Project number: 874662
Project Acronym: HEAP

Project Title: Human Exposome Assessment Platform
Project Website URL:
Project Coordinator: Joakim Dillner
Organisation: KI
E-mail: Joakim.Dillner@ki.se

Work Package 6
Informatics platform and Knowledge Engine (PaaS, SaaS)

Work package Leader: Jim Dowling
Organisation: Logical Clocks
E-Mail: jim@hopsworks.ait

Project Deliverable

D6.5: Advanced statistics analysis on Finnish Maternity Cohort and Swedish Cervical Screening Cohort (part A)

Deliverable due date: 2023-12-31
Deliverable due month: M48

Document history

Version	Date	Changes	By	Reviewed
1.0	2023-12-29	First draft	Mark Clements	Joakim Dillner Roxana Merino

Executive Summary

The research question is whether we can use detailed cervical screening data to predict the risk for cervical cancer and the risk for high-grade lesions. The broad analytical approach is an adaptation of the screening models developed by Day and Walter during the 1980s. To our knowledge, our approach is novel for cervical cancer screening. We provide (a) SQL and SAS code for the data extraction from the Swedish Cervical Cancer Screening Register and (b) R and C++ code for the likelihood construction, optimisation, and predictions. As a proof of concept, we fit the model to the cohort of women born in 1960 who were living in Sweden on their fifteenth birthday. We found some issues with fitting the model due to the lack of identifiability of the model parameters. We propose some extensions to the current approach, including scaling up the computations to more birth cohorts and extending the model to include negative biopsies. Finally, we conclude that such a mathematical approach could be compared with predictions based on machine learning and evaluate whether the machine learning algorithm gives sufficient weight to different screening histories.

Table of Contents

1 Introduction

The research question is whether we can use detailed cervical screening data to predict the risk for cervical cancer and the risk for high-grade intraepithelial (HSIL) lesions. The broad analytical approach is an adaptation of the screening models developed by Day and Walter (1984). The adaptation includes an addition of an additional state to allow for HSIL, together with a screening path from HSIL after treatment. To our knowledge, our approach is novel for cervical cancer screening. We also provide code for: (a) the data extraction from the Swedish Cervical Cancer Screening Register (the register name in Swedish is abbreviated "NKCx") and (b) R and C++ code for the likelihood construction, optimisation, and predictions.

For screening, we assume primary cytology testing. A positive cytology test would typically be followed by a colposcopy and, where indicated, a biopsy. The biopsy would then provide the disease status.

In the outline, we describe the available data and develop the mathematical model. We then provide some early results from the model fitting and give some discussion on the path forward. We also include appendices for further mathematical developments. The full report is provided as a PDF document, including diagrams and a full mathematical development. This document provides an outline of the full report.

2 Methods

2.1 Data

Individual-level data from NKCx were linked with Swedish health and population registers. For each individual, we had data on: the month and year of birth; the date of first immigration to Sweden and the date of first emigration from Sweden; the date of death; dates of cytology testing; for histology, the date of sampling, the topography (cervix) and morphology (HSIL or cancer). These data were combined to create screening histories that included: year of birth; a maximum follow-up time, which was the earliest of the date of first cancer based on histology, date of emigration, date of death, and 2022-12-31; the sample dates for cytology; the date of first and second HSIL histology; and the date of first cervical cancer diagnosis. We classified the screening histories into seven categories for which we will likelihood components: see Table 1 of the full report. We restricted to women who were living in Sweden on their 15th birthday. For a preliminary analysis, we further restricted the analysis to those who were born in 1960.

For data management, the NKCx data were stored on an SQL Server database. Data manipulation used a combination of SQL on the server and SAS datasteps. The SQL and SAS code used for this manipulation is provided on https://github/mclements/nkcx_analysis.

3 Summary

We have developed and implemented a mathematical model for cervical cancer screening and fitted that model individual-level data from the Swedish National Cervical Cancer Screening Register. The model fit demonstrated issues with identifiability of the model parameters and should be considered as a first step in the application of the proposed model framework.

As a strength of our approach, we have used a detailed mathematical model to use individual-level screening data. To our knowledge, this is the first mathematical model fitted to individual-level data from the Swedish Cervical Cancer Screening Register. There are several limitations to our approach. First, we were not able to resolve the lack of identifiability for the model parameters. Consequently, our fitted model A lacked plausibility and, in our opinion, would not provide reasonable predictions. As a next step, we would like to further investigate the identifiability of the model parameters. Another consequence of the lack of identifiability is that the covariance

matrix for the parameters was not invertible, so that it was difficult to represent the uncertainty in the model parameters. We could investigate model uncertainty using the bootstrap, which would add a further layer of computational complexity. Second, there have been changes in cervical cancer screening tests over time, particularly the change from conventional cytology to liquid-based cytology around 2011, and the introduction of primary HPV testing from age 30 years at around 2017. Moreover, younger birth cohorts will be partially vaccinated for HPV, which will change the onset distribution for HSIL. Any applications of the model should take those factors into consideration.

3.1 Extensions

An obvious extension would be to fit these models to more birth cohorts. This would be an increasingly large computational task and would require the use of a computer cluster. The C++ code is well suited for multi-core computations on Linux servers. It is possible that the C++ code could be adapted to use general purpose graphics processing units (GPUs). The use of automatic differentiation libraries such as PyTorch of JAX may be useful for any calculations on GPUs -- and would have the advantage of calculating gradients for optimisation of the log-likelihood.

The implementation could also be simplified by noting that the current fitted models are also Markov models. We could then use Kolmogorov's forward differential equation to solve the model, with step changes at the screening occasions. This will be simpler, and may be faster, than the current more general implementation.

We have provided a very specific implementation for cervical cancer screening. The R and C++ code could be generalised to: (a) the model due to Day and Walter (1984) that includes individual-based screening patterns, for use with traditional cancer screening, where pre-cancerous lesions can not be treated; and (b) prostate cancer screening, which could include a sub-model for prostate-specific antigen. It is unclear whether this class of screening models are potentially useful for the Finnish Maternity Cohort.

As a possible extension, we could copy the NKCx datasets to Hopsworks. DuckDB is a columnar-based embedded database that is available on Hopsworks. The version of SQL on DuckDB is very similar to that for SQL Server, which would aid in the translation. The main challenge in the translation is that the SAS datastep code would need to be translated to

either Python or R. To our knowledge, no version of the NKCx database has been put on a Hopsworks instance. Hopsworks would have some clear advantages if a GPU-based implementation were available.

