In equation (4), $g_{hi}(\alpha_{\ell - r})$ is not identified because all choices of $g_{hi}(\alpha_{\ell - r})$ are compatible with the law of observed data. Hence, no statistical test can reject any specific choice of censoring bias function.
3. In addition, by specifying $g_{hi}(\alpha_{\ell - r})$ we do not place any restrictions on the law of the observed data.

In our analysis, we use $g_{hi}(\alpha_{\ell - r}) = \alpha_{\ell - r}(i - (h+1))$. Parameter $\alpha_{\ell - r}$ in the censoring bias function can facilitate sensitivity analysis and is interpreted as the conditional log odds ratio of dropping out between time h and $h+1$ per one round increase in i for subjects who received $r < \ell$ false positive results at or before round h. Since in the case of screening tests, S is always observed, $\alpha_{\ell - r}$ is estimable. In our analysis, we use $\alpha_{\ell - r} = (\ell - r)\alpha$. We estimate α directly from the data. Using (5) and Bayes' rule we have

$$\text{logit}\,(P(S = h \mid S \geqslant h, w_\ell = j, n_h = r)) = g_{hj}(\alpha_{\ell - r}) + C_h \tag{6}$$

where C_h is a constant. Note that when $j \leqslant k$ we have

$$P(W_\ell = j \mid S = k, n_h = r) = \sum_{C_{\ell j}} p_{i1}^{(k,1,r)} \prod_{m=2}^{j} p_{i(m-1)im}^{(k,m,r)}$$

where $p_{y_1y_2}^{(k,m,r)} = P(y_m = y_2 \mid y_{m-1} = y_1, S = k, n_h = r)$.

All three models discussed in this section and Web Appendix B can easily be extended to allow for estimation of the predicted cumulative probability of multiple false positive results individualized for a given set of patient characteristics by including patient covariates in the regression model for $p_{y_1y_2}^{(h,j)}$. Hubbard et al. (2010), Xu et al. (2004) and Hubbard and Miglioretti (2013) provided variance estimators in the case of at least one false positive result in discrete survival, population average and censoring bias models, respectively. For the case of multiple false positive screening tests, variance estimates can be derived via the delta method. In simulations below, delta method standard errors were used. Since the analytic variance formulas are only available for models without covariates, in our application to BCSC data, bootstrap standard errors were used.

5 Simulation study

We conducted simulations to evaluate the performance of our three proposed models for cumulative risk of multiple false positive screening tests under a variety of scenarios for the relationship between false positive risk, screening round, censoring time, and the number of prior false positive test results. We considered 10 different scenarios that include potential deviations from the multi-state model assumptions. The target of inference is the cumulative probability of at least two false positive results after 10 and 5 rounds of screening, that is $P(W_2 \leqslant 10)$ and $P(W_2 \leqslant 5)$ respectively. The small sample properties of these models are important to understand because, even in a large sample, the number of subjects observed across many rounds may be small. For the censoring bias model we estimated a using equation (6). We compared bias and efficiency of these models for cumulative risk of multiple false positives under 10 scenarios. The 10 simulation scenarios are described in Table 1.

In scenarios 3 - 6 and 9 the assumptions of all three models are violated. Table 2 provides a summary of the dependence relationships assumed by the 10 simulation scenarios and whether the data generating mechanism satisfies the modeling assumptions.

We selected values for transition probabilities and other parameters of the data generating model based on values estimated from real data on repeat screening mammography provided by the BCSC. Please see Web Appendix C for detailed description of transition probabilities and parameter values used for each simulation scenario. In all simulation scenarios we generated a cohort of 50,000 subjects. Estimates for our simulation study are based on 5000 simulated data sets for each scenario. For each simulation scenario we calculated bias relative to the true cumulative probability of two false positives after 10 and 5 rounds of screening, theoretical standard errors, and empirical standard errors computed across the 5000 simulation iterations.

