Master's thesis

Master's Programme in Mathematics and Statistics

A multistate analysis of ulcerative colitis and colorectal cancer

Tomas Tanskanen

December 2022

Supervisors: Janne Pitkäniemi, Sangita Kulathinal

University of Helsinki

Faculty of Science

P. O. Box 68 (Pietari Kalmin katu 5)
00014 University of Helsinki

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty Faculty of Science	Koulutusohjelma — Utbildningsprogram — Degree programme Master's Programme in Mathematics and Statistics

Tekijä — Författare — Author

Tomas Tanskanen

Työn nimi — Arbetets titel — Title

A multistate analysis of ulcerative colitis and colorectal cancer

Työn laji — Arbetets art — Level Master's thesis	Aika — Datum — Month and year December 2022	Sivumäärä — Sidantal — Number of pages 37

Tiivistelmä — Referat — Abstract

Colorectal cancer (CRC) accounts for one in 10 new cancer cases worldwide. CRC risk is determined by a complex interplay of constitutional, behavioral, and environmental factors. Patients with ulcerative colitis (UC) are at increased risk of CRC, but effect estimates are heterogeneous, and many studies are limited by small numbers of events. Furthermore, it has been challenging to distinguish the effects of age at UC diagnosis and duration of UC. Multistate models provide a useful statistical framework for analyses of cancers and premalignant conditions. This thesis has three aims: to review the mathematical and statistical background of multistate models; to study maximum likelihood estimation in the illness-death model with piecewise constant hazards; and to apply the illness-death model to UC and CRC in a population-based cohort study in Finland in 2000–2017, considering UC as a premalignant state that may precede CRC.

A likelihood function is derived for multistate models under noninformative censoring. The multistate process is considered as a multivariate counting process, and product integration is reviewed. The likelihood is constructed by partitioning the study time into subintervals and finding the limit as the number of subintervals tends to infinity. Two special cases of the illness-death model with piecewise constant hazards are studied: a simple Markov model and a non-Markov model with multiple time scales. In the latter case, the likelihood is factorized into terms proportional to Poisson likelihoods, which permits estimation with standard software for generalized linear models.

The illness-death model was applied to study the relationship between UC and CRC in a population-based sample of 2.5 million individuals in Finland in 2000–2017. Dates of UC and CRC diagnoses were obtained from the Finnish Care Register for Health Care and the Finnish Cancer Registry, respectively. Individuals with prevalent CRC were excluded from the study cohort. Individuals in the study cohort were followed from January 1, 2000, to the date of first CRC diagnosis, death from other cause, emigration, or December 31, 2017, whichever came first. A total of 23,533 incident CRCs were diagnosed during 41 million person-years of follow-up. In addition to 8,630 patients with prevalent UC, there were 19,435 cases of incident UC. Of the 23,533 incident CRCs, 298 (1.3%) were diagnosed in patients with pre-existing UC. In the first year after UC diagnosis, the HR for incident CRC was 4.67 (95% CI: 3.07, 7.09) in females and 7.62 (95% CI: 5.65, 10.3) in males. In patients with UC diagnosed 1–3 or 4–9 years earlier, CRC incidence did not differ from persons without UC. When 10–19 years had passed from UC diagnosis, the HR for incident CRC was 1.63 (95% CI: 1.19, 2.24) in females and 1.29 (95% CI: 0.96, 1.75) in males, and after 20 years, the HR was 1.61 (95% CI: 1.13, 2.31) in females and 1.74 (95% CI: 1.31, 2.31) in males. Early-onset UC (age <40 years) was associated with a markedly increased long-term risk of CRC. The HR for CRC in early-onset UC was 4.13 (95% CI: 2.28, 7.47) between 4–9 years from UC diagnosis, 4.88 (95% CI: 3.46, 6.88) between 10–19 years, and 2.63 (95% CI: 2.01, 3.43) after 20 years.

In this large population-based cohort study, we estimated CRC risk in persons with and without UC in Finland in 2000–2017, considering both the duration of UC and age at UC diagnosis. Patients with early-onset UC are at increased risk of CRC, but the risk is likely to depend on disease duration, extent of disease, attained age, and other risk factors. Increased CRC risk in the first year after UC diagnosis may be in part due to detection bias, whereas chronic inflammation may underlie the long-term excess risk of CRC in patients with UC.

Avainsanat — Nyckelord — Keywords

ulcerative colitis, colorectal cancer, multistate model, epidemiology

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Contents

Abbreviations

CI	Confidence interval
CRC	Colorectal cancer
HR	Hazard ratio
IBD	Inflammatory bowel disease
MLE	Maximum likelihood estimate
SIR	Standardized incidence ratio
UC	Ulcerative colitis

1. Introduction

1.1 Epidemiological research

Epidemiology has been defined as “the study of the occurrence and distribution of health-related events, states, and processes in specified populations, including the study of the determinants influencing such processes, and the application of this knowledge to control relevant health problems” (Porta et al., 2014).

Time-to-event analysis is widely used in epidemiological research. For example, the interest may be in estimating the distribution of time from birth to diagnosis, or from diagnosis to death. However, the study period may end before all patients have experienced the event of interest, and some patients may be lost to follow-up, leading to right-censored data. A common measure of disease occurrence is the incidence rate, which is defined as the number of new cases divided by the person-years of follow-up (Rothman et al., 2021). If the underlying hazard of event occurrence is constant over time, the incidence rate provides an estimate of the hazard rate. Hazard ratios (HRs) are often used to compare rates of occurrence between groups or degrees of exposure.

A cohort is a group of individuals who share a common characteristic such as the same year of birth, lifestyle habit, or disease. In cohort studies, a cohort is followed over time, and one or more event types are observed. The study objective may be to estimate rates of health-related events, or to relate risk factors to the event rates. The design of a cohort study allows the investigator to determine the timing and temporal order of exposure and outcome, which are important criteria for a possible causal relationship.

1.2 Colorectal cancer and ulcerative colitis

Colorectal cancer (CRC) accounts for one in 10 new cancer cases worldwide (Sung et al., 2021). The risk of developing CRC is determined by a complex interplay of constitutional, behavioral, and environmental factors. The incidence of CRC is higher in older people. In the Nordic countries, the age-standardized incidence is higher in men than in women (Engholm et al., 2010; Larønningen et al., 2022). Modifiable risk factors for CRC include obesity, physical inactivity, dietary factors, alcohol consumption, and smoking (Dekker et al., 2019). Hereditary risk factors for CRC include Lynch syndrome, polyposis syndromes, and low-penetrance alleles (Peters et al., 2015). Medical risk factors for CRC include ulcerative colitis (UC), Crohn's disease, and type 2 diabetes mellitus (Dekker et al., 2019).

UC is an inflammatory bowel disease (IBD) that primarily affects the rectum and colon. Patients with UC are at increased risk of CRC, but effect estimates are heterogeneous, and many studies are limited by small numbers of events (Jess et al., 2012; Lutgens et al., 2013). Recently, a large population-based cohort study in Denmark and Sweden confirmed UC as a risk factor for CRC, although the excess risk decreased over time (Olén et al., 2020). Risk factors for CRC in patients with UC include early age at diagnosis, longer disease duration, extensive colitis, coexisting primary sclerosing cholangitis, and a family history of CRC (Annese et al., 2013). Based on the available data, it is challenging to distinguish the effects of age at UC diagnosis and duration of UC (Lutgens et al., 2013; Annese et al., 2013). Reliable estimates of CRC risk in UC are important for designing preventive measures such as endoscopic screening and surveillance protocols.

1.3 Multistate models

Multistate models (see, e.g., Andersen and Keiding (2002)) provide a useful statistical framework for analyses of cancers and premalignant conditions. The illness-death model is a simple yet flexible multistate model with one initial state ("Healthy"), one intermediate state ("Ill"), and one final state ("Dead"). It is applicable to semicompeting risks data involving a terminal and a nonterminal event (Rothman et al., 2021). Other examples of multistate models include models for competing risks and recurrent events.

1.4 Organization of the thesis

In this population-based cohort study, we use the illness-death model to study the relationship between UC and CRC in Finland in 2000–2017. Chapter 1 provides a brief introduction to epidemiological research, CRC, UC, and multistate models. Chapter 2 defines the aims of the study. In Chapter 3, we construct a likelihood function for multistate models. In Chapter 4, we study maximum likelihood estimation in the illness-death model with piecewise constant hazards. In Chapter 5, we apply the illness-death model to data on UC and CRC. In Chapter 6, the results are discussed and interpreted in the context of earlier research.

2. Aims of the study

I. To review the mathematical and statistical background of multistate models. A likelihood function is derived for multistate models under noninformative censoring.

II. To study maximum likelihood estimation in the illness-death model with piecewise constant hazards. Two models are considered: a simple Markov model and a non-Markov model with multiple time scales.

III. To apply the illness-death model to UC and CRC in a population-based cohort in Finland in 2000–2017, regarding UC as a premalignant state that may precede CRC.

3. Multistate models

Section 3.1 introduces stochastic processes, right-censoring, and left-truncation. In Sections 3.2, 3.3, and 3.4, we derive a likelihood function for multistate models following Cook and Lawless (2007, 2018) and Aalen et al. (2008).

3.1 Basic concepts

3.1.1 Stochastic processes

Consider a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, where Ω is a sample space, $\mathscr{F}$ is a σ-algebra, and $\mathbb{P}$ is a probability measure. A random variable is a measurable function from a probability space to a measurable space, such as the real line $(-\infty, \infty)$ equipped with the Borel σ-algebra $(\mathbb{R}, \mathscr{B})$. A filtration $\{\mathscr{F}_t : t \geq 0\}$, also called a history, is an increasing set of sub-σ-algebras ($\mathscr{F}_t \subset \mathscr{F}$), meaning that if $s < t$, then $\mathscr{F}_s \subset \mathscr{F}_t$. The limit $\mathscr{F}_{t-}$ is the smallest σ-algebra containing all the sets in $\bigcup_{h>0} \mathscr{F}_{t-h}$ (Fleming and Harrington, 1991).

A stochastic process is an ordered set of random variables defined in the same probability space. If ω is an outcome in the probability space Ω, a continuous-time stochastic process can be written as $\{X_t(\omega) : t \geq 0\}$, where t has a continuous set of values. A continuous-time stochastic process may or may not have continuous sample paths. The process $\{X_t(\omega) : t \geq 0\}$ has continuous sample paths if $X_t(\omega)$ is a continuous function of t for almost all $\omega \in \Omega$. Left-continuous and right-continuous stochastic processes are defined similarly. We will use the notation $\{X(t) : t \geq 0\}$ for stochastic processes.

A multistate process is a right-continuous stochastic process $\{Z(t) : t \geq 0\}$ that takes values in a finite state space $S = \{1, 2, ..., K\}$. The state space consists of transient and absorbing states. Transient states are those from which exit is possible, whereas absorbing states cannot be left once entered. A multistate process is a Markov process if its transition hazards depend on the process history only through the current state. In a semi-Markov process, the transition hazards depend on the process history only through time from entry to the current state (Andersen, 1993).

3.1.2 Right-censoring and left-truncation

Right-censoring and left-truncation are common types of incomplete observation in time-to-event studies. In the context of multistate models, right-censoring means that observation ends before the multistate process has reached an absorbing state. Left-truncation, also called late entry, occurs when observation begins after the start of the multistate process. For example, if the multistate process starts at birth, late entry occurs when an individual enters the study some time after birth.