3.2 Conclusions

We have provided a proof-of-concept for mathematical modelling of cervical cancer screening. The predictions from a well-fitted model could be compared with a prediction model based on machine learning. For such a comparison, one could evaluate whether the machine learning approach provides different weight for the different screening histories.

4 References

Day, N. E. and Walter, S. D. (1984). Simplified models of screening for chronic disease: estimation procedures from mass screening programmes. *Biometrics*, 40:1–13.

Mathematical modelling for cervical cancer screening: Analysis using the Swedish National Cervical Cancer Register

Mark Clements

December 2023

Abstract

The research question is whether we can use detailed cervical screening data to predict the risk for cervical cancer and the risk for high-grade lesions. The broad analytical approach is an adaptation of the screening models developed by Day and Walter during the 1980s. To our knowledge, our approach is novel for cervical cancer screening. We provide: (a) SQL and SAS code for the data extraction from the Swedish Cervical Cancer Screening Register; and (b) R and C++ code for the likelihood construction, optimisation, and predictions. As a proof of concept, we fit the model to the cohort of women born in 1960 who were living in Sweden on their fifteen birthday. We found some issues with fitting the model due to lack of identifiability of the model parameters. We propose some extensions to the current approach, including: scaling up the computations to more birth cohorts; and extending the model to include negative biopsies. Finally, we conclude that such a mathematical approach could be compared with predictions based on machine learning and evaluate whether the machine learning algorithm gives sufficient weight to different screening histories.

1 Introduction

The research question is whether we can use detailed cervical screening data to predict the risk for cervical cancer and the risk for high-grade intraepithelial (HSIL) lesions. The broad analytical approach is an adaptation of the screening models developed by Day and Walter [1]. The adaptation includes an addition of an additional state to allow for HSIL, together with a screening path from HSIL after treatment. To our knowledge, our approach is novel for cervical cancer screening. We also provide code for: (a) the data extraction from the Swedish Cervical Cancer Screening Register (the register name in Swedish is abbreviated "NKCx"); and (b) R and C++ code for the likelihood construction, optimisation, and predictions.

For screening, we assume primary cytology testing. A positive cytology test would typically be followed by a colposcopy and, where indicated, a biopsy. The biopsy would then provide the disease status.

In outline, we describe the available data and develop the mathematical model. We then provide some early results from the model fitting and give some discussion on the path forward. We also provide appendices for further mathematical developments.

2 Methods

2.1 Data

Individual-level data from NKCx were linked with Swedish health and population registers. For each individual, we had data on: the month and year of birth; the date of first immigration to Sweden

and the date of first emigration from Sweden; the date of death; dates of cytology testing; for histology, the date of sampling, the topography (cervix) and morphology (HSIL or cancer). These data were combined to create screening histories that included: year of birth; a maximum follow-up time, which was the earliest of the date of first cancer based on histology, date of emigration, date of death, and 2022-12-31; the sample dates for cytology; the date of first and second HSIL histology; and the date of first cervical cancer diagnosis. We classified the screening histories into seven categories for which we will likelihood components: see Table 1. We restricted to women who were living in Sweden on their 15th birthday. For a preliminary analysis, we further restricted the analysis to those who were born in 1960.

For data management, the NKCx data were stored on an SQL Server database. Data manipulation used a combination of SQL on the server and SAS datasteps. The SQL and SAS code used for this manipulation is provided on `https://github.com/mclements/nkcx_analysis`.

2.2 Natural history model

Let the modelled states at age t be represented by: $W(t)$ for the general "healthy" population, that also includes women with active infections due to human papillomavirus (HPV) and low-grade squamous intraepithelial lesions; $X(t)$ for HSIL; $Y(t)$ for preclinical cancer; and $Z(t)$ for diagnosed cervical cancer. Following Day and Walter [1], we assume that the transitions for the natural history model have an ordering, such that $W \to X \to Y \to Z$, and that the transitions are semi-Markov, based on time in state rather than age. Let the probability density functions and survival functions: for $W \to X$ be f_1 and S_1, respectively; for $X \to Y$ be f_2 and S_2, respectively; and for $Y \to Z$ be f_3 and S_3, respectively. In the absence of screening and using the subscript "NH" to denote natural history, we then have:

$$
\begin{aligned}
W_{\mathrm{NH}}(t) &= S_1(t) \\
X_{\mathrm{NH}}(t) &= \int_0^t f_1(s) S_2(t-s) \mathrm{d}s \\
Y_{\mathrm{NH}}(t) &= \int_0^t \int_s^t f_1(s) f_2(u-s) S_3(t-u) \mathrm{d}u \mathrm{d}s \\
Z_{\mathrm{NH}}(t) &= \int_0^t \int_s^t \int_u^t f_1(s) f_2(u-s) f_3(v-u) \mathrm{d}v \mathrm{d}u \mathrm{d}s \qquad (1) \\
&= 1 - (W(t) + X(t) + Y(t)) \qquad (2)
\end{aligned}
$$

The integral on the right-hand-side of Equation 1 for $Z(t)$ is useful to ensure that the implementations for the other equations are correct, whereas Equation 2 is trivial to calculate given the other values. We also have an incidence rate

$$
I_{\mathrm{NH}}(t) = \int_0^t \int_s^t f_1(s) f_2(u-s) f_3(t-u) \mathrm{d}u \mathrm{d}s
$$

2.3 Screening

For a given individual, consider a screening history (or filtration) $\mathcal{F}$ at a series of times $t_1 < \ldots < t_n$ for n screens and for some observation time $t > t_n$. It is useful to define $t_0 = 0$ and $t_{n+1} = t$. Note that we may have an earlier observation time, which would then have a shorter screening history.

