Date of acceptance: Grade

Instructor: Sangita Kulathinal

Bayesian hidden Markov model for overdiagnosis in colorectal cancer screening

Aapeli Nevala

Helsinki April 18, 2020

UNIVERSITY OF HELSINKI
Department of Mathematics and Statistics
For Finnish Cancer Registry

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Faculty of Science	Department of Mathematics and Statistics

Tekijä — Författare — Author
Aapeli Nevala

Työn nimi — Arbetets titel — Title
Bayesian hidden Markov model for
overdiagnosis in colorectal cancer screening

Oppiaine — Läroämne — Subject
Statistics

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
	April 18, 2020	49 pages + 11 appendix pages

Tiivistelmä — Referat — Abstract

Thanks to modern medical advances, humans have developed tools for detecting diseases so early, that a patient would be better off had the disease gone undetected. This is called overdiagnosis. Overdiagnosis is a problem especially common in acts, where the target population of an intervention consists of mostly healthy people.

Colorectal cancer (CRC) is a relatively rare disease. Thus screening for CRC affects mostly cancer-free population. In this thesis I evaluate overdiagnosis in guaiac faecal occult blood test (gFOBT) based CRC screening programme.

In gFOBT CRC screening there are two goals: to detect known predecessors of cancers called adenomas and to remove them (cancer prevention), and to detect malign CRCs early enough to be still treatable (early detection). Overdiagnosis can happen when detecting adenomas, but also when detecting cancers. This thesis focuses on overdiagnosis due to detection of adenomas that are non-progressive in their nature.

Since there is no clinical means to make distinction between progressive and non-progressive adenomas, statistical methods must be applied. Classical methods to estimate overdiagnosis fail in quantifying this type of overdiagnosis for couple of reasons: incidence data of adenomas is not available, and adenoma removal results in lowering cancer incidence in screened population. While the latter is a desired effect of screening, it makes it impossible to estimate overdiagnosis by just comparing cancer incidences among screened and control populations.

In this thesis a Bayesian Hidden Markov model using HMC NUTS algorithm via software Stan is fitted to simulate the natural progression of colorectal cancer. The five states included in the model were healthy (1), progressive adenoma (2), screen-detectable CRC (3), clinically apparent CRC (4) and non-progressive adenoma (5). Possible transitions are from 1 to 2, 1 to 5, 2 to 3 and 3 to 4. The possible observations are screen-negative (1), detected adenoma (2), screen-detected CRC (3), clinically manifested CRC (3).

Three relevant estimands for evaluating this type of overdiagnosis with a natural history model are presented. Then the methods are applied to estimate overdiagnosis proportion in guaiac faecal occult blood test (gFOBT) based CRC screening programme conducted in Finland between 2004 and 2016.

The resulting mean overdiagnosis probability for all the patients that had an adenoma detected for programme is 0.48 (0.38, 0.56, 95-percent credible interval). Different estimates for overdiagnosis in sex and age-specific stratas of the screened population are also provided.

In addition to these findings, the natural history model can be used to gain more insight about natural progression of colorectal cancer.

Avainsanat — Nyckelord — Keywords
Cancer screening, Bayesian statistics, Hidden Markov model, Statistics, Cancer, Epidemiology

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Contents

Appendices

1 Introduction

1.1 Biological and epidemiological background

Cancers are the second to most common cause of death in Finland, behind only artery and hearth diseases. For people in working ages 15-64, cancer is the leading cause of death. According to Finnish Cancer Registry, 12788 people died due to cancer in 2017. 1368 of them died due to colorectal cancer (CRC), which is the second-to-most common cancer for both women and men.

According to Syöpätaudit (Cancer Diseases, Duodecim, 2013) the most typical CRC symptoms are stomach pain, constipation and blood in stool. The patients are often relatively old and the symptoms have been sporadically present in their lives through years. This often leads to postponing the visit to a doctor. Additionally, the general nature of the symptoms may lead to even more delays in the diagnostic confirmation of the disease.

This makes CRC a tempting subject for screening. Additionally there exists a known pre-cancerous phase for CRC. Benign tumors called adenomas are known to precede malign adenocarcinomas, which, According to IARC (2019), consist of 85% of the CRCs worldwide.

The problem with huge screening programmes is that an organised screening affects mostly healthy people. Mass-screening of asymptomatic inviduals can lead to excessive detection of such diseases, that the patient would be better off without diagnosis. This is called overdiagnosis.

The phenomena of overdiagnosis in cancer screening has been illustrated in many screening programmes and in multiple studies with different study designs. Overview of these are provided by, for example, Houssami (2017) for breast cancer screening and Loeb et al. (2015) for prostate cancer screening. It's hypothesised that practically all of the cancer screening programmes come with some degree of overdiagnosis.

Estimates of overdiagnosis rates and proportions tend to vary even within a certain type of cancer screening programme. Marcus et al. (2015) cite figures ranging even from 0 to 50% in a same type of screening programme for same cancer. One study can find no overdiagnosis in a programme and another study that half true positive findings are due to overdiagnosis.

This thesis will focus on overdiagnosis in CRC screening programme organised in Finland between 2004 and 2016. More precisely, the focus will be on overdiagnosis caused by unnecessary detection of pre-cancerous lesions, namely non-progressive adenomas.

1.2 Colorectal cancer screening

The eventual end goal of organised cancer screening programme is simple: extension of life.

Cancers have the ugly feature of being treacherous residents in human body, sometimes for years, before showing any clinical symptoms. When symptoms appear, the disease may be too advanced for successful treatment. Thus earlier detection of a cancer is often beneficial. Organised cancer screening programmes are a widely used public health policy aiming for this.

The general idea behind an organised screening programme is to invite whole population of a certain area, often in a certain feasible age, to some kind of systematic procedure that is aiming to spot an asymptomatic cancer or its predecessor.

The benefit of cancer screening can be achieved via two different ways: by prevention of cancer and by early detection of cancer. For example, cervical cancer screening extends life by preventing cancers by detecting pre-cancerous lesions and treating them, while breast cancer screening focuses on detecting cancers while they are still local and thus more treatable. This distinction is due to both biological traits of certain cancer type and the availability of tests for their detection. In cervical cancer screening there exists both a known pre-cancerous stages with long sojourn times and tests for detecting that pre-cancerous lesion (papa-test).

CRC screening can be organised in variety of ways. One review of different programmes is given by Patel and Ahnen in the book Epidemiologic Studies in Cancer screening and prevention (2013, chapter 16). Options can be rougly separated to two categories: stool tests and imaging (radiographic or endoscopic). This thesis will forcus on a programme that used the former. To be more specific, Finnish health service study concluded between 2004 and 2016 used a test called gFOBT, that stands for guaiac fecal occult blood test.

In gFOBT screening both prevention of cancer (finding adenomas and removing them) and early detection of the cancer are achieved. Patel and Ahnen write that blood-based tests are currently more sensitive at detecting cancers that colonic polyps, yet most of the detected lesions are indeed polyps, not cancers. This is because pre-cancerous lesions are more prevalent and have higher incidence than cancers, but cancers tend to bleed more.

For example, estimates of sensitivity of gFOBT for CRC are around 50% (Hakama and Malila, 2019, Patel and Ahnen, 2013), while estimates for adenoma sensitivity are closer to 10% according to Brenner (2015) and Morikawa (2005). Specificity estimates for gFOBT range from 87% to 98% according to Patel and Ahnen (2013).

Screening procedure is following: an invited person gets a gFOBT kit via mail. If the person decides to attend, he or she sends their own stool sample to a laboratory for analysis. If blood is found, person is further invited for colonoscopy for diagnostic confirmation of the disease (adenoma or cancer). Most of the gFOBT-positive lesions are not adenomas or CRCs: from the total of 17440 colonoscopy confirmations

conducted in the whole Finnish RHS study 10532 were neither adenomas nor CRCs.

Patients with confirmed diagnosis receive a proper treatment, while both initially gFOBT-negative and colonoscopy-negative persons return to screening programme for the next screening round. A screening interval is two years. People who are treated will also return to programme, but since reliable data on treatments is unavailable, events after return will be ignored in this thesis.

1.3 Definition of overdiagnosis in cancer screening

Overdiagnosis is loosely defined as detecting a disease when a patient with the disease would be better off had the disease been undetected, but a more rigorous conceptual definition is required in order to proceed to mathematical treatment of the subject.

First, to have overdiagnosis, one must have a true-positive finding from the screening. In this study, a true positive is a screen-found adenoma. Here the term "disease" is adopted to mean the whole path from adenoma to cancer. Screen-found malign cancers in real world are also "true positives", but since the focus is on overdiagnosis due to detection of non-progressive adenomas, it's practical to forget about malign cancers for now.

Miller (2010) has divided true positive screening findings into four categories:

1. Disease detected by screening that would have been incurable after waiting until clinical diagnosis.
2. Disease detected by screening that could be cured even after waiting until clinical diagnosis.
3. Disease detected by screening that are incurable either way.
4. Disease that would never appear if one had kept waiting until clinical diagnosis.

Now category 1 means adenomas that would progress into eventually deadly cancers, category 2 are the adenomas that, when progressed into cancers, would have still been able to be treated (5-year relative survival for colorectal cancer is around 70%). Category 3 does not practically exists for adenomas and is thus not relevant in this thesis.

The definition of category 4 corresponds to commonly used term overdiagnosis. To decompose things even further, the framework proposed by Miller has been extended by Marcus et al. (2015). Miller was a part of their group that further categorised these overdiagnosed (category 4) diseases into three different types:

A Disease that would regress by itself even if left untreated.

B Disease that stops growing or progresses so slow that it would not threaten even oldest of people.

C Patient dies due to another disease before the cancer would become evident.

Now this thesis focuses on types A and B - the non-progressive adenomas.

Statistically the question with types A and B is about clustering diseases into non-progressive or regressive cancers and type C about calculating residual life distribution of the patient (from detection to death) and residual life distribution of the disease (from detection to hypothetical clinical appearance). These are two different problems, and I will only address the first one.

Additionally I will assume all the malign screen findings to be progressive in their nature. This is because, as Kalager et al. (2018) point out, there is no evidence for stagnation or regression of the malign disease, and because malign diseases account for very small proportion of the screen-positive findings and won't affect the estimates that much, but it would indeed complicate the model.

1.4 Earlier research

Kalager et al. (2018) point out that there is not a lot of literature about overdiagnosis in specific domain of colorectal cancer screening.

The natural history of colorectal cancer has been studied earlier with the data from the Finnish colorectal cancer screening programme by Chiu et al. [9]. The research, though, only included data from 2004 to 2007, while this thesis will have the screening trial data from 2004 to 2016 and follow-up data up until 2017. Also, they focused on mortality reduction and detection of malign disease.

Overdiagnosis using similar kind of Hidden Markov model that will later be applied in this thesis has been studied by Wu et al. (2019) in breast cancer screening, without very conclusive results, though.

Luo, Cambon and Wu (2011) developed a three-state semi-Markov model for cancer progression, that they used to classify screen-attending inviduals into four categories based on their disease status. One of these categories was overdiagnosed patients, where the overdiagnosed patient was defined as a patient who had a pre-clinical cancer in screening, but would not have had a clinically emergent disease before dying in case of no screening. Then they reported the posterior conditional probabilities for being overdiagnosed patient given they had a pre-clinical finding in screening. They eventually reported an overdiagnosis percentage between 6 to 9 percent for females and males respectively, with relatively wide confidence intervals of (2,5%, 23,1%) and (1,9%, 44,7%). Their data was based on Minnesota trial concluded in the US in the 1970s and 1980s.

Their work was focused on lifetime distribution of a person compared to cancer incidence probability of a person. Practically they would have, in context of this work, had focused on malign cancers and overdiagnosis of category 4 type C.

Hakama et al. (2019) give an indicator of overdiagnosis in Finnish randomised health service study in CRC screening, but their focus is on difference between two sensitivities for malign cancers and is thus very different in focus from the focus of this thesis, where the focus is on adenoma. Also their approach differs fundamentally.

1.5 Terminology and notations

A brief summary of cancer screening terminology is required for concluding a study of the subject.

Screening round refers to one set of actions, where people are invited to screening, those who attend are screened, evaluated and then directed to proper successive care, that can be for example an oncological treatment (if a cancer is diagnosed) or no treatment (if cancer is not found).

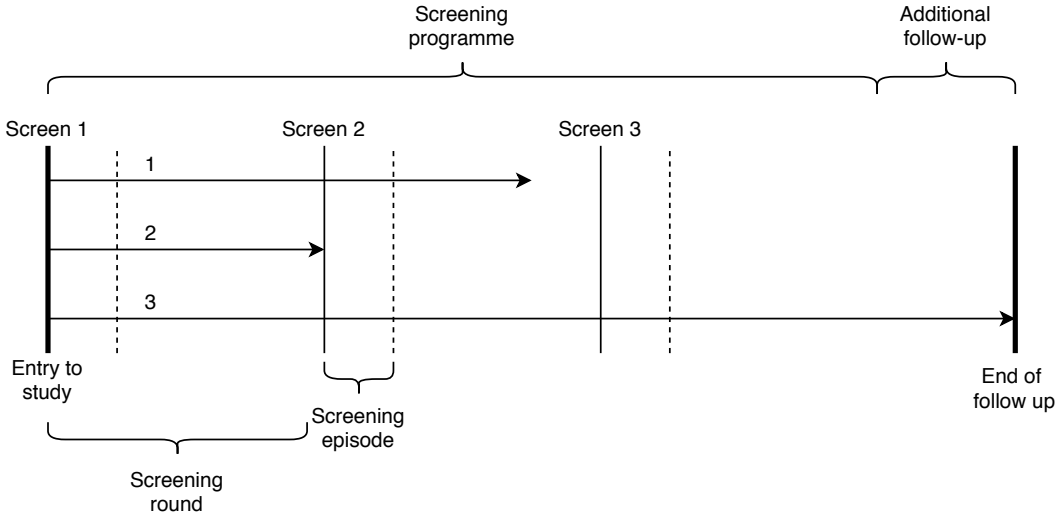

Figure 1: Illustration of some definitions used in this thesis. Person enters the study during first screening event. Space between screen and following dashed line is screening episode, a term used for inital screening test and possible diagnostic confirmation. Screening round includes one screening and time up until the next screen. Screening interval, the length between two screens, is two years. Three different follow-ups are illustrated in the figure as well: follow-up of a person can end during a screening interval (if an interval cancer is found or person dies (1)), in a screen (if an adenoma or cancer is found (2)) or at the latest three years after last screen (3).

CRC screening programme studied in this thesis is periodic, which means it consists of multiple rounds.

This study focuses on one screening programme that consists of multiple screening rounds. One round includes test and two-year screening interval.

Pre-clinical cancer refers to a colorectal cancer, that is possible to detect in a cancer screening (but may not be, due to lack of sensitivity of a screening test).

Screening cycle refers to period $(t_k, t_{k+1}]$ where t_k, t_{k+1} are screening times. Interval cancers are cancers diagnosed during period (t_k, t_{k+1}). In this notation screening round includes $[t_k, t_{k+1})$.

Screening episode is the period from the positive gFOBT test to clinical confirmation. Episode-negative refers to a screening round of an invidual, where the end result for the person is negative finding. These cases are thus gFOBT-negative persons who are not invited to colonoscopy and gFOBT-positive persons who have no disease found in clinical confirmation. They may have other malady than cancer, but as the terminology above suggests, those are not labeled diseases. Terms screen-negative and episode-negative will be used interexchangeably from now on.

Finally, below is a table where mathematical notations and common abbreviations of this thesis are introduced.

Symbol/abbreviation	Meaning
(CT-)HMM	(Continuous Time) Hidden Markov model
$p(x\|\theta)$	$\mathbb{P}(X \in A \mid \theta)$, where A is an event.
R.V.	Random variable
MCMC	Markov Chain Monte Carlo
$\mathbf{P} \in \mathbf{R}^{N \times M}$	Matrix P with N rows and M columns.
$\mathbf{exp}(\mathbf{Q})$	Matrix exponential
(g)FOBT	(guaiac) Fecal Occult Blood Test
MSM	Multi-state model

Table 1: Abbreviations and mathematical notations

1.6 Outline

This thesis has six sections.

- Chapter 1 is introduction.
- Chapter 2 provides a few examples of how to mathematically quantify over-diagnosis and illustrate why some approaches may or may not work.
- Chapter 3 is about reviewing the necessary mathematical basis for building a Hidden Markov model, starting from very basic definition of a Markov chain and generalizing it to a continuous time Hidden Markov Model. Additionally

the chapter contains a general framework for making statistical inference in those models.

- Chapter 4 will focus on the data and likelihood details of this specific model.
- Chapter 5 provides the results.

Final chapter discusses the relevance of the findings and limitations of the model.

2 Estimands for overdiagnosis

The conceptual definition of overdiagnosis is counterfactual in its nature: in order to detect cancers in cancer screening, screening must be applied, and thus one cannot know for certain if the disease had progressed to a clinically significant cancer in case of no screening.

In ethically very questionable setting it would be possible to apply only screening and no treatments for found lesions and find which ones progress, but use of mathematical modelling is often more feasible solution.

Often this counterfactual quantity is tackled using methods that estimate either:

1 the excess cumulative incidence between screened population and expected incidence of screened population in case of no screening (for example: by estimating it from control population of a RCT), or

2 the natural history of colorectal cancer and derive overdiagnosis using the attained natural history model (for example: a hidden Markov model).

I'll start with expected incidence methods and their limitations and proceed to natural history methods.