5.1 Simulation Study Results

Simulation study results for cumulative risk of two false positive results after 10 rounds of screening are presented in Table 3. We first considered scenarios where false positive risk was independent of censoring. When false positive risk was also independent of number of prior false positives and was constant across screening rounds (Scenario 1), all three models were unbiased. With variable false positive risk across screening rounds but false positive risk independent of the number of prior false positives (Scenario 2), the discrete survival model and censoring bias model exhibited lower bias than the population average model. When false positive risk was constant across screening rounds but false positive risk depended on the number of prior false positive results (Scenario 3), the population average model had low bias while the other two models had moderate bias under both moderate and strong dependency. With variable risk across screening rounds and false positive risk dependent on the number of prior false positives (Scenario 4) the censoring bias model was unbiased while the discrete survival model showed moderate bias and the population average model demonstrated higher bias. We next considered scenarios with dependent censoring. When false positive risk was also dependent on number of prior false positives with variable risk across screening rounds (Scenario 5), the censoring bias model had very little bias under both moderate and strong dependency, compared to the other two models. With constant risk across screening rounds but with false positive risk dependent on the number of prior false positives (Scenario 6), the discrete survival model performed poorly while the population average model and censoring bias model had very little bias and both performed better under moderate dependency. In Scenario 7, when false positive risk was constant across screening rounds but depended on the number of prior false positives, the population average model and censoring bias model demonstrated significantly lower bias compared to the discrete survival model. Under strong dependency, the population average model and censoring bias model performed similarly. However, under moderate dependency, the censoring bias model performed significantly better. Under variable risk across screening rounds with false positive risk independent of number of prior false positives (Scenario 8), the censoring bias model demonstrated reasonably low bias compared to the other models especially when the dependency was strong. In scenario 9, when false positive risk was constant across screening round but censoring time depended on the number prior false positive results, all three models exhibited low bias. Finally, in scenario 10, when we generated the data assuming second order Markov model but fitted the models assuming a first order Markov model, censoring bias model had lower bias compared to the other two models.

In summary, the discrete survival and population average models performed well when their assumptions were satisfied but poorly when their assumptions were violated. In comparison, the censoring bias model performed well across all scenarios. In all scenarios, theoretical standard errors tended to underestimate empirical standard errors. This is likely due to the fact that all estimates rely heavily on the false positive risk among subjects who are observed across all screening rounds (Hubbard and Miglioretti (2013)). This group will be small if the proportion censored at each screening round is large. The censoring bias model was also generally less efficient than the other models, although this difference was small for most scenarios. In almost all scenarios, the censoring bias model demonstrated low bias.

Simulation study results for cumulative risk of two false positive results after 5 rounds of screening which are shown in Web Table 3 suggest similar conclusions as Table 3.

The goodness of fit assessment of the multi-state models are provided in Web Appendix D.

6 Application to the BCSC

We illustrate the performance of our methods using data collected by six breast imaging registries in the BCSC: (1) The North Carolina Mammography Registry, (2) the New Hampshire Mammography Network, (3) the San Francisco Mammography Registry, (4) Kaiser Permanente Washington Registry, (5) the Vermont Breast Cancer Surveillance System, (6) the Metro Chicago Breast Cancer Registry. These registries link information on women who receive a mammogram at a participating facility to state cancer registries and pathology databases to determine breast cancer outcomes. We included women who had their first screening mammogram between the ages of 40 and 74 years at a participating BCSC facility. We included this first screening mammogram along with subsequent screening mammograms meeting inclusion criteria performed from 2000 to the most recent year with complete breast cancer capture, which varied from 2010 - 2014 across the six mammography registries. Mammograms were classified as positive or negative using standard BCSC definitions (BCSC (2018)) based on the initial Breast Imaging Reporting and Data Systems (BI-RADS) assessment and recommendations assigned by the radiologist. A positive mammogram was considered to be false positive if the woman was not diagnosed with breast cancer within 1 year after the index mammogram and prior to the next screening mammogram. A negative mammogram was considered to be true negative if the woman was not diagnosed with breast cancer within 1 year after the index mammogram and prior to the next screening mammogram. This negative mammogram would be considered a true negative and any subsequent mammogram for a women with true negative would also be included in the analysis. The interval between screening mammograms was categorized as 9 – 18 months (approximately annual), 19 – 30 months (approximately biennial), or no prior mammogram within 30 months. Breast cancer diagnoses, including invasive carcinoma or ductal carcinoma in situ, were treated as competing events given we are interested in evaluating breast cancer screening in women without a history of breast cancer.