We introduce two screening probabilities. Let β_3 represent the probability that an individual who has a screening test and who factually has pre-clinical cancer (that is, state Y) will then *not*

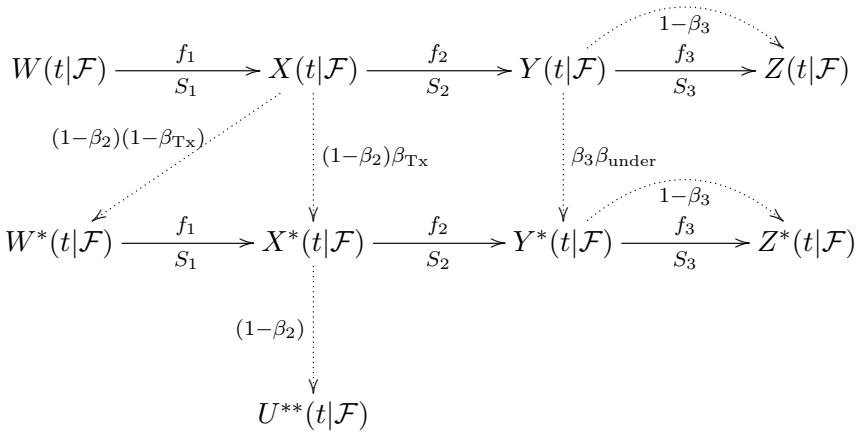

Figure 1: Natural history and screening model for cervical cancer. In the absence of high-grade squamous intraepithelial lesions (HSIL) being detected: W is for the population without HSIL or cancer; X is for HSIL; Y is for preclinical cancer; and Z is for diagnosed cancer. W^*, X^*, Y^* and Z^* are similar states after having had one HSIL detected. U^{**} is for two HSILs detected. The solid lines are continuous time natural history transitions and the dashed lines are discrete event screening transitions. f_i and S_i are density and survival functions. β_2 and β_3 are false positive values. β_{Tx} is the probability of treatment success for HSIL. β_{under} is the probability that a preclinical cervical cancer is under-classified as HSIL alone. $\mathcal{F}$ denotes the screening history.

be diagnosed following the screening test and will remain in the current state. This probability will be a combination of either a false negative screening test, not having a colposcopy or a biopsy, a false negative colposcopy or a false negative biopsy. The value $1 - \beta_3$ is then the probability of being diagnosed (moving from state Y to state Z). This parameter is closely related to the β parameter used by Day and Walter [1].

Let β_2 represent the probability that an individual who has a screening and factually has HSIL will *not* be detected for HSIL following the screening test. This probability will be a combination of either a false negative screening test, not having a colposcopy or a biopsy, a false negative colposcopy or a false negative biopsy. The value $1 - \beta_2$ is then the probability of moving from HSIL to the general healthy population (moving from state X to state W). It is well recognised that women who have been treated for HSIL have a higher risk of subsequent cervical cancer; however, although the current model allows for women who have been treated to cycle back to the general healthy population, then model assumes a similar probability of HSIL.

We also introduce a treatment probability β_{Tx}, which is the probability that a screen-detected HSIL case will *not* be cured following treatment. We now introduce five additional states: let W^* be the HSIL cases who were successfully treated after their first HSIL detection; let X^* be HSIL cases who had been treated after their first HSIL treatment, either due to unsuccessful treatment or disease progression after a successful treatment; let U^{**} be cases in X^* who are detected as HSIL at a subsequent time; let Y^* be preclinical cancers are detection for HSIL; and let Z^* be diagnosed cancers after detection for HSIL.

In the presence of screening and at observation time $t > t_n$ (preceding any screen and preceding

any detection of HSIL), we then have:

$$
\begin{aligned}
W(t|\mathcal{F}) &= S_1(t) \\
X(t|\mathcal{F}) &= \sum_{i=0}^{n} \int_{t_i}^{t_{i+1}} f_1(s) S_2(t-s) \beta_2^{n-i} \, \mathrm{d}s \\
Y(t|\mathcal{F}) &= \sum_{i=0}^{n} \sum_{j=i}^{n} \int_{t_i}^{t_{i+1}} \int_{\max(s,t_j)}^{t_{j+1}} f_1(s) f_2(u-s) S_3(t-u) \beta_2^{j-i} \beta_3^{n-j} (1-\beta_{\text{under}})^{n-j} \, \mathrm{d}u \, \mathrm{d}s
\end{aligned}
$$

Let us define $W^*(t|t_j,\mathcal{F}), X^*(t|t_j,\mathcal{F})$ and $Y^*(t|t_j,\mathcal{F})$ as the densities for being in states W^*, X^* and Y^* at observation time t given an HSIL detected at time t_j. In the presence of screening, we then have:

$$
\begin{aligned}
W^*(t|t_j,\mathcal{F}) &= \sum_{i=0}^{j-1} \int_{t_i}^{t_{i+1}} f_1(s) S_2(t_j-s) \beta_2^{j-i-1} (1-\beta_2)(1-\beta_{\text{Tx}}) S_1(t-t_j) \, \mathrm{d}s \\
X^*(t|t_j,\mathcal{F}) &= \sum_{i=0}^{j-1} \int_{t_i}^{t_{i+1}} f_1(s) S_2(t_j-s) \beta_2^{j-i-1} (1-\beta_2) \Bigg\{ \beta_{\text{Tx}} \beta_2^{n-j} S_2(t-t_j) + \\
&\qquad (1-\beta_{\text{Tx}}) \sum_{k=j}^{n} \int_{t_k}^{t_{k+1}} f_1(u-t_j) S_2(t-u) \beta_2^{n-k} \, \mathrm{d}u \Bigg\} \, \mathrm{d}s \\
Y^*(t|t_j,\mathcal{F}) &= \sum_{i=0}^{j-1} \int_{t_i}^{t_{i+1}} f_1(s) S_2(t_j-s) \beta_2^{j-i-1} (1-\beta_2) \Bigg\{ \\
&\qquad \beta_{\text{Tx}} \sum_{k=j}^{n} \int_{t_k}^{t_{k+1}} f_2(u-t_j) S_3(t-u) \beta_2^{k-j} \beta_3^{n-k} \, \mathrm{d}u + \\
&\qquad (1-\beta_{\text{Tx}}) \sum_{k=j}^{n} \sum_{m=k}^{n} \int_{t_k}^{t_{k+1}} \int_{\max(u,t_m)}^{t_{m+1}} f_1(u-t_j) f_2(v-u) S_3(t-v) \beta_2^{m-k} \beta_3^{n-m} \, \mathrm{d}v \, \mathrm{d}u \Bigg\} \, \mathrm{d}s + \\
&\qquad \sum_{i=0}^{j-1} \sum_{k=i}^{j-1} \int_{t_i}^{t_{i+1}} \int_{\max(s,t_k)}^{t_{j+1}} f_1(s) f_2(u-s) S_3(t-u) \beta_2^{k-i} \beta_3^{n-k} (1-\beta_{\text{under}})^{j-k-1} \beta_{\text{under}} \, \mathrm{d}u \, \mathrm{d}s
\end{aligned}
$$

The interval cancer incidence density with no HSIL detected is then

$$
I(t|\mathcal{F}) = \sum_{i=0}^{n} \sum_{j=i}^{n} \int_{t_i}^{t_{i+1}} \int_{\max(s,t_j)}^{t_{j+1}} f_1(s) f_2(u-s) f_3(t-u) \beta_2^{j-i} \beta_3^{n-j} (1-\beta_{\text{under}})^{n-j} \, \mathrm{d}u \, \mathrm{d}s
$$