2.1 Expected incidence methods

Most classical approach is using an approach where the expected incidence for population in case of no screening comes from a control population of RCT or similar study design.

To be more formal, let $\lambda_i(t)$ and $\lambda_u(t)$ be the CC incidence rates for invited population (not necessarily participicated) and uninvited (control population, not invited to screening) population starting from $t = 0$ beginning at the screening programme.

Then overdiagnosis can be defined as:

$$O_1 := 1 - \frac{\int_0^T \lambda_u(v)dv}{\int_0^T \lambda_i(v)dv}, \tag{1}$$

where T is some follow-up time extending hopefully far beyond the screening programme.

The need for long follow-up time comes from the effect screening has on cancer incidence. First, at the start of the programme, the incidence rate is bigger for the screening population since, as wished, the cancers are detected earlier. Later the effect should wear out a bit, since the cancers in control population will eventually (if they are not very slowly progressive or regressive) became evident clinically. This behavior is illustrated in figure 2.

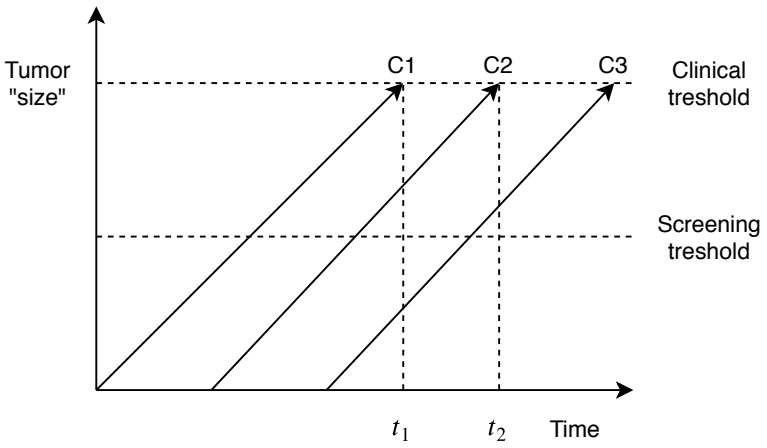

Figure 2: Effect of early detection. Y-axis represents the size of cancer tumour (in quotes since this is a more abstract "detectableness" but it should be very much correlated with size), x-axis time. Now at time t_1 cancers C1 and C2 would have been found if screened, but without screening C2 would not have been found. At t_2 similarly, all the three cancers would have been found, but C3 would have not in case of no screening. Yet the excess incidence caused by this phenomena should not be counted as overdiagnosis, since all the earlier detections were indeed favorable (assuming all the patients lived long enough to reach the clinical "barrier").

The approach of O_1, though, can be a bit misleading, since some of the population invited to programme will never attend. Thus the person-years accumulated by the non-attended population will smoothen out the incidence rate a bit, and the effects of screening in excess incidence will not be as evident as they should be.

One might be more interested in comparing only screen-attenders (given a certain screening round) to an estimate derived from the control population, but then using just incidence rate from control population yields biased estimates since screening attendance is voluntary and thus elective.

One method for correcting this phenomena is as follows. Assuming uninvited control population has the incidence rate $\lambda_u = \alpha\lambda_{u_1} + (1-\alpha)\lambda_{u_2}$, where α is the proportion of person years of screen-attenders (and λ_{u2} their incidence rate), λ_{u_1} has the interpretation as the unbiased expected incidence rate of screen-attenders in case of no screening. Then after solving the equation for λ_{u1} overdiagnosis becomes:

$$O_2 := 1 - \frac{\int_0^T \lambda_{u_1}(v)dv}{\int_0^T \lambda_a(v)dv}, \tag{2}$$

where λ_a is the incidence rate for screen-attenders.

While easy to interpret and conceptually sound, these approaches are are very much dependent on having a control population. If no control population exists, one has to do some kind of model (such as age-period-cohort model as in Hakama et al.

(2015) or another extrapolation) to calculate the counterfactual cancer incidence. These approaches, though, are very sensitive for assumptions made in the modelling process.

Additionally, since focus of this thesis is on overdiagnosis due to detecting pre-cancerous adenomas, these methods are destined to fail for two reasons.

First, people usually don't need medical care for regular adenomas, which means there's no data on adenoma incidence for control population. Additionally there exists the effect of reducing incidence. In colorectal cancer screening programme where almost 90% of the found lesions are indeed pre-cancerous stages that can be removed and thus the incidence can be reduced. If the incidence is reduced more than there is cancer overdiagnosis, these methods fail to account overdiagnosis for cancer as well.

The effect is illustrated in figure 3.

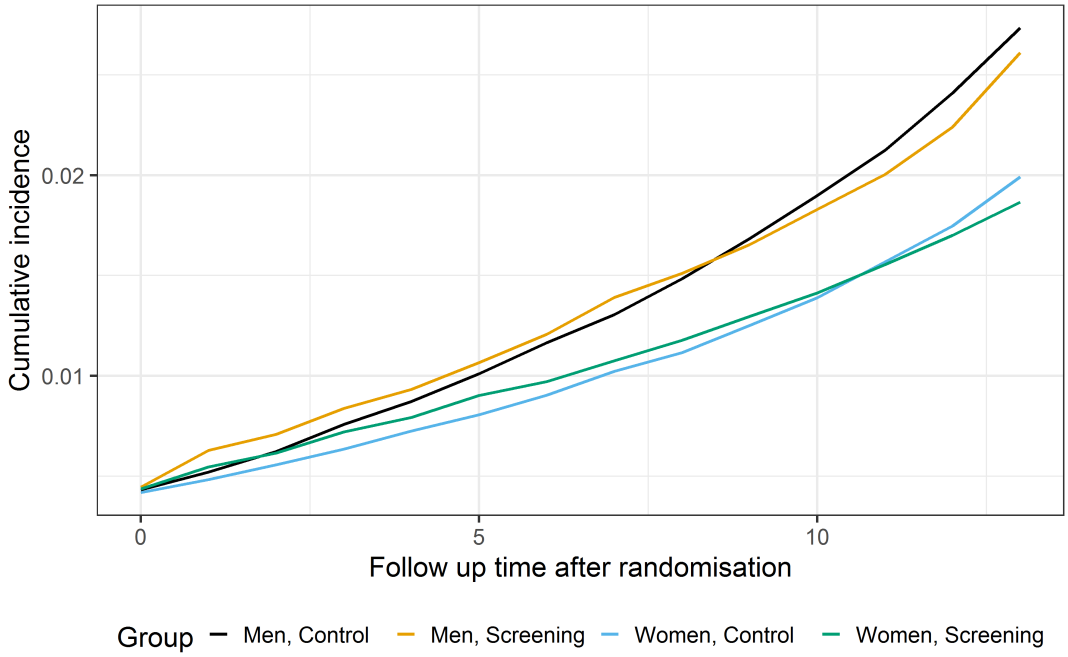

Figure 3: Cumulative incidence of colorectal cancer screening first increases due to early detection and overdiagnosis, then starts to decrease when the effect of incidence reduction starts to show. Eventually cumulative incidence in control population exceeds the cumulative incidence in screening population due to this effect in both groups of men and women.

Calculating O_1 and O_2 from this type of figure for long follow up time would yield negative estimates. This is not necessarily a contradiction with the heuristic that all the cancer programmes come with some kind of overdiagnosis, since the indicators are dealing with only incident cancers and overdiagnosis can be apparent in cancers

that are not yet in cancerous stage.

Additionally, since gFOBT CRC screening both reduces the incidence and detects cancers early, this just shows that the effect of incidence reduction is larger than the effect of overdiagnosis of malign cancers (category 4 type C in terms of introductory section).

Yet it should now be apparent that simply looking at the difference of the cohorts is not enough, when there are both effects of increasing incidence and reducing incidence. This opens up the field for another kind of estimands.

2.2 Natural progression based methods

Methods of type 2 provide very different approach for modelling overdiagnosis. The basic idea is to simulate the progression of colorectal cancer using some kind of multi-state model (MSM) and calculate overdiagnosis using the estimated MSM parameters. MSMs have been used extensively for cancer progression, and there are lot of different versions about how the cancer progression should be modelled.

Typical choice is to focus on states that are apparent on the data or that are relevant with respect to the quantities that are to be estimated.

One model used in this kind of setting was proposed by Wu et al. (2019) for breast cancer screening, where they defined four latent states that describe a natural progression breast cancer, namely: (1) healthy, (2) progressive pre-clinical cancer, (3) clinical disease and (4) non-progressive pre-clinical cancer. The approach of this thesis is conceptually similar, but due to different biological characteristics of breast cancer and CRC the model has to be adjusted a bit.

For CRC let's adopt five states: (1) healthy, (2) progressive pre-cancerous lesion, (3) pre-clinical cancer, (4) clinical cancer and (5) non-progressive pre-cancerous lesion. This multi-state stucture is illustrated in figure 4.

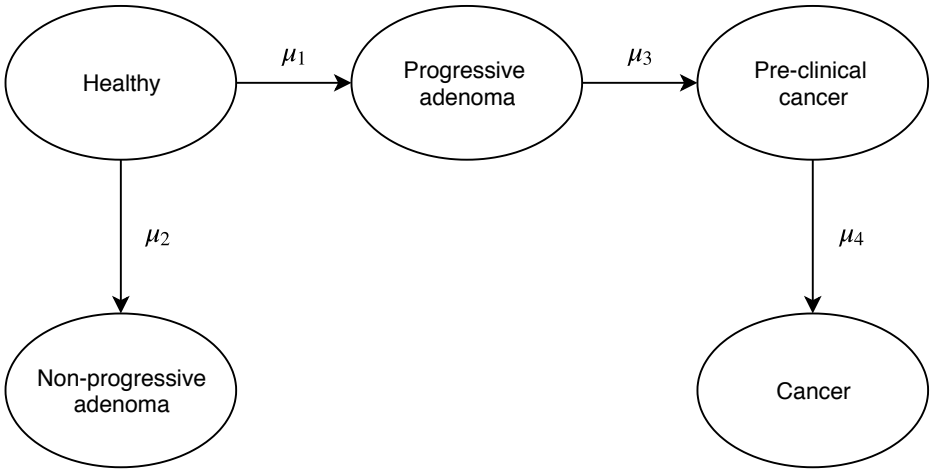

Figure 4: Transition structure of the colorectal cancer overdiagnosis model. One must note that the model has only transitions forward.

Now transitions between these states are described using matrix $\mathbf{Q}$, where element i, j describes transition intensity from state i to state j:

$$\mathbf{Q}(t;z) = \begin{bmatrix} -(\mu_1(t;z) + \mu_2(t;z)) & \mu_1(t;z) & 0 & 0 & \mu_2(t;z) \\ 0 & -\mu_3(t;z) & \mu_3(t;z) & 0 & 0 \\ 0 & 0 & -\mu_4(t;z) & \mu_4(t;z) & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \quad (3)$$

Theory behind these MSM models will be addressed in chapter 3, so without going into details: t denotes time and z the covariates associated with the person (sex, eating habits, and so on). $\mu_1(t;z)$ is the transition rate from being free of CC and adenoma to a progressive adenoma and $\mu_3(t;z)$ the transition rate from progressive pre-clinical disease to pre-clinical cancer, and $\mu_4(t;z)$ is the transition from pre-clinical cancer to clinical cancer. This is the assumed progression path from adenoma-free to cancerous.

Additionally $\mu_2(t;z)$ represents the transition intensity from free of CC to non-progressive adenomas (these are the diseases, that are labeled as overdiagnosed ones, in case they are detected). This state transition structure is visualised in 4.

From the transition intensity matrix one can construct transition probability matrix $\mathbf{P}(t;z)$, and that will be addressed in chapter 3.

The problem in screening setting is that we don't know when the transitions are made, nor if there has really happened a transition (gFOBT-based CRC screening comes with relatively low sensitivity). Thus one needs to have one more layer of abstraction: assume that given state X_t the screening observations come from certain probability distribution $\mathbb{P}(Y_t \mid X_t)$ where Y_t represents the observation. These possible observations are (1) no pre-clinical finding, (2) adenoma, (3) screen-detected CC. Clinical cancers (4), in this model, are defined to become clinical the precise moment they are clinically diagnosed, and thus clinical disease is observed if and only if in clinical state: $\mathbb{P}(Y_t = 4 \mid X_t = 4) = 1, \mathbb{P}(Y_t = 4 \mid X_t \neq 4) = 0$.

Since our observations are categorical, this behavior is best characterised using a matrix $\mathbf{E}$ that is called emission matrix in some machine learning lingo:

$$\mathbf{E}(t,z) = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} e_{11}(t;z) & 0 & 0 & 0 \\ e_{21}(t;z) & e_{22}(t;z) & 0 & 0 \\ e_{31}(t;z) & 0 & e_{33}(t;z) & 0 \\ 0 & 0 & 0 & 1 \\ e_{51}(t;z) & e_{52}(t;z) & 0 & 0 \end{pmatrix} \end{array} \tag{4}$$

Each row $e_i(t;z)$ of this emission matrix $\mathbf{E}(t;z)$ defines a conditional distribution $Y_t \mid X_t \sim \text{Categorical}(e_i(t;z))$, where z again represents covariates associated with the person. In probability sense this means that:

$$e_{ij}(t;z) = \mathbb{P}(Y_t = j \mid X_t = i, Z_t = z). \tag{5}$$

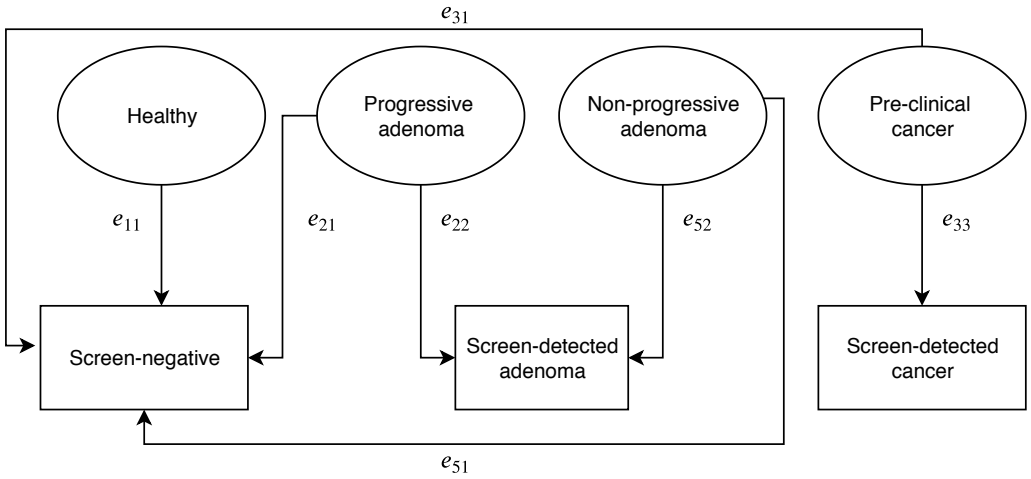

Figure 5: Emission stucture of the hidden Markov model. If person is healthy, it is assumed that there will be no colonscopy-found adenoma or cancer. Progressive and non-progressive adenomas output episode-negatives (if no blood is or when the clinical colonoscopy confirmation of the disease is negative) and screen-positive findings (when both gFOBT and colonoscopy are positive). The emission for clinical appearance of the cancer is omitted from the figure, since it is assumed to follow identity distribution (only clinical cancers are found outside screening programme and there is no misclassification assumed).

Here many of the e_{ij} have common epidemiological interpretations: e_{11} is the specificity, e_{22} is the sensitivity for progressive pre-clinical disease and e_{33} is the sensitivity for non-progressive pre-clinical disease. The notation e comes from word "error", since the matrix is often called the misclassification matrix. To simplify model, I will use episode sensitivity explained in chapter 1.5. The emission structure is visualised in the figure 5.

Parameter estimation for models based on these two matrices are discussed in chapter 3, but using them once can define overdiagnosis in multiple ways.

For one screening round overdiagnosis can be thought as the ratio of cumulative incidence of non-progressive pre-clinical disease and the cumulative incidence of all pre-clinical diseases, T being the length of the screening round:

$$O_3 = \frac{\int_0^T e_{52}(v;z)\mu_2(v;z)dv}{\int_0^T [e_{22}(v;z)\mu_1(v;z) + e_{52}(v;z)\mu_2(v;z)]dv}. \tag{6}$$

Assuming constant rates and sensitivities this simplifies to:

$$O_3 = \frac{e_{52}\mu_2}{\mu_1 e22 + \mu_2 e_{52}} \tag{7}$$

Another approach that's easier to generalise over a programme with multiple screening rounds would be looking at a conditional probability instead. Overdiagnosis can be thought as the probability of being in non-progressive state conditioned on getting a screen-positive result. This is dependent on not only covariates and time but also the number of screening rounds (every screening round some adenomas are found and removed and thus no longer detectable).

Generally this quantity is:

$$O_4 = \mathbb{P}(X_k = 5 \mid Y_k = 2) \tag{8}$$

Usually we are interested on overdiagnosis in some part of population and some part of screening programme. For example, quantity of interest might be overdiagnosis in women who entered screening programme at the age of 60. Additional interest could be overdiagnosis in given screening round. In that case one must condition the quantity O_4 on not having earlier adenomas detected (since then a treament is applied and the person is no longer included in the model).

