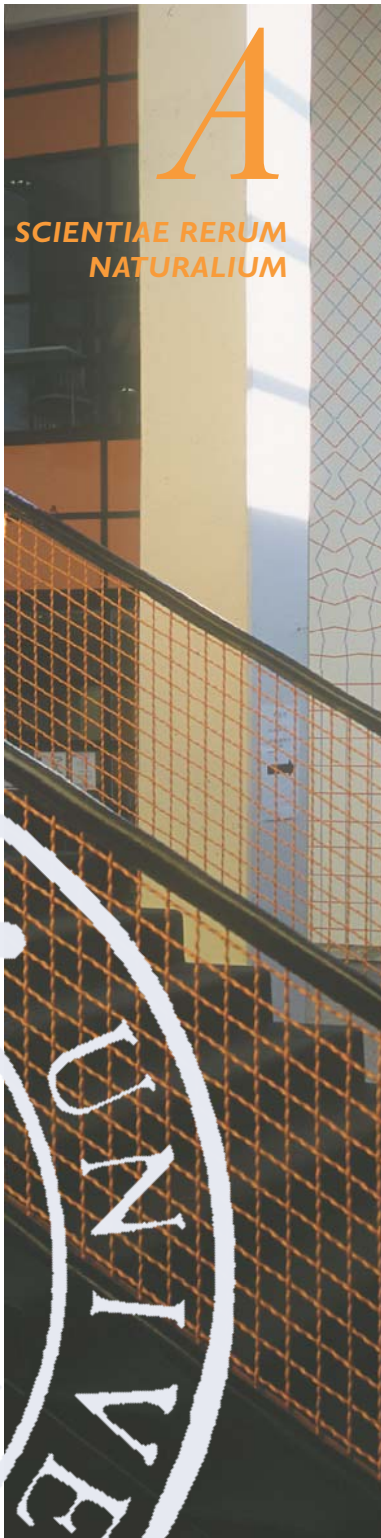
OULU 2012
A 603

ACTA UNIVERSITATIS OULUENSIS



Karri Seppä

QUANTIFYING REGIONAL VARIATION IN THE SURVIVAL OF CANCER PATIENTS

UNIVERSITY OF OULU GRADUATE SCHOOL;
UNIVERSITY OF OULU,
FACULTY OF SCIENCE,
DEPARTMENT OF MATHEMATICAL SCIENCES;
FINNISH CANCER REGISTRY,
INSTITUTE FOR STATISTICAL AND EPIDEMIOLOGICAL CANCER RESEARCH

ACTA UNIVERSITATIS OULUENSIS
A Scientiae Rerum Naturalium 603

KARRI SEPPÄ

QUANTIFYING REGIONAL VARIATION IN THE SURVIVAL OF CANCER PATIENTS

Academic dissertation to be presented with the assent of the Doctoral Training Committee of Technology and Natural Sciences of the University of Oulu for public defence in OP-sali (Auditorium L10), Linnanmaa, on 15 December 2012, at 12 noon

UNIVERSITY OF OULU, OULU 2012

Acta Univ. Oul. A 603, 2012

Supervised by
Professor Esa Läärä
Professor Timo Hakulinen
Doctor Hyon-Jung Kim-Ollila

Reviewed by
Professor Jukka Corander
Professor Antti Penttinen

ISBN 978-952-62-0010-1 (Paperback)
ISBN 978-952-62-0011-8 (PDF)

ISSN 0355-3191 (Printed)
ISSN 1796-220X (Online)

Cover Design
Raimo Ahonen

JUVENES PRINT
TAMPERE 2012

Seppä, Karri, Quantifying regional variation in the survival of cancer patients.
University of Oulu Graduate School; University of Oulu, Faculty of Science, Department of Mathematical Sciences, P.O. Box 3000, FI-90014 University of Oulu, Finland; Finnish Cancer Registry, Institute for Statistical and Epidemiological Cancer Research, Pieni Roobertinkatu 9, FI-00130 Helsinki, Finland
Acta Univ. Oul. A 603, 2012
Oulu, Finland

Abstract

Monitoring regional variation in the survival of cancer patients is an important tool for assessing realisation of regional equity in cancer care. When regions are small or sparsely populated, the random component in the total variation across the regions becomes prominent. The broad aim of this doctoral thesis is to develop methods for assessing regional variation in the cause-specific and relative survival of cancer patients in a country and for quantifying the public health impact of the regional variation in the presence of competing hazards of death using summary measures that are interpretable also for policy-makers and other stakeholders.