3.2 Transition hazard functions

Let $\{Z^c(t) : t \geq 0\}$ denote an uncensored (complete) right-continuous multistate process with state space $S = \{1, 2, ..., K\}$, and let $\{X^c(t) : t \geq 0\}$ denote an uncensored left-continuous covariate process. The hazard function for the transition $k \to l$ ($k, l \in S$, $k \neq l$) is defined as

$$\alpha_{kl}(t|\mathscr{F}^c_{t-}) = \lim_{\Delta t \to 0+} \frac{\mathbb{P}(Z^c(t + \Delta t-) = l|Z^c(t-) = k, \mathscr{F}^c_{t-})}{\Delta t}, \tag{3.2.1}$$

where $\mathscr{F}^c_t = \sigma\{Z^c(s), X^c(s) : 0 \leq s \leq t\}$, and $\mathscr{F}^c_{t-}$ is the history up to but not including time t. The numerator is the conditional probability that a $k \to l$ transition occurs over the interval $[t, t + \Delta t)$ given the history $\mathscr{F}^c_{t-}$ and given that the multistate process is in state k just before time t. From definition (3.2.1), it follows that

$$\mathbb{P}(Z^c(t + \Delta t-) = l|Z^c(t-) = k, \mathscr{F}^c_{t-}) = \alpha_{kl}(t|\mathscr{F}^c_{t-})\Delta t + o(\Delta t), \tag{3.2.2}$$

where $o(\cdot)$ is a function such that $\frac{o(x)}{x} \to 0$ as $x \to 0$.

3.2.1 Uncensored counting process

A multistate process can be conveniently expressed as a multivariate counting process. Let $\{N^c_{kl}(t) : t \geq 0\}$ denote an uncensored right-continuous counting process that records the number of $k \to l$ transitions over time, and let $\Delta N^c_{kl}(t) = N^c_{kl}(t + \Delta t-) - N^c_{kl}(t-)$ denote the number of $k \to l$ transitions over the interval $[t, t+\Delta t)$. Also, let $Y_k(t) = I(Z^c(t) = k)$ indicate whether the multistate process is in state k at time t.

The intensity function of the uncensored counting process N_{kl}^c is defined as

$$\lambda_{kl}^c(t|\mathscr{F}_{t-}^c) = \lim_{\Delta t \to 0+} \frac{\mathbb{P}(\Delta N_{kl}^c(t) = 1|\mathscr{F}_{t-}^c)}{\Delta t}, \tag{3.2.3}$$

and the intensity and hazard functions are related by $\lambda_{kl}^c(t|\mathscr{F}_{t-}^c) = Y_k(t-)\alpha_{kl}(t|\mathscr{F}_{t-}^c)$ (Aalen et al., 2008, Section 1.4).

3.2.2 Observed counting process

Consider an individual who is observed over the period $[E, C]$, where $E \geq 0$ is an entry time and C is a censoring time. C is defined as $\min\{C^R, C^A\}$, where C^R is a random censoring time and C^A is a fixed administrative censoring time. In case of late entry, $E > 0$. Then $Y(t) = I(E < t \leq C)$ is an indicator of whether the individual is under observation at time t, and $\{Y(t) : t \geq 0\}$ is a left-continuous observation ("at risk") process. The observed number of $k \to l$ transitions over $[0, t]$ is given by the observed counting process

$$N_{kl}(t) = \int_0^t Y(s) dN_{kl}^c(s) = \sum_{t_j \in \mathscr{D}_{kl}^c} I(t_j \leq t) Y(t_j)(N_{kl}^c(t_j) - N_{kl}^c(t_j^-)), \tag{3.2.4}$$

where $\mathscr{D}_{kl}^c$ is the set of $k \to l$ transition times over $[0, C^A]$. The increment of the observed counting process over the interval $[t, t + \Delta t)$ is denoted by $\Delta N_{kl}(t) = N_{kl}(t + \Delta t-) - N_{kl}(t-)$. The observed multivariate counting process can be written shortly as $N(t) = (N_1(t)^T, ..., N_K(t)^T)^T$, where $N_k(t) = (N_{kl}(t), l \neq k, l = 1, ..., K)^T$. The observed counting process N_{kl} has the intensity function

$$\lambda_{kl}(t|\mathscr{F}_{t-}) = \lim_{\Delta t \to 0+} \frac{\mathbb{P}(\Delta N_{kl}(t) = 1|\mathscr{F}_{t-})}{\Delta t}, \tag{3.2.5}$$

where $\mathscr{F}_{t-}$ is the history of the observed processes.

3.3 A product integral

This section defines a product integral (Cook and Lawless, 2007, p. 28) that will be used to construct a likelihood function for multistate analysis in Section 3.4.2. Let $a = u_0 < u_1 < \cdots < u_R = b$ define a partition of the interval $[a, b]$ into R subintervals of length $\Delta u_r = u_{r+1} - u_r$. If g is a continuous function over the interval $[a, b]$, its product integral over $[a, b]$ is defined as

$$\prod_{[a,b]} \{1 + g(u)du\} = \lim_{R\to\infty} \prod_{r=0}^{R} \{1 + g(u_r)\Delta u_r\}, \tag{3.3.1}$$

where $u_{R+1} = u_R^+$. For small Δu_r, the sum $1 + g(u_r)\Delta u_r$ is positive, and we can write

$$\begin{aligned}\prod_{[a,b]} \{1 + g(u)du\} &= \lim_{R\to\infty} \prod_{r=0}^{R} \exp\{\log(1 + g(u_r)\Delta u_r)\} \\ &= \lim_{R\to\infty} \exp\left\{\sum_{r=0}^{R} \log(1 + g(u_r)\Delta u_r)\right\} \\ &= \exp\left\{\lim_{R\to\infty} \sum_{r=0}^{R} \log(1 + g(u_r)\Delta u_r)\right\},\end{aligned} \tag{3.3.2}$$

where the last equality follows from the continuity of the exponential function. When $|x| < 1$, the function $\log(1+x)$ has the Maclaurin expansion $\log(1+x) = x + x\epsilon(x)$, where ϵ is a function such that $\epsilon(x) \to 0$ as $x \to 0$. By substituting $x = g(u_r)\Delta u_r$ into $\log(1+x)$ and using the fact that g is continuous and therefore bounded over $[a, b]$, we get

$$\begin{aligned}\log(1 + g(u_r)\Delta u_r) &= g(u_r)\Delta u_r + g(u_r)\Delta u_r\epsilon(g(u_r)\Delta u_r) \\ &= g(u_r)\Delta u_r + o(\Delta u_r).\end{aligned} \tag{3.3.3}$$

Assuming that $\max_r \Delta u_r \to 0$ as $R \to \infty$, the product integral of g over $[a, b]$ is

$$\begin{aligned}\prod_{[a,b]} \{1 + g(u)du\} &= \exp\left\{\lim_{R\to\infty} \sum_{r=0}^{R} \Big(g(u_r)\Delta u_r + o(\Delta u_r)\Big)\right\} \\ &= \exp\left\{\lim_{R\to\infty} \sum_{r=0}^{R} \left(g(u_r) + \frac{o(\Delta u_r)}{\Delta u_r}\right)\Delta u_r\right\} \\ &= \exp\left\{\int_a^b g(u)du\right\},\end{aligned} \tag{3.3.4}$$

where $\int_a^b g(u)du$ is a Riemann integral.

3.4 Likelihood function

3.4.1 Model assumptions

The likelihood construction is based on the following assumptions.

1. The observation and covariate processes are noninformative (Cook and Lawless, 2018, p. 33).
2. The probability of two or more transitions over $[t, t + \Delta t)$ is of the order $o(\Delta t)$, which means that two or more transitions do not occur simultaneously.
3. The hazard functions are continuous.

By the first assumption, modeling the observation and covariate processes does not provide information on the parameters of the multistate process, and we can restrict attention to the conditional likelihood of the multistate process given the observed late entry times, censoring times, and covariate paths. The probability of a $k \to l$ transition over $[t, t+\Delta t)$ for an individual followed over $[t, t + \Delta t)$ is

$$\begin{aligned}\mathbb{P}(\Delta N_{kl}(t) = 1 | E \leq t < t + \Delta t \leq C, \mathscr{F}_{t-}) &= \mathbb{P}(\Delta N^c_{kl}(t) = 1 | \mathscr{F}^c_{t-}) \\ &= \lambda^c_{kl}(t|\mathscr{F}^c_{t-}) + o(\Delta t) \\ &= Y_k(t-)\alpha_{kl}(t|\mathscr{F}^c_{t-}) + o(\Delta t).\end{aligned} \tag{3.4.1}$$

The probability of not observing any transition out of state k for an individual followed over $[t, t+\Delta t)$ can be derived similarly. Let $\Delta N_{k\cdot}(t) = \sum_{l\neq k=1}^{K} \Delta N_{kl}(t)$ denote the observed number of transitions out of state k over $[t, t+\Delta t)$. By the second assumption and equation (3.4.1),

$$\begin{aligned}&\mathbb{P}(\Delta N_{k\cdot}(t) = 0 | E \leq t < t + \Delta t \leq C, \mathscr{F}_{t-}) \\ &= 1 - \mathbb{P}(\Delta N_{k\cdot}(t) \geq 1 | E \leq t < t + \Delta t \leq C, \mathscr{F}_{t-}) \\ &= 1 - \sum_{l\neq k=1}^{K} \mathbb{P}(\Delta N_{kl}(t) \geq 1 | E \leq t < t + \Delta t \leq C, \mathscr{F}_{t-}) \\ &= 1 - \sum_{l\neq k=1}^{K} (\mathbb{P}(\Delta N_{kl}(t) = 1 | E \leq t < t + \Delta t \leq C, \mathscr{F}_{t-}) + o(\Delta t)) \\ &= 1 - \sum_{l\neq k=1}^{K} \Big(Y_k(t-)\alpha_{kl}(t|\mathscr{F}^c_{t-})\Delta t + o(\Delta t) + o(\Delta t)\Big) \\ &= 1 - Y_k(t-)\alpha_{k\cdot}(t|\mathscr{F}^c_{t-})\Delta t + o(\Delta t),\end{aligned} \tag{3.4.2}$$

where $\alpha_{k\cdot}(t|\mathscr{F}^c_{t-}) = \sum_{l\neq k=1}^{K} \alpha_{kl}(t|\mathscr{F}^c_{t-})$.

3.4.2 Likelihood construction

Here we construct a likelihood function for multistate models allowing late entry and right-censoring (Cook and Lawless, 2018, Section 2.2). Let $0 = u_0 < u_1 < ... < u_R = C^A$ denote a partition of the study time $[0, C^A]$ into R subintervals. In cases of late entry and random censoring, let $u_m = E$ and $u_n = C^R$ for some $0 < m < n < R$. Let $\Delta N_{kl}(u_r) = N_{kl}(u_{r+1}-) - N_{kl}(u_r-)$ denote the observed number of $k \to l$ transitions over the interval $[u_r, u_{r+1})$. When the partition is sufficiently dense, there will be at most one transition per subinterval, and the likelihood contribution of an individual can be written as

$$\begin{aligned}\prod_{r=0}^{R}\prod_{k=1}^{K}\prod_{l\neq k=1}^{K} \Big[& \mathbb{P}\left(\Delta N_{kl}(u_r) = 1 | E \leq u_r < u_{r+1} \leq C, \mathscr{F}_{u_r-}\right)^{\Delta N_{kl}(u_r)} \\ & \times \mathbb{P}\left(\Delta N_{k\cdot}(u_r) = 0 | E \leq u_r < u_{r+1} \leq C, \mathscr{F}_{u_r-}\right)^{1-\Delta N_{k\cdot}(u_r)}\Big]^{Y(u_{r+1})},\end{aligned} \tag{3.4.3}$$

where 0^0 is defined as 1. Using equations (3.4.1) and (3.4.2), the likelihood can be written in terms of the hazard functions as

$$\begin{aligned}\prod_{r=0}^{R}\prod_{k=1}^{K}\prod_{l\neq k=1}^{K} \Bigg[& \Big(Y_k(u_r-)\alpha_{kl}(u_r|\mathscr{F}^c_{u_r-})\Delta u_r + o(\Delta u_r)\Big)^{\Delta N_{kl}(u_r)} \\ & \times \Big(1 - Y_k(u_r-)\alpha_{k\cdot}(u_r|\mathscr{F}^c_{u_r-})\Delta u_r + o(\Delta u_r)\Big)^{1-\Delta N_{k\cdot}(u_r)}\Bigg]^{Y(u_{r+1})},\end{aligned} \tag{3.4.4}$$