For first screening round at time 0 given covariate-dependent initial distribution $\theta(z)$, we get:

$$\begin{aligned} O_4 &= \frac{\mathbb{P}(Y_0 = 2 \mid X_0 = 5)\mathbb{P}(X_0 = 5)}{\mathbb{P}(Y_0 = 2)} \\ &= \frac{e_{52}(z)\theta_5(z)}{e_{22}(z)\theta_2(z) + e_{52}(z)\theta_5(z)}, \end{aligned} \tag{9}$$

where the last expression comes from

$$\begin{aligned} \mathbb{P}(Y_0 = 2) &= \mathbb{P}(X_0 = 2, Y_0 = 2) + \mathbb{P}(X_0 = 5, Y_0 = 2) \\ &= \mathbb{P}(Y_0 = 2 \mid X_0 = 2)\mathbb{P}(X_0 = 2) + \mathbb{P}(Y_0 = 2 \mid X_0 = 5)\mathbb{P}(X_0 = 5). \end{aligned}$$

Additionally probabilities $\mathbb{P}(X_0 = 2)$ and $\mathbb{P}(X_0 = 5)$ come just from initial distribution $\theta(z)$.

As said, overdiagnosis is dependent on both covariates and consecutive number of screening round. Setting a fixed length for screening round one can estimate this using an estimand such as:

$$O_4(z, r) = \mathbb{P}(X_{2r} = 5 \mid Y_{2r} = 2) \tag{10}$$

for 0th (so-called prevalence round), 1st, 2nd round, and so on.

Another approach, that could be more of interest for public health policy makers, is using actual persons observed in the programme instead of hypothetical screening sequences.

Then the estimand of interest is again similar to O_4, but instead of conditioning on hypothetical inviduals with constant screening round times, we calculate average of single-invidual i probabilities

$$\mathbb{P}(X_{K_i,i} = 5 \mid Y_{1,i}, \dots Y_{K_i,i} = 2) \tag{11}$$

over the actual screened population. This quantity can be calculated, given a model, using Forward algorithm that will be addressed in chapter 3.

Overdiagnosis probability in the whole screening programme can then be calculated as the average of probabilities over all screen-found adenomas:

$$O_5 = \frac{1}{N} \sum_{i=1}^{N} \mathbb{P}(X_{K_i,i} = 5 \mid Y_{0,i}, \dots Y_{K_i,i} = 2) \tag{12}$$

Next chapters will focus on building a model to address overdiagnosis definitions O_4 and O_5.

3 Markov models

3.1 Markov chains in discrete and continuous time

Some theoretical basics are necessary for building a hidden Markov model that is needed to address overdiagnosis definitions O_3, O_4 and O_5 in previous chapter.

Starting from the very basic first order Markov assumption: it can be heuristically stated as: "the future is independent of the past, given the present". Markov process, in general, is a process that satisfies the Markov assumption.

Definition 13. A Markov chain

Let $\{X_n\}_{n\in\mathbb{N}}$ be a discrete time stochastic process. Let K be its state space, that means $\forall n X_n \in K$. $\{X_n\}_{n\in\mathbb{N}}$ is a Markov chain, if $\forall i, j \in K, n \in \mathbb{N}$ the equation

$$\mathbb{P}(X_{n+1} = i | X_0 = x_0, \ldots, X_n = j) = \mathbb{P}(X_{n+1} = i | X_n = j)$$

holds.

Letter S stands for state space of the Markov chain/process. The state space is assumed to be finite for the rest of the thesis, since the main focus is to build a model suitable for disease modelling.

Discrete time Markov chains are often characterised using transition probability matrix.

Definition 14. Transition probability matrix

Consider a discrete time Markov chain $\{X_n\}_{n\in\mathbb{N}}$. Let S be the state-space. Let $\mathbf{P} \in \mathbb{R}^{|S|\times|S|}$, where $|S|$ is the cardinality of the state space. $\mathbf{P}$ is a transition probability matrix if $\forall i, j, n$:

$$p_{ij} = \mathbb{P}(X_{n+1} = j \mid X_n = i),$$

that is the elements of the matrix p_{ij} represent probabilities of moving from state i to state j.

This implies that $\sum_{j\in S} p_{ij} = 1$ for all i.

Discrete time Markov chains are especially useful in solving problems where observations come in sequential form. This would be great when considering only observations from cancer screening programme. In cancer screening, though, it is required to move on to continuous time processes, since the observed interval cancers defined in chapter 1.5 appear in continuous time.

For continuous time Markov process the definition 13 doesn't make sense anymore since the set of time-indexes is uncountable. The usual way of getting rid of this problem is by using a transition intensity matrix. One example of those was presented in equation 3.

Definition 15. Transition intensity matrix

Let S be a finite state space. Transition intensity matrix of a continuous-time Markov process is a matrix $\mathbf{Q} \in R^{|S|\times|S|}$ where $\forall i \neq j$ it holds that $q_{ij} \geq 0$ and $\forall i \sum_{j\in S} q_{ij} = 0$.

This implies also that $q_{ii} = -\sum_{j\neq i} q_{ij}$

The elements of the transition intensity matrix define a limit:

$$q_{ij} = \lim_{\delta h \to 0} \frac{\mathbb{P}(X_{t+\delta h} = j \mid X_t = i)}{\delta h}$$

When $\mathbf{Q}$ is constant with respect to time, one can compute probability transition matrix $\mathbf{P}$ using so called Kolmogorov equation $\mathbf{P}'(t) = \mathbf{P}(t)\mathbf{Q}$ with initial condition $\mathbf{P}(0) = \mathbf{I}$, and get very neat closed-form equations.

Continuous time Markov chain with a constant transition intensity matrix is called time-homogeneous. When $\mathbf{Q}$ varies over time, it's not guaranteed there exists a closed-form solution to this equation. These chains are called time-inhomogeneous or time-heterogeneous. For the rest of this thesis I will make time-homogeneity assumption to make calculations more tractable.

Solution to Kolmogorov equation is given by matrix exponential: $\mathbf{P}(t) = \mathbf{exp}(\mathbf{Q}t) = \sum_{k=0}^{\infty} \frac{\mathbf{Q}^k t^k}{k!}$. I will refer to matrix exponential from now on with **exp** (note the bolded font).

Definition 16. Continuous time Markov chain (Markov process)

Let $Q \in R^{|S|\times|S|}$, and let $\mathbf{P}(t) = \mathbf{exp}(\mathbf{Q}t)$. $\{X_t\}_{t\in\mathbb{R}}$ is a continuous time Markov process, if for all $T \in K$ and for all times $t_i < t_j \in \mathbb{R}$:

$$\mathbb{P}(X_{t_j} = j \mid X_{t_i} = k_i) = p_{ij}(t_j - t_i)$$

Theorem 17. *If* $\mathbf{Q}$ *is a transition intensity matrix with constant intensities, then corresponding probability matrix* $\mathbf{P}$ *has the property for* $t+s > 0$ $\mathbf{P}(t+s) = \mathbf{P}(s)\mathbf{P}(t)$.

Proof. $\mathbf{P}(t+s) = \mathbf{exp}(\mathbf{Q}(t+s)) = \mathbf{exp}(\mathbf{Q}t)\mathbf{exp}(\mathbf{Q}s) = \mathbf{P}(s)\mathbf{P}(t)$ □

The Kolmogorov equation also has the extremely useful property that can be used for modelling missing observations, that is highlighted in the next example.

Example 18. Let $\mathbf{Q} \in \mathbb{R}^{3\times3}$ be a transition intensity matrix and $S = \{1, 2, 3\}$ its state space. Assume further that possible transitions are $1 \to 2$ and $2 \to 3$ with intensities λ_1 and λ_2. Now:

$$\mathbf{P}(t) = \exp(\mathbf{Q}t) = \begin{pmatrix} e^{-\lambda_1 t} & \frac{\lambda_1\left(e^{-\lambda_2 t} - e^{-\lambda_1 t}\right)}{\lambda_1 - \lambda_2} & 1 - \frac{\lambda_1(e^{-\lambda_2 t} - e^{-\lambda_1 t})}{\lambda_1 - \lambda_2} \\ 0 & e^{-\lambda_2 t} & 1 - e^{-\lambda_2 t} \\ 0 & 0 & 1 \end{pmatrix}$$

There are no possibility for direct transformation between 1 and 3, but if, for some reason, the transition $1 \to 2$ is missed, the model is still useful. This is the most typical thing in cancer screening setting. If we assume states to be such that 1 represents being healthy, 2 having a pre-clinical cancer and 3 having a clinical cancer, we often notice transfers from 1 to 3, even though it is assumed that the cancer always passes through a pre-clinical state.

But matrix exponential accounts for this and gives probability for passing to state 3 throught state 2:

$$\mathbb{P}(X_{t_1} = 3 \mid X_{t_0} = 1) = 1 - \frac{\lambda_1(e^{-\lambda_2 t} - e^{-\lambda_1 t})}{\lambda_1 - \lambda_2}.$$

It has to be noted that even this very simple example yields quite complex equations.

3.2 Hidden Markov models

As in this thesis, the underlying state of a Markov process is not itself fully observable due to some misclassification error. In these situations one makes inference via outputs that are dependent on the underlying state. This setting is called a hidden Markov model. In Continuous time Hidden Markov model (abbreviated HMM from now on) there is an underlying continuous time Markov process $\{X_t\}_{t\in\mathbb{R}}$ that is not observed. The chain, though, outputs observed data with the distribution of the observed variable differing based on the hidden state.

The very basic structure of a HMM is illustrated in figure 6.

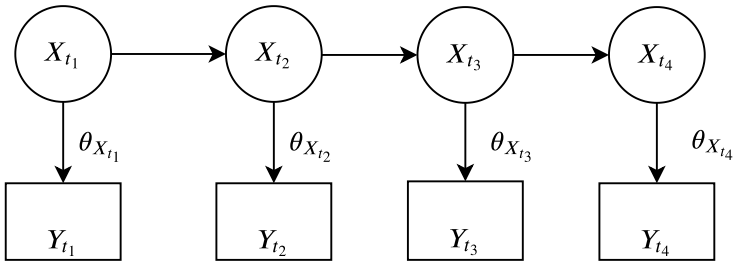

Figure 6: Hidden Markov Model with observations Y, hidden states X and parameters θ_X for distribution $Y \mid X$.

Outputs Y_t are often called emissions in HMM literature. Assuming $Y_t \mid X_t$ has a parametric distribution, their relation can be determined by state-specific emission parameter θ_X. Example of this was already addressed in equation 4, where we noted $e_X = \theta_X$.

Definition 19. Hidden Markov process Let $\{X_t\}_{t\in\mathbb{R}_+}$ be a continuous time Markov process. Then the pair of sequences $\{\{X_t\}_{t\in\mathbb{R}_+}, \{Y_t\}_{t\in\mathbb{R}}\}$ is a Hidden Markov process, if only $\{Y_t\}_{t\in\mathbb{R}}$ is observed and if $\forall t$:

$$\mathbb{P}(Y_t \mid X_t) = \mathbb{P}(Y_t \mid X_{s\in T}, Y_{s\in T\setminus\{t\}}),$$

where T represents the set of times Y_t was observed.

In application of this thesis, $Y_t \mid X_t \sim \text{Categorical}(\theta_{X_t})$.

For sake of simplicity, I will use notation $\mathbb{P}(Y_t \mid X_t) := p(y_t \mid x_t)$ from now on.

3.3 Adding covariates

Both the emissions and state-transitions are often thought to be dependent on certain covariates associated with a person. Adding covariates to a hidden Markov model is quite straight-forward, as long as the covariates are assumed to be time-invariant.

Example for setting the covariates in this setting is found from, for example, Duffy et al. (2003).

Covariates z can be added to state transitions using logarithmic link function. This yields a new transition intensity matrix where elements q_{ij} are of form:

$$q_{ij}(z) = q_{ij} \exp(\beta z). \tag{20}$$

Similarly covariates w that can be same as z but don't have to, can be added to emission probabilities using logistic link:

$$\log(\frac{e_{ij}}{1 - e_{ij}}) = \gamma w. \tag{21}$$

3.4 Inference in Hidden Markov models

When making inference of HMM, one has two different problems: parameter inference (parameter estimation) and state sequence inference (state sequence estimation). This chapter goes through them both.

I'll discuss first how to estimate the parameters and then proceed to "online learning" that is estimation of state at time t given observations up until time point t, namely $p(x_k \mid y_{1:k})$.

3.4.1 Parameter estimation

All inference in Hidden Markov models requires writing down the likelihood of the model that is the conditional density of the data given model parameter. In frequentist domain this likelihood is then usually maximized to get an maximum likelihood estimate for parameter. Typical way to do this is the expectation-maximisation algorithm.

In the Bayesian setting, if $\mathbf{Z}$ is the data and θ is the parameter of the model, the task is to estimate posterior distribution, that is the distribution of parameter given data:

$$p(\theta \mid \mathbf{Z}) = \frac{p(\mathbf{Z}|\theta)p(\theta)}{p(\mathbf{Z})}. \tag{22}$$

Here $p(\mathbf{Z}|\theta)$ is the model likelihood and $p(\theta)$ the prior distribution of the parameter. Since these two have a closed form solution, the task reduces to calculating so-called normalizing coefficient $p(\mathbf{Z}) = \int_\Theta p(\mathbf{Z}, \theta)d\theta$. This is done using Markov Chain Monte Carlo integration (MCMC), since the integral has no tractable solution. In this thesis the integration is done using software Stan, that is introduced by Carpenter et al. (2014).

Stan uses a certain variant of the Hamiltonian Monte Carlo (HMC) No-U-Turn-Sampler (NUTS) algorithm. The mathematics behind HMC have been investigated by Betancourt (2014) and conceptually described by Betancourt (2018). The very basic naive HMC algorithm as described by Hoffmann and Gelman (2012) in their NUTS-paper is attached to Appendix 1.

Details of the likelihood $p(\mathbf{Z}|\theta)$ are addressed in chapter four.

3.4.2 State sequence estimation

Another type of estimation problem one may face with Hidden Markovian models is the estimation of the hidden state sequence. This procedure is concerned with the probabilities $p(x_t \mid y_{0:t})$, where $t \leq n$. One is often very much interested what was the hidden state at given state t (for example, was the disease non-progressive during a screen, given current and past obsevations).

As one may note, this is directly associated with the defintion O_5 of overdiagnosis in chapter 2.

Practically there exists two possibilities for this state estimation: hard and soft classification. Hard classification is interested in maximum a-posteriori (MAP) estimates of state given observations, while soft classification is more concerned with probabilities. Hard classification in HMM is often done using Viterbi algorithm, while the common procedure for soft classification is the forward filtering and backward smoothing. All of these algorithms require a successful parameter estimation in advance, since they make use of the fitted statistical model.

Forward filtering yields probabilities $p(x_k \mid y_{0:k})$. So-called forward-backward algorithm would extend this to $p(x_k \mid y_{0:K})$ for $K > k$, but forward algorithm is sufficient for purposes of this thesis. Forward algorithm it also provides a way for specifying the likelihood in practical terms (see section 4.3).

These algorithms and their correctness is discussed in detail in chapters 3 (their correctness) and 5 (their application for models with discrete state space) of the book Inference in Hidden Markov Models by Cappé et al. (2005).

Algorithm 1: Forward algorithm

Result: ν

Let S be the state space of the Hidden Markov chain. Let $p(x_k|x_{k-1})$ be transition probabilities from state x_{k-1} to x_k, $p(y_k \mid x_k)$ the emission probabilities, K be the number of observations and $\mathbf{y} \in S^{|K|}$ be the observed sequence and k = 1.
$\forall x_0 \in S$ set $\nu_0(x_0) = p(x_0)$ that's the initial state distribution.;

while $k <= K$ **do**
  for $x_k = 1, \ldots, S$ **do**
    $\nu_k(x_k) = p(y_k \mid x_k) \sum_{x_{k-1} \in S} \nu_{k-1}(x_{k-1}) p(x_k \mid x_{k-1})$;
    $k = k + 1$;
  end
end

The forward algorithm's result are the probabilities $p(x_k, y_{0:k}) = \nu_k(x_k)$. From this one can calculate $p(x_k \mid y_{0:k}) = \frac{\nu_k(x_k)}{\sum_{x_k \in S} \nu_k(x_k)}$, that are what is the probability of being in state x_k given all the information we have at time t_k. This is enough for purposes of this thesis. The resulting probabilities $p(x_k \mid y_{0:k})$ are the probabilities that the hidden state was x_k given current observations and all the pats observations (at time t_k corresponding to k). This, given that observation y_k was screen-found adenoma, is the single person O_5 from chapter 2.

3.4.3 Uncertainty of the estimates

Whole point of computational Bayesian statistics is to generate observations from posterior distribution expressed in equation (22).

So assuming the sampling works, the end product of the analysis is a set of observations $(\Theta_i)_{i \in \mathbb{N}} \sim p(\Theta \mid \text{data}, \text{prior distribution parameters})$.

Thus a general algorithm for producing estimates for uncertainty of the model is:

Algorithm 2: Uncertainty estimation algorithm

Result: U

Let $(\theta_i)_{i \in \mathbb{N}} \sim p(\theta \mid \text{data}, \text{prior parameters})$ be a sample of the model parameter from the posterior distribution. Let $f : \theta \to \xi$ be a function that maps the parameter to desired estimand. For some K:;

for $k = 1, \ldots, K$ **do**
  Sample $\theta_k \sim p(\theta \mid \text{data}, \text{prior parameters})$
  Calculate $f_k(\theta_k) = f_k$
end

With $f_{1:K}$ calculate desired uncertainty estimate U.

Few examples of uncertainty estimates U are:

U1 Visualised empirical distribution of the estimand or parameter.

U2 Sample variance of the $f_{1:K}$.