6.1 Summary of observed multiple false positive rates

We included 168,716 women who each received between 1 and 10 screening mammograms over the study period for a total of 359,842 mammograms. Across the six mammography registries, the number of women received screening mammograms varied from 7,420 - 66,835 and the total number of screening mammograms varied from 19,247 - 123,031. Some characteristics of this cohort such as the distribution of baseline age, race/ethnicity, family history of breast cancer and breast density and number of observed rounds of screening along with the proportion of women within each category with a false positive mammography result at their first mammogram are presented in Table 4. A summary of these characteristic across the six mammography registries is provided in Web Table 4. The majority of women were observed for one or two rounds of screening (73.8%) while only 10.9% were observed to receive five or more mammograms. The number of women

observed to receive five or more mammograms varied from 8,289 - 39,439 across the six registries. The probability of receiving a false positive result at the first screening round was somewhat lower for women with five or more screening rounds compared to women with fewer screening rounds. This was also the case for receiving false positive results at subsequent screens. Table 4 shows that black women had a higher probability of receiving false positive results at their first and subsequent screenings. The probability of receiving a false positive at the first screen was also higher for women with a family history of breast cancer and women with heterogeneously dense breasts at baseline. For each woman in the study we identified the reason study follow-up had ended. Follow-up ended due to loss to follow-up (no further screening) (93.5%), end of study (3.3%), breast cancer diagnosis (2.2%) or death (1.0%).

The relationship between false positive risk, screening round, and censoring time is illustrated in Figure 1. Risk decreases substantially between the first and second screening rounds, regardless of the number of observations available per subject. At any individual screening round, false positive risk appears lower for women with more observed screening rounds and higher for women with fewer observed rounds.

Web Figure 2 (a) and (b) show the observed probability of multiple false positive results across screening round, $P(W_i = 1 | S \geq i)$, and the empirical cumulative probability of multiple false positive test results, $P(W_i \leq 1 | S \geq i)$ respectively. The probabilities of first and second false positive results were highest at the first (24.6%) and second (3.6%) screening round respectively and decreased at subsequent screens.

Adjusted odds ratios for baseline characteristics are provided in Web Appendix E.

We applied each of the three estimators for cumulative risk of multiple false positive results to the BCSC cohort to estimate cumulative risk after 10 rounds of annual and 5 rounds of biennial screening. To illustrate how covariates can be incorporated into each model, we provide personalized risk estimates. Similar to the simulation studies, we used different transition matrices for the first and subsequent screening rounds. However, in the analysis of BCSC, we also incorporate patients characteristics into the transitional matrices.

We modeled false positive risk conditional on baseline age, family history of breast cancer, breast density, race/ethnicity and interval between screening mammograms. We also assumed false positive risk depends on the number of prior false positive exams. Under the same settings, we also utilized the three proposed models to calculate the probability of receiving at least two consecutive false positive results after 10 rounds of annual and 5 rounds of biennial screening. Table 5 shows the estimated cumulative probability of multiple false positive mammography results and bootstrap confidence intervals after 10 rounds of annual screens and 5 rounds of biennial screens for groups at high and low risk of a false positive result. We defined the high risk group for receiving a false positive result as 40 year old, non-Hispanic black women with a family history of breast cancer and heterogeneously dense breasts. The low risk group for receiving a a false positive result was defined as 50 year old, Asian women with no baseline family history of breast cancer and almost entirely fat breasts. For the censoring bias model, a was estimated for each number of

false positive results using equation (6). For the cumulative risk of one false positive, the estimated risk was lowest using the discrete survival model whereas estimated risk based on the population average model was the highest, with the censoring bias model intermediate between the two. However, the pattern was different for estimates of multiple false positive results. For estimates of the cumulative probability of multiple false positive results, a large proportion of the sample is censored and this proportion increases as the target number of false positives increases. In addition, the probability of receiving a greater number of false positives increases with greater number of screening rounds. As a result, estimates of the cumulative probability of four and five false positives are larger based on the censoring bias model compared to the population average model due to its assumption of constant risk following censoring.