The interval cancer incidence density with an HSIL detected at time t_j and an interval cancer diagnosed at time t (without a second HSIL being detected) is then

$$
\begin{aligned}
I^*(t|t_j,\mathcal{F}) = I(n \geq 1) \sum_{i=0}^{j-1} \int_{t_i}^{t_{i+1}} f_1(s) S_2(t_j - s) \beta_2^{j-i-1} (1-\beta_2) \Bigg\{ & \\
\beta_{\text{Tx}} \sum_{k=j}^{n} \int_{t_k}^{t_{k+1}} f_2(u-t_j) f_3(t-u) \beta_2^{k-j} \beta_3^{n-k} \, \mathrm{d}u + & \\
(1-\beta_{\text{Tx}}) \sum_{k=j}^{n} \sum_{m=k}^{n} \int_{t_k}^{t_{k+1}} \int_{\max(u,t_m)}^{t_{m+1}} f_1(u-t_j) f_2(v-u) f_3(t-v) \beta_2^{m-k} \beta_3^{n-m} \, \mathrm{d}v \, \mathrm{d}u & \\
\Bigg\} \mathrm{d}s + & \\
\sum_{i=0}^{j-1} \sum_{k=i}^{j-1} \int_{t_i}^{t_{i+1}} \int_{\max(s,t_k)}^{t_{k+1}} f_1(s) f_2(u-s) f_3(t-u) \beta_2^{k-i} \beta_3^{n-k} (1-\beta_{\text{under}})^{j-k-1} \beta_{\text{under}} \, \mathrm{d}u \, \mathrm{d}s &
\end{aligned}
$$

2.4 Likelihood

We are now in the position to define expressions for several likelihood terms (see Table 1). Note that we censor observations after the second HSIL has been detected due to the increased mathematical complexity; within the analysis dataset, we will describe whether this outcome is common.

Screening history	Likelihood expression
1. No cancer detected and no HSIL detected	$W(t\|\mathcal{F}) + X(t\|\mathcal{F}) + Y(t\|\mathcal{F})$
2. First HSIL detected at t_j with no subsequent detection	$W^*(t\|t_j,\mathcal{F}) + X^*(t\|t_j,\mathcal{F}) + Y^*(t\|t_j,\mathcal{F})$
3. Second HSIL detected with first HSIL at t_j	$X^*(t\|t_j,\mathcal{F})(1-\beta_2)$
4. Interval cancer with no HSIL detected	$I(t\|\mathcal{F})$
5. Interval cancer with HSIL detected at time t_j	$I^*(t\|t_j,\mathcal{F})$
6. Screen-detected cancer with no previous HSIL	$Y(t\|\mathcal{F})(1-\beta_3)$
7. Screen-detected cancer with one previous HSIL detected	$Y^*(t\|t_j,\mathcal{F})(1-\beta_3)$

Table 1: Likelihood terms for different screening histories

The full log-likelihood is the sum of the log of the likelihood terms for the observations. We optimised the log-likelihood using the bobyqa algorithm in the minqa package on CRAN. The model likelihood was implemented in C++ and optimised using R.

To reduce the computational task, we took a 10% sample of the event history type 1 and included weights of 10 for the sampled individuals and weights of 1 for the other screening history types. Given this weighting, the variance calculations used the sandwich estimator, which required the calculation of scores (gradients) for each log-likelihood component using finite differences.

Estimand := Estimator
$\Pr(\text{Preclinical cancer} \mid \text{no HSIL}, \mathcal{F}) := \dfrac{Y(t\|\mathcal{F})}{W(t\|\mathcal{F}) + X(t\|\mathcal{F}) + Y(t\|\mathcal{F})}$
$\Pr(\text{Preclinical cancer} \mid \text{HSIL at } t_j, \mathcal{F}) := \dfrac{Y^*(t\|t_j, \mathcal{F})}{W^*(t\|t_j, \mathcal{F}) + X^*(t\|t_j, \mathcal{F}) + Y^*(t\|t_j, \mathcal{F})}$
$\Pr(\text{Screen-detected cancer at } t \mid \text{no HSIL}, \mathcal{F}) := \dfrac{Y(t\|\mathcal{F})(1-\beta_3)}{W(t\|\mathcal{F}) + X(t\|\mathcal{F}) + Y(t\|\mathcal{F})}$
$\Pr(\text{Screen-detected cancer at } t \mid \text{HSIL at } t_j, \mathcal{F}) := \dfrac{Y^*(t\|t_j, \mathcal{F})(1-\beta_3)}{W^*(t\|t_j, \mathcal{F}) + X^*(t\|t_j, \mathcal{F}) + Y^*(t\|t_j, \mathcal{F})}$
$\text{Risk of interval cancer by } t \mid \text{no HSIL at } t_n, \mathcal{F} := 1 - \exp\left(-\int_{t_n}^{t} I(u\|\mathcal{F})\,\mathrm{d}u\right)$
$\text{Risk of interval cancer by } t \mid \text{previous HSIL at } t_j, \mathcal{F} := 1 - \exp\left(-\int_{t_n}^{t} I^*(u\|t_j, \mathcal{F})\,\mathrm{d}u\right)$

Table 2: Estimands and estimators based on the fitted model

2.5 Estimands from the fitted model

Some estimands of interest, including their estimators from the fitted model, are shown in Table 2. We can also calculate marginal estimators, such as:

$$\text{Incidence rate of interval cancers} := \frac{\sum_i \left(I(t|\mathcal{F}_i) + I^*(t|\mathcal{F}_i)\right)}{\sum_i \left(W(t|\mathcal{F}_i) + X(t|\mathcal{F}_i) + Y_i(t) + W^*(t|\mathcal{F}_i) + X^*(t|\mathcal{F}_i) + Y^*(t|\mathcal{F}_i)\right)}$$

where we use a subscript i for individuals in the population who have had not had two HSILs detected.

3 Results

A frequency table for the types of screening histories by the number of previous HSILs is given in Table 3. We see that most of the 1960 birth cohort never had an HSIL histology or a cancer histology ($51592/51807 = 99.6\%$).