U3 Quantiles of $f_{1:K}$

This theoretical background will be put to use in the following section.

4 Hidden Markov model for colorectal cancer

4.1 Data

This thesis uses the data from two different sources: 1) The Finnish Cancer Registry (FCR) and 2) Randomised health service study (RHS) completed in Finland between 2004 and 2016, where there exists over 400 000 persons randomised into screening or control groups. The RHS population consists of 60 to 69 years old men and women living in municipalities that participated in the program.

Persons randomised into these two cohorts enter to the study during varying times between 2004 and 2016. This is all done according to randomisation scheme described in Malila et al. (2005), where the municipalities that attended the program are also listed.

Basic outline of the study in multi-state model perspective is described in table 2.

Analysis specification	
Inclusion criteria	All screen-attenders excluding prevalent CRCs
Events	Episode-negative (1), Adenoma (2), Screen-CRC (3), interval CRC (4)
Time origin	First attended screening
Entry time	First attended screening
Censoring rule	Death without CRC or 3 years after final screen-participation
Follow-up time	Until 3 years after final screen-participation

Table 2: Overview of the multi state model details.

The study population is formed from the RHS population as follows:

- The control population is excluded.
- All the persons who have never attended the screening programme are excluded from the study.
- Persons are linked to FCR data using social security number and those who have had a CRC before first screening participation are excluded from the model.

For these patients an event-history is formed from following events: screen-found adenomas, negative gFOBT-results, positive gFOBT-results with negative colonoscopy, screen-found CRCs and interval cancers.

The data in RHS is structured as follows: every person has a specified screening round, that includes gFOBT test. If the test is positive, there exists a colonoscopy event that corresponds to the same screening round. From this data event history is specified as follows:

- If person has a screening round without completed gFOBT test, that round is removed from the data.
- If person was over 70 years old when the screening took place, that round is removed from the data.
- If person has a screening round with negative gFOBT, that round is added to data as screen-negative event on the date the gFOBT result was recorded.
- If person has a screening round with positive gFOBT but no colorectal polyps found in colonoscopy, that round is added to data as screen-negative event on the date colonoscopy exam was conducted.
- If person has CRC/adenoma in the colonoscopy succeeding a positive gFOBT, screen-detected CRC/adenoma are added to event history of the person.
- Interval cancers (from Cancer Registry) are added to data using rolling join procedure of R-package data.table. All cancers from the FCR data are linked to the strictly preceding screening round of a person. If a round has both screen-detected and clinical CRC, clinical CRC is removed from the data.
- Additionally, all the events after first not-screen-negative observation are removed from the data.
- If an interval cancer is found and person would still have screening rounds left after finding, then the cancer is assumed to be found as clinical cancer at the screen check-up time, that is set to be two years after preceding screen.
- If an interval cancer is found and person does not have any scheduled screening rounds left, the cancer is assumed to be found as clinical cancer three years after previous screening round.
- If person exists the screening programme and doesn't develop cancer in following three years, it's assumed the person leaves the follow up in one of states (1), (2) or (3).
- If person dies during a screening round, the follow up stops at the start of that screening round.

The rationale behind this is to create a standard structure for the screening process. I assume that persons are observed in approximately two year intervals while in the programme, and force clinical cancers into this form even though the diagnosis dates for those are known. This is because 1) clinical cancers are known to be diagnosed with certain delay, Syöpätaudit (2016) cites mean delay from symptoms to detection to be as high as 10 months, and because it's easier to build model where all types of event have similar observation structure.

In real world all persons are invited to participate again in the screening programme independent of the screening results. Since the focus is now in overdiagnosis, this thesis is only interested in the very first event.

For people with no events beside screen-negatives, follow-up is extended up to either three years beyond final screening round or 31. December 2017, whichever is the first. This is due to cancer registry's close date, and after that there are no data of cancer diagnosed outside screening programme. If no clinical cancer appears between final screening and that time point, person is assumed to be either in healthy state, in some of two adenoma states or in pre-clinical cancer state. If a cancer is found during that period, it is assumed to be found three years after final round or 31.12.2017, whichever the first.

Now all this has been combined into a time histories, figures 7 and 8 illustrate some of them in two different time scales.

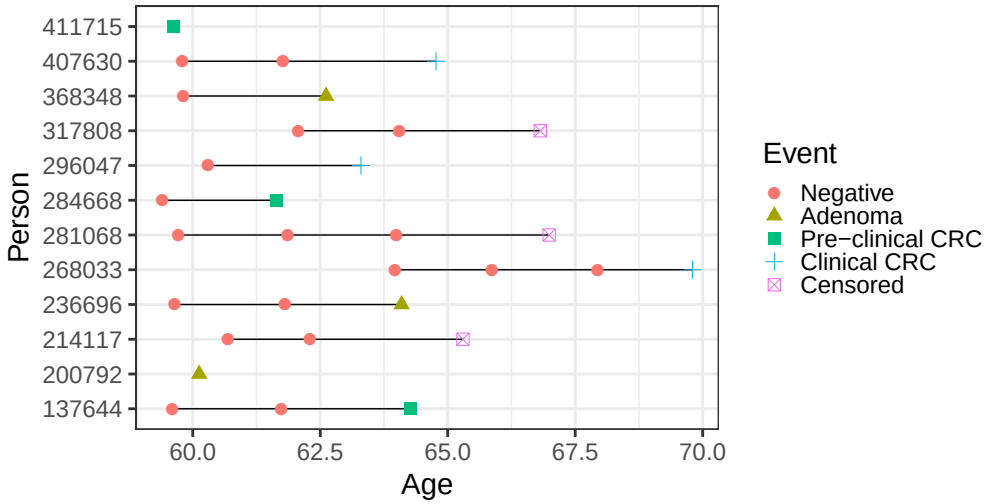

Figure 7: Some randomly picked up event histories for given subjects with age as a time scale. Lines represent time the person is being followed up. If there is no line, person only contributes to initial distributions and sensitivities.

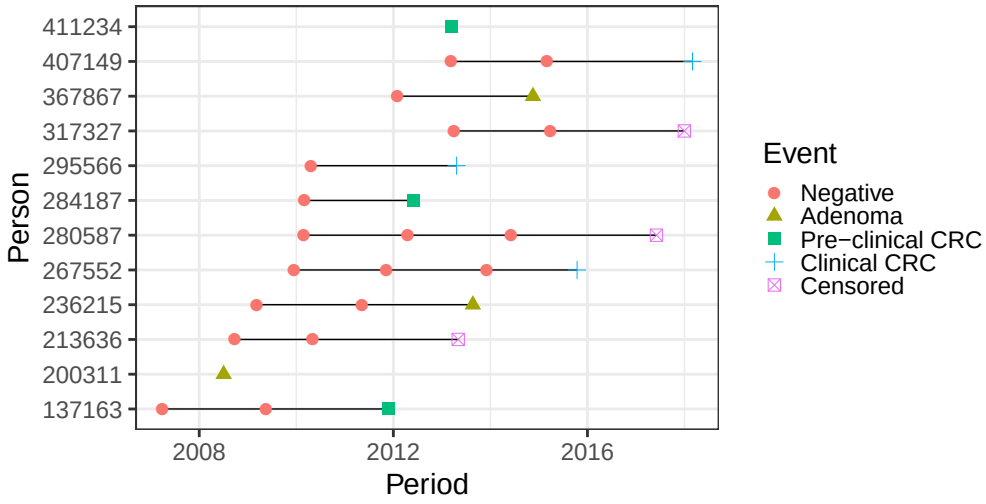

Figure 8: Some randomly picked up event histories for given subjects with calendar time as a time scale. Lines represent time the person is being followed up. If there is no line, person only contributes to initial distributions and sensitivities.

As the emission matrix (4) in chapter 2 suggests, the data is partially observed. It means that there exists observations when, according to model, the hidden state is actually revealed. These are 1) clinical observations and 2) pre-clinical malign observations. Also, given observed adenoma, the hidden state is known to be either progressive or non-progressive adenoma.

In the table 3 observation counts by observation rank are tabularised:

sex	Event	1	2	3	4	5	6
Men	1	76393	62227	38968	21207	7497	0
Men	2	1167	778	465	211	64	0
Men	3	136	60	46	19	9	0
Men	4		90	77	58	43	16
Women	1	93575	80471	52367	29176	10430	3
Women	2	492	425	258	120	42	0
Women	3	81	51	39	17	7	0
Women	4		108	93	74	48	26

Table 3: Rows present events and columns their subsequent numbers in the event history of the person. Here one can see that 93575 women started the screening programme with episode-negative screening result. Observation rank corresponds to screening round with exception for clinical findings (4). Persons who have had six and some who have had five observations may have two episode-negatives for one screening round, but this does not effect the model or estimands.

4.2 Likelihood under the model

As described in chapter 2, the probability model for CT-HMM is built from emission matrix $\mathbf{E}$ and transition probability matrix $\mathbb{P}(t)$. Additionally, a initial distribution of hidden state θ at time of the entry is required.

Let's assume the hidden state X_0 at time of entry, that is the time person first attends the screening programme, has a Categorical(θ) -distribution, where:

$$\theta = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 & 0 & \theta_5 \end{bmatrix}' \tag{23}$$

θ_4 is assumed to be zero, since it would not make sense to enter people already in a clinical phase of the cancer.

In reality θ varies depending on sex and age. Thus for the model I'll stratify θ to eight different stratas based on four age groups ($<$ 60 years old, 60-61, 62-64 and 65+) and two sexes (male, female).

Then, to make estimation of transition intensity matrix easier, I will make simplifying assumption that the transition rates are sex-dependent, but piecewise constant with respect to age. So I will fit the model using four transition rates for each for transitions: rate for young and old males ($<=$ 65, $>$ 65), rate for young and old females ($<=$ 65, $>$ 65).

Now let $\mathbf{Q}$ be the transition intensity matrix of the Markovian model:

$$\mathbf{Q}(t) = \begin{bmatrix} -(\mu_1(z) + \mu_2(z)) & \mu_1(z) & 0 & 0 & \mu_2(z) \\ 0 & -\mu_3(z) & \mu_3(z) & 0 & 0 \\ 0 & 0 & -\mu_4(z) & \mu_4(z) & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{24}$$

It is the same as the matrix (3) in chapter 2, but with the exception that the rates are assumed to be constant over time.

Here it is to be noted that possible transitions are from healthy to progressive adenoma, from healthy to non-progressive adenoma, from progressive adenoma to screen-detectable cancer and from screen-detectable maling cancer to clinically apparent cancer.

Using matrix exponential $\mathbf{P}(t) = \mathbf{exp}(\mathbf{Q}t)$ one can solve the Kolmogorov equation to get the time-dependent transition probability matrix. The analytical approach solves some computational issues that may arise when computing, for example, derivatives of the numerical matrix exponential.

The elements of solved matrix equation are quite complex and thus they are moved to Appendix 3. Thus, I will refer to probability matrix as the following matrix:

$$\mathbf{P}(t) = \begin{bmatrix} p_{11}(t;z) & p_{12}(t;z) & p_{13}(t;z) & p_{14}(t;z) & p_{15}(t;z) \\ 0 & p_{22}(t;z) & p_{23}(t;z) & p_{24}(t;z) & 0 \\ 0 & 0 & p_{33}(t;z) & p_{34}(t;z) & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} = (P_{ij})(t;z). \tag{25}$$

The transition probability matrix is both time- and covariate-dependent. Here covariates are as in the transition rate matrix, the age group at beginning of screening round and sex.

Next it's needed to address the emission probabilities. They are stratified by sex.

First assume that $p(y_k = 1 \mid x_k = 1) = 1$. This is due to the fact that the model will use the whole diagnostic episode as the screening test. It is reasonable to assume that if a healthy person has blood in their stool and is further taken into colonoscopy, there won't be any adenomas or cancers found if there aren't any.

Additionally parameters representing following quantities are required:

1) $S_e = p(y_k = 2 \mid x_k = 2)$, the sensitivity for progressive adenoma.

2) $S_d = p(y_k = 3 \mid x_3 = 3)$, the sensitivity for pre-clinical screen-detectable CC.

3) $S_{od} = p(y_k = 2 \mid x_k = 5)$, the sensitivity for non-progressive adenoma.

Here indices e, d, od refer to words "episode", "detection" and "overdiagnosis". Now using these elements to build the emission matrix we get:

$$\mathbf{E} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ (1 - S_e(z)) & S_e(z) & 0 & 0 \\ (1 - S_d(z)) & 0 & S_d(z) & 0 \\ 0 & 0 & 0 & 1 \\ (1 - S_{od}(z)) & S_{od}(z) & 0 & 0 \end{pmatrix} \end{array} = (e)_{ij}. \tag{26}$$

For the sake of notational simplicity I will refer to elements of $\mathbf{E}$ as e_{ij}. This is a special case of the matrix (4).

Now, for a given individual i, we have complete data likelihood for observations made during screening process:

$$\begin{aligned} p(y_{0:K}, x_{0:K} \mid \Theta) &= p(x_0)p(y_0 \mid x_0) \prod_{k=1}^{K} p(x_k \mid x_{k-1}, y_{k-1} \dots y_0, x_0) \times \\ &\quad p(y_k \mid x_k, x_{k-1}, y_{k-1} \dots y_0, x_0) \\ &= p(x_0)p(y_0 \mid x_0) \prod_{k=1}^{K} p(x_k \mid x_{k-1}, y_{k-1} \dots x_0, y_0)p(y_k \mid x_k) \\ &= p(x_0)p(y_0 \mid x_0) \prod_{k=1}^{K} p(x_k \mid x_{k-1}, y_{k-1} \dots y_0)p(y_k \mid x_k) \\ &= p(x_0)p(y_0 \mid x_0) \prod_{k=1}^{K} p(x_k \mid x_{k-1})p(y_k \mid x_k). \end{aligned} \tag{27}$$

Here indices k refer to order of observations, but the observations come with certain observation times. The observation times are assumed to be non-informative, since they refer to deterministic administrative decisions that define the screening times. Thus every observation k is associated with a time t_k, so the likelihood could we written in terms of initial distribution, probability transition matrix and emission matrix as:

$$
\begin{aligned}
&p(x_0)p(y_0 \mid x_0)\prod_{k=1}^{K} p(x_k \mid x_{k-1})p(y_k \mid x_k) = \\
&\theta(x_0)(z)e_{x_0,y_0}(z)\prod_{k=1}^{K} p_{x_k,x_{k-1}}(t_k - t_{k-1}; z)e_{x_k,y_k}(z).
\end{aligned}
\tag{28}
$$

When the underlying states of the process are not observed, one has to take sums over hidden states. Then the observed likelihood, data given parameters, becomes:

$$
\begin{aligned}
p(y_{0:K} \mid \Theta) &= \sum_{x_K \in S} p(y_{0:K}, x_K \mid \Theta) \\
&= \sum_{x_K \in S} p(y_K \mid x_K, \Theta)p(x_K \mid y_{0:K-1}, \Theta)\times \\
&\quad p(y_{0:K-1} \mid \Theta) \\
&= \sum_{x_K \in S} p(y_K \mid x_K, \Theta)\times \\
&\quad \sum_{x_{K-1} \in S} p(x_K \mid x_{K-1}, \Theta)\times \\
&\quad p(x_{K-1}, y_{0:K-1} \mid \Theta)
\end{aligned}
\tag{29}
$$

Now recalling Forward algorithm from chapter 3, one may note that the likelihood can be written using recursive manner using so-called forward probabilities, namely $\nu_k(x_k) = p(x_k \mid y_{0:k})$. Then denoting all parameters of the model as single parameter Θ:

$$
\begin{aligned}
\nu_k(x_k) = p(x_k, y_{0:k} \mid \Theta) &= \sum_{x_{k-1}} p(x_{k-1}, x_x, y_{0:K} \mid \Theta) \\
&= p(y_k \mid x_k, \Theta)\sum_{x_{k-1}} p(x_k \mid x_{k-1}, \Theta)p(x_{k-1}, y_{0:K-1} \mid \Theta) \\
&= p(y_k \mid x_k, \Theta)\sum_{x_{k-1}} p(x_k \mid x_{k-1}, \Theta)\nu_{k-1}(x_{k-1})
\end{aligned}
\tag{30}
$$

simplifies the observed data likelihood into

$$
\begin{aligned}
\sum_{x_K \in S} \nu_K(x_K) = \sum_{x_K \in S} &p(y_K \mid x_K, \Theta)\times \\
&\sum_{x_{K-1} \in S} p(x_K \mid x_{K-1}, \Theta)\nu_{k-1}(x_{k-1})
\end{aligned}
\tag{31}
$$

The recursive nature of defining likelihood is typical for Hidden Markov model especially in computational point of view, since it drastically reduces the computational

cost of the fitting by this dynamic programming approach. Time-complexity of forward algorithm is $O(KS^2)$, where it is $O(KS^K)$ for naive approach, where K is the length of the sequence and S the cardinality of the state space.

Finally, as learned from the data description, all of the observations don't come from screening process. Since data has follow-up beyond last screening observation, the last observed data point y_K is either censoring (death or end of follow-up) or clinical appearance of cancer.

Let $S_{x_T}(t_+)$ be the certain survival function, where the outcome is clinical appearance of cancer and T is the time of final cancer screening (whether or not person attended), and $t_+ = t - T$ where t is the time of last observation of the person - either clinical manifestation of the disease or end of follow up. In this model $S_{x_T}(t - T) = p_{x_T,4}(t_+)$, so the likelihood (still for given invidual) will become:

$$\sum_{x_K \in S} S_{x_T}(t_+)^{1-\mathbb{I}\{C\}}(1 - S_{x_T}(t_+))^{\mathbb{I}\{C\}}\nu_K(x_K) \tag{32}$$

where $\mathbb{I}\{C\}$ is the indicator function of whether the last observation at final observation time t was appearance of cancer or censoring.