Web Table 7 shows the estimated risk of receiving at least two consecutive false positive mammography results and bootstrap confidence intervals after 10 rounds of annual and 5 rounds of biennial screenings for the defined low and high risk group. The estimated risk of receiving at least two consecutive false positives was lowest using the discrete survival model whereas the estimated risk based on the population average model was the highest, with the censoring bias model intermediate between the two.

7 Discussion

We proposed a non-homogeneous Markov chain to represent the process of repeat screening and used this framework to develop three models for estimating the cumulative risk of multiple false positive screening tests. We compared the performance of these three models using simulations under 10 different scenarios varying our assumptions about the censoring mechanism, variation in risk across screening rounds and dependence of false positive risk on prior false positive results. In most scenarios, the censoring bias approach was nearly unbiased and had lower bias and best goodness of fit than other approaches. In particular, when false positive risk depends on censoring time and number of prior false positive results and varies across screening rounds (Scenario 5), the discrete survival and population average models both showed substantial bias while the censoring bias model eliminated most of the bias. In addition, when censoring time depends on false positive risk and risk varies across screening rounds, the censoring bias model is nearly unbiased while the other two models exhibit substantial bias especially under strong dependency. A strength of the Markov chain approach to describing the screening process is its straightforward ability to incorporate competing risks. By specifying an absorbing state, we can easily distinguish between patients who have been censored due to drop-out from screening, loss to follow-up, or the end of study follow-up and those who have ceased screening due to cancer diagnosis or death. In the context of breast cancer screening where relatively few women will be diagnosed with cancer compared to the number censored, accounting for competing risks has relatively little practical impact on estimates of cumulative false positive risk (Hubbard and Miglioretti (2013)). However, if applied in contexts where competing events have higher prevalence such as screening for high blood pressure, appropriately distinguishing between competing events and censoring is critical. It is also straightforward to incorporate subject characteristics into each estimator in order to provide personalized risk prediction. We applied this approach to the mammography setting by comparing estimates of cumulative

risk of multiple false positive results after 10 rounds of annual and 5 rounds of biennial exams with developing breast cancer as a competing risk. We reported cumulative false positive risks for women with high and low risk profiles. The estimated cumulative risk of a single false positive is lowest using the discrete survival model and highest using the population average model. The discrete survival model is likely to underestimate risk since it assumes that women observed over more screening rounds are representative of those censored earlier. In contrast, the population average model assumes that risk is constant across screening rounds which is unlikely to be true in the mammography setting. However, for the cumulative risk of a larger number of false positives, the population average model tends to underestimate risk, especially when the proportion of censored individuals is significantly higher than uncensored. In addition, for the cumulative risk of a larger number of false positives, the risk among biennial screeners becomes significantly smaller and the relative risk in biennial compared to annual screeners decreases significantly.

In this paper we reported the cumulative risk after 10 years of screening mammograms. Women who are breast cancer free through age 79 and comply with screening guidelines would undergo up to 40 screening mammograms. The proposed Markov model can be used to estimate the cumulative risk of multiple false positive results over longer screening periods. However, this would require either data observed over a longer period or extrapolation.

Understanding the risk of receiving multiple false positive test results has important implications evaluating effectiveness of regimens consisting of many rounds of repeat screening. Individuals who begin screening at younger ages are at risk for experiencing a greater total number of false positive test results due to the greater total number of screening tests they will be exposed to and possible dependence of false positive risk on the number of prior false positive results. This may cause individuals to become less interested in future screening, decreasing adherence to screening recommendations (Klompenhouwer et al. (2014)) and, consequently, reducing screening regimen effectiveness. Our new approach makes an important contribution to evaluating screening strategies by providing methods to estimate multiple false positive results as a function of screening round, censoring time, number of prior false positives and other women’s characteristics in the presence of competing risks.