By dividing the likelihood by $\prod_{r=0}^{R}\prod_{k=1}^{K}\prod_{l\neq k=1}^{K}(\Delta u_r)^{\Delta N_{kl}(u_r)}$, which only depends on the observed data and the partition points, we get

$$\begin{aligned}\prod_{r=0}^{R}\prod_{k=1}^{K}\prod_{l\neq k=1}^{K} \Bigg[& \left(Y_k(u_r-)\alpha_{kl}(u_r|\mathscr{F}^c_{u_r-}) + \frac{o(\Delta u_r)}{\Delta u_r}\right)^{\Delta N_{kl}(u_r)} \\ & \times \Big(1 - Y_k(u_r-)\alpha_{k\cdot}(u_r|\mathscr{F}^c_{u_r-})\Delta u_r + o(\Delta u_r)\Big)^{1-\Delta N_{k\cdot}(u_r)}\Bigg]^{Y(u_{r+1})},\end{aligned} \tag{3.4.5}$$

and by changing the order of multiplication, the likelihood becomes

$$\begin{aligned}\prod_{k=1}^{K}\prod_{l\neq k=1}^{K} \Bigg[& \prod_{r=0}^{R}\left(Y_k(u_r-)\alpha_{kl}(u_r|\mathscr{F}^c_{u_r-}) + \frac{o(\Delta u_r)}{\Delta u_r}\right)^{\Delta N_{kl}(u_r)Y(u_{r+1})} \\ & \times \prod_{r=0}^{R}\Big(1 - Y_k(u_r-)\alpha_{k\cdot}(u_r|\mathscr{F}^c_{u_r-})\Delta u_r + o(\Delta u_r)\Big)^{(1-\Delta N_{k\cdot}(u_r))Y(u_{r+1})}\Bigg].\end{aligned} \tag{3.4.6}$$

To find the limit of (3.4.6) as $R \to \infty$, assume that $\max_r \Delta u_r \to 0$ as $R \to \infty$. Then

$$\begin{aligned}&\lim_{R\to\infty} \prod_{r=0}^{R} \left(Y_k(u_r-)\alpha_{kl}(u_r|\mathscr{F}^c_{u_r-}) + \frac{o(\Delta u_r)}{\Delta u_r} \right)^{\Delta N_{kl}(u_r)Y(u_{r+1})} \\ &= \prod_{t_j \in \mathscr{D}_{kl}} \alpha_{kl}(t_j|\mathscr{F}^c_{t_j-}),\end{aligned} \tag{3.4.7}$$

where $\mathscr{D}_{kl}$ is the set of observed $k \to l$ transition times over $[0, C^A]$. Note that we only had to consider intervals that contain a transition time. On the other hand, from the relation between the product integral and the Riemann integral (Section 3.3), it follows that

$$\begin{aligned}&\lim_{R\to\infty} \prod_{r=0}^{R} \Big(1 - Y_k(u_r-)\alpha_{k\cdot}(u_r|\mathscr{F}^c_{u_r-})\Delta u_r + o(\Delta u_r)\Big)^{(1-\Delta N_{k\cdot}(u_r))Y(u_{r+1})} \\ &= \exp\left\{-\int_0^\infty Y(u)Y_k(u-)\alpha_{k\cdot}(u|\mathscr{F}^c_{u-})du\right\}\end{aligned} \tag{3.4.8}$$

because $\Delta N_{k\cdot}(u_r) = 0$ for all but a finite number of subintervals. Using the results (3.4.7) and (3.4.8), we find that the likelihood contribution is

$$\prod_{k=1}^{K} \prod_{l\neq k=1}^{K} \prod_{t_j \in \mathscr{D}_{kl}} \alpha_{kl}(t_j|\mathscr{F}^c_{t_j-}) \exp\left\{-\int_0^\infty Y(u)Y_k(u-)\alpha_{kl}(u|\mathscr{F}^c_{u-})du\right\}. \tag{3.4.9}$$

To extend the model to n independent individuals $i \in \{1, ..., n\}$ and parametric hazard functions, we can write the likelihood function L as

$$L(\theta) = \prod_{i=1}^{n} \prod_{k=1}^{K} \prod_{l\neq k=1}^{K} \prod_{t_{ij} \in \mathscr{D}_{ikl}} \alpha_{ikl}(t_{ij}|\theta, \mathscr{F}^c_{i,t_{ij}-}) \exp\left\{-\int_0^\infty Y_i(u)Y_{ik}(u-)\alpha_{ikl}(u|\theta, \mathscr{F}^c_{i,u-})du\right\}, \tag{3.4.10}$$

where θ is a parameter vector.

4. The illness-death model

4.1 Model definition

In this section, we consider the unidirectional illness-death model. The state space is $S = \{1, 2, 3\}$, and the possible transitions are $1 \to 2$, $1 \to 3$, and $2 \to 3$. The states 1, 2, and 3 are often labeled as "Healthy", "Ill", and "Dead", respectively. States 1 and 2 are transient, whereas state 3 is absorbing. A state diagram is shown in Figure 1.

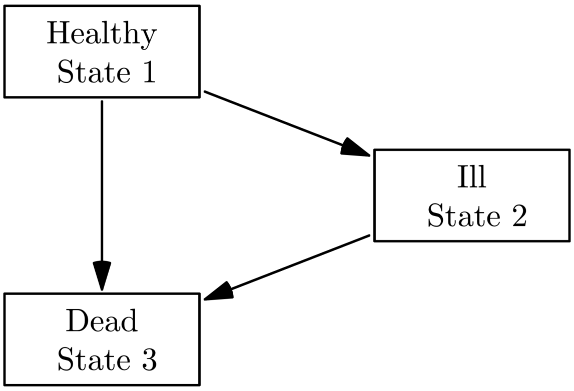


Figure 1: The unidirectional illness-death model.

Using the likelihood expression (3.4.10) from the previous chapter, the likelihood for n independent individuals is $L(\theta) = L_{12}(\theta)L_{13}(\theta)L_{23}(\theta)$, where

$$\begin{aligned}
L_{12}(\theta) =& \prod_{i=1}^{n} \prod_{t_{ij} \in \mathscr{D}_{i12}} \alpha_{i12}(t_{ij}|\theta, \mathscr{F}^c_{i,t_{ij}-}) \exp\left\{ -\int_0^\infty Y_i(u)Y_{i1}(u-)\alpha_{i12}(u|\theta, \mathscr{F}^c_{i,u-})du \right\}, \\
L_{13}(\theta) =& \prod_{i=1}^{n} \prod_{t_{ij} \in \mathscr{D}_{i13}} \alpha_{i13}(t_{ij}|\theta, \mathscr{F}^c_{i,t_{ij}-}) \exp\left\{ -\int_0^\infty Y_i(u)Y_{i1}(u-)\alpha_{i13}(u|\theta, \mathscr{F}^c_{i,u-})du \right\}, \text{ and} \\
L_{23}(\theta) =& \prod_{i=1}^{n} \prod_{t_{ij} \in \mathscr{D}_{i23}} \alpha_{i23}(t_{ij}|\theta, \mathscr{F}^c_{i,t_{ij}-}) \exp\left\{ -\int_0^\infty Y_i(u)Y_{i2}(u-)\alpha_{i23}(u|\theta, \mathscr{F}^c_{i,u-})du \right\}.
\end{aligned} \tag{4.1.1}$$

There are five types of likelihood contributions. Consider individual i with the entry time E_i and a possible censoring time C_i. The occurrence times of the transitions $1 \to 2$, $1 \to 3$, and $2 \to 3$ are denoted by t_{i12}, t_{i13}, and t_{i23}, respectively.

(i) If the individual enters the study in state 1 and stays in state 1 until censoring, the likelihood contribution is

$$L_i(\theta) = \exp\left\{-\int_{E_i}^{C_i}(\alpha_{i12}(u|\theta, \mathscr{F}^c_{i,u-}) + \alpha_{i13}(u|\theta, \mathscr{F}^c_{i,u-}))du\right\}. \tag{4.1.2}$$

(ii) If the individual enters the study in state 2 and stays in state 2 until censoring, the likelihood contribution is

$$L_i(\theta) = \exp\left\{-\int_{E_i}^{C_i}\alpha_{i23}(u|\theta, \mathscr{F}^c_{i,u-})du\right\}. \tag{4.1.3}$$

(iii) If the individual enters the study in state 1 and makes a $1 \to 3$ transition, the likelihood contribution is

$$\begin{aligned} L_i(\theta) =& \exp\left\{-\int_{E_i}^{t_{i13}}(\alpha_{i12}(u|\theta, \mathscr{F}^c_{i,u-}) + \alpha_{i13}(u|\theta, \mathscr{F}^c_{i,u-}))du\right\} \\ &\times \alpha_{i13}(t_{i13}|\theta, \mathscr{F}^c_{i,t_{i13}-}). \end{aligned} \tag{4.1.4}$$

(iv) If the individual enters the study in state 1, makes a $1 \to 2$ transition and stays in state 2 until censoring, the likelihood contribution is

$$\begin{aligned} L_i(\theta) =& \exp\left\{-\int_{E_i}^{t_{i12}}(\alpha_{i12}(u|\theta, \mathscr{F}^c_{i,u-}) + \alpha_{i13}(u|\theta, \mathscr{F}^c_{i,u-}))du\right\} \\ &\times \alpha_{i12}(t_{i12}|\theta, \mathscr{F}^c_{i,t_{i12}-}) \\ &\times \exp\left\{-\int_{t_{i12}}^{C_i}\alpha_{i23}(u|\theta, \mathscr{F}^c_{i,u-}))du\right\}. \end{aligned} \tag{4.1.5}$$

(v) If the individual enters the study in state 1 and makes the transitions $1 \to 2$ and $2 \to 3$, the likelihood contribution is

$$\begin{aligned} L_i(\theta) =& \exp\left\{-\int_{E_i}^{t_{i12}}(\alpha_{i12}(u|\theta, \mathscr{F}^c_{i,u-}) + \alpha_{i13}(u|\theta, \mathscr{F}^c_{i,u-}))du\right\} \\ &\times \alpha_{i12}(t_{i12}|\theta, \mathscr{F}^c_{i,t_{i12}-}) \\ &\times \exp\left\{-\int_{t_{i12}}^{t_{i23}}\alpha_{i23}(u|\theta, \mathscr{F}^c_{i,u-}))du\right\} \\ &\times \alpha_{i23}(t_{i23}|\theta, \mathscr{F}^c_{i,t_{i23}-}). \end{aligned} \tag{4.1.6}$$

4.2 Piecewise constant hazards

4.2.1 A Markov model

Suppose that t represents the attained age of the individual, and that the interest is in comparing mortality rates between healthy and ill individuals. Let $\theta = (\theta_{12}, \theta_{13}, \lambda) \in \mathbb{R}^3$ denote a parameter vector, and for all $i \in \{1, ..., n\}$, define the constant hazard functions $\alpha_{i12}(t|\theta, \mathscr{F}^c_{i,t-}) = \theta_{12}$, $\alpha_{i13}(t|\theta, \mathscr{F}^c_{i,t-}) = \theta_{13}$, and $\alpha_{i23}(t|\theta, \mathscr{F}^c_{i,t-}) = \lambda\theta_{13}$, where λ is the HR for death. Since the transition hazards do not depend on the history of the process, the model satisfies the Markov property. The likelihood for n independent individuals is $L(\theta) = L_{12}(\theta)L_{13}(\theta)L_{23}(\theta)$, where

$$\begin{aligned} L_{12}(\theta) =& \prod_{i=1}^{n} \prod_{t_{ij} \in \mathscr{D}_{i12}} \theta_{12} \exp\left\{- \int_0^\infty Y_i(u)Y_{i1}(u-)\theta_{12} du\right\}, \\ L_{13}(\theta) =& \prod_{i=1}^{n} \prod_{t_{ij} \in \mathscr{D}_{i13}} \theta_{13} \exp\left\{- \int_0^\infty Y_i(u)Y_{i1}(u-)\theta_{13} du\right\}, \text{ and} \\ L_{23}(\theta) =& \prod_{i=1}^{n} \prod_{t_{ij} \in \mathscr{D}_{i23}} \lambda\theta_{13} \exp\left\{- \int_0^\infty Y_i(u)Y_{i2}(u-)\lambda\theta_{13} du\right\}. \end{aligned} \tag{4.2.1}$$