Methods for summarising the survival of a patient population with incomplete follow-up in terms of the mean and median survival times are proposed. A cure fraction model with two sets of random effects for regional variation is fitted to cause-specific survival data in a Bayesian framework using Markov chain Monte Carlo simulation. This hierarchical model is extended to the estimation of relative survival where the expected survival is estimated by region and considered as a random quantity. The public health impact of regional variation is quantified by the extra survival time and the number of avoidable deaths that would be gained if the patients achieved the most favourable level of relative survival.

The methods proposed were applied to real data sets from the Finnish Cancer Registry. Estimates of the mean and the median survival times of colon and thyroid cancer patients, respectively, were corrected for the bias that was caused by the inherent selection of patients during the period of diagnosis with respect to their age at diagnosis. The cure fraction model allowed estimation of regional variation in cause-specific and relative survival of breast and colon cancer patients, respectively, with a parsimonious number of parameters yielding reasonable estimates also for sparsely populated hospital districts.

Keywords: cancer, competing risks, cure fraction, mixture model, random effect, regional variation, relative survival

Seppä, Karri, Syöpäpotilaiden elossaolon alueellisen vaihtelun kvantifiointi.
Oulun yliopiston tutkijakoulu; Oulun yliopisto, Luonnontieteellinen tiedekunta, Matemaattisten tieteiden laitos, PL 3000, 90014 Oulun yliopisto; Suomen Syöpärekisteri, Syöpätautien tilastollinen ja epidemiologinen tutkimuslaitos, Pieni Roobertinkatu 9, 00130 Helsinki
Acta Univ. Oul. A 603, 2012
Oulu

Tiivistelmä

Syöpäpotilaiden elossaolon alueellisen vaihtelun seuraaminen on tärkeää arvioitaessa syövänhoidon oikeudenmukaista jakautumista alueittain. Kun alueet ovat pieniä tai harvaan asuttuja, alueellisen kokonaisvaihtelun satunnainen osa kasvaa merkittäväksi. Tämän väitöstutkimuksen tavoitteena on kehittää menetelmiä, joilla pystytään arvioimaan maan sisäistä alueellista vaihtelua lisäkuolleisuudessa, jonka itse syöpä potilaille aiheuttaa, ja tiivistämään alueellisen vaihtelun kansanterveydellinen merkitys mittalukuihin, jotka ottavat kilpailevan kuolleisuuden huomioon ja ovat myös päättäjien tulkittavissa.

Ehdotetuilla menetelmillä voidaan potilaiden ennustetta kuvailla käyttäen elossaoloajan keskiarvoa ja mediaania, vaikka potilaiden seuruu olisi keskeneräinen. Potilaiden syykohtaiselle kuolleisuudelle sovitetaan bayesiläisittäin MCMC-simulaatiota hyödyntäen malli, jossa parantuneiden potilaiden osuuden kuvaamisen lisäksi alueellinen vaihtelu esitetään kahden satunnaisefektijoukon avulla. Tämä hierarkkinen malli laajennetaan suhteellisen elossaolon estimointiin, jossa potilaiden odotettu elossaolo estimoidaan alueittain ja siihen liittyvä satunnaisvaihtelu otetaan huomioon. Alueellisen vaihtelun kansanterveydellistä merkitystä mitataan elossaoloajan keskimääräisellä pidentymällä sekä vältettävien kuolemien lukumäärällä, jotka voitaisiin saavuttaa, mikäli suotuisin suhteellisen elossaolon taso saavutettaisiin kaikilla alueilla.