Taking all of the N inviduals into account, the observed data likelihood becomes:

$$\prod_{i=1}^{N}\left(\sum_{x_{K_i} \in S} Sx_{K_i}(t_+)^{1-\mathbb{I}\{C_i\}}(1 - S_{x_{K_i}}(t_+))^{\mathbb{I}\{C_i\}}\nu_{i,K_i}(x_{K_i})\right), \tag{33}$$

where $\nu_{i,K_i}(x_{K_i}) = p(x_{K_i}, y_{i,0:K_i})$ and $\nu_{i,0}(x_0) = p(x_0)p(y_{i,0} \mid x_0)$

4.3 Prior distributions for parameters

As in typical Bayesian setting one needs to figure out also the prior distributions for the parameters. These priors are constructed by my own expertise, that is not much, and a more rigorous investigation is required to write a scientific publication.

Things we know for certain are that transition parameters μ must be somewhere close to zero, but not infinitesimally close (μ is in the presence of zero but as we get closer to zero the pdf of μ should also decrease). For this in both models all the transition rates are given Gamma(2,1) prior.

Additional modelling thing I do is that I'll assume adenomas to have one progression rate $\tilde{\mu}$ and from this μ_1 and μ_2 are constructed via relation $\tilde{\mu} = (1 - \alpha)\mu_1 + \alpha\mu 2$, where α is the proportion of progressive adenomas out of non-progressive. α is assumed to have 1/2-centered Beta(10,10). This is again same in both models, and is assumed to be same independent of age group. The prior distribution has a standard deviation of $\approx 0,11$. Thus we assume a-priori that most likely 30-70% of the incident detected adenomas are non-progressive.

Emissions distribution parameters can be either estimated separately, as in paper by Wu (2018) or modelled simultaneously from the observations as Duffy et al. (2013). I will use an approach adopted by Duffy, but since there exists literature about the sensitivities of the blood-based stool tests for colorectal cancers and adenomas, it's reasonable to construct priors from those to help the convergence of the model. In general somewhat strict priors in HMM are essential, since the model description is vague and obtained estimates can be impossible with respect to previous research if there are no restrictive priors put in place.

According to Patel and Ahnen the sensitivity of gFOBT test depends on both number of taken samples and wether the samples are rehydrated or not. The Finnish RHS used multiple samples but no rehydration, that according to Patel and Ahnen yield sensitivities around 50%. This is on par with Finnish results published in for example Hakamaa and Malila (2019).

The priors that differ in the two models are specified in the table 4.3.

Parameter	Prior
S_d	Beta(500, 500)
S_e	Beta(60, 600)
S_{od}	Beta(60, 600)
θ	Dirichlet(100, 10, 1, 10)
μ	Gamma(2,1)

Sensitivity of gFOBT for adenomas is cited to be around 8% by Brenner et al. (2015), but ranging from 9 to 36% Morikawa et al. (2005). The data from Brenner's study comes from Bavaria with very similar setting as the Finnish RHS, and thus the prior adapted for both sensitity for progressive adenoma and sensitivity for non-progressive adenoma is assumed to be in line with these. The Beta(60, 600) prior has mean 0.91 and standard deviation of 0.011.

Here I assume same prior for all age-stratified θ, since idea is to give just a ball-park estimate of the phenomena. The Dirichlet priors of form Dirichlet$(100a, 10a, a, 10a)$ imply the expected value is such that there are roughly 83% healthy people and 17% people with adenomas (progressive and non-progressive combined), and very few with prevalent cancers. This is justifiable by looking at table 3, where less than a percent of the persons are diagnosed with CRC in the initial screen. The bigger the sum of parameters, the lower variance -principle applies here as well, so model has very much stricter prior distribution for initial distribution as well. Dirichlet prior parameters can be interpreted as pseudo-counts, so now the prior distribution assumes probabilites that correspond for sampling 121 "persons" from a population and detecting 100 healthy persons, 10 with non-progressive adenomas, 1 with a pre-clinical cancer and 10 with progressive adenomas.

Additionally I have added some constraints to parameters: S_e and S_{od} are assumed

to be at most half of S_d for given strata. There exists no studies where the sensitivity of gFOBT test for adenoma would be in the same ballpark with sensitivity for cancer. Also I assume that S_e and S_{od} are at least 0.05, so there is at least some sentivitity for all kinds of adenomas. S_d is constrained to be between 0.3 and 0.7. Finnish RHS used multiple non-rehydrated version of the test, that is cited by Patel and Ahnen to have sensitivity of 0.5, while single non-rehydrated test is cited to have sensitivity of 0.3 and multiple rehydrated 0.7.

Transition rates are obviously constrained to be more than zero, except for μ_1 and μ_2 have additional constraint $\mu_{1,2} \geq 0.001$ to avoid the sampler crashing to zero. Thus $\mu_1 = 0.001 + (1 - \alpha)\mu_{1,raw}$ for all stratas where α is as descibed earlier.

5 Results

In this section I'll go through model convergence, the estimates for model parameters and finally the quantification of overdiagnosis estimands described in chapter 2.

5.1 Natural history model

Stan and its R-interface package rstan provide a wide range of tools for diagnosing and evaluating a Bayesian model fit such as $\hat{R}$ values and warnings on divergent transitions.

For both models there exists no divergences, all $\hat{R}$ values of the parameters are less than 1.05 and HMC energy indicates no pathological behavior. These are basic requirements for a possibility that a MCMC simulation has succeeded. Also traceplot checks and autocorrelation plot checks did not indicate any pathologies.

After that it is important to take a look at the parameter posterior distributions. All of the figures in this section are drawn using R-package bayesplot. It is available on CRAN and described in paper Gabry, Simpson et al. (2019). Full tables of estimated parameters with some statistics and confidence intervals can be found from Appendix 4, but here it's sensible to take a graphical look at the posteriors as well since there may be pathologies that plain numerics don't reveal (such as multimodality of the posterior).

Let's first take a look at the sensitivies of the model. Here the effect of different priors for sensitivities is illustrated in figure 5.1. More informative priors lead to more concentrated and well-behaved posteriors, when model description is otherwise similar. In this case, though, one is in the risk of making inference out of priors only.

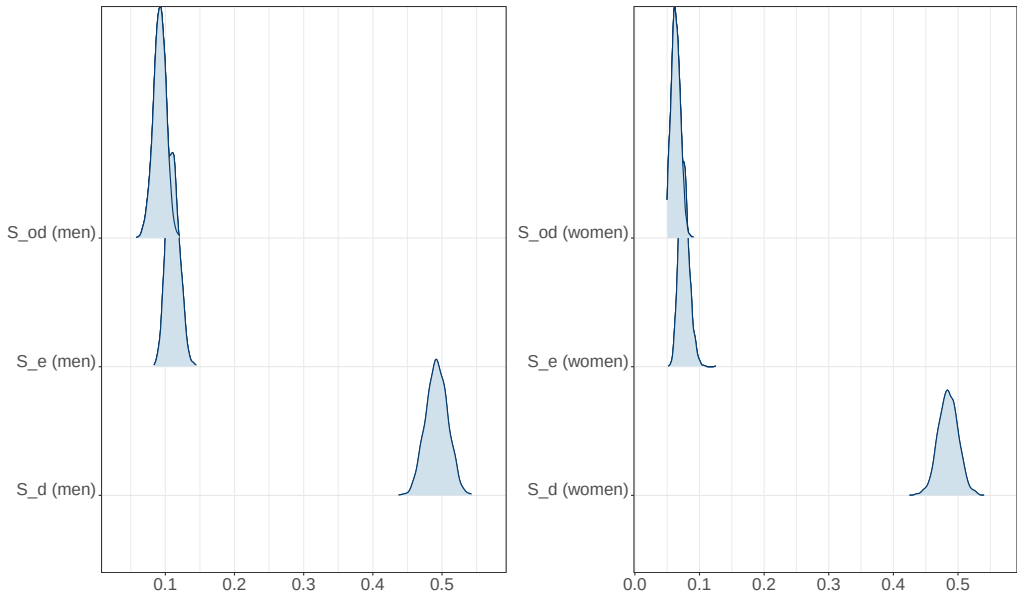

Figure 9: Posterior sensitivities. Sensitivity for non-progressive adenomas seems to be a bit lower than senstivity for progressive adenomas, and adenomas sensitivities for women seem to be less than adenoma sensitivities for men.

In general all sensitivity posteriors have quite nice form and there is no reason to suspect the convergence.

Then let's proceed to the initial state distributions.

No pathologies are apparent when looking at initial state posterior distributions of the two models. They are illustrated in figures 10, 11.

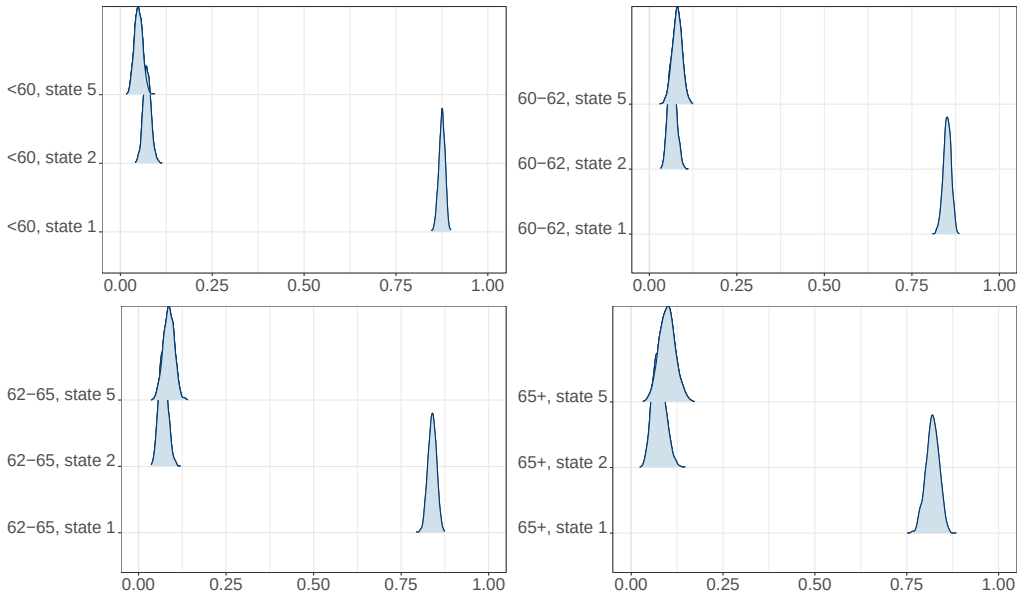

Figure 10: Initial state posterior distributions for men.

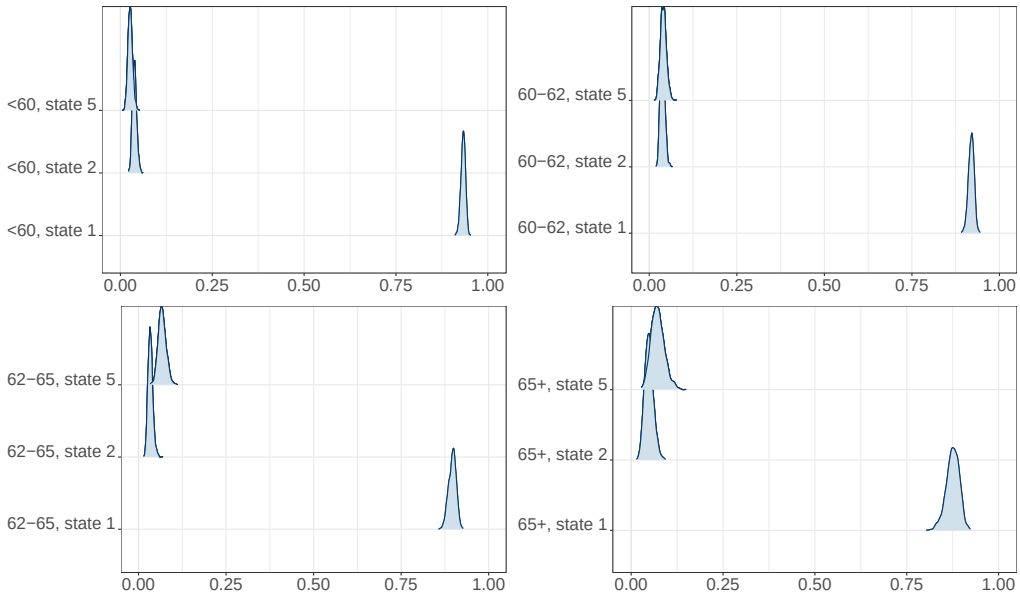

Figure 11: Initial state posterior distributions for women.

These posteriors also seem to be well-behaved. Additionally an interesting feature is that prevalence for non-progressive adenoma seems to grow with age. This can be explained, since progressive-adenomas indeed progress and are thus removed from the bowel naturally by turning into cancers, but non-progressive adenomas should not and thus they should be more prevalent in later ages.

This behavior is inevitable in progressive model that was used with transition rates, but the initial distributions are kind of independent of the progressive behavior, which indicates that the phenomena could be real instead of being just a model technicality.

More interesting phenomena becomes apparent when looking at posterior distributions for transition rates.

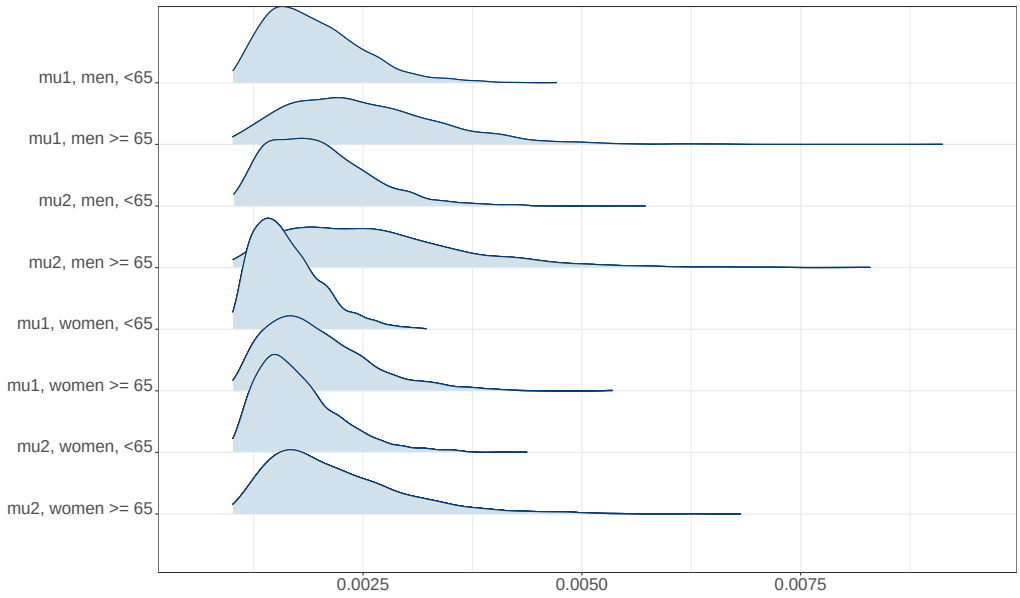

Figure 12: Transition rate parameter posteriors from healthy to different adenomas.

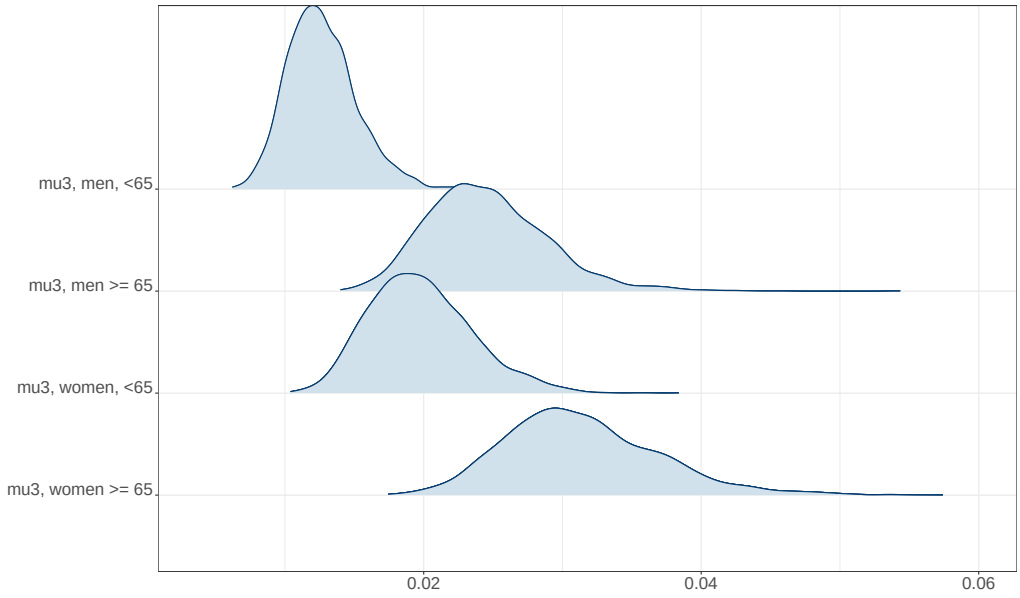

Figure 13: Transition rate parameter posteriors from adenoma to pre-clinical cancer. The rates seem to be higher for older people, as could be expected.