It should be noted that, in the proposed approach transition probabilities are described conditional on the total number of rounds of screening that an individual participates in which is fundamentally future information.

Software in the form of R code scripts used to produce the results of this paper has been deposited to Github (Golmakani (2020)).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This research was funded by the National Cancer Institute (P01CA154292). Data collection for this work was additionally supported, in part, by funding from the National Cancer Institute (U54CA163303) and the Agency for Health Research and Quality (R01 HS018366-01A1). The collection of cancer and vital status data used in this study was supported in part by several state public health departments and cancer registries throughout the U.S (BCSCregistries (2018)). We thank the participating women, mammography facilities and radiologists for the data they provided.

Data Availability Statement

Research data are not shared. Readers may request a subset of the data by contacting the Statistical Coordinating Center for the Breast Cancer Surveillance Consortium at kpwa. scc@KP.org. More information can be found visiting the BCSC website at https://www.bcsc-research.org.

References

BCSC (2018). BCSC Glossary of Terms, Version 2. http://www.bcsc-research.org/data/bcsc_data_definitions_version2_final__2017.pdf. [Online; accessed 25-April-2021].

BCSCdata (2018). Statistical Coordinating Center for the Breast Cancer Surveillance Consortium. kpwa.scc@KP.org.

BCSCregistries (2018). public health departments and cancer registries. http://www.bcsc-research.org/work/acknowledgement.html.

BCSCwebsite (2018). BCSC Website. https://www.bcsc-research.org.

Burman ML, Taplin SH, Herta DF, and Elmore JG (1999). Effect of false-positive mammograms on interval breast cancer screening in a health maintenance organization. Annals of internal medicine 131, 1–6. [PubMed: 10391809]

Gelfand AE and Wang F (2000). Modelling the cumulative risk for a false-positive under repeated screening events. Statistics in medicine 19, 1865–1879. [PubMed: 10867676]

Golmakani MK (2020). Cumulative risk. https://github.com/kgolmakani/CumulativeRisk.git.

Hubbard RA and Miglioretti DL (2013). A semiparametric censoring bias model for estimating the cumulative risk of a false-positive screening test under dependent censoring. Biometrics 69, 245–253. [PubMed: 23383717]

Hubbard RA, Miglioretti DL, and Smith RA (2010). Modelling the cumulative risk of a false-positive screening test. Statistical methods in medical research 19, 429–449. [PubMed: 20356857]

Klompenhouwer EG, Duijm LE, Voogd AC, den Heeten GJ, Strobbe LJ, Louwman MW, Coebergh JW, Venderink D, and Broeders MJ (2014). Reattendance at biennial screening mammography following a repeated false positive recall. Breast cancer research and treatment 145, 429–437. [PubMed: 24748569]

Lehman CD, Arao RF, Sprague BL, Lee JM, Buist DS, Kerlikowske K, Henderson LM, Onega T, Tosteson AN, Rauscher GH, et al. (2016). National performance benchmarks for modern screening digital mammography: update from the breast cancer surveillance consortium. Radiology 283, 49–58. [PubMed: 27918707]

Miglioretti DL, Lange J, Van Den Broek JJ, Lee CI, Van Ravesteyn NT, Ritley D, Kerlikowske K, Fenton JJ, Melnikow J, De Koning HJ, et al. (2016). Radiation-induced breast cancer incidence and mortality from digital mammography screening: a modeling study. Annals of internal medicine 164, 205–214. [PubMed: 26756460]

Oeffinger KC, Fontham ET, Etzioni R, Herzig A, Michaelson JS, Shih Y-CT, Walter LC, Church TR, Flowers CR, LaMonte SJ, et al. (2015). Breast cancer screening for women at average risk: 2015 guideline update from the american cancer society. Jama 314, 1599–1614. [PubMed: 26501536]