This simplifies to

$$L(\theta) = \theta_{12}^{|\mathscr{D}_{12}|}\theta_{13}^{|\mathscr{D}_{13}|}(\lambda\theta_{13})^{|\mathscr{D}_{23}|} \exp\left\{-(\theta_{12} + \theta_{13})T_1 - \lambda\theta_{13}T_2\right\}, \tag{4.2.2}$$

where $|\mathscr{D}_{kl}| = \sum_{i=1}^n |\mathscr{D}_{ikl}|$ is the total number of observed $k \to l$ transitions, $|\mathscr{D}_{ikl}|$ is the number of observed $k \to l$ transitions for individual i, and $T_k = \sum_{i=1}^n \int_0^\infty Y_i(u)Y_{ik}(u)du$ is the total follow-up time in state k. The log-likelihood is

$$\begin{aligned} l(\theta) =& |\mathscr{D}_{12}| \log \theta_{12} + |\mathscr{D}_{13}| \log \theta_{13} + |\mathscr{D}_{23}| \log(\lambda\theta_{13}) \\ & - (\theta_{12} + \theta_{13})T_1 - \lambda\theta_{13}T_2. \end{aligned} \tag{4.2.3}$$

Since the hazard rates are positive but their logarithms are unrestricted, the model may be reparameterized by the parameter vector $\beta = (\beta_{12}, \beta_{13}, \beta_\lambda) = (\log\theta_{12}, \log\theta_{13}, \log\lambda)$ (Clayton and Hills, 1993, p. 81). The reparameterized log-likelihood is

$$\begin{aligned} l(\beta) =& |\mathscr{D}_{12}|\beta_{12} + |\mathscr{D}_{13}|\beta_{13} + |\mathscr{D}_{23}|(\beta_\lambda + \beta_{13}) \\ & - (\exp\{\beta_{12}\} + \exp\{\beta_{13}\})T_1 - \exp\{\beta_\lambda + \beta_{13}\} T_2, \end{aligned} \tag{4.2.4}$$

and the corresponding score function is

$$\nabla l(\beta) = \left(\frac{\partial}{\partial \beta_{12}} l(\beta), \frac{\partial}{\partial \beta_{13}} l(\beta), \frac{\partial}{\partial \beta_{\lambda}} l(\beta) \right)^T, \tag{4.2.5}$$

where

$$\begin{aligned} \frac{\partial}{\partial \beta_{12}} l(\beta) &= |\mathscr{D}_{12}| - \exp\{\beta_{12}\} T_1, \\ \frac{\partial}{\partial \beta_{13}} l(\beta) &= |\mathscr{D}_{13}| + |\mathscr{D}_{23}| - \exp\{\beta_{13}\} T_1 - \exp\{\beta_{13} + \beta_{\lambda}\} T_2, \text{ and} \\ \frac{\partial}{\partial \beta_{\lambda}} l(\beta) &= |\mathscr{D}_{23}| - \exp\{\beta_{13} + \beta_{\lambda}\} T_2. \end{aligned} \tag{4.2.6}$$

By solving the equation $\nabla l(\beta) = (0,0,0)^T$, we find the maximum likelihood estimate (MLE)

$$\hat{\beta} = (\hat{\beta}_{12}, \hat{\beta}_{13}, \hat{\beta}_{\lambda}) = \left(\log \frac{|\mathscr{D}_{12}|}{T_1}, \log \frac{|\mathscr{D}_{13}|}{T_1}, \log \left(\frac{|\mathscr{D}_{23}|}{T_2} \bigg/ \frac{|\mathscr{D}_{13}|}{T_1} \right) \right). \tag{4.2.7}$$

The sampling covariance matrix of $\hat{\beta}$ can be estimated from the observed information matrix $\mathcal{J}(\beta) = -\nabla \nabla^T l(\beta)$, which is

$$\begin{aligned} \mathcal{J}(\beta) &= \begin{bmatrix} -\frac{\partial^2}{\partial \beta_{12}^2} l(\beta) & -\frac{\partial^2}{\partial \beta_{12} \beta_{13}} l(\beta) & -\frac{\partial^2}{\partial \beta_{12} \beta_{\lambda}} l(\beta) \\ -\frac{\partial^2}{\partial \beta_{13} \beta_{12}} l(\beta) & -\frac{\partial^2}{\partial \beta_{13}^2} l(\beta) & -\frac{\partial^2}{\partial \beta_{13} \beta_{\lambda}} l(\beta) \\ -\frac{\partial^2}{\partial \beta_{\lambda} \beta_{12}} l(\beta) & -\frac{\partial^2}{\partial \beta_{\lambda} \beta_{13}} l(\beta) & -\frac{\partial^2}{\partial \beta_{\lambda}^2} l(\beta) \end{bmatrix} \\ &= \begin{bmatrix} \exp\{\beta_{12}\} T_1 & 0 & 0 \\ 0 & \exp\{\beta_{13}\} T_1 + \exp\{\beta_{13} + \beta_{\lambda}\} T_2 & \exp\{\beta_{13} + \beta_{\lambda}\} T_2 \\ 0 & \exp\{\beta_{13} + \beta_{\lambda}\} T_2 & \exp\{\beta_{13} + \beta_{\lambda}\} T_2 \end{bmatrix}. \end{aligned} \tag{4.2.8}$$

The inverse of the observed information matrix is

$$\mathcal{J}^{-1}(\beta) = \begin{bmatrix} \frac{1}{\exp\{\beta_{12}\} T_1} & 0 & 0 \\ 0 & \frac{1}{\exp\{\beta_{13}\} T_1} & -\frac{1}{\exp\{\beta_{13}\} T_1} \\ 0 & -\frac{1}{\exp\{\beta_{13}\} T_1} & \frac{T_1 + \exp\{\beta_\lambda\} T_2}{\exp\{\beta_{13} + \beta_\lambda\} T_1 T_2} \end{bmatrix}, \qquad (4.2.9)$$

and therefore an estimate of the covariance matrix of $\hat{\beta}$ is given by

$$\mathcal{J}^{-1}(\hat{\beta}) = \begin{bmatrix} \frac{1}{|\mathscr{D}_{12}|} & 0 & 0 \\ 0 & \frac{1}{|\mathscr{D}_{13}|} & -\frac{1}{|\mathscr{D}_{13}|} \\ 0 & -\frac{1}{|\mathscr{D}_{13}|} & \frac{1}{|\mathscr{D}_{13}|} + \frac{1}{|\mathscr{D}_{23}|} \end{bmatrix}. \qquad (4.2.10)$$

Let $q_{1-\alpha}$ denote the $(1-\alpha)$ quantile of the standard normal distribution. Using the $(1-\alpha)$% Wald confidence interval (CI) for the log HR β_λ, we find that a $(1-\alpha)$% CI for the HR λ is

$$\exp\left\{\hat{\beta}_\lambda \pm q_{1-\alpha}\sqrt{\mathcal{J}^{-1}_{3,3}(\hat{\beta})}\right\} = \left(\frac{|\mathscr{D}_{23}|}{T_2} \Big/ \frac{|\mathscr{D}_{13}|}{T_1}\right) \exp\left\{\pm q_{1-\alpha}\sqrt{\frac{1}{|\mathscr{D}_{13}|} + \frac{1}{|\mathscr{D}_{23}|}}\right\}, \qquad (4.2.11)$$

where $\mathcal{J}^{-1}_{3,3}(\hat{\beta})$ is the third diagonal element of $\mathcal{J}^{-1}(\hat{\beta})$. CIs for the hazard rates θ_{12} and θ_{13} can be obtained similarly.

4.2.2 A non-Markov model

It may be useful to model transition rates on multiple time scales, such as age, calendar period, and duration of illness. Here we extend the model of Section 4.2.1 to multiple time scales. Let t denote the attained age of the individual, B_i denote the calendar time of birth, and $P_i(t) = B_i + t$ denote the attained calendar period. Duration of illness is generally defined only after the onset of illness, but for likelihood derivation, it can be defined as $D_i(t) = t - t_{i12}$ when $Y_{i2}(t) = 1$ and $D_i(t) = -\infty$ otherwise.

Categorize age into p intervals $\mathscr{A} = \{A_1, ..., A_p\}$ that partition the interval $[0, \infty)$, calendar time into q intervals $\mathscr{P} = \{P_1, ..., P_q\}$ that partition the study period, and duration of illness into r categories $\mathscr{S} = \{S_1, ..., S_r\}$, where $S_1 = \{-\infty\}$, and $S_2, ..., S_r$ partition the interval $[0, \infty)$. The transition hazard functions are defined as

$$
\begin{aligned}
\alpha_{i12}(t|\theta,\mathscr{F}^c_{i,t-}) &= \sum_{(A,P)\in\mathscr{A}\times\mathscr{P}} \theta^{(A)}_{12\mathscr{A}}\theta^{(P)}_{12\mathscr{P}} I(t\in A, P_i(t)\in P),\\
\alpha_{i13}(t|\theta,\mathscr{F}^c_{i,t-}) &= \sum_{(A,P)\in\mathscr{A}\times\mathscr{P}} \theta^{(A)}_{13\mathscr{A}}\theta^{(P)}_{13\mathscr{P}} I(t\in A, P_i(t)\in P), \text{ and}\\
\alpha_{i23}(t|\theta,\mathscr{F}^c_{i,t-}) &= \sum_{(A,P,S)\in\mathscr{A}\times\mathscr{P}\times\mathscr{S}} \theta^{(A)}_{13\mathscr{A}}\theta^{(P)}_{13\mathscr{P}}\theta^{(S)}_{\mathscr{S}} I(t\in A, P_i(t)\in P, D_i(t)\in S),
\end{aligned}
\tag{4.2.12}
$$

where $\theta^{(A)}_{12\mathscr{A}}$ and $\theta^{(A)}_{13\mathscr{A}}$ are age-specific baseline hazards for age group A, $\theta^{(P)}_{12\mathscr{P}}$ and $\theta^{(P)}_{13\mathscr{P}}$ are HRs for calendar period P with respect to period P_1, and $\theta^{(S)}_{\mathscr{S}}$ is the HR for duration of illness S with respect to duration of illness S_1. To define P_1 and S_1 as the reference levels, let $\theta^{(P_1)}_{12\mathscr{P}} = \theta^{(P_1)}_{13\mathscr{P}} = \theta^{(S_1)}_{\mathscr{S}} = 1$. The full parameter vector is $\theta = (\theta^T_{12\mathscr{A}}, \theta^T_{12\mathscr{P}}, \theta^T_{13\mathscr{A}}, \theta^T_{13\mathscr{P}}, \theta^T_{\mathscr{S}})^T$, where $\theta_{12\mathscr{A}} = (\theta^{(A_1)}_{12\mathscr{A}}, ..., \theta^{(A_p)}_{12\mathscr{A}})^T$, $\theta_{12\mathscr{P}} = (\theta^{(P_1)}_{12\mathscr{P}}, ..., \theta^{(P_q)}_{12\mathscr{P}})^T$, $\theta_{13\mathscr{A}} = (\theta^{(A_1)}_{13\mathscr{A}}, ..., \theta^{(A_p)}_{13\mathscr{A}})^T$, $\theta_{13\mathscr{P}} = (\theta^{(P_1)}_{13\mathscr{P}}, ..., \theta^{(P_q)}_{13\mathscr{P}})^T$, and $\theta_{\mathscr{S}} = (\theta^{(S_1)}_{\mathscr{S}}, ..., \theta^{(S_r)}_{\mathscr{S}})^T$. This framework corresponds to nonparametric regression in the Lexis diagram (Keiding, 1990).