As a proof of concept, Model 1 was defined using exponential distributions with constant hazards for f_2 and f_3, and assuming a log-logistic distribution with cure for f_1. Prior to the introduction of the βunder parameter, we found that 25 individuals had numerically zero likelihoods and who had an HSIL biopsy with no cancer less than one month before a biopsy with cancer. These individuals were defined as interval cancers. We interpret their history as being more consistent with an under-classified biopsy finding, where they were likely to already have had preclinical cervical cancer rather than that they transitioned rapidly from HSIL to preclinical cancer to diagnosed cancer within a month. Based on this observation, we added the path from Y to Y^* with probability $\beta_2\beta_{\text{under}}$.

The main challenge with the model was whether the model parameters were *identifiable*. Model A included logs of the log-logistic parameters (with cure fixed at 0.9); Model B included the log-

	Number of previous HSILs detected					
Screening history type	0	1	2	3	4	Total
1	51592	0	0	0	0	51592
2	0	2110	0	0	0	2110
3	0	0	267	41	7	315
4	163	0	0	0	0	163
5	0	32	2	1	0	35
6	52	0	0	0	0	52
7	0	8	2	0	0	10
Total	51807	2150	271	42	7	51807

Table 3: Description of the 1960 birth cohort from NKCx by event history type and number of previous HSILs detected at end of follow-up

logistic and logit cure parameters, together with the logs of the rates for f_2 and f_3 and logits for β_2 and β_3. The initial and fitted values are shown in Figure 4.

Parameter name	Symbol	Initial value	Model A	Model B
HSIL onset shape	a_1	10.00	5.68 (5.41,5.96)	419.37
HSIL onset scale	b_1	30.00	33.51 (31.93,35.16)	22.15
HSIL onset cure	π_1	0.90		0.172
Preclinical onset rate	λ_2	$-\log(1-0.05)$		1.73
Clinical onset rate	λ_3	$-\log(1-0.5)$		0.0002
False negative | HSIL	β_2	0.20		0.951
False negative | preclinical cancer	β_3	0.05		0.999
Pr(HSIL treatment failure | HSIL)	β_{Tx}	0.20		
Pr(biopsy misclassified as HSIL | preclinical cancer)	β_{under}	0.05		

Table 4: Model parameters for models A and B

The individual likelihood components were bimodal: women with a negative screening history provide comparatively information on whether they have HSIL or preclinical cancer, whereas having a biopsy-confirmed HSIL or a cancer diagnosis provides substantial information to the likelihood.

The predicted onset distribution for HSIL from each model is shown in Figure 2. The predicted shape for Model A is broadly consistent with HPV infection and then subsequent onset of HSIL. Model B predicts a sharp spike in onset at age 22. This predicted pattern is consistent with a lack of model identifiability.

3.1 Predictions

Using the parameters from Model B, we provide some predictions as a proof-of-concept. We show the state probabilities for Model A for no screening and for a single screen at age 30 years (see Figure 3). *Importantly, these predictions are based on a model that is only fitted for HSIL shape and scale and are presented for illustrative purposes only.*

4 Discussion

We have developed and implemented a mathematical model for cervical cancer screening and fitted that model individual-level data from the Swedish National Cervical Cancer Screening Register.