Kehitettyjä menetelmiä käytettiin Suomen Syöpärekisterin aineistojen analysointiin. Paksusuoli- ja kilpirauhassyöpäpotilaiden elinaikojen keskiarvojen ja mediaanien estimaatit oikaistiin harhasta, joka aiheutui potilaiden luontaisesta valikoitumisesta diagnosointijakson aikana iän suhteen. Parantuneiden osuuden satunnaisefektimalli mahdollisti rintasyöpäpotilaiden syykohtaisen kuolleisuuden ja paksusuolisyöpäpotilaiden suhteellisen elossaolon kuvaamisen vähäisellä määrällä parametreja ja antoi järkeenkäyvät estimaatit myös harvaan asutuille sairaanhoitopiireille.

Asiasanat: alueellinen vaihtelu, kilpailevat riskit, parantuneiden osuus, satunnaisefekti, sekoitemalli, suhteellinen elossaolo, syöpä

Acknowledgements

This study was carried out at the Department of Mathematical Sciences in the University of Oulu and at the Finnish Cancer Registry, Institute for Statistical and Epidemiological Cancer Research, in Helsinki during the years 2006–2012. I greatly acknowledge the financial support of this work provided by the Academy of Finland and the Cancer Society of Finland.

I wish to express my deepest gratitude to my supervisors. I thank Professor Esa Läärä for his excellent guidance already in my Master's thesis and through my studies in statistics, and for providing me employment at the Department over two semesters including some nice teaching duties. I thank Professor Timo Hakulinen for offering me the opportunity to work at the Registry and participate in several interesting research projects outside my thesis, and above all, for his enthusiasm for research. They both have encouraged me over these years, taken care of funding and provided great facilities for completing this thesis. Their strong experience in statistics and cancer epidemiology has been essential in designing the studies and revising my drafts. I also wish to thank my third supervisor Dr. Hyon-Jung Kim-Ollila for her time and advice on the issues of Bayesian inference.

I am grateful to Professors Jukka Corander and Antti Penttinen, who carefully reviewed the manuscript of this thesis and gave quite positive assessments and constructive comments.

I have had a great opportunity to use the high quality data collected by the Finnish Cancer Registry. This study would not have been possible without the contributions of many other people working at the Registry. I would like to thank all the staff members, especially my closest workmates Arun Pokhrel, Tadek Dyba, Tapio Luostarinen and Timo Hakulinen, for their help, support and good company. I am also most grateful to my teachers and workmates at the Department of Mathematical Sciences. I want to express my appreciation to Luanne Siliämaa who revised the language of this thesis.

Finally, I want to thank my family and friends. I am fortunate to have great parents Hillevi and Matti, and sister Henriikka, who have supported my life and studies all the way from the beginning. My warmest thanks go to my dear wife Annina for her patience and love.

Karri Seppä Helsinki, October 2012

Abbreviations

The following abbreviations are used throughout this thesis:

Functions of follow-up time

h	Overall hazard of death
h_c	Cause-specific hazard of death due to cause c (1=cancer, 2=other causes)
h_c^*	Net (marginal) hazard of death due to cause c (1=cancer, 2=other causes)
h_D	Excess hazard of death due to cancer in non-cured patients
h_D^*	Net (marginal) hazard of death due to cancer in non-cured patients
h_P	Expected hazard of death on the basis of a relevant reference population
h_R	Excess hazard of death due to cancer
S	Cumulative overall survival
S_c^*	Cumulative net (marginal) survival associated with death from cause c (1=cancer, 2=other causes)
S_D	Cumulative relative survival in non-cured patients
S_D^*	Cumulative net (marginal) survival associated with death from cancer in non-cured patients
S_P	Cumulative expected survival on the basis of a relevant reference population
S_R	Cumulative relative survival

Other abbreviations

π	Theoretical proportion of statistically cured patients
τ	Theoretical mean (mathematical expectation) of survival time
CI	95% confidence interval
MCMC	Markov chain Monte Carlo
PI	95% equal tail posterior interval

List of original articles

The thesis consists of the summary part and the following four articles, which are referred to in the text by their Roman numerals (I–IV):

I Seppä K & Hakulinen T (2009) Mean and median survival times of cancer patients should be corrected for informative censoring. Journal of Clinical Epidemiology 62: 1095–1102.