The likelihood is $L(\theta) = L_{12}(\theta)L_{13}(\theta)L_{23}(\theta)$, where the first term is

$$
\begin{aligned}
L_{12}(\theta) = &\prod_{i=1}^{n}\prod_{t_{ij}\in\mathscr{D}_{i12}} \left\{ \sum_{(A,P)\in\mathscr{A}\times\mathscr{P}} \theta^{(A)}_{12\mathscr{A}}\theta^{(P)}_{12\mathscr{P}} I(t_{ij}\in A, P_i(t_{ij})\in P) \right\}\\
&\times \exp\left\{ -\int_0^\infty Y_i(u)Y_{i1}(u-) \left(\sum_{(A,P)\in\mathscr{A}\times\mathscr{P}} \theta^{(A)}_{12\mathscr{A}}\theta^{(P)}_{12\mathscr{P}} I(u\in A, P_i(u)\in P) \right) du \right\}.
\end{aligned}
\tag{4.2.13}
$$

To simplify $L_{12}(\theta)$, define the total number of observed $k \to l$ transitions for age group A, calendar period P, and duration of illness S as

$$
\left|\mathscr{D}^{(A,P,S)}_{kl}\right| = \sum_{i=1}^{n}\sum_{t_{ij}\in\mathscr{D}_{ikl}} I(t_{ij}\in A, P_i(t_{ij})\in P, D_i(t_{ij})\in S) \tag{4.2.14}
$$

and the total follow-up time in state k by age, period, and duration of illness as

$$
T^{(A,P,S)}_{k} = \sum_{i=1}^{n}\int_0^\infty Y_i(u)Y_{ik}(u-)I(u\in A, P_i(u)\in P, D_i(u)\in S)du. \tag{4.2.15}
$$

The sums of $\left|\mathscr{D}^{(A,P,S)}_{kl}\right|$ and $T^{(A,P,S)}_{k}$ over $S\in\mathscr{S}$ are denoted by $\left|\mathscr{D}^{(A,P)}_{kl}\right|$ and $T^{(A,P)}_{k}$. Then

$$
L_{12}(\theta) = \prod_{(A,P)\in\mathscr{A}\times\mathscr{P}} \left(\theta^{(A)}_{12\mathscr{A}}\theta^{(P)}_{12\mathscr{P}}\right)^{\left|\mathscr{D}^{(A,P)}_{12}\right|} \exp\left\{ -\sum_{(A,P)\in\mathscr{A}\times\mathscr{P}} \theta^{(A)}_{12\mathscr{A}}\theta^{(P)}_{12\mathscr{P}} T^{(A,P)}_{1} \right\}, \tag{4.2.16}
$$

which is proportional to a Poisson likelihood. Correspondingly, the second term is

$$L_{13}(\theta) = \prod_{(A,P)\in\mathscr{A}\times\mathscr{P}} \left(\theta_{13\mathscr{A}}^{(A)}\theta_{13\mathscr{P}}^{(P)}\right)^{\left|\mathscr{D}_{13}^{(A,P)}\right|} \exp\left\{-\sum_{(A,P)\in\mathscr{A}\times\mathscr{P}} \theta_{13\mathscr{A}}^{(A)}\theta_{13\mathscr{P}}^{(P)}T_1^{(A,P)}\right\}, \qquad (4.2.17)$$

and the third term is

$$\begin{aligned} L_{23}(\theta) = &\prod_{(A,P,S)\in\mathscr{A}\times\mathscr{P}\times\mathscr{S}} \left(\theta_{13\mathscr{A}}^{(A)}\theta_{13\mathscr{P}}^{(P)}\theta_{\mathscr{S}}^{(S)}\right)^{\left|\mathscr{D}_{23}^{(A,P,S)}\right|} \\ &\times \exp\left\{-\sum_{(A,P,S)\in\mathscr{A}\times\mathscr{P}\times\mathscr{S}} \theta_{13\mathscr{A}}^{(A)}\theta_{13\mathscr{P}}^{(P)}\theta_{\mathscr{S}}^{(S)}T_2^{(A,P,S)}\right\}. \end{aligned} \qquad (4.2.18)$$

The likelihood can now be written as

$$L(\theta) = L_{12}(\theta_{12\mathscr{A}}, \theta_{12\mathscr{P}})L_{13}(\theta_{13\mathscr{A}}, \theta_{13\mathscr{P}})L_{23}(\theta_{13\mathscr{A}}, \theta_{13\mathscr{P}}, \theta_{\mathscr{S}}), \qquad (4.2.19)$$

which shows that the parameter vectors $(\theta_{12\mathscr{A}}^T, \theta_{12\mathscr{P}}^T)^T$ and $(\theta_{13\mathscr{A}}^T, \theta_{13\mathscr{P}}^T, \theta_{\mathscr{S}}^T)^T$ are orthogonal. Therefore, $(\theta_{12\mathscr{A}}^T, \theta_{12\mathscr{P}}^T)^T$ can be estimated using the Poisson regression model

$$\log\left(\mu_{A,P}\right) = \log T_1^{(A,P)} + \beta_{12\mathscr{A}}^{(A)} + \beta_{12\mathscr{P}}^{(P)}, \qquad (4.2.20)$$

where $\mu_{A,P}$ is the mean number of $1 \to 2$ transitions at age A and calendar period P, $\log T_1^{(A,P)}$ is an offset, and $\beta_{12\mathscr{A}}^{(A)}$ and $\beta_{12\mathscr{P}}^{(P)}$ are the logarithms of $\theta_{12\mathscr{A}}^{(A)}$ and $\theta_{12\mathscr{P}}^{(P)}$, respectively. A Poisson regression model may also be used to estimate $(\theta_{13\mathscr{A}}^T, \theta_{13\mathscr{P}}^T, \theta_{\mathscr{S}}^T)^T$. Since $\left|\mathscr{D}_{13}^{(A,P,S)}\right|$ and $T_1^{(A,P,S)}$ are both equal to 0 for $S \in \{S_2, ..., S_r\}$, and $\theta_{\mathscr{S}}^{(S_1)} = 1$, we can rewrite $L_{13}(\theta)$ as

$$\begin{aligned} L_{13}(\theta) = &\prod_{(A,P,S)\in\mathscr{A}\times\mathscr{P}\times\mathscr{S}} \left(\theta_{13\mathscr{A}}^{(A)}\theta_{13\mathscr{P}}^{(P)}\theta_{\mathscr{S}}^{(S)}\right)^{\left|\mathscr{D}_{13}^{(A,P,S)}\right|} \\ &\times \exp\left\{-\sum_{(A,P,S)\in\mathscr{A}\times\mathscr{P}\times\mathscr{S}} \theta_{13\mathscr{A}}^{(A)}\theta_{13\mathscr{P}}^{(P)}\theta_{\mathscr{S}}^{(S)}T_1^{(A,P,S)}\right\}. \end{aligned} \qquad (4.2.21)$$

Now $L_{13}(\theta)$ and $L_{23}(\theta)$ have a similar form, and

$$\begin{aligned} L_{13}(\theta)L_{23}(\theta) = & \prod_{(A,P,S)\in\mathscr{A}\times\mathscr{P}\times\mathscr{S}} \left(\theta_{13\mathscr{A}}^{(A)}\theta_{13\mathscr{P}}^{(P)}\theta_{\mathscr{S}}^{(S)}\right)^{\left|\mathscr{D}_{13}^{(A,P,S)}\right|+\left|\mathscr{D}_{23}^{(A,P,S)}\right|} \\ & \times \exp\left\{-\sum_{(A,P,S)\in\mathscr{A}\times\mathscr{P}\times\mathscr{S}} \theta_{13\mathscr{A}}^{(A)}\theta_{13\mathscr{P}}^{(P)}\theta_{\mathscr{S}}^{(S)}\left(T_1^{(A,P,S)}+T_2^{(A,P,S)}\right)\right\}. \end{aligned} \tag{4.2.22}$$

An equivalent Poisson model is

$$\log\left(\mu_{A,P,S}^{*}\right) = \log\left(T_1^{(A,P,S)}+T_2^{(A,P,S)}\right)+\beta_{13\mathscr{A}}^{(A)}+\beta_{13\mathscr{P}}^{(P)}+\beta_{\mathscr{S}}^{(S)}, \tag{4.2.23}$$

where $\mu_{A,P,S}^{*}$ is the mean number of deaths at age A, calendar period P, and duration of illness S, and the parameters $\beta_{13\mathscr{A}}^{(A)}$, $\beta_{13\mathscr{P}}^{(P)}$, and $\beta_{\mathscr{S}}^{(S)}$ are the logarithms of $\theta_{13\mathscr{A}}^{(A)}$, $\theta_{13\mathscr{P}}^{(P)}$, and $\theta_{\mathscr{S}}^{(S)}$, respectively.

5. Application

5.1 Data sources

The population sample comprised 2,549,992 individuals who were randomly selected from the Population Information System of Finland on January 1, 2000, covering nearly one half of the total population of Finland. The Population Information System, maintained by the Digital and Population Data Services Agency, provided data on date of birth, gender, first date of emigration after January 1, 2000, and date of death. Personal identity codes were used to link the data to the Finnish Care Register for Health Care and the Finnish Cancer Registry.

The Finnish Care Register for Health Care provided nationwide data on inpatient hospital care in 1970–1997 and both inpatient and outpatient hospital care in 1998–2017. UC diagnoses were identified by the ICD-8 codes 563.10 and 569.02, the ICD-9 code 556, and the ICD-10 code K51. For a systematic review of studies on the quality of the Finnish Care Register for Health Care, see Sund (2012).

The Finnish Cancer Registry provided data on CRC diagnoses in Finland 1953–2017. The definition of CRC was based on the International Classification of Diseases for Oncology topography codes C18, C19, and C20 (World Health Organization, 2013). Hematolymphoid neoplasms of the colorectum were excluded. The Finnish Cancer Registry collects data on all cancer diagnoses and deaths among cancer patients in Finland since 1953 and has nearly complete coverage of solid tumors (Leinonen et al., 2017). Based on special legislation, health care providers, hospitals, and laboratories are required to report all newly diagnosed cancers without permission of the patient.

The required permissions for the study were obtained from the Finnish Institute for Health and Welfare (THL/118/6.02.00/2019) and from the Digital and Population Data Services Agency (VRK/4504/2019-2).

5.2 Statistical analysis

Participants were followed from January 1, 2000, to the date of first CRC diagnosis, death from other cause, emigration, or December 31, 2017, whichever came first. Patients diagnosed with CRC before January 1, 2000, were excluded from the study cohort, whereas patients diagnosed with UC before January 1, 2000, were included. CIs for binomial proportions were computed using the Clopper-Pearson method (Clopper and Pearson, 1934).

The illness-death model (Chapter 4) was applied to study the incidence rates of UC and CRC in Finland in 2000–2017. People with neither UC nor CRC were defined as being in state 1 (“Healthy”), those diagnosed with UC but not CRC in state 2 (“Ill”), and those diagnosed with CRC in state 3 (“Dead”). A state diagram is shown in Section 4.1.

To describe the follow-up data, we used the illness-death model with constant transition hazards (Section 4.2.1). MLEs and 95% CIs were computed using the analytical solutions.

To implement the piecewise constant hazard model of Section 4.2.2, age was categorized into 5-year intervals (0–4, 5–9, ..., 90–94, and ≥ 95), calendar time into 6-year intervals (2000–2005, 2006–2011, and 2012–2017), and time from UC diagnosis into <1, 1–3, 4–9, 10–19, or ≥ 20 years. Transitions $1 \rightarrow 2$ and $1 \rightarrow 3$ (“Healthy” → “UC” and “Healthy” → “CRC”) were modeled using attained age and calendar time as the time scales. For transition $2 \rightarrow 3$ (“UC” → “CRC”), UC duration was used as an additional time scale. Separate models were fit for females and males. To estimate HRs for CRC by age at UC diagnosis (<40 or ≥ 40 years), we combined both genders into a single data set and extended the model of Section 4.2.2 by adjusting for gender and letting the HRs for CRC depend on both UC duration and age at UC diagnosis. MLEs and 95% CIs were computed using the Poisson likelihood and the log link function as explained in Section 4.2.2 (McCullagh and Nelder, 1998).