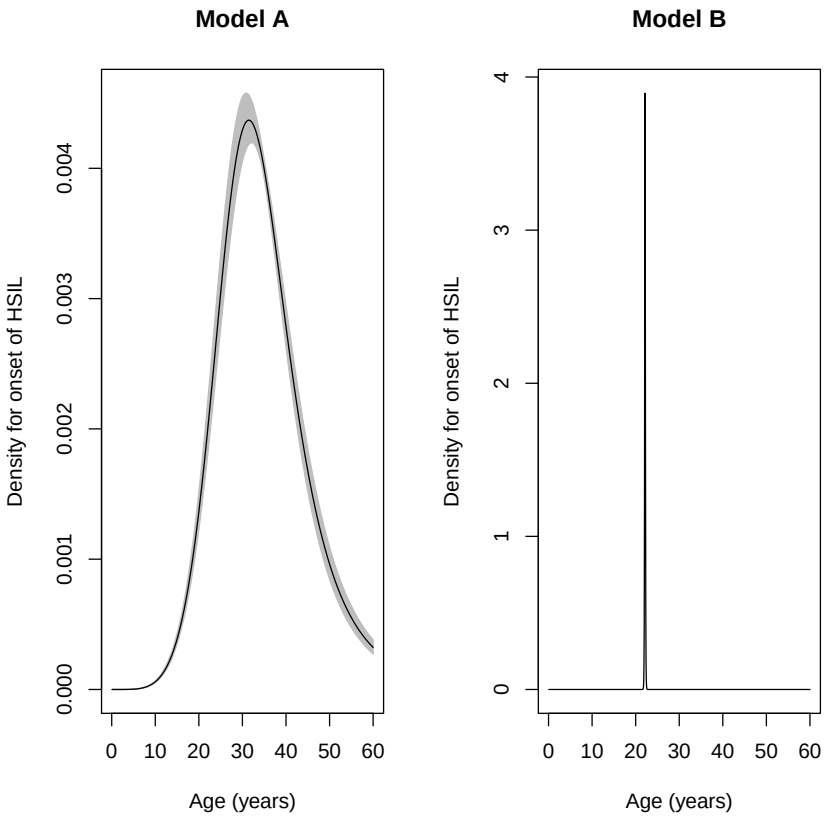

Figure 2: Fitted density for onset of HSIL, by model fit

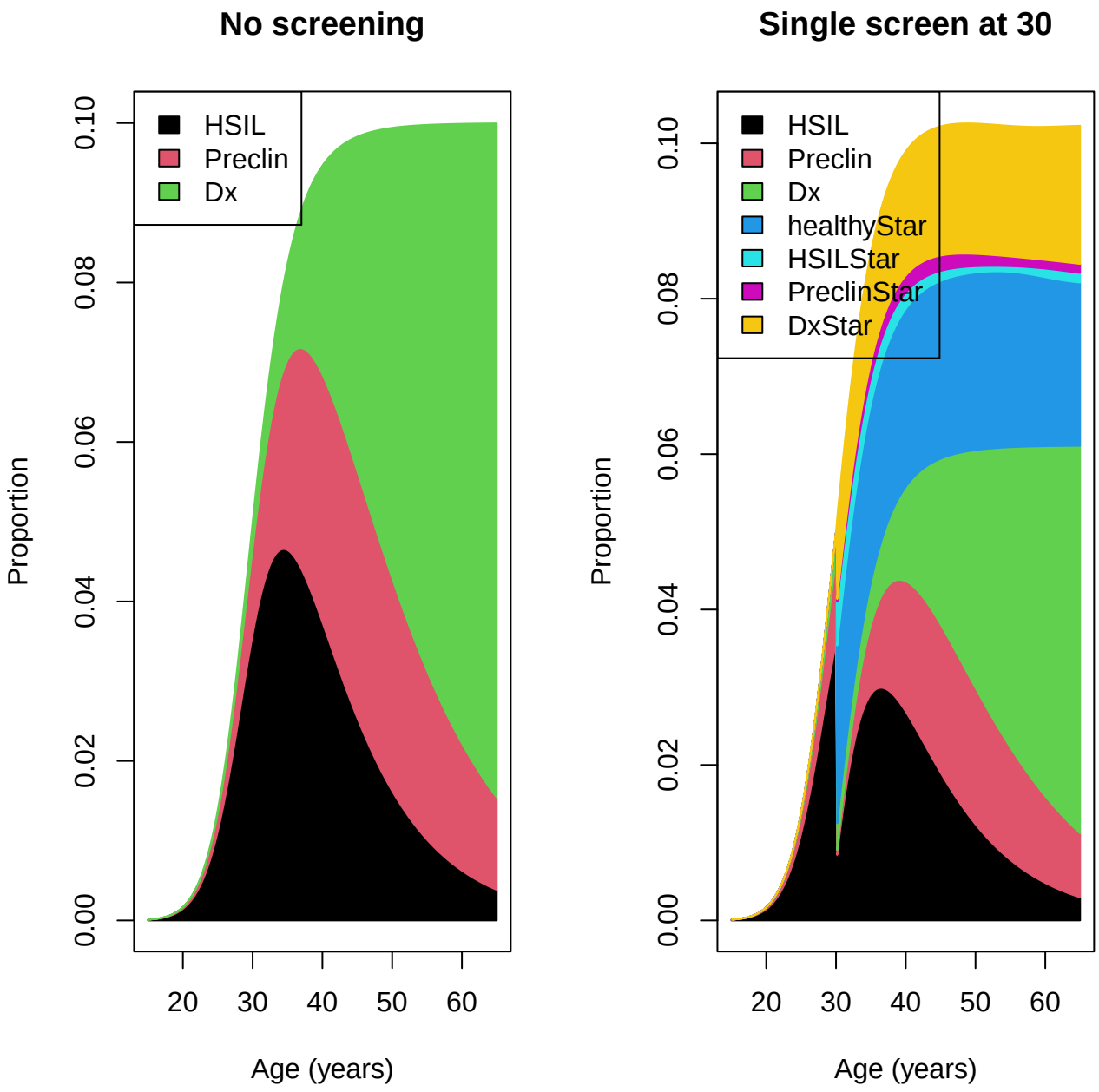

Figure 3: State probabilities for no screening and for a single screen at age 30 years, Model A parameters

The model fit demonstrated issues with identifiability of the model parameters and should be considered as a first step in the application of the proposed model framework.

As a strength of our approach, we have used a detailed mathematical model to use individual-level screening data. To our knowledge, this is the first mathematical model fitted to individual-level data from the Swedish Cervical Cancer Screening Register. There are several limitations to our approach. First, we were not able to resolve the lack of identifiability for the model parameters.

Consequently, our fitted model A lacked plausibility and, in our opinion, would not provide reasonable predictions. As a next step, we would like to further investigate the identifiability of the model parameters. Another consequence of the lack of identifiability is that the covariance matrix for the parameters was not invertible, so that it was difficult to represent the uncertainty in the model parameters. We could investigate model uncertainty using the bootstrap, which would add a further layer of computational complexity. Second, there have been changes in cervical cancer screening tests over time, particularly the change from conventional cytology to liquid-based cytology around 2011, and the introduction of primary HPV testing from age 30 years at around 2017. Moreover, younger birth cohorts will be partially vaccinated for HPV, which will change the onset distribution for HSIL. Any applications of the model should take those factors into consideration.

4.1 Extensions

An obvious extension would be to fit these models to more birth cohorts. This would be an increasingly large computational task and would require the use of a computer cluster. The C++ code is well suited for multi-core computations on Linux servers. It is possible that the C++ code could be adapted to use general purpose graphics processing units (GPUs). The use of automatic differentiation libraries such as PyTorch of JAX may be useful for any calculations on GPUs – and would have the advantage of calculating gradients for optimisation of the log-likelihood.

The implementation could also be simplified by noting that the current fitted models are also Markov models. We could then use Kolmogorov's forward differential equation to solve the model, with step changes at the screening occasions. This will be simpler, and may be faster, than the current more general implementation.

We have provided a very specific implementation for cervical cancer screening. The R and C++ code could be generalised to: (a) the model due to Day and Walter [1] that includes individual-based screening patterns, for use with traditional cancer screening, where pre-cancerous lesions can not be treated; and (b) prostate cancer screening, which could include a sub-model for prostate-specific antigen. It is unclear whether this class of screening models are potentially useful for the Finnish Maternity Cohort.

As a possible extension, we could copy the NKCx datasets to HopsWorks. DuckDB is a columnar-based embedded database that is available on HopsWorks. The version of SQL on DuckDB is very similar to that for SQL Server, which would aid in the translation. The main challenge in the translation is that the SAS datastep code would need to be translated to either Python or R. To our knowledge, no version of the NKCx database has been put on a HopsWorks instance. HopsWorks would have some clear advantages if a GPU-based implementation were available.

4.2 Conclusion

We have provided a proof-of-concept for mathematical modelling of cervical cancer screening. The predictions from a well-fitted model could be compared with a prediction model based on machine learning. For such a comparison, one could evaluate whether the machine learning approach provides different weight for the different screening histories.

References

[1] Day, N. E. and Walter, S. D. (1984). Simplified models of screening for chronic disease: estimation procedures from mass screening programmes. *Biometrics*, 40:1–13.