Scharfstein D, Robins JM, Eddings W, and Rotnitzky A (2001). Inference in randomized studies with informative censoring and discrete time-to-event endpoints. Biometrics 57, 404–413. [PubMed: 11414563]

Xu J-L, Fagerstrom RM, Prorok PC, and Kramer BS (2004). Estimating the cumulative risk of a false-positive test in a repeated screening program. Biometrics 60, 651–660. [PubMed: 15339287]

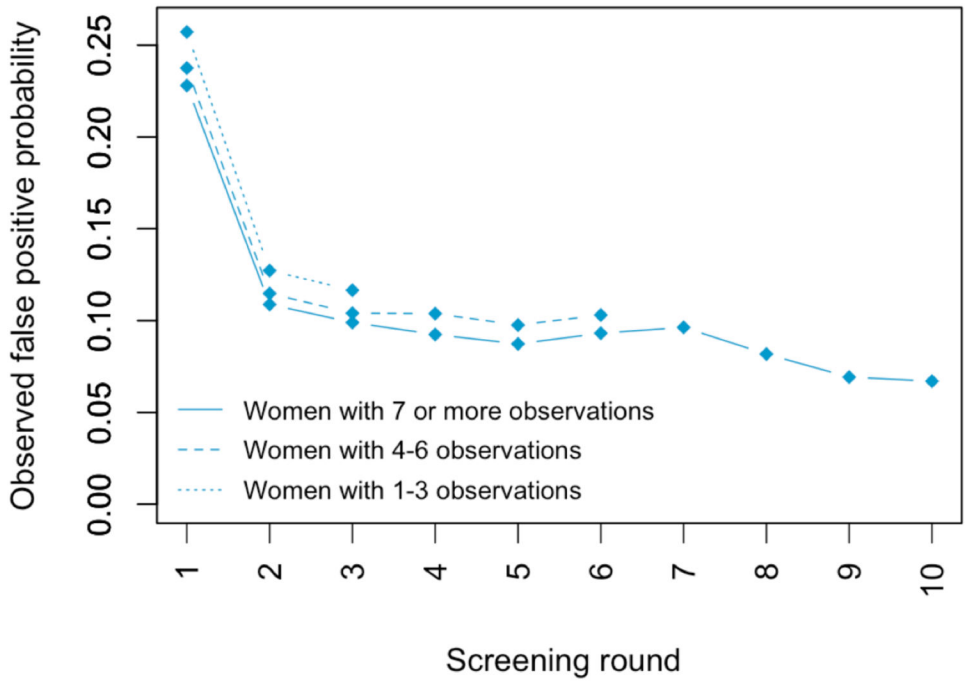

Figure 1:
False positive risk at screening rounds 1-10 by censoring time. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

Table 1:

Descriptions of the simulation scenarios used to evaluate the performance of the three proposed models.

Scenario	Description
1	False positive risk is independent of censoring time and number of prior false positive results and is constant across screening rounds. This scenario satisfies the assumptions of all three models.
2	False positive risk depends on screening round but is independent of censoring time and number of prior false positive results. This scenario violates the assumption of the population average model.
3	False positive risk depends on the number of prior false positive results but is independent of screening round and censoring time.
4	False positive risk depends on the number of prior false positive results and screening rounds but is independent of censoring time.
5	False positive risk depends on the number of prior false positive results, censoring time and screening round.
6	False positive risk depends on the number of prior false positive results and censoring time but is constant across screening rounds.
7	False positive risk depends on censoring time and is independent of screening round and number of prior false positive results. This scenario violates the assumptions of the discrete survival model.
8	False positive risk depends on censoring time and screening round but is independent of number of prior false positive results. This scenario violates the assumptions of the discrete survival model and population average model.
9	False positive risk depends on the censoring time and independent of screening round. Further, we assume that censoring time depends on the number of prior false positives.
10	Similar to the first scenario, false positive risk is independent of censoring time and number of prior false positive results and is constant across screening rounds. In this scenario, the data was generated assuming second order Markov chain which means the result of the current screen depends on the result of the prior two screens. However, we fitted the models assuming a first order Markov chain (first order Markov model assumption was violated).