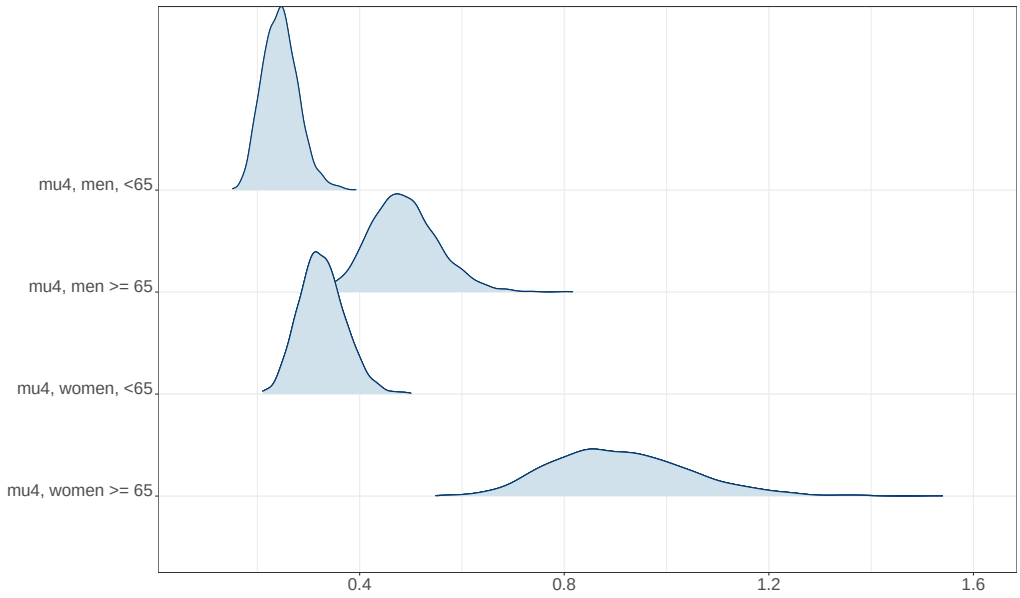

Figure 14: Transition rate parameter posteriors from pre-clinical to clinical cancer. The rates seem to be higher for older people, as could be expected.

Looking at figures 12, 13 and 14 it seems that men have on average a bit higher adenoma rate, but women have both higher progression rate from adenoma to cancer and especially from pre-clinical cancer to clinical cancer. This is not directly related to overdiagnosis, but could indicate why women are found to benefit less from CRC screening, as noted by for example Koskenvuo et al. (2018).

5.2 Overdiagnosis

Let's first take a look at overdiagnosis O_3 described in chapter 2. Since transition rates were parametrised as constants and additionally transition rates μ_1 and μ_2 were restricted to follow relation $\tilde{\mu} = (1-\alpha)\mu_1 + \alpha\mu_2$ where $\alpha \in (0,1)$, O_3 remains constant through age, the overdiagnosis of type O_3 remains constant through age for men and women.

From this we get following estimates for male and female visualised in figure 15 and tabularised in table 5.2.

Weakness of this estimator is that it doesn't take into account the prevalence at start of the study, and is only focused on estimated incidence rates. Following results will come to opposite conclusions, since the estimated prevalences of adenomas were higher for male than female (see figures 10, 11). This is why point estimates for women are lower than for men, even though estimates O_4 and O_5 will have a different opinion.

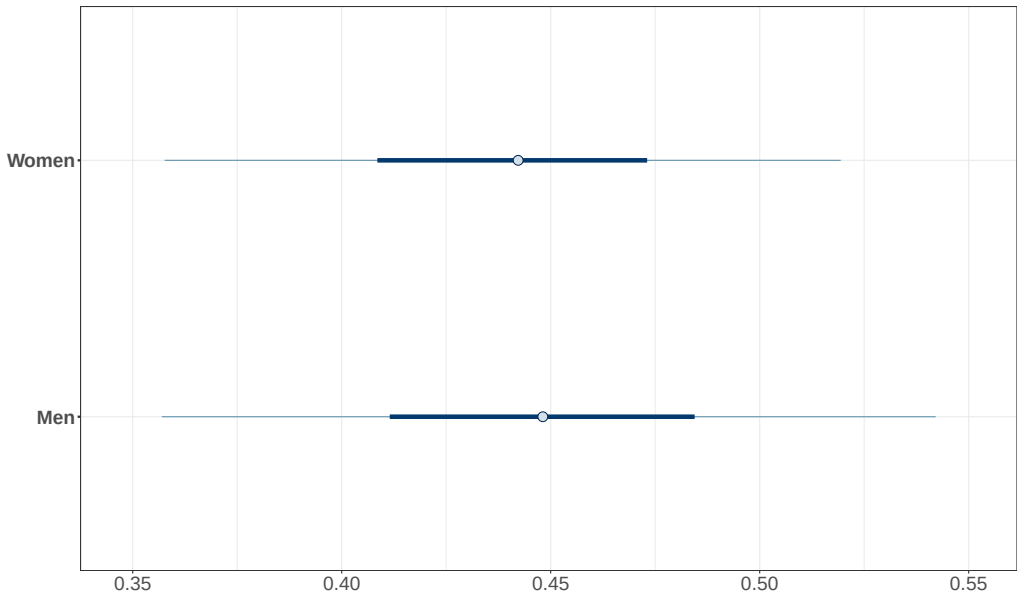

Figure 15: Overdiagnosis using cumulative incidence ratio of progressive and non-progressive adenomas, with 50 and 90% credible intervals (thin and deep blue) and median point estimate..

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%
O3, men	0.45	0.00	0.06	0.34	0.41	0.45	0.48	0.56
O3, women	0.44	0.00	0.06	0.33	0.40	0.44	0.47	0.55

Table 4: Method O_3 overdiagnosis for men and women.

Then I calculated the overdiagnosis using method $O_4(z) = \mathbb{P}(X_t = 5 \mid Y_t = 2)$ by conditioning on history of a hypothetical invidual, where covariates z include sex, age at the initial screen and number of consecutive screening round.

I used inviduals for both sexes with four age groups: persons starting the programme at age 60 (thus having $\theta_{<60}$ as the initial state distribution, person starting at 61 (θ_{60-62} as the initial distribution), and persons in ages 63 and 66. Then the hypothetical persons went through 5, 5, 4 and 3 screening rounds (five rounds was the maximum in data and no screenings were conducted for persons over 70 years old). These are described in tables 5.2, 5.2 and figures 16, 17.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%
Men, 60-, round 1	0.36	0.00	0.09	0.20	0.30	0.36	0.42	0.53
Men, 60-, round 2	0.38	0.00	0.08	0.22	0.32	0.38	0.43	0.54
Men, 60-, round 3	0.39	0.00	0.08	0.24	0.34	0.39	0.45	0.54
Men, 60-, round 4	0.41	0.00	0.08	0.26	0.35	0.41	0.46	0.55
Men, 60-, round 5	0.43	0.00	0.07	0.29	0.38	0.43	0.48	0.57
Men, 61, round 1	0.50	0.00	0.09	0.32	0.44	0.51	0.56	0.66
Men, 61, round 2	0.51	0.00	0.08	0.34	0.46	0.51	0.57	0.66
Men, 61, round 3	0.52	0.00	0.08	0.36	0.47	0.52	0.57	0.66
Men, 61, round 4	0.53	0.00	0.07	0.38	0.48	0.54	0.58	0.66
Men, 61, round 5	0.54	0.00	0.07	0.40	0.50	0.55	0.59	0.67
Men, 63, round 1	0.50	0.00	0.09	0.32	0.44	0.51	0.57	0.67
Men, 63, round 2	0.51	0.00	0.08	0.34	0.46	0.52	0.57	0.67
Men, 63, round 3	0.53	0.00	0.08	0.37	0.47	0.53	0.58	0.67
Men, 63, round 4	0.54	0.00	0.08	0.39	0.49	0.54	0.59	0.67
Men, 66, round 1	0.52	0.00	0.11	0.30	0.44	0.52	0.60	0.71
Men, 66, round 2	0.53	0.00	0.10	0.33	0.46	0.53	0.61	0.71
Men, 66, round 3	0.54	0.00	0.10	0.35	0.48	0.55	0.61	0.72

Table 5: Overdiagnosis using method O_4 for hypothetical men starting screening at given age (corresponds to given initial distribution) and given screening round (corresponds to round - 1 predecessing episode-negative findings). 60- means starting at "just before" age 60, that is the hypothetical person still gets the same initial distribution as everyone under 60.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%
Women, 60-, r 1	0.37	0.00	0.08	0.22	0.32	0.37	0.42	0.52
Women, 60-, r 2	0.39	0.00	0.07	0.25	0.34	0.39	0.44	0.53
Women, 60-, r 3	0.41	0.00	0.07	0.28	0.36	0.41	0.46	0.54
Women, 60-, r 4	0.43	0.00	0.06	0.31	0.39	0.43	0.48	0.56
Women, 60-, r 5	0.46	0.00	0.06	0.34	0.41	0.46	0.50	0.57
Women, 61, r 1	0.47	0.00	0.08	0.31	0.42	0.47	0.53	0.63
Women, 61, r 2	0.48	0.00	0.08	0.33	0.43	0.48	0.53	0.63
Women, 61, r 3	0.49	0.00	0.07	0.35	0.45	0.50	0.54	0.63
Women, 61, r 4	0.51	0.00	0.07	0.38	0.47	0.51	0.56	0.63
Women, 61, r 5	0.53	0.00	0.06	0.40	0.48	0.53	0.57	0.64
Women, 63, r 1	0.62	0.00	0.07	0.46	0.58	0.63	0.68	0.75
Women, 63, r 2	0.63	0.00	0.07	0.48	0.58	0.63	0.67	0.75
Women, 63, r 3	0.63	0.00	0.06	0.49	0.59	0.64	0.68	0.74
Women, 63, r 4	0.64	0.00	0.06	0.51	0.60	0.64	0.68	0.74
Women, 66, r 1	0.55	0.00	0.10	0.34	0.49	0.55	0.63	0.74
Women, 66, r 2	0.57	0.00	0.10	0.37	0.50	0.57	0.63	0.74
Women, 66, r 3	0.58	0.00	0.09	0.39	0.52	0.58	0.64	0.74

Table 6: Overdiagnosis using method O_4 for hypothetical women starting screening at given age (corresponds to given initial distribution) and given screening round (corresponds to round - 1 predecessing episode-negative findings). 60- means starting at "just before" age 60, that is the hypothetical person still gets the same initial distribution as everyone under 60.

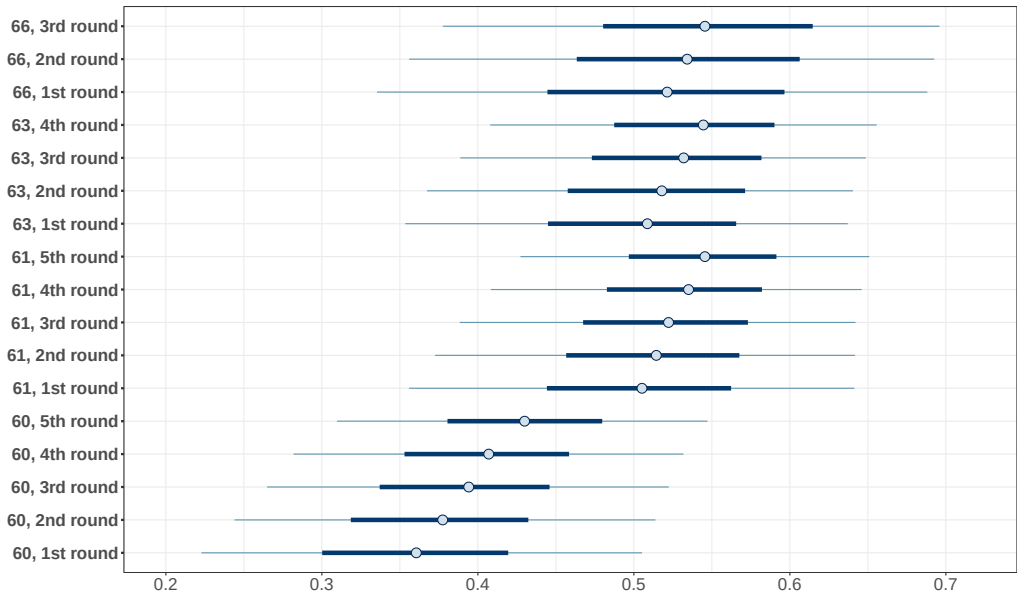

Figure 16: Overdiagnosis O_4 for hypothetical male patients starting at given ages at given screening rounds with estimated median and 50 and 90-percent credible intervals (thin line, thick line). 60 means "60-", that is just below sixty so the initial distribution is still $\theta_{<60}$.

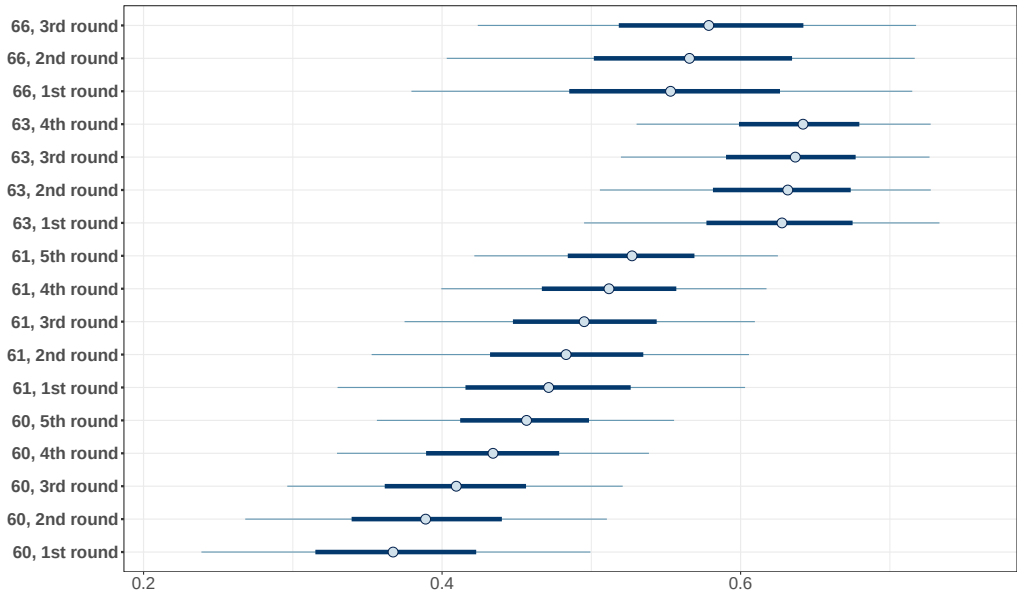

Figure 17: Overdiagnosis O_4 for hypothetical woman patients starting at given ages at given screening rounds with estimated median and 50 and 90-percent credible intervals (thin line, thick line). 60 means "60-", that is just below sixty so the initial distribution is still $\theta_{<60}$.

Here the overdiagnosis percentage shows interesting behavior, since it is usually thought that the overdiagnosis is highest during the first screening round. Yet here it seems that it just increases with age, with some non-linearity caused by stratified initial state distributions. This is addressed in discussion.

Additionally I calculated overdiagnosis percentage averages (O_5) for whole programme taking average over all the state occupancy probabilities of non-progressive adenoma for the persons who had an adenoma diagnosed.

The estimated posterior median for overdiagnosis in the whole programme was 48% (38%-56% 95-percent credible interval).

Additionally I calculated O_5 for both sexes and in different strata by age at start of the screening programme. This is just O_5 again but instead averaging over whole programme, I took the average over certain given strata.

The results for these are presented in figures 18 and 19 and in table 7.

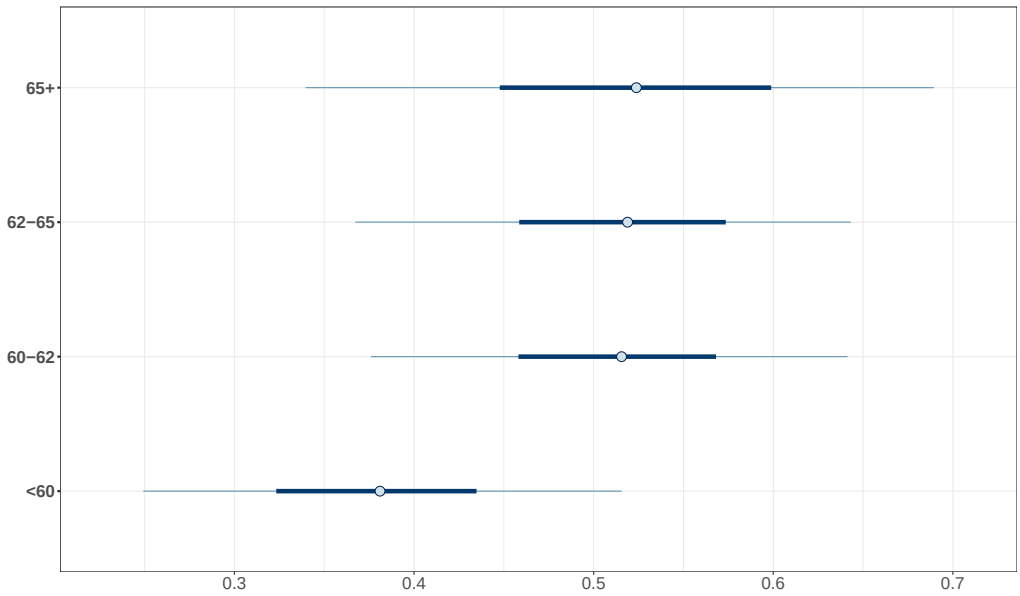

Figure 18: Average state 5 occupancy probability given eventually observed adenoma for men, stratified by age at start of the screening programme.