A Mathematical equations for the state occupation probabilities

$$
\begin{aligned}
Z(t|\mathcal{F}) = &\sum_{i=0}^{n}\sum_{j=i}^{n}\sum_{k=j}^{n}\int_{t_i}^{t_{i+1}}\int_{\max(s,t_j)}^{t_{j+1}}\int_{\max(u,t_k)}^{t_{k+1}} f_1(s)f_2(u-s)f_3(v-u)\beta_2^{j-i}\beta_3^{k-j}\,\mathrm{d}v\,\mathrm{d}u\,\mathrm{d}s+\\
&\sum_{i=0}^{n-1}\sum_{j=i}^{n-1}\sum_{k=j+1}^{n-1}\int_{t_i}^{t_{i+1}}\int_{\max(s,t_j)}^{t_{i+j}} f_1(s)f_2(u-s)S_3(t_k-u)\beta_2^{j-i}\beta_3^{k-j-1}(1-\beta_3)\,\mathrm{d}u\,\mathrm{d}s\\
W^*(t|\mathcal{F}) = &\sum_{j=1}^{n} W^*(t|t_j,\mathcal{F})\\
X^*(t|\mathcal{F}) = &\sum_{j=1}^{n} X^*(t|t_j,\mathcal{F})\\
Y^*(t|\mathcal{F}) = &\sum_{j=1}^{n} Y^*(t|t_j,\mathcal{F})\\
U^{**}(t|\mathcal{F}) = &\sum_{i=0}^{n-2}\sum_{j=i+1}^{n-1}\int_{t_i}^{t_{i+1}} f_1(s)S_2(t_j-s)\beta_2^{j-i-1}(1-\beta_2)^2\Bigg\{\beta_{\mathrm{Tx}}\sum_{k=j+1}^{n}\beta_2^{k-j-1}S_2(t_k-t_j)+\\
&(1-\beta_{\mathrm{Tx}})\sum_{k=j}^{n-1}\sum_{m=k+1}^{n}\int_{t_k}^{t_{k+1}} f_1(u-t_j)S_2(t_m-u)\beta_2^{m-k-1}\,\mathrm{d}u\Bigg\}\,\mathrm{d}s\\
Z^*(t|\mathcal{F}) = &\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\int_{t_i}^{t_{i+1}} f_1(s)S_2(t_j-s)\beta_2^{j-i-1}(1-\beta_2)\Bigg\{\\
&\beta_{\mathrm{Tx}}\sum_{k=j}^{n}\sum_{m=k}^{n}\int_{t_k}^{t_{k+1}}\int_{\max(u,t_m)}^{t_{m+1}} f_2(u-t_j)f_3(v-u)\beta_2^{k-j}\beta_3^{m-k}\,\mathrm{d}v\,\mathrm{d}u+\\
&\beta_{\mathrm{Tx}}\sum_{k=j}^{n-1}\sum_{m=k+1}^{n}\int_{t_k}^{t_{k+1}} f_2(u-t_j)S_3(t_m-u)\beta_2^{k-j}\beta_3^{m-k-1}(1-\beta_3)\,\mathrm{d}u+\\
&(1-\beta_{\mathrm{Tx}})\sum_{k=j}^{n}\sum_{l=k}^{n}\sum_{m=l}^{n}\int_{t_k}^{t_{k+1}}\int_{\max(u,t_l)}^{t_{l+1}}\int_{\max(v,t_m)}^{t_{m+1}} f_1(u-t_j)f_2(v-u)f_3(x-v)\beta_2^{l-k}\beta_3^{m-l}\,\mathrm{d}x\,\mathrm{d}v\,\mathrm{d}u+\\
&(1-\beta_{\mathrm{Tx}})\sum_{k=j}^{n-1}\sum_{l=k}^{n-1}\sum_{m=l+1}^{n}\int_{t_k}^{t_{k+1}}\int_{\max(u,t_l)}^{t_{l+1}} f_1(u-t_j)f_2(v-u)S_3(t_m-v)\beta_2^{l-k}\beta_3^{m-l-1}(1-\beta_3)\,\mathrm{d}v\,\mathrm{d}u\\
&\Bigg\}\,\mathrm{d}s\\
= &\,1-(W(t|\mathcal{F})+X(t|\mathcal{F})+Y(t|\mathcal{F})+Z(y|\mathcal{F})+W^*(t|\mathcal{F})+X^*(t|\mathcal{F})+Y^*(t|\mathcal{F})+Z^*(t|\mathcal{F})+U^{**}(t|\mathcal{F}))
\end{aligned}
$$

B Mathematical equations for negative biopsies

The following development includes states for having a negative biopsy. Note that this development does not currently include mis-classification of preclinical cancers as HSIL.

As a further model extension, we could use data on negative cervical biopsies (see Figure 4). Expert opinion suggests that the negative predictive value for a negative biopsy is approximately

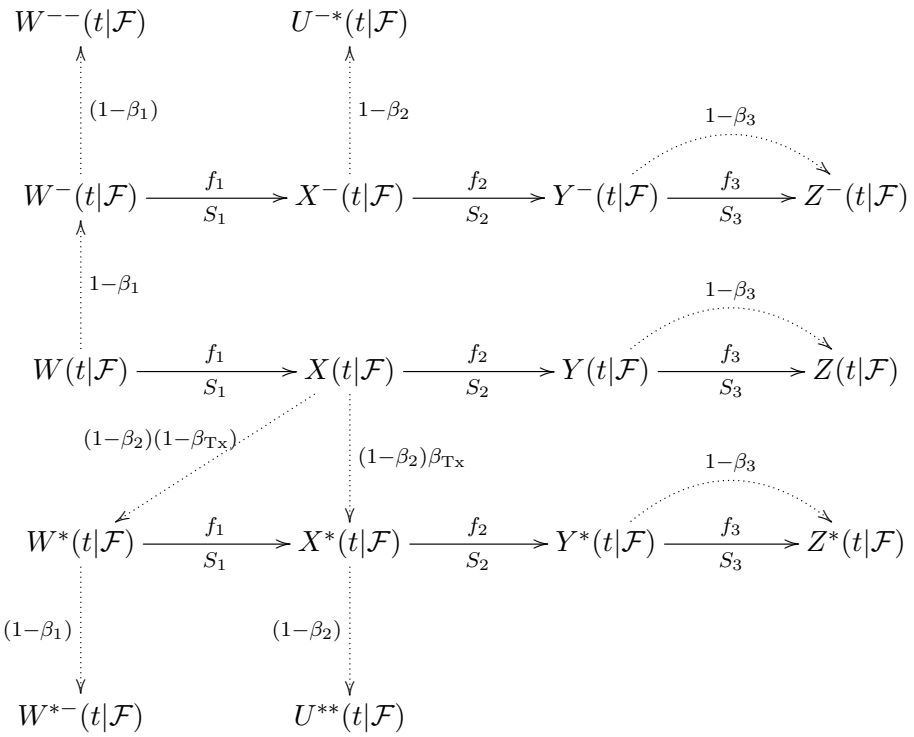

Figure 4: Extended natural history and screening model for cervical cancer. In the absence of high-grade squamous intraepithelial lesions (HSIL) being detected: W is for the population without HSIL or cancer; X is for HSIL; Y is for preclinical cancer; and Z is for diagnosed cancer. W^*, X^*, Y^* and Z^* are similar states after having had one HSIL detected. W^-, X^-, Y^- and Z^- are similar states after having a negative biopsy. U^{**} is for two HSILs detected. U^{-*} is for having a negative biopsy followed by an HSIL detected, W^{--} is for two negative biopsies, and W^{*-} is for a successfully treated HSIL followed by a negative biopsy. The solid lines are continuous time natural history transitions and the dashed lines are discrete event screening transitions. f_i and S_i are density and survival functions. β_2 and β_3 are false positive values. β_{Tx} is the probability of treatment success for HSIL. $1-\beta_1$ is the probability that an individual in state W will have a biopsy, which is assumed to be negative. $\mathcal{F}$ denotes the screening history.