No adjustments were made for multiple comparisons. Statistical analyses were performed using R version 4.0.2 (R Foundation for Statistical Computing, Vienna, Austria) with the Epi package version 2.46.

5.3 Results

5.3.1 Baseline characteristics

Of the 2,549,992 individuals included in the population sample on January 1, 2000, 1,306,261 (51%) were female and 1,243,731 (49%) were male. There were 8,720 individuals with prevalent UC, 6,392 individuals with prevalent CRC, and 90 individuals with both conditions before the sampling date. The baseline prevalence of UC was 321 per 100,000 in females (95% CI: 312, 331) and 364 per 100,000 in males (95% CI: 353, 374), whereas the baseline prevalence of CRC was 277 per 100,000 in females (95% CI: 268, 287) and 223 per 100,000 in males (95% CI: 214, 231). After excluding patients with prevalent CRC, 2,543,600 individuals remained in the study cohort, including 8,630 patients with prevalent UC. Characteristics of the study cohort are shown in Table 1. The median age at the start of follow-up was 41 years in females (range: 0, 108) and 38 years in males (range: 0, 105).

Table 1: Baseline characteristics of the study cohort on January 1, 2000.

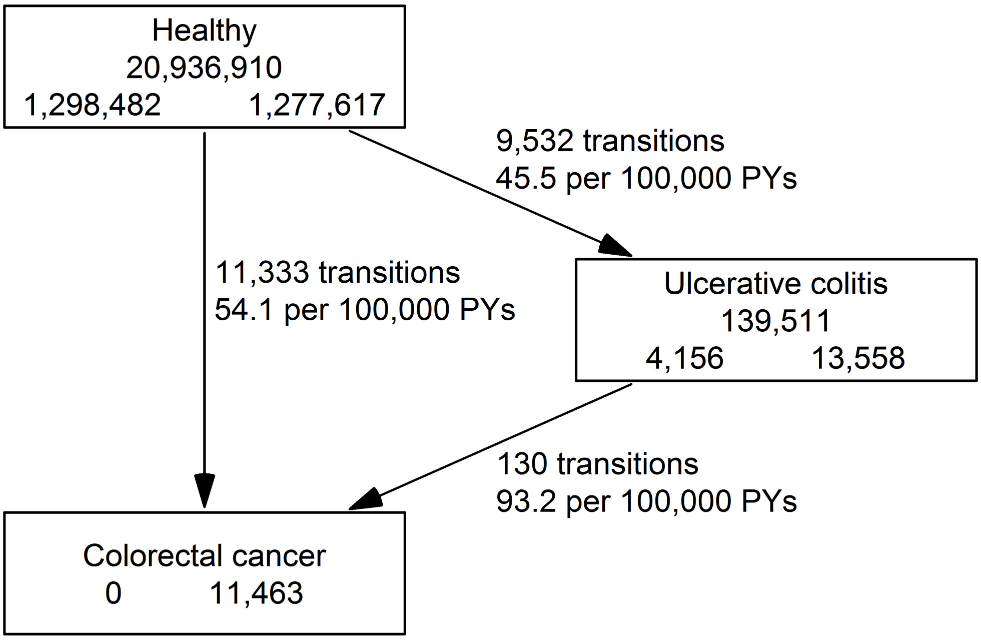
Characteristic	Female (N = 1,302,638)[a]	Male (N = 1,240,962)[a]
Age, years	41 (21, 58)	38 (19, 53)
Ulcerative colitis	4,156 (0.3%)	4,474 (0.4%)

[a] Median (interquartile range); n (%)

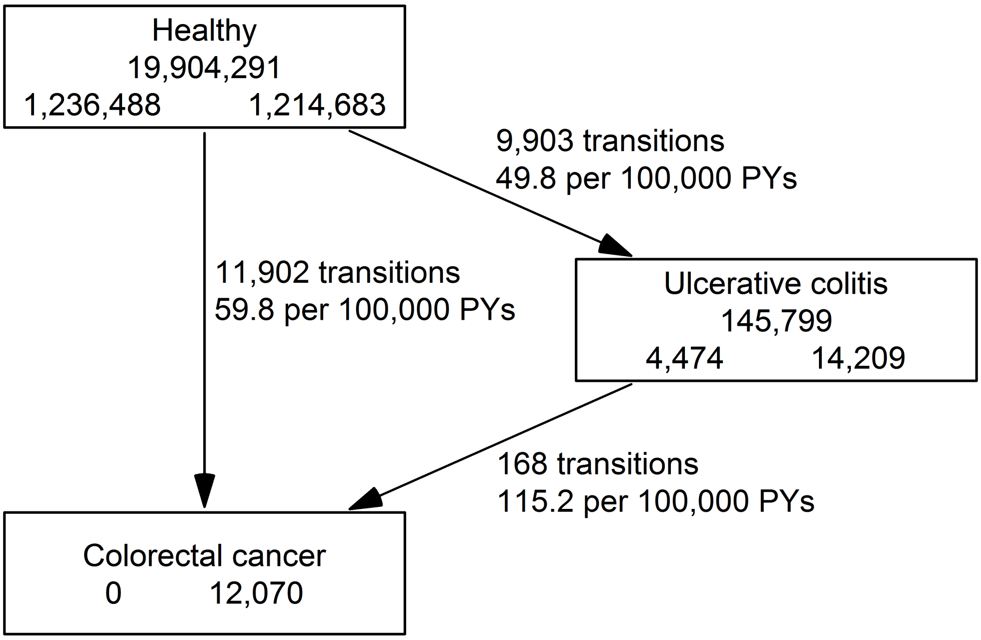
5.3.2 Follow-up data

State diagrams for females and males are shown in Figures 2 (a) and 2 (b). A total of 23,533 incident CRCs were diagnosed during 41 million person-years of follow-up. In addition to the 8,630 patients with prevalent UC, there were 19,435 cases of incident UC. Of the 23,533 incident CRCs, 298 (1.3%) were diagnosed in patients with pre-existing (prevalent or incident) UC.

Crude CRC incidence was 54.1 (95% CI: 53.1, 55.1) per 100,000 person-years in females and 59.8 (95% CI: 58.7, 60.9) per 100,000 person-years in males. In patients with UC, CRC incidence was increased by a crude HR of 1.72 (95% CI: 1.45, 2.05) in females and 1.93 (95% CI: 1.65, 2.24) in males. Crude UC incidence was 45.5 (95% CI: 44.6, 46.5) per 100,000 person-years in females and 49.8 (95% CI: 48.8, 50.7) per 100,000 person-years in males.



(a) Female



(b) Male

Figure 2: State diagrams. The numbers in the boxes indicate person-years of follow-up (middle), persons at the start of follow-up (lower left), and persons at the end of follow-up (lower right). PYs, person-years.

5.3.3 Multistate analysis

Results of the multistate analysis, using the model of Section 4.2.2, are shown in Figure 2 and Tables 2 and 3. CRC incidence increased monotonically with age up to a peak age of 85–89 in men and 85–94 years in women. Before age 40 years, CRC incidence was low in both genders. In people aged 55 years and older, CRC incidence was higher in men than in women, but in the younger age groups, there were no marked gender differences. There was an increasing trend of CRC incidence over calendar time in both genders. The HR for 2012–2017 compared to 2000–2005 was 1.08 (95% CI: 1.03, 1.13; P<0.001) in females and 1.08 (95% CI: 1.04, 1.13; P<0.001) in males.

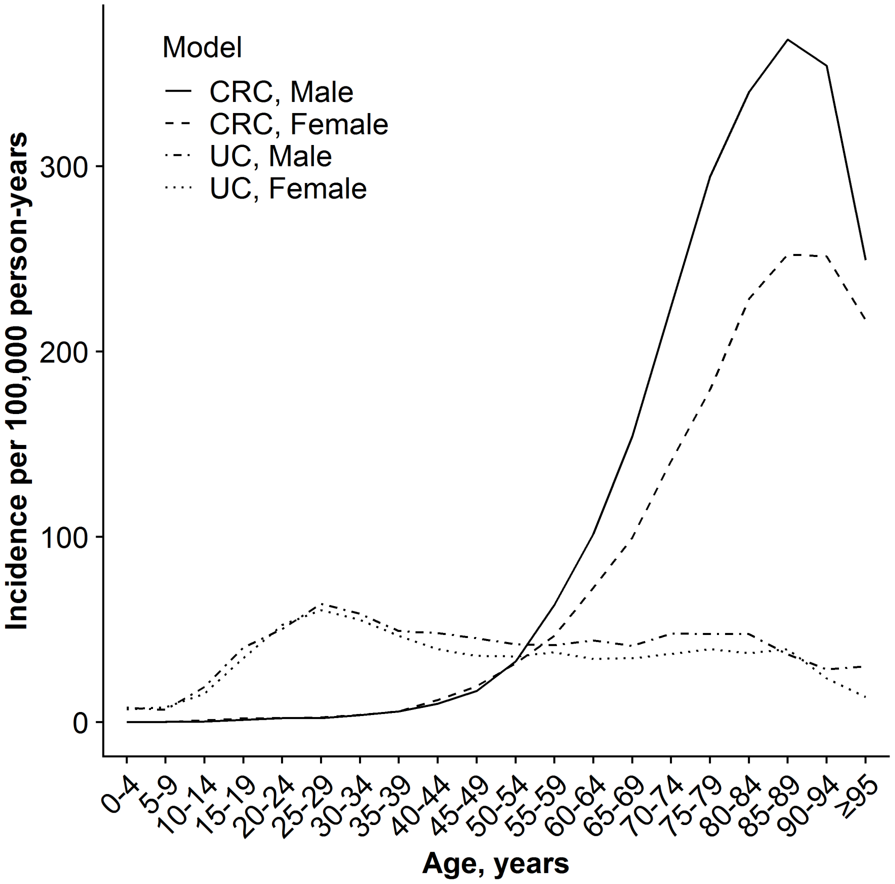


Figure 3: Age-specific baseline hazards (incidence per 100,000 person-years) for colorectal cancer and ulcerative colitis. The estimates apply to individuals without ulcerative colitis or colorectal cancer in Finland in 2000–2005. CRC, colorectal cancer; UC, ulcerative colitis.

Table 2: Incidence of ulcerative colitis in a population-based cohort in Finland in 2000–2017. Adjusted hazard ratios for calendar period are shown together with age-specific baseline hazard rates per 100,000 person-years. The symbols $\theta_{12\mathscr{P}}$ and $\theta_{12\mathscr{A}}$ refer to parameter vectors defined in Section 4.2.2.