Table 2:

Data generating mechanism for the 10 simulation scenarios. For each scenario, models with assumptions satisfied by the data generating mechanism are listed under Model. Abbreviations: false positive (FP), discrete survival (DS), population average (PA), censoring bias (CB).

Scenario	FP risk depends on:			Model
	censoring	round	prior FPs	
1	X	X	X	DS, PA, CB
2	X	✓	X	DS, CB
3	X	X	✓	———
4	X	✓	✓	———
5	✓	✓	✓	———
6	✓	X	✓	———
7	✓	X	X	PA, CB
8	✓	✓	X	CB
9	✓	X	✓	———
10 *	X	X	X	DS, PA, CB

* First order Markov model assumption was violated for this scenario.

Table 3:

Relative bias, theoretical standard errors (TSE), and empirical standard errors (ESE) for three estimators for cumulative risk of two false positive after ten rounds of screening for simulated data.

Models	Strong dependence			Moderate dependence		
	Relative Bias	TSE	ESE	Relative Bias	TSE	ESE
Scenario 1 *						
Discrete survival	−0.053	0.001	0.004			
Population average	−0.001	0.001	0.005			
Censoring bias	0.003	0.004	0.005			
Scenario 2						
Discrete survival	−0.057	0.001	0.003	−0.054	0.001	0.004
Population average	0.338	0.001	0.005	0.150	0.001	0.005
Censoring bias	0.039	0.003	0.004	0.017	0.004	0.004
Scenario 3						
Discrete survival	−0.037	0.001	0.004	−0.048	0.001	0.004
Population average	−0.006	0.001	0.005	−0.006	0.001	0.005
Censoring bias	0.012	0.004	0.005	0.016	0.004	0.005
Scenario 4						
Discrete survival	−0.050	0.001	0.004	−0.049	0.001	0.004
Population average	0.165	0.001	0.005	0.082	0.001	0.005
Censoring bias	0.007	0.003	0.004	0.003	0.004	0.004
Scenario 5						
Discrete survival	−0.420	0.003	0.009	−0.335	0.004	0.010
Population average	0.152	0.001	0.007	0.063	0.001	0.007
Censoring bias	0.011	0.003	0.017	0.019	0.005	0.017
Scenario 6						
Discrete survival	−0.329	0.004	0.013	−0.244	0.005	0.017
Population average	−0.095	0.001	0.007	−0.063	0.001	0.007
Censoring bias	−0.044	0.006	0.017	0.029	0.007	0.019
Scenario 7						
Discrete survival	−0.333	0.004	0.014	−0.242	0.005	0.017
Population average	−0.049	0.001	0.007	−0.039	0.001	0.007
Censoring bias	−0.042	0.005	0.018	0.006	0.007	0.019
Scenario 8						
Discrete survival	−0.418	0.003	0.009	−0.333	0.004	0.013
Population average	0.149	0.001	0.007	0.060	0.001	0.007
Censoring bias	−0.004	0.004	0.017	−0.009	0.006	0.018
Scenario 9						
Discrete survival	−0.005	0.001	0.004	−0.035	0.001	0.004
Population average	0.017	0.001	0.005	0.084	0.001	0.004
Censoring bias	0.041	0.004	0.008	0.018	0.003	0.008
Scenario 10 *						

Models	Strong dependence			Moderate dependence		
	Relative Bias	TSE	ESE	Relative Bias	TSE	ESE
Discrete survival	−0.202	0.001	0.004			
Population average	−0.148	0.001	0.005			
Censoring bias	−0.069	0.004	0.009			

* There is no dependency for this scenario.

Table 4:

The distribution of baseline age, race/ethnicity, family history of breast cancer, and breast density and number of observed rounds of screening for 168,716 women along with the proportion of women within each category with false positive mammography results at their first and subsequent mammograms.