98%. As a simplification, we could assume that the negative predictive value is 100%, such that negative biopsies accurately identify individuals who do not have cervical cancer and do not have HSIL. Under this assumption, if t_k is the time for the negative biopsy at the k-th screen, then the

likelihood at the time of the negative biopsy is $W(t_k-)(1-\beta_1)$. The state probabilities are then

$$
\begin{aligned}
W(t|\mathcal{F}) &= S_1(t)\beta_1^n \\
X(t|\mathcal{F}) &= \sum_{i=0}^{n} \int_{t_i}^{t_{i+1}} f_1(s) S_2(t-s) \beta_1^i \beta_2^{n-i} \,\mathrm{d}s \\
Y(t|\mathcal{F}) &= \sum_{i=0}^{n} \sum_{j=i}^{n} \int_{t_i}^{t_{i+1}} \int_{\max(s,t_j)}^{t_{j+1}} f_1(s) f_2(u-s) S_3(t-u) \beta_1^i \beta_2^{j-i} \beta_3^{n-j} \,\mathrm{d}u \,\mathrm{d}s \\
W^-(t|t_k,\mathcal{F}) &= W(t_k-|\mathcal{F})(1-\beta_1) S_1(t)/S_1(t_k) \beta_1^{n-k} = S_1(t)\beta_1^{n-1}(1-\beta_1) \\
X^-(t|t_k,\mathcal{F}) &= W(t_k-|\mathcal{F})(1-\beta_1) \sum_{i=k}^{n} \int_{t_i}^{t_{i+1}} f_1(s) S_2(t-s) \beta_1^{i-k} \beta_2^{n-i} \,\mathrm{d}s \\
Y^-(t|t_k,\mathcal{F}) &= W(t_k-|\mathcal{F})(1-\beta_1) \sum_{i=k}^{n} \sum_{j=i}^{n} \int_{t_i}^{t_{i+1}} \int_{\max(s,t_j)}^{t_{j+1}} f_1(s) f_2(u-s) S_3(t-u) \beta_1^{i-k} \beta_2^{j-i} \beta_3^{n-j} \,\mathrm{d}u \,\mathrm{d}s \\
W^*(t|t_j,\mathcal{F}) &= \sum_{i=0}^{j-1} \int_{t_i}^{t_{i+1}} f_1(s) S_2(t_j-s) \beta_1^i \beta_2^{j-i-1} (1-\beta_2)(1-\beta_{\mathrm{Tx}}) S_1(t-t_j) \,\mathrm{d}s \\
X^*(t|t_j,\mathcal{F}) &= \sum_{i=0}^{j-1} \int_{t_i}^{t_{i+1}} f_1(s) S_2(t_j-s) \beta_1^i \beta_2^{j-i-1} (1-\beta_2) \Bigg\{ \beta_{\mathrm{Tx}} \beta_2^{n-j} S_2(t-t_j) + \\
&\qquad (1-\beta_{\mathrm{Tx}}) \sum_{k=j}^{n} \int_{t_k}^{t_{k+1}} f_1(u-t_j) S_2(t-u) \beta_1^k \beta_2^{n-k} \,\mathrm{d}u \Bigg\} \,\mathrm{d}s \\
Y^*(t|t_j,\mathcal{F}) &= \sum_{i=0}^{j-1} \int_{t_i}^{t_{i+1}} f_1(s) S_2(t_j-s) \beta_1^i \beta_2^{j-i-1} (1-\beta_2) \Bigg\{ \\
&\qquad \beta_{\mathrm{Tx}} \sum_{k=j}^{n} \int_{t_k}^{t_{k+1}} f_2(u-t_j) S_3(t-u) \beta_2^{k-j} \beta_3^{n-k} \,\mathrm{d}u + \\
&\qquad (1-\beta_{\mathrm{Tx}}) \sum_{k=j}^{n} \sum_{m=k}^{n} \int_{t_k}^{t_{k+1}} \int_{\max(u,t_m)}^{t_{m+1}} f_1(u-t_j) f_2(v-u) S_3(t-v) \beta_1^k \beta_2^{m-k} \beta_3^{n-m} \,\mathrm{d}v \,\mathrm{d}u \Bigg\} \,\mathrm{d}s
\end{aligned}
$$

The interval cancer incidence density with a negative biopsy at t_k and no HSIL detected is then

$$
I^-(t|t_k,\mathcal{F}) = W(t_k-|\mathcal{F})(1-\beta_1) \sum_{i=k}^{n} \sum_{j=i}^{n} \int_{t_i}^{t_{i+1}} \int_{\max(s,t_j)}^{t_{i+j}} f_1(s) f_2(u-s) f_3(t-u) \beta_1^{i-k} \beta_2^{j-i} \beta_3^{n-j} \,\mathrm{d}u \,\mathrm{d}s
$$

There are now five new likelihood terms; see Table 5.

Screening history	Likelihood expression
8. No cancer/HSIL detected with negative biopsy at t_k, $\mathcal{F}$	$W^-(t\|t_k,\mathcal{F}) + X^-(t\|t_k,\mathcal{F}) + Y^-(t\|t_k,\mathcal{F})$
9. Second negative biopsy at t with negative biopsy at t_k, $\mathcal{F}$	$W^-(t - \|t_k,\mathcal{F})(1-\beta_1)$
10. First HSIL detected, negative biopsy at t_k	$X^-(t\|t_k,\mathcal{F})(1-\beta_2)$
11. Interval cancer with negative biopsy at t_k and no HSIL, $\mathcal{F}$	$I^-(t\|t_k,\mathcal{F})$
12. Screen-detected cancer with negative biopsy at t_k, $\mathcal{F}$	$Y^-(t\|t_k,\mathcal{F})(1-\beta_3)$

Table 5: Likelihood terms for different screening histories modelling for negative biopsies