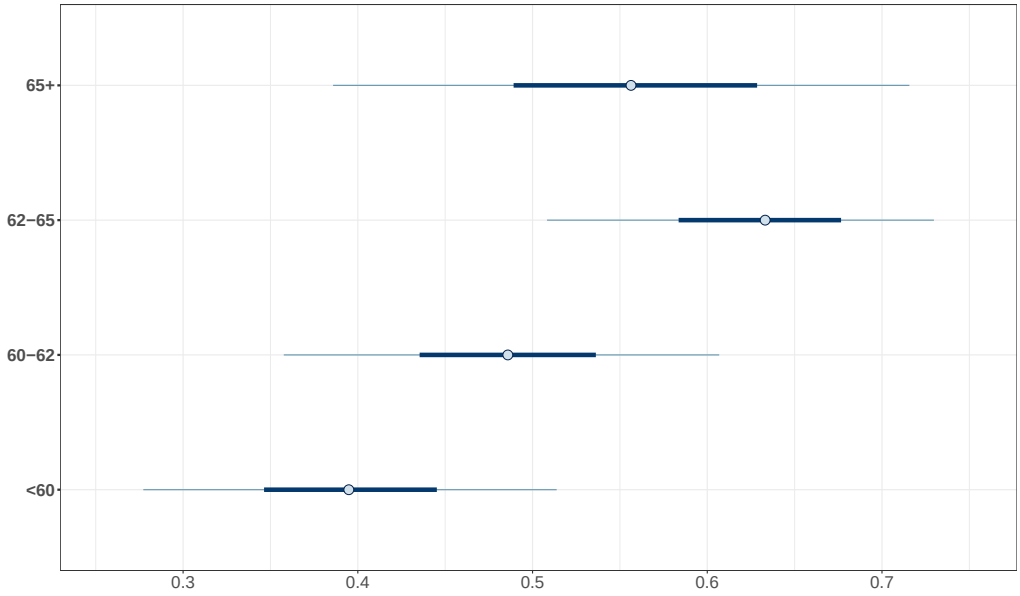

Figure 19: Average state 5 occupancy probability given eventually observed adenoma for women, stratified by age at start of the screening programme.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%
Men, <60	0.38	0.00	0.08	0.22	0.32	0.38	0.43	0.54
Men, 60-62	0.51	0.00	0.08	0.34	0.46	0.52	0.57	0.66
Men, 62-65	0.51	0.00	0.08	0.34	0.46	0.52	0.57	0.67
Men, 65+	0.52	0.00	0.11	0.31	0.45	0.52	0.60	0.71
Women, <60	0.40	0.00	0.07	0.26	0.35	0.39	0.45	0.53
Women, 60-62	0.48	0.00	0.07	0.34	0.44	0.49	0.54	0.63
Women, 62-65	0.63	0.00	0.07	0.48	0.58	0.63	0.68	0.75
Women, 65+	0.56	0.00	0.10	0.34	0.49	0.56	0.63	0.74
Programme total	0.48	0.00	0.05	0.38	0.45	0.48	0.51	0.56

Table 7: Overdiagnosis using method O_5 in different sex-starting age strata and in the whole screening programme.

These results seem to be quite well in line with O_4 but different from O_3, that is, as said, due to O_3 modelling only incidence of adenoma and ignoring the prevalence.

6 Discussion

This thesis has provided two different things: 1) investigation of a specific kind of overdiagnosis in CRC screening, and 2) an investigation of natural history models for colorectal cancer progression.

In literary review I found no similar study made for CRC screening, so it's hard to compare this to existing research. The main result is that around half of the screen-detected adenomas are overdiagnosed ones. This wasn't unexpected, since similar results have been found when analysing overdiagnosis in screening for pre-cancerous lesions.

One comparison point, to get an idea of in which ballpark these estimates should be in, could be found from studies conducted on overdiagnosis in cervical cancer screening, where the overdiagnosis comes from similar source (from detecting non-progressive pre-cancerous forms) as in this study. Hamashima et al. (2018) cite lifetime overdiagnosis frequencies (the same thing estimated here) of over 50% for cervical cancer screening. This is on par with this study - but two different cancers are of course not comparable directly.

Wu et al. (2019) concluded, using a very similar HMM, that there is practially no overdiagnosis in breast cancer screening programme (less than 3%), while RCT-based for excess detection of cancers estimates start from 10%. For breast cancer there is no similar known pre-cancerous state as adenoma, which could explain why HMM could work differently for CRC screening.

Also they stated that a model misspecification cannot be ruled out. In Bayesian framework, though, there exists a lot of possible ways to check that. A medical expert can take a look at multiple things in this kind of model - for example there exists posterior distributions for all of the 60 parameters that can be evaluated critically with respect to biological reality. If the model is not biologically realistic, it must be restructured.

The thesis illustrates an interesting phenomena that contradicts a common heuristic: usually it is thought that the prevalence round (first screening round) comes with the most overdiagnosis, while the results show that overdiagnosis due to detection of non-progressive adenomas might increase when more screening rounds is applied. This is an inevitable consequence of the model: only forward transitions are allowed, the proportion of prevalence for progressive adenomas decreases as age increases, since these adenomas move forward to next states in the model, but the non-progressive adenoma state is an absorbing state. Of course both kinds of adenomas are removed when found, but since there is no reason to suspect gFOBT test having a higher sensitivity for non-progressive adenomas, the model inevitably converges to this situation.

Yet the increasing prevalance proportion of non-progressive adenomas is also apparent in the initial state distributions, which are estimated in different framework. This could indicate, that the phenomena is not only a model property.

Overdiagnosis in CRC screening is a very subtle problem - there are many aspects to it and all of them cannot be answered using a single model. These results estimate overdiagnosis of very specific kind. That is both a strength and a limitation of this kind of model.

Natural history models for cancer enable researcher to do many kinds of inference even without control population,. That is essential in CRC screening in Finland right now, since the ongoing programme has no control population at all. Developing more methods of this kind or different to overcome this is essential in order to evaluate properties of that kind of programme. This model shows some really interesting features such as that men have more adenomas but women in general have preclinical diseases progressing more rapidly into cancers. Also according to model women have lower sensitivities for adenomas. These are things that, if deemed realistic, need to be accounted when thinking about fine-tuning a CRC screening programme.

I used informative priors, but they were not constructed in co-operation with a medical expert. That would be definitely needed in order to write a scientific publication about a subject with this much importance. When good informative priors and their critical evaluation is done properly, the framework of this thesis provides a great way for doing open science about this subject.

The model could be extended in multiple ways, such as moving from piece-wise constant rates to continuously time-dependent rates and adding covariates. Also death could be modelled simultaneously in order to account for more types of overdiagnosis described in chapter 2.

One assumption that needs to be critically evaluated when using HMM for CRC screening is the conditional indepencence of a screening tests given underlying state. It's reasonable to assume that some people in their nature just bleed less, and this would make screening sensitivities for one invidual come from very different distributions. Using some kind of hierachical model for sensitivities could be more realistic for modelling repeating episode-negative observations. Also Markov switching models could be used to model the dependence between observations of a person.

Another improvement could be modelling the screening programme using multiple processes, for example using a different process for screen-attendance. This, though, goes beyond the question of overdiagnosis

Additionally a lot of data available in the Mass Screening Registry's database was not used at all. For example, the database containing screening results has a lot more information such as histological behavior of detected adenomas and different covariates about living habits for the screened population. Taking these into account could yield more interesting results.

References

1 Michael Betancourt. Hamiltonian monte carlo - a conceptual introduction. *ArXiv*, 2018.

2 Olivier Cappé, Eric Moulines, and Tobias Rydén. *Inference in Hidden Markov Models*. Springer, 2005.

3 Nicholas E. Day. *Advances in Cancer Screening, Chapter 2: The theoretical basis for cancer screening*. Kluwer Academic Publishers, 1996.

4 Bob Carpenter et al. Stan: A probabilistic programming language. *Journal of Statistical Software*, 2017.

5 Hermann Brenner et al. Which adenomas are detected by fecal occult blood testing? a state-wide analysis from Bavaria, Germany. *International Journal of Cancer*, 2015.

6 Laura Koskenvuo et al. Sex differences in faecal occult blood test screening for colorectal cancer. *BJS*, 2018.

7 Mette Kalager et al. Overdiagnosis in colorectal cancer screening: Time to acknowledge a blind spot. *Gastroenterology*, 2018.

8 Pamela M. Marcus et al. Conceptualizing overdiagnosis in cancer screening. *JNCI J Natl Cancer Inst*, 2015.

9 Sherry Yueh-Hsia Chiu et. al. Predicting the effectiveness of the Finnish population-based colorectal cancer screening programme. *Journal of Medical Screening*, 2015.

10 Tamiya Morikawa et al. A comparison of the immunochemical fecal occult blood test and total colonoscopy in the asymptomatic population. *Gastroenterology*, 2005.

11 Tony H. H. Chen et. al. Estimation of sojourn time in chronic disease screening without data on interval cases. *Biometrics*, 2000.

12 Jonah Gabry and Daniel Simpson et al. Visualization in Bayesian workflow. *J. R. Stat. Soc. A*, 2019.

13 Matti Hakama and Nea Malila. Estimation of overdiagnosis by detection to incidence ratio in cancer screening. *Journal of Medical Screening*, 2019.

14 Matti Hakama, Arun Pokhrel, Nea Malila, and Timo Hakulinen. Sensitivity, effect and overdiagnosis in screening for cancers with detectable pre-invasive phase. *International Journal of Cancer*, 2015.

15 Matthew D. Hoffman and Andrew Gelman. The No-U-Turn Sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 2014.

16 Nehmat Houssami. Overdiagnosis of breast cancer in population screening: does it make breast screening worthless? *Cancer Biology and Medicine*, 2017.

17 IARC. *IARC Handbooks of Cancer Prevention Volume 17.* IARC, 2019.

18 Oliver C. Ibe. *Markov Processes for Stochastic Modeling.* Elsevier, London, 2nd edition, 2013.

19 Stacy Loeb. Overdiagnosis and overtreatment of prostate cancer. *European urology*, 2017.

20 Dianhong Luo, Alexander C. Cambon, and Dongfeng Wu. Evaluating the long-term effect of fobt in colorectal cancer screening. *Cancer Epidemiology*, 2011.

21 Nea Malila, Ahti Anttila, and Matti Hakama. Colorectal cancer screening in Finland: details of the national screening programme implemented in autumn 2004. *Journal of Medical Screening*, 2004.

22 Anthony B. Miller. Conundrums in screening for cancer. *International Journal of Cancer*, 2010.

23 Wendy Wu and et. al. Overdiagnosis in the population-based organized breast cancer screening program estimated by a non-homogeneous multi-state model: a cohort study using individual data with long-term follow-up. *Breast Cancer Research*, 2018.

Appendix 1. Naive HMC algorithm

Algorithm 3: Hamiltonian Monte Carlo

Result: $\theta \in \mathbb{R}^{M \times N}$

Let $\theta_0 \in \mathbb{R}^N$ be initial value of parameter, where N is the dimension of parameter space. $\mathcal{L}$ the log-likelihood of the model, $M \in \mathbb{N}$ the number of iterations and $L \in \mathbb{N}$ the number of leapfrog steps. $I \in \mathbb{R}^{N \times N}$ is the identity matrix. r and $\nabla_\theta \mathcal{L}(\theta)$ have the same dimension as the parameter space so the calculations make sense. ϵ is a tuning-parameter admitting values in $(0, 1)$.;

for $m = 1, \ldots, M$ **do**

  Sample $r_0 \sim \mathcal{N}(0, I)$;

  Set $\theta^m = \theta_{m-1}, \tilde{\theta} = \theta_{m-1}, \tilde{r} = r_0$;

  for $i = 1, \ldots, L$ **do**

    Set $\tilde{\theta}, \tilde{r} = \text{Leapfrog}(\tilde{\theta}, \tilde{r}, \epsilon)$;

  end

  With probability $\alpha = \min\left\{1, \frac{\exp(\mathcal{L}(\tilde{\theta}) - \frac{1}{2}\tilde{r} \cdot \tilde{r})}{\exp(\mathcal{L}(\theta_{m-1}) - \frac{1}{2} r_0 \cdot r_0)}\right\}$ set $\theta_m = \tilde{\theta}, r_m = -\tilde{r}$;

end

function Leapfrog$(\tilde{\theta}, \tilde{r}, \epsilon)$

Set $\tilde{r} = r + (\epsilon/2)\nabla_\theta \mathcal{L}(\theta)$;

Set $\tilde{\theta} = \theta + \epsilon \tilde{r}$;

Set $\tilde{r} = \tilde{r} + (\epsilon/2)\nabla_{\tilde{\theta}} \mathcal{L}(\tilde{\theta})$;