	Female			Male		
Characteristic	exp(Est.)[a]	95% CI[b]	P-value	exp(Est.)[a]	95% CI[b]	P-value
Calendar period ($\theta_{12\mathscr{P}}$)[c]						
2000–2005	1.00	–	–	1.00	–	–
2006–2011	1.26	1.20, 1.32	<0.001	1.17	1.12, 1.23	<0.001
2012–2017	1.27	1.21, 1.33	<0.001	1.18	1.13, 1.24	<0.001
Age, years ($\theta_{12\mathscr{A}}$)[d]						
0–4	6.89	3.91, 12.1	–	7.73	4.58, 13.0	–
5–9	7.94	5.95, 10.6	–	6.74	4.95, 9.19	–
10–14	15.4	13.2, 18.0	–	19.1	16.6, 21.9	–
15–19	34.7	31.7, 38.0	–	40.3	37.0, 43.9	–
20–24	52.4	48.7, 56.5	–	50.3	46.6, 54.2	–
25–29	60.4	56.2, 64.9	–	63.8	59.6, 68.4	–
30–34	55.2	51.3, 59.4	–	58.4	54.4, 62.7	–
35–39	46.6	43.1, 50.4	–	49.2	45.6, 53.0	–
40–44	39.5	36.4, 42.8	–	48.1	44.6, 51.8	–
45–49	35.7	32.9, 38.8	–	45.5	42.2, 49.0	–
50–54	35.3	32.6, 38.3	–	41.9	38.8, 45.2	–
55–59	37.7	34.8, 40.8	–	41.6	38.5, 45.0	–
60–64	34.1	31.3, 37.2	–	44.2	40.8, 47.9	–
65–69	34.6	31.6, 37.9	–	41.1	37.6, 45.0	–
70–74	36.8	33.4, 40.5	–	47.8	43.4, 52.6	–
75–79	39.4	35.7, 43.5	–	47.5	42.5, 53.1	–
80–84	37.3	33.2, 41.8	–	47.6	41.4, 54.8	–
85–89	39.3	34.2, 45.2	–	36.6	29.1, 46.1	–
90–94	23.7	18.0, 31.2	–	28.6	17.7, 46.0	–
$\geq$ 95	13.7	6.51, 28.7	–	30.1	9.71, 93.4	–

[a] exp(Est.), exponentiated parameter estimate

[b] CI, confidence interval

[c] Adjusted hazard ratio

[d] Baseline hazard rate per 100,000 person-years

Table 3: Incidence of colorectal cancer in a population-based cohort in Finland in 2000–2017. Adjusted hazard ratios for ulcerative colitis and calendar period are shown together with age-specific baseline hazards per 100,000 person-years. The symbols $\theta_{\mathscr{S}}$, $\theta_{13\mathscr{P}}$, and $\theta_{13\mathscr{A}}$ refer to parameter vectors defined in Section 4.2.2.

	Female			Male		
Characteristic	exp(Est.)^a	95% CI^b	P-value	exp(Est.)^a	95% CI^b	P-value
Ulcerative colitis ($\theta_{\mathscr{S}}$)^c						
Absent	1.00	–	–	1.00	–	–
<1 year	4.67	3.07, 7.09	<0.001	7.62	5.65, 10.3	<0.001
1–3 years	0.85	0.49, 1.51	0.6	0.66	0.36, 1.19	0.2
4–9 years	1.14	0.78, 1.66	0.5	0.77	0.51, 1.17	0.2
10–19 years	1.63	1.19, 2.24	0.002	1.29	0.96, 1.75	0.092
$\geq$ 20 years	1.61	1.13, 2.31	0.009	1.74	1.31, 2.31	<0.001
Calendar period ($\theta_{13\mathscr{P}}$)^c						
2000–2005	1.00	–	–	1.00	–	—
2006–2011	1.02	0.97, 1.07	0.4	1.02	0.98, 1.07	0.3
2012–2017	1.08	1.03, 1.13	<0.001	1.08	1.04, 1.13	<0.001
Age, years ($\theta_{13\mathscr{A}}$)^d						
0–4	0.00	na^e	–	0.00	na^e	–
5–9	0.00	na^e	–	0.17	0.02, 1.23	–
10–14	1.05	0.57, 1.96	–	0.30	0.10, 0.94	–
15–19	2.03	1.39, 2.96	–	1.36	0.87, 2.14	–
20–24	2.11	1.47, 3.02	–	2.27	1.62, 3.18	–
25–29	2.52	1.81, 3.52	–	2.18	1.54, 3.08	–
30–34	4.03	3.10, 5.25	–	3.75	2.88, 4.90	–
35–39	5.79	4.67, 7.18	–	5.74	4.64, 7.09	–
40–44	11.9	10.3, 13.8	–	10.0	8.57, 11.7	–
45–49	19.4	17.4, 21.7	–	16.9	15.0, 19.0	–
50–54	31.8	29.1, 34.7	–	32.8	30.1, 35.8	–
55–59	46.6	43.3, 50.3	–	63.4	59.3, 67.7	–
60–64	72.7	68.1, 77.6	–	102	96.0, 108	–
65–69	99.7	93.8, 106	–	155	146, 163	–
70–74	141	133, 149	–	224	213, 236	–
75–79	179	170, 189	–	294	279, 310	–
80–84	229	216, 242	–	340	320, 362	–
85–89	253	237, 269	–	368	340, 399	–
90–94	251	229, 276	–	354	307, 409	–
$\geq$ 95	217	177, 265	–	249	166, 376	–

[a] exp(Est.), exponentiated parameter estimate

[b] CI, confidence interval

[c] Adjusted hazard ratio

[d] Baseline hazard rate per 100,000 person-years

[e] No incident CRCs were observed in this age group

In the first year after UC diagnosis, the HR for incident CRC was 4.67 (95% CI: 3.07, 7.09; P<0.001) in females and 7.62 (95% CI: 5.65, 10.3; P<0.001) in males. When 1–3 or 4–9 years had passed from UC diagnosis, CRC incidence did not differ from persons without UC. In patients with UC diagnosed 10–19 years earlier, the HR for incident CRC was 1.63 (95% CI: 1.19, 2.24; P=0.002) in females and 1.29 (95% CI: 0.96, 1.75; P=0.092) in males, and after 20 years from UC diagnosis, the HR was 1.61 (95% CI: 1.13, 2.31; P=0.009) in females and 1.74 (95% CI: 1.31, 2.31; P<0.001) in males.

UC incidence was highest at ages 25–29 years and lowest in the first decade of life. In most age groups, UC incidence was slightly higher in males than in females. There was also an increasing trend of UC incidence over calendar time in both genders. The HR for 2012–2017 compared to 2000–2005 was 1.27 (95% CI: 1.21, 1.33; P<0.001) in females and 1.18 (95% CI: 1.13, 1.24; P<0.001) in males.

Early-onset UC (defined here as age <40 years) was associated with an increased long-term risk of CRC (Table 4). In late-onset UC, an increased risk of CRC was observed only in the first year after UC diagnosis. In the first year after UC diagnosis, the HR for CRC was 28.6 in early-onset UC (95% CI: 14.2, 57.6; P<0.001) and 5.66 in late-onset UC (95% CI: 4.36, 7.34; P<0.001). After the first year after UC diagnosis, the HR for CRC in early-onset UC increased over time up to 10–19 years (HR 4.88; 95% CI: 3.46, 6.88; P<0.001) and then declined after 20 years (HR 2.63; 95% CI: 2.01, 3.43; P<0.001).

Table 4: Hazard ratios for colorectal cancer by duration and age at diagnosis of ulcerative colitis. The estimates were adjusted for age (0–4, 5–9, ..., 90–94, $\geq$ 95), calendar period (2000–2005, 2006–2011, 2012–2017), and gender.

	Early-onset UC[a] (<40 years)			Late-onset UC[a] ($\geq$ 40 years)		
Time from UC[a] diagnosis	HR[b]	95% CI[c]	P-value	HR[b]	95% CI[c]	P-value
No UC[a]	1.00	–	–	1.00	–	–
<1 year	28.6	14.2, 57.6	<0.001	5.66	4.36, 7.34	<0.001
1–3 years	0.97	0.14, 6.92	>0.9	0.74	0.49, 1.12	0.2
4–9 years	4.13	2.28, 7.47	<0.001	0.77	0.56, 1.05	0.10
10–19 years	4.88	3.46, 6.88	<0.001	0.98	0.74, 1.29	0.9
$\geq$ 20 years	2.63	2.01, 3.43	<0.001	0.95	0.63, 1.41	0.8

[a] UC, ulcerative colitis

[b] HR, hazard ratio

[c] CI, confidence interval

6. Discussion

In this large population-based cohort study, including more than 28,000 patients with UC and more than 23,000 incident CRCs, we studied the risk of CRC in persons with and without UC, considering both the duration of UC and age at UC diagnosis. Patients with UC overall, and especially those with early-onset UC, had an increased long-term risk of CRC. Multistate modeling of UC and CRC in a single population-based cohort provided a useful framework for studying the relationship between the two disease processes.

Patients with long-standing ($\geq$ 10 years) UC were at increased risk of CRC. Estimated HRs ranged from 1.3 to 1.7 depending on UC duration (10–19 or $\geq$ 20 years) and gender. A recent population-based cohort study in Denmark and Sweden reported HRs of approximately 1.9 for incident CRC in patients with UC diagnosed 10–19 or $\geq$ 20 years earlier (Olén et al., 2020). Meta-analyses of population-based cohort studies have reported pooled standardized incidence ratios (SIRs) of 1.7 and 2.4 for CRC in unselected patients with IBD and UC, respectively (Jess et al., 2012; Lutgens et al., 2013). The risk of CRC in UC appears to have decreased over calendar time, which may be due to improved treatments or surveillance protocols (Castaño-Milla et al., 2014; Olén et al., 2020). The long-term excess risk of CRC in patients with UC has been primarily attributed to chronic inflammation (Beaugerie and Itzkowitz, 2015). Colitis-associated CRC may develop through the dysplasia-carcinoma sequence, which differs from the classical adenoma-carcinoma sequence of colorectal carcinogenesis.

Increased CRC risk in the first year after UC diagnosis may be in part due to detection bias. Patients with suspected or newly diagnosed UC routinely undergo colonoscopy, which may reveal CRC shortly before or after the diagnosis of UC. In persons with 1–3 or 4–9 years from UC diagnosis, the incidence of CRC did not differ from persons without UC.

Early-onset UC (defined here as age <40 years) was associated with an increased long-term risk of CRC. The HRs for incident CRC were 1.0, 4.1, 4.9, and 2.6 in patients with early-onset UC diagnosed 1–3, 4–9, 10–19, or $\geq$ 20 years ago, respectively. The late decline in the HR may reflect selection bias among those who remain under follow-up beyond 20 years from UC diagnosis. Olén et al. (2020) reported HRs of 37, 4.1, 1.4, and 1.0 for incident CRC in patients diagnosed with UC at ages <18, 18–39, 40–59, and

$\geq$ 60 years, respectively, while adjusting for years of follow-up and other risk factors. In late-onset UC (age $\geq$ 40 years), we observed an increased incidence of CRC only in the first year after UC diagnosis. In patients diagnosed with late-onset UC more than 20 years ago, the 95% CI for the HR ranged from 0.6 to 1.4. Overall, the excess risk of CRC in UC may be largely attributable to patients with early-onset UC. Adjustment for disease extent (e.g., using the Montreal classification) and other clinical characteristics might clarify whether age at UC diagnosis is an independent risk factor for CRC in UC. Childhood-onset UC, which often presents with extensive colitis, may be biologically and clinically distinct from UC in young adults.

UC incidence was highest at ages 25–29 years. Some studies suggest a second peak at older age, but in this regard, our study is inconclusive (Bernstein et al., 2006). The decline in CRC incidence after a peak age of 85–94 years may be due to challenges in diagnosing cancer in very old patients, the healthy survivor effect, or age-related biological changes.

Increasing trends of both UC and CRC incidence over calendar time have been observed worldwide, although in some developed countries, screening may have stabilized or reduced CRC incidence (Ungaro et al., 2017; Dekker et al., 2019). We did not assess whether the incidence trends of UC and CRC differed between age groups.

The main strengths of the study are the population-based cohort design and the large sample size. Tertiary referral center studies are likely to overestimate CRC risk in UC because patients with less severe disease are typically underrepresented. In the population-based cohort of 2.5 million individuals, hospital diagnoses were available for 1970–2017. The quality of the data on UC relies on the Finnish Care Register for Health Care, which covers both outpatient and inpatient hospital care since 1998. Data on outpatient care in 1970–1997 are largely missing, and therefore UC cases recorded in 1970–1997 may be more severe than those recorded in 1998–2017. This may have introduced an upward bias into the HR estimates and caused misclassification of time from UC diagnosis in some patients. Diagnoses of CRC are recorded with high quality and nearly complete coverage in the Finnish Cancer Registry (Leinonen et al., 2017).