Factor	N	%	% FP at first screen	% FP at subsequent screens
Baseline age				
40 — 49 years	130,921	77.6	24.4	11.1
50 — 59 years	37,795	22.4	28.0	9.3
Race/ethnicity				
White	88,390	52.4	25.6	11.0
Black	27,711	16.4	30.2	11.90
Asian/Pacific Islander	16,931	10.0	18.3	8.5
American Indian/Alaska Native	597	0.4	26.6	8.4
Hispanic	14,423	8.5	23.9	9.8
Mixed/other	20,664	12.2	23.1	11.1
Baseline family history of breast cancer				
Yes	11,047	6.5	28.7	10.9
No	157669	93.4	24.9	10.8
Baseline BI-RADS breast density				
Almost entirely fat	9,112	5.4	15.9	6.8
Scattered fibroglandular	59,492	35.3	25.6	10.1
Heterogen dense	82,147	48.7	27.7	11.8
Extremely dense	17,965	10.6	17.1	10.3
Screening round per woman				
1	95,601	56.7	25.6	
2	28,958	17.2	25.8	13.1
3	15,890	9.4	24.6	11.7
4	9,933	5.9	24.1	11.1
⩾ 5	18,334	10.9	23.2	9.7

Table 5:

Estimated cumulative probability of multiple false positive mammography results after 10 rounds of annual screening and 5 rounds of biennial screening (95% confidence intervals) for high and low risk groups

Model	High risk group		Low risk group	
	Annual	Biennial	Annual	Biennial
One FP				
Discrete survival	72.72 (72.07, 73.35)	59.85 (59.32, 60.39)	36.50 (35.82, 37.18)	27.34 (26.86, 27.81)
Population average	83.77 (83.43, 84.14)	66.98 (66.52, 67.43)	49.38 (48.68, 50.10)	31.76 (31.20, 32.32)
Censoring bias	75.02 (74.30, 75.77)	60.75 (60.19, 61.31)	38.15 (37.39, 38.92)	27.76 (27.25, 28.26)
Two FP				
Discrete survival	39.73 (38.89, 40.57)	22.50 (21.99, 22.98)	8.63 (8.16, 9.11)	4.45 (4.21, 4.70)
Population average	51.99 (50.98, 52.99)	24.90 (24.22, 25.58)	13.61 (12.88, 14.30)	4.85 (4.56, 5.17)
Censoring bias	40.38 (39.61, 41.15)	23.94 (23.45, 24.44)	11.28 (10.38, 12.28)	5.47 (5.16, 5.79)
Three FP				
Discrete survival	16.59 (15.90, 17.26)	5.61 (5.32, 5.91)	1.55 (1.33, 1.78)	0.50 (0.42, 0.59)
Population average	25.73 (24.81, 26.65)	6.07 (5.74, 6.41)	2.81 (2.52, 3.11)	0.47 (0.41, 0.54)
Censoring bias	20.96 (19.75, 22.17)	8.11 (7.30, 8.97)	3.28 (2.80, 3.73)	0.89 (0.70, 1.12)
Four FP				
Discrete survival	5.82 (5.35, 6.28)	0.94 (0.82, 1.08)	0.25 (0.15, 0.35)	0.04 (0.02, 0.07)
Population average	9.91 (9.30, 10.51)	0.87 (0.77, 0.97)	0.43 (0.35, 0.51)	0.03 (0.02, 0.04)
Censoring bias	11.16 (10.24, 12.09)	1.97 (1.63, 2.38)	1.50 (1.23, 1.59)	0.14 (0.09, 0.21)
Five FP				
Discrete survival	1.79 (1.52, 2.07)	0.09 (0.05, 0.13)	0.04 (0.02, 0.06)	< 0.001
Population average	2.96 (2.67, 3.26)	0.06 (0.05, 0.07)	0.05 (0.04, 0.06)	< 0.001
Censoring bias	6.04 (5.43, 6.73)	0.30 (0.20, 0.40)	0.80 (0.63, 1.02)	< 0.001