return $\tilde{\theta}, \tilde{r}$

Appendix 2. Stan code for parameter estimation

```
functions {


  matrix transition_matrix(
    real mu1, real mu2,
    real mu3, real mu4,
    real ts
  ) {

    matrix[5,5] P = rep_matrix(0, 5, 5);

    P[1,1] = exp((-mu1-mu2) * ts);
    P[1,2] = mu1 * (exp(-mu3*ts) - exp(-(mu1+mu2) * ts)) / (mu1+mu2-mu3);
    P[1,3] = mu1 * mu3 * (
      exp(-mu4*ts) * (mu1+mu2-mu3) +
      exp(-mu3*ts) * (mu4-mu1-mu2) +
      exp(-(mu1+mu2) * ts) * (mu3 - mu4)
      )/(
        (mu1+mu2-mu3) * (mu1+mu2-mu4) * (mu3-mu4)
        );
    P[1,4] = mu1/(mu1+mu2) +
        exp(-mu3*ts) * mu1 * mu4 / ((mu1+mu2-mu3)*(mu3-mu4)) +
        exp(-mu4*ts) * mu1 * mu3 / ((mu3-mu4) * (mu4-mu1-mu2)) +
        exp(-(mu1+mu2) * ts) * (
          - mu1 / (mu1+mu2) + mu1/(mu1+mu2-mu3)
          - mu1 *mu3 * (1.0 /
              ((mu1+mu2-mu3)*(mu3-mu4))+1.0/((mu3-mu4)*(mu4-mu1-mu2)))
          );
    P[1,5] = mu2*(1.0-exp(-(mu1+mu2)*ts))/(mu1+mu2);
    P[2,2] = exp(-mu3 *ts);
    P[2,3] = mu3 *(exp(-mu4*ts)-exp(-mu3*ts))/(mu3-mu4);
    P[2,4] = 1.0-(mu3*exp(-mu4 *ts))/(mu3-mu4)+mu4*exp(-mu3*ts)/(mu3-mu4);
    P[3,3] = exp(-mu4 *ts);
    P[3,4] = 1.0-exp(-mu4*ts);
    P[4,4] = 1.0;
    P[5,5] = 1.0;

    return(P);
  }

  vector efficient_forward_loop(
    matrix phi,
    matrix P,
    int obs,
```

```
  vector gamma_prev
) {


  int current_state[4] = {1,2,3,5};
  int prev_state[4] = {1,2,3,5};
  vector[5] gamma;
  vector[5] acc;

  for(i in current_state) {
    if(obs != 5 && phi[i, obs] == 0) {
      gamma[i] = negative_infinity();
    } else {
      for(j in prev_state) {
        acc[j] = gamma_prev[j] + log(P[j,i]);
        if(obs != 5) {
          acc[j] += log(phi[i, obs]);
        }
      }
      gamma[i] = log_sum_exp(acc[prev_state]);

    }

  }

  return gamma;

}

vector hmm_seq_lr(
  vector pars,
  vector covs,
  real[] ts,
  int[] obs
) {
  vector[15] all_pars;
  vector[4] mu;
  matrix[5,4] phi;
  int M;
  row_vector[5] gamma;
  matrix[5,5] P;
  matrix[4,4] prev_matr;
  real res;
  real k;
  vector[5] acc;
  vector[5] gamma_safe;

  if(covs[1] == 0) {
```

```
  if(covs[2] < 61) {
    all_pars = pars[{1,2,3, 4, 5, 6, 7, 12, 13, 14, 15, 8, 9,10,11}];
  } else if(covs[2] < 63) {
    all_pars = pars[{1,2,3, 4, 5, 6, 7, 16, 17, 18, 19, 8, 9, 10,11}];

  } else if(covs[2] < 65) {
    all_pars = pars[{1,2,3, 4, 5, 6, 7, 20, 21, 22, 23, 8, 9, 10,11}];

  } else {
    all_pars = pars[{1,2,3,8,9,10,11,24,25,26,27,8,9,10,11}];
  }
} else if(covs[1] == 1) {
  if(covs[2] < 61) {
    all_pars = pars[{1 + 27,2 + 27,3 + 27, 4 + 27, 5 + 27, 6 + 27, 7 +
        27,
    12 + 27, 13 + 27, 14 + 27, 15 + 27, 8+27, 9+27, 10+27,11+27}];
  } else if(covs[2] < 63) {
    all_pars = pars[{1 + 27,2 + 27,3 + 27, 4 + 27, 5 + 27, 6 + 27, 7 +
        27,
    16 + 27, 17 + 27, 18 + 27, 19 + 27, 8+27, 9+27, 10+27,11+27}];

  } else if(covs[2] < 65) {
    all_pars = pars[{1 + 27,2 + 27,3 + 27, 4 + 27, 5 + 27, 6 + 27, 7 +
        27,
    20 + 27, 21 + 27, 22 + 27, 23 + 27, 8+27, 9+27, 10+27,11+27}];

  } else {
    all_pars = pars[{1 + 27,2 + 27,3 + 27, 8 + 27, 9 + 27, 10 + 27,
    11 + 27, 24 + 27, 25 + 27, 26 + 27, 27 + 27, 8+27, 9+27,
        10+27,11+27}];
  }
}


M = dims(ts)[1];
phi = rep_matrix(0, 5, 4);
phi[1,1] = 1;
phi[2,1] = 1-all_pars[1];
phi[2,2] = all_pars[1];
phi[3,1] = 1-all_pars[2];
phi[3,3] = all_pars[2];
phi[4,4] = 1;
phi[5,1] = 1-all_pars[3];
phi[5,2] = all_pars[3];
mu = all_pars[4:7];
gamma[{1,2,3,5}] = log(all_pars[8:11]') + log(phi[{1,2,3,5}, obs[1]]');

for (m in 2:M) {
```

```
      if(ts[m] + covs[2] > 65) {
        mu = all_pars[12:15];
      }
      if(obs[m] != -1) {
        P = transition_matrix(
          mu[1],
          mu[2],
          mu[3],
          mu[4],
          ts[m]-ts[m-1]
        );
        if(ts[m] < ts[m-1]) {
          reject("Mita helv.. negat. aika for obs: ", obs[m]);
        }
      }

      if(obs[m] == 6 || obs[m] == 4 || obs[m] == -1) {
        M = m;
        break;
      }

      if(obs[m] == 2 || obs[m] == 3 || obs[m] == 1) {
        gamma = efficient_forward_loop(phi, P, obs[m], gamma')';
      }

    }

    if(obs[M] == 6) {
      gamma[{1,2,3,5}] += log(1-P[{1,2,3,5}, 4]');
    }

    if(obs[M] == 4) {
      gamma[{1,2,3,5}] += log(P[{1,2,3,5}, 4]');
    }

    res = log_sum_exp(gamma[{1,2,3,5}]);

    return [res]';
  }
}

data {
  int<lower=1> N;
  int<lower=1> M;
  int<lower=-1,upper=6> w[N, M];
  real<lower = 0> times[N, M];
  vector[2] covs[N];
  vector[4] xi;
```

```
}


parameters {
  // Sex-specific parameters
  // Men
  real<lower = 0.3, upper = 0.7> S_d_men;
  real<lower = 0.05, upper = S_d_men/2.0> S_e_men;
  real<lower = 0.05, upper = S_e_men> S_od_men;
  real<lower = 0, upper = 1> mix_par_men;
  // Women
  real<lower = 0.3, upper = 0.7> S_d_women;
  real<lower = 0.05, upper = S_d_women/2.0> S_e_women;
  real<lower = 0.05, upper = S_e_women> S_od_women;
  real<lower = 0, upper = 1> mix_par_women;

  // Sex-age-group-stratified parameters
  real<lower = 0> mu1_raw_old_men;
  real<lower = 0> mu3_old_men;
  real<lower = 0> mu4_old_men;

  real<lower = 0> mu1_raw_young_men;
  real<lower = 0> mu3_young_men;
  real<lower = 0> mu4_young_men;

  real<lower = 0> mu1_raw_old_women;
  real<lower = 0> mu3_old_women;
  real<lower = 0> mu4_old_women;

  real<lower = 0> mu1_raw_young_women;
  real<lower = 0> mu3_young_women;
  real<lower = 0> mu4_young_women;

  simplex[4] theta_u60_men;
  simplex[4] theta_60_62_men;
  simplex[4] theta_62_65_men;
  simplex[4] theta_65_plus_men;

  simplex[4] theta_u60_women;
  simplex[4] theta_60_62_women;
  simplex[4] theta_62_65_women;
  simplex[4] theta_65_plus_women;
}

transformed parameters {
  real<lower = 0> mu1_old_men = 0.001 +(1-mix_par_men) * mu1_raw_old_men;
  real<lower = 0> mu2_old_men = 0.001 + mix_par_men * mu1_raw_old_men;
```

```
  real<lower = 0> mu1_young_men = 0.001 +(1-mix_par_men) *
      mu1_raw_young_men;
  real<lower = 0> mu2_young_men = 0.001 + mix_par_men * mu1_raw_young_men;

  real<lower = 0> mu1_old_women = 0.001 +(1 - mix_par_women) *
      mu1_raw_old_women;
  real<lower = 0> mu2_old_women = 0.001 + mix_par_women *
      mu1_raw_old_women;

  real<lower = 0> mu1_young_women = 0.001 +(1-mix_par_women) *
      mu1_raw_young_women;
  real<lower = 0> mu2_young_women = 0.001 + mix_par_women *
      mu1_raw_young_women;
}

model {

  theta_u60_men ~ dirichlet(xi);
  theta_60_62_men ~ dirichlet(xi);
  theta_62_65_men ~ dirichlet(xi);
  theta_65_plus_men ~ dirichlet(xi);

  theta_u60_women ~ dirichlet(xi);
  theta_60_62_women ~ dirichlet(xi);
  theta_62_65_women ~ dirichlet(xi);
  theta_65_plus_women ~ dirichlet(xi);


  mix_par_men ~ beta(10,10);
  mix_par_women ~ beta(10,10);

  mu1_raw_young_men ~ gamma(2,1);
  mu1_raw_young_women ~ gamma(2,1);
  mu1_raw_old_men ~ gamma(2,1);
  mu1_raw_old_women ~ gamma(2,1);

  mu3_young_men ~ gamma(2,1);
  mu3_young_women ~ gamma(2,1);
  mu3_old_men ~ gamma(2,1);
  mu3_old_women ~ gamma(2,1);

  mu4_young_men ~ gamma(2,1);
  mu4_young_women ~ gamma(2,1);
  mu4_old_men ~ gamma(2,1);
  mu4_old_women ~ gamma(2,1);

  S_od_men ~ beta(60, 600);
  S_d_men ~ beta(500, 500);
```

```
  S_e_men ~ beta(60, 600);

  S_od_women ~ beta(60, 600);
  S_d_women ~ beta(500, 500);
  S_e_women ~ beta(60, 600);


  target += sum(
    map_rect(
      hmm_seq_lr,
      to_vector([
        S_e_men, S_d_men, S_od_men,
        mu1_young_men, mu2_young_men, mu3_young_men, mu4_young_men,
        mu1_old_men, mu2_old_men, mu3_old_men, mu4_old_men,
        theta_u60_men[1], theta_u60_men[2], theta_u60_men[3],
            theta_u60_men[4],
        theta_60_62_men[1], theta_60_62_men[2], theta_60_62_men[3],
            theta_60_62_men[4],
        theta_62_65_men[1], theta_62_65_men[2], theta_62_65_men[3],
            theta_62_65_men[4],
        theta_65_plus_men[1], theta_65_plus_men[2], theta_65_plus_men[3],
            theta_65_plus_men[4],
        S_e_women, S_d_women, S_od_women,
        mu1_young_women, mu2_young_women, mu3_young_women, mu4_young_women,
        mu1_old_women, mu2_old_women, mu3_old_women, mu4_old_women,
        theta_u60_women[1], theta_u60_women[2], theta_u60_women[3],
            theta_u60_women[4],
        theta_60_62_women[1], theta_60_62_women[2], theta_60_62_women[3],
            theta_60_62_women[4],
        theta_62_65_women[1], theta_62_65_women[2], theta_62_65_women[3],
            theta_62_65_women[4],
        theta_65_plus_women[1], theta_65_plus_women[2],
            theta_65_plus_women[3], theta_65_plus_women[4]
        ]),
      covs,
      times,
      w
    )
  );

}
```

Appendix 3. Solution to Kolmogorov equation

The non-zero elements for transition probability matrix are given in the table below.

$$\begin{pmatrix} p_{11} & p_{12} & p_{13} & p_{14} & p_{15} \\ 0 & p_{22} & p_{23} & p_{24} & 0 \\ 0 & 0 & p_{33} & p_{34} & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{34}$$

Element	Equation
p_{11}	$\exp(-(\mu_1+\mu_2)t)$
p_{12}	$\frac{\mu_1(exp(-\mu_3 t)-\exp(-(mu1+mu2)t))}{(\mu_1+\mu_2-\mu_3)}$
p_{13}	$\frac{\mu_1\mu_3(\exp(-\mu_4 t)(\mu_1+\mu_2-\mu_3)+\exp(-\mu_3 t)(\mu_4-\mu_1-\mu_2)+\exp(-(\mu_1+\mu_2)t)(\mu_3-\mu_4))}{(\mu_1+\mu_2-\mu_3)(-\mu_4+\mu_1+\mu_2)(\mu_3-\mu_4)}$
p_{14}	$\frac{\mu_1}{\mu_1+\mu_2} + \frac{\exp(-\mu_3 t)\mu_1\mu_4}{(\mu_1+\mu_2-\mu_3)(\mu_3-\mu_4)} + \frac{\exp(-\mu_4 t)\mu_1\mu_3}{(\mu_3-\mu_4)(\mu_4-\mu_1-\mu_2)} +$ $\exp(-(\mu_1+\mu_2)t)(-\frac{\mu_1}{\mu_1+\mu_2}+\frac{\mu_1}{\mu_1+\mu_2-\mu_3}-\mu_1\mu_3(\frac{1}{(\mu_1+\mu_2-\mu_3)(\mu_3-\mu_4)}+\frac{1}{(\mu_3-\mu_4)(\mu_4-\mu_1-\mu_2)}))$
p_{15}	$\frac{\mu_2(1-\exp(-(\mu_1+\mu_2)t))}{\mu_1+\mu_2}$
p_{22}	$\exp(-\mu_3 t)$
p_{23}	$\frac{\mu_3(\exp(-\mu_4 t)-\exp(-\mu_3 t))}{(\mu_3-\mu_4)}$
p_{24}	$1 - \frac{\mu_3\exp(-\mu_4 t)+\mu_4\exp(-\mu_3 t)}{\mu_3-\mu_4}$
p_{33}	$\exp(-\mu_4 t)$
p_{34}	$1-\exp(-\mu_4 t)$

Table 8: Solution to equation $\mathbf{P}(t) = \mathbf{exp}(\mathbf{Q}t)$ where $\mathbf{Q}$ is as in equation 24.

Appendix 4. Parameter estimates, $\hat{R}$-values and effective sample sizes.

Table 9: Sensitivity parameter estimates in model $\mathcal{M}_1$

Parameter	Rhat	n_eff	mean	sd	2.5%	50%	97.5%
$S_{d,\text{men}}$	0.999	2666	0.492	0.016	0.462	0.492	0.522
$S_{e,\text{men}}$	1.000	1266	0.111	0.010	0.093	0.110	0.130
$S_{od,\text{men}}$	1.003	1667	0.092	0.009	0.073	0.092	0.109
$S_{d,\text{women}}$	0.998	2997	0.485	0.015	0.455	0.485	0.514
$S_{e,\text{women}}$	1.002	1117	0.076	0.008	0.062	0.076	0.094
$S_{od,\text{women}}$	1.001	1175	0.064	0.007	0.051	0.063	0.077

Table 10: Transition rate parameter estimates in model $\mathcal{M}_1$

Parameter	$\hat{R}$	n_eff	mean	sd	2.5%	50%	97.5%
$\mu_{1,\text{men},<65y}$	0.999	2872	0.002	0.001	0.001	0.002	0.003
$\mu_{1,\text{men},\geq 65y}$	0.999	2414	0.003	0.001	0.001	0.002	0.005
$\mu_{2,\text{men},<65y}$	0.999	2765	0.002	0.001	0.001	0.002	0.003
$\mu_{2,\text{men},\geq 65y}$	0.998	2294	0.003	0.001	0.001	0.002	0.005
$\mu_{3,\text{men},<65y}$	1.003	1341	0.013	0.002	0.008	0.012	0.018
$\mu_{3,\text{men},\geq 65y}$	0.999	1166	0.025	0.004	0.017	0.024	0.034
$\mu_{4,\text{men},<65y}$	0.999	3227	0.245	0.035	0.185	0.244	0.321
$\mu_{4,\text{men},\geq 65y}$	0.999	3117	0.487	0.067	0.368	0.482	0.629
$\mu_{1,\text{women},<65y}$	0.998	2341	0.002	0.000	0.001	0.002	0.003
$\mu_{1,\text{women},\geq 65y}$	0.999	3550	0.002	0.001	0.001	0.002	0.003
$\mu_{2,\text{women},<65y}$	0.998	3272	0.002	0.000	0.001	0.002	0.003
$\mu_{2,\text{women},\geq 65y}$	0.999	2368	0.002	0.001	0.001	0.002	0.004
$\mu_{3,\text{women},<65y}$	1.001	1253	0.020	0.004	0.014	0.019	0.028
$\mu_{3,\text{women},\geq 65y}$	1.000	878	0.031	0.006	0.022	0.031	0.044
$\mu_{4,\text{women},<65y}$	0.999	3071	0.327	0.045	0.248	0.324	0.421
$\mu_{4,\text{women},\geq 65y}$	0.999	3057	0.912	0.135	0.685	0.902	1.196
α_{men}	0.999	3278	0.514	0.113	0.294	0.520	0.731
α_{women}	0.999	3543	0.543	0.110	0.319	0.547	0.747

Table 11: Age stratified initial state distribution parameter estimates for men in model $\mathcal{M}_1$

Parameter	$\hat{R}$	n_eff	mean	sd	2.5%	50%	97.5%
$\theta_{1,<60,\text{men}}$	1.002	1119	0.875	0.009	0.858	0.875	0.891
$\theta_{2,<60,\text{men}}$	1.001	1113	0.073	0.011	0.052	0.072	0.095
$\theta_{3,<60,\text{men}}$	0.999	1887	0.003	0.000	0.002	0.003	0.004
$\theta_{4,<60,\text{men}}$	1.001	1263	0.049	0.011	0.029	0.049	0.073
$\theta_{1,60-62,\text{men}}$	1.003	1112	0.851	0.011	0.828	0.851	0.871
$\theta_{2,60-62,\text{men}}$	0.999	1339	0.066	0.012	0.045	0.065	0.089
$\theta_{3,60-62,\text{men}}$	0.999	2251	0.004	0.001	0.003	0.004	0.005
$\theta_{4,60-62,\text{men}}$	1.000	1333	0.080	0.014	0.052	0.080	0.107
$\theta_{1,62-65,\text{men}}$	1.004	1212	0.839	0.012	0.815	0.839	0.862
$\theta_{2,62-65,\text{men}}$	1.002	1531	0.071	0.013	0.048	0.070	0.097
$\theta_{3,62-65,\text{men}}$	0.999	2828	0.004	0.001	0.003	0.004	0.005
$\theta_{4,62-65,\text{men}}$	0.999	1692	0.087	0.015	0.056	0.087	0.116
$\theta_{1,65+,\text{men}}$	1.002	1599	0.820	0.018	0.784	0.820	0.853
$\theta_{2,65+,\text{men}}$	0.999	2185	0.076	0.018	0.043	0.075	0.113
$\theta_{3,65+,\text{men}}$	0.999	2906	0.006	0.001	0.004	0.006	0.009
$\theta_{4,65+,\text{men}}$	0.999	2177	0.098	0.022	0.058	0.098	0.142

Table 12: Age stratified initial state distribution parameter estimates for women in model $\mathcal{M}_1$

Parameter	$\hat{R}$	n_eff	mean	sd	2.5%	50%	97.5%
$\theta_{1,<60,\text{women}}$	1.002	1053	0.933	0.006	0.920	0.933	0.944
$\theta_{2,<60,\text{women}}$	0.999	1168	0.038	0.006	0.028	0.038	0.051
$\theta_{3,<60,\text{women}}$	1.000	2041	0.002	0.000	0.001	0.002	0.002
$\theta_{4,<60,\text{women}}$	1.003	1124	0.027	0.006	0.015	0.027	0.040
$\theta_{1,60-62,\text{women}}$	1.002	1114	0.919	0.008	0.903	0.920	0.934
$\theta_{2,60-62,\text{women}}$	1.000	1378	0.038	0.007	0.026	0.038	0.051
$\theta_{3,60-62,\text{women}}$	1.000	2629	0.002	0.000	0.001	0.002	0.003
$\theta_{4,60-62,\text{women}}$	1.002	1419	0.040	0.008	0.025	0.040	0.058
$\theta_{1,62-65,\text{women}}$	1.002	1143	0.896	0.011	0.874	0.897	0.915
$\theta_{2,62-65,\text{women}}$	0.999	1430	0.034	0.007	0.022	0.033	0.049
$\theta_{3,62-65,\text{women}}$	1.000	3401	0.002	0.000	0.001	0.002	0.003
$\theta_{4,62-65,\text{women}}$	1.002	1369	0.068	0.011	0.047	0.067	0.091
$\theta_{1,65+,\text{women}}$	1.000	2004	0.875	0.018	0.836	0.876	0.905
$\theta_{2,65+,\text{women}}$	1.000	2539	0.049	0.012	0.027	0.048	0.074
$\theta_{3,65+,\text{women}}$	0.999	3778	0.004	0.001	0.002	0.004	0.007
$\theta_{4,65+,\text{women}}$	1.001	2537	0.072	0.018	0.041	0.071	0.113