Other possible sources of bias should also be considered. Although the confounding effects of age, gender, and calendar period were controlled in the analysis, there is a risk of unmeasured confounding by smoking, dietary factors, physical activity, non-steroidal anti-inflammatory drugs, and other possible common risk factors of UC and CRC. In addition, comparisons of persons with and without medical conditions are susceptible to surveillance bias. Patients with UC are likely to use health services more frequently than healthy individuals and are offered regular medical examinations and colonoscopies, which may reveal asymptomatic or otherwise undetected CRCs. Studies of CRC mortality in

UC are at low risk of surveillance bias (Olén et al., 2020). Finally, some patients with UC are treated surgically, which can greatly reduce CRC risk. Because of the register-based data collection, we did not have detailed data on surgical treatments and therefore did not censor patients at the time of colectomy. This may have led to underestimation of CRC risk.

Statistical methods in previous population-based cohort studies of CRC in UC include the Cox model and indirect standardization (Cox, 1972; Armitage and Colton, 2005). The Cox model is often used in cohort designs that include a reference group, whereas indirect standardization compares observed event counts to expected counts derived from a large standard population. Both methods provide measures of CRC risk in UC (i.e., HRs or SIRs), but the incidence of UC is not necessarily modeled. We used the illness-death model to estimate incidence rates by age, gender, calendar period, age at UC diagnosis, and time from UC diagnosis. The estimated rates can also be used to derive transition probabilities for a given individual and time period. The semiparametric Cox model is also applicable to multistate models. In this setting, the cumulative baseline hazards can be estimated using the Breslow estimator (Putter et al., 2007). However, parametric models provide parsimonious summaries of the data and may be more convenient to use for prediction (Reid, 1994).

The piecewise constant hazard model can approximate a variety of parametric models, and extensions to multiple time scales are straightforward. The time scales enter the model in the same way as other covariates. Factorization of the likelihood function into one or more Poisson likelihoods permits estimation using standard software for generalized linear models. A disadvantage is the need to categorize the time variables. However, smooth effects can be estimated using splines or other suitable functions (Carstensen, 2021). Data preparation involves splitting individual follow-up time along one or more time scales, which produces multiple data rows per individual and may be computationally demanding. Royston and Parmar proposed a class of flexible parametric survival models that avoid the need to split the time scale (Royston and Parmar, 2002).

A useful property of multistate models is the possibility to model different transitions with shared parameters (Putter et al., 2007). This approach may provide more precise estimates when the data are sparse. In this study, attained age and calendar period were assumed to have the same effect on CRC risk in persons with and without UC because only a small proportion of all CRCs are diagnosed in patients with pre-existing UC. In this study, the proportion of UC-associated CRCs was 1.3%.

In conclusion, this large-population-based cohort study provides estimates of CRC risk in persons with and without UC in Finland in 2000–2017, considering both the duration of UC and age at UC diagnosis. Patients with early-onset UC are at increased risk of CRC, but the risk is likely to depend on disease duration, extent of disease, attained age, and other risk factors. Increased CRC risk in the first year after UC diagnosis may be in part due to detection bias, whereas chronic inflammation may underlie the long-term excess risk of CRC in patients with UC.

7. Acknowledgments

I am sincerely grateful to Professor Janne Pitkäniemi for supervising this thesis and giving me the opportunity to work in his excellent team at the Finnish Cancer Registry. Thank you for your guidance, mentorship, and continued support.

I would like to thank Professor Sangita Kulathinal for her supervision and insightful comments and suggestions that helped improve the thesis considerably. I am thankful for the knowledge you shared with me.

Many thanks to all my colleagues at the Finnish Cancer Registry for creating a positive and supportive work environment.

Lastly, I would like to thank my family for their wonderful support and encouragement.

Bibliography

Aalen, O. O., Borgan, Ø. and Gjessing, S. (2008), *Survival and event history analysis: a process point of view*, Statistics for biology and health, Springer, New York, NY. OCLC: ocn213855657.

Andersen, P. K., ed. (1993), *Statistical models based on counting processes*, Springer series in statistics, Springer-Verlag, New York, NY.

Andersen, P. K. and Keiding, N. (2002), 'Multi-state models for event history analysis', *Statistical Methods in Medical Research* 11(2), 91–115. URL: `http://journals.sagepub.com/doi/10.1191/0962280202SM276ra`

Annese, V., Daperno, M., Rutter, M. D., Amiot, A., Bossuyt, P., East, J., Ferrante, M., Götz, M., Katsanos, K. H., Kießlich, R., Ordás, I., Repici, A., Rosa, B., Sebastian, S., Kucharzik, T., Eliakim, R. and European Crohn's and Colitis Organisation (2013), 'European evidence based consensus for endoscopy in inflammatory bowel disease', *Journal of Crohn's & Colitis* 7(12), 982–1018.

Armitage, P. and Colton, T., eds (2005), *Encyclopedia of biostatistics*, 2nd edn, John Wiley, Chichester, West Sussex, England ; Hoboken, NJ. OCLC: ocm57168526.

Beaugerie, L. and Itzkowitz, S. H. (2015), 'Cancers Complicating Inflammatory Bowel Disease', *New England Journal of Medicine* 372(15), 1441–1452. URL: `http://www.nejm.org/doi/10.1056/NEJMra1403718`

Bernstein, C. N., Wajda, A., Svenson, L. W., MacKenzie, A., Koehoorn, M., Jackson, M., Fedorak, R., Israel, D. and Blanchard, J. F. (2006), 'The epidemiology of inflammatory bowel disease in Canada: a population-based study', *The American Journal of Gastroenterology* 101(7), 1559–1568.

Carstensen, B. (2021), *Epidemiology with R*, Oxford University Press, Oxford. OCLC: on1162988032.

Castaño-Milla, C., Chaparro, M. and Gisbert, J. P. (2014), 'Systematic review with meta-analysis: the declining risk of colorectal cancer in ulcerative colitis', *Alimentary Pharmacology & Therapeutics* 39(7), 645–659.

Clayton, D. and Hills, M. (1993), *Statistical models in epidemiology*, Oxford University Press, Oxford; New York, NY.

Clopper, C. J. and Pearson, E. S. (1934), 'The use of confidence or fiducial limits illustrated in the case of the binomial', *Biometrika* 26(4), 404–413. URL: `https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/26.4.404`

Cook, R. J. and Lawless, J. F. (2007), *The statistical analysis of recurrent events*, Statistics for biology and health, Springer, New York, NY. OCLC: ocn124025386.

Cook, R. J. and Lawless, J. F. (2018), *Multistate models for the analysis of life history data*, CRC Press, Boca Raton, FL.

Cox, D. R. (1972), 'Regression Models and Life-Tables', *Journal of the Royal Statistical Society: Series B (Methodological)* 34(2), 187–202. URL: `http://doi.wiley.com/10.1111/j.2517-6161.1972.tb00899.x`

Dekker, E., Tanis, P. J., Vleugels, J. L. A., Kasi, P. M. and Wallace, M. B. (2019), 'Colorectal cancer', *The Lancet* 394(10207), 1467–1480. URL: `https://linkinghub.elsevier.com/retrieve/pii/S0140673619323190`

Engholm, G., Ferlay, J., Christensen, N., Bray, F., Gjerstorff, M. L., Klint, A., Køtlum, J. E., Olafsdóttir, E., Pukkala, E. and Storm, H. H. (2010), 'NORDCAN–a Nordic tool for cancer information, planning, quality control and research', *Acta Oncologica (Stockholm, Sweden)* 49(5), 725–736.

Fleming, T. R. and Harrington, D. P. (1991), *Counting processes and survival analysis*, Wiley series in probability and mathematical statistics, Wiley, New York, NY.

Jess, T., Rungoe, C. and Peyrin-Biroulet, L. (2012), 'Risk of colorectal cancer in patients with ulcerative colitis: a meta-analysis of population-based cohort studies', *Clinical Gastroenterology and Hepatology: The Official Clinical Practice Journal of the American Gastroenterological Association* 10(6), 639–645.

Keiding, N. (1990), 'Statistical inference in the Lexis diagram', *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences* 332(1627), 487–509. URL: `https://royalsocietypublishing.org/doi/10.1098/rsta.1990.0128`

Larønningen, S., Ferlay, J., Beydogan, H., Bray, F., Engholm, G., Ervik, M., Gulbrandsen, J., Hansen, H., Hansen, H., Johannesen, T., Kristensen, S., Kristiansen, M., Kønig, S., Lam, F., Laversanne, M., Miettinen, J., Mørch, L., Ólafsdóttir, E., Óskarsson, O., Pejicic, S., Petterson, D., Skog, A., Skovlund, C., Tanskanen, T., Tian, H., Virtanen, A., Aagnes, B. and Storm, H. (2022), 'NORDCAN: Cancer Incidence, Mortality, Prevalence and Survival in the Nordic Countries, Version 9.2 (June 23, 2022). Available from: https://nordcan.iarc.fr/, accessed on October 11, 2022.'. URL: `https://nordcan.iarc.fr/en`

Leinonen, M. K., Miettinen, J., Heikkinen, S., Pitkäniemi, J. and Malila, N. (2017), 'Quality measures of the population-based Finnish Cancer Registry indicate sound data quality for solid malignant tumours', *European Journal of Cancer (Oxford, England: 1990)* 77, 31–39.

Lutgens, M. W. M. D., van Oijen, M. G. H., van der Heijden, G. J. M. G., Vleggaar, F. P., Siersema, P. D. and Oldenburg, B. (2013), 'Declining risk of colorectal cancer in inflammatory bowel disease: an updated meta-analysis of population-based cohort studies', *Inflammatory Bowel Diseases* 19(4), 789–799.

McCullagh, P. and Nelder, J. A. (1998), *Generalized linear models*, number 37 *in* 'Monographs on statistics and applied probability', 2nd edn, Chapman & Hall/CRC, Boca Raton, FL.

Olén, O., Erichsen, R., Sachs, M. C., Pedersen, L., Halfvarson, J., Askling, J., Ekbom, A., Sørensen, H. T. and Ludvigsson, J. F. (2020), 'Colorectal cancer in ulcerative colitis: a Scandinavian population-based cohort study', *The Lancet* 395(10218), 123–131. URL: `https://linkinghub.elsevier.com/retrieve/pii/S0140673619325450`

Peters, U., Bien, S. and Zubair, N. (2015), 'Genetic architecture of colorectal cancer', *Gut* 64(10), 1623–1636.

Porta, M. S., Greenland, S., Hernán, M., Silva, I. d. S. and Last, J. M., eds (2014), *A dictionary of epidemiology*, 6th edn, Oxford University Press, Oxford.

Putter, H., Fiocco, M. and Geskus, R. B. (2007), 'Tutorial in biostatistics: competing risks and multi-state models', *Statistics in Medicine* 26(11), 2389–2430. URL: `https://onlinelibrary.wiley.com/doi/10.1002/sim.2712`

Reid, N. (1994), 'A Conversation with Sir David Cox', *Statistical Science* 9(3). URL: `https://projecteuclid.org/journals/statistical-science/volume-9/issue-3/A-Conversation-with-Sir-David-Cox/10.1214/ss/1177010394.full`

Rothman, K. J., Lash, T. L., VanderWeele, T. J. and Haneuse, S. (2021), *Modern epidemiology*, 4th edn, Wolters Kluwer, Philadelphia, PA.

Royston, P. and Parmar, M. K. B. (2002), 'Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects', *Statistics in Medicine* 21(15), 2175–2197.

Sund, R. (2012), 'Quality of the Finnish Hospital Discharge Register: a systematic review', *Scandinavian Journal of Public Health* 40(6), 505–515.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A. and Bray, F. (2021), 'Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries', *CA: a cancer journal for clinicians* 71(3), 209–249.

Ungaro, R., Mehandru, S., Allen, P. B., Peyrin-Biroulet, L. and Colombel, J.-F. (2017), 'Ulcerative colitis', *The Lancet* 389(10080), 1756–1770. URL: `https://linkinghub.elsevier.com/retrieve/pii/S0140673616321262`

World Health Organization (2013), *International classification of diseases for oncology (ICD-O)*, 3rd ed., 1st revision edn, World Health Organization, Geneva. Journal Abbreviation: ICD-O Publication Title: ICD-O Section: viii, 242 p. URL: `https://apps.who.int/iris/handle/10665/96612`