II Seppä K, Hakulinen T & Läärä E (2012) Avoidable deaths and random variation in patients' survival. British Journal of Cancer 106: 1846–1849.

III Seppä K, Hakulinen T, Kim H-J & Läärä E (2010) Cure fraction model with random effects for regional variation in cancer survival. Statistics in Medicine 29: 2781–2793.

IV Seppä K, Hakulinen T & Läärä E (2012) Regional variation in relative survival — Quantifying the effects of the competing risks of death using cure fraction model with random effects. Manuscript.

Contents

1 Introduction

A national health service system should ensure the same level of cancer care to all people in the country. Assessing survival of cancer patients by region is important, because variation in the survival may reflect regional differences in the effectiveness of cancer care. Apart from the possible real differences in cancer care affecting the survival, variations in survival may also be due to confounding variables or chance (Karjalainen 1990).

Cancer patients are on average quite old at diagnosis. Because of this and the lengthening of life expectancy over time, it is increasingly common that many patients die due to causes unrelated to the cancer of interest. When assessing the effect of the cancer itself on the survival across regions, regional differences in the mortality due to other causes have to be taken into account.

The net survival is defined as the hypothetical survival of the patients in the absence of other causes of death than the cancer itself (Estève *et al.* 1990). It is often estimated in terms of cause-specific survival, based on division of deaths by cause into those from the target cancer and those from other causes. This approach leans heavily on the assumption of independence between the two competing causes of death. However, if the independence assumption is violated, the net survival is a functional of cause-specific hazards without any proper probability interpretation (Andersen & Keiding 2012).

The analysis of relative survival is often preferred over the cause-specific survival in population-based cancer studies (Sant *et al.* 2009). It describes the overall survival of the patients in relation to the survival of a comparable group in a relevant reference population and does not rely on cause of death information that may be unreliable or even unavailable (Ederer *et al.* 1961). During the follow-up, if the hazard of death in a group of patients obtains the same level as in a comparable group in the reference population, the patient group can be regarded as statistically cured and their proportion can be estimated using cure fraction models (Verdecchia *et al.* 1998).

When the regional units used in survival comparisons become more numerous and smaller in size, their estimates will be less precise, at least if the survival is estimated separately for each region. Even the mortality rates in small reference populations cannot be estimated precisely. Random effects may be used to capture the regional

variation in survival and to increase the precision of estimation within each region by reducing the effective number of parameters being estimated (Ohlssen *et al.* 2007a).

When keeping in mind the issues of interpretation, the concept of net survival is very useful for comparing the survival across regions. If regional variation exists in the net survival across regions, it is important to quantify the effects of the variation in the presence of competing mortality. However, methods for estimating regional variation in the cause-specific and relative survival in a country and quantifying the public health impact of the regional variation with realistic error margins have not been fully developed.

This thesis develops methods for estimating regional variation in the net survival of cancer patients in a country and for summarising the public health impact of the regional differences in terms of the extra survival time per patient and the number of avoidable deaths within a given period after diagnosis. The remaining chapters are organised as follows: In chapter 2, a review of the relevant literature and the basic concepts of the survival analysis with competing risks are presented. The aims of this thesis are listed in chapter 3. Chapter 4 summarises the methods developed and their implementations. Chapter 5 presents the cancer registry data to which the methods were applied and summarises the empirical results. The methods and their applications are discussed in chapter 6, and general conclusions are presented in chapter 7.

2 Methodological background

2.1 Brief review of the literature

The concept of relative survival was introduced by Berkson & Gage (1950) for estimating the mortality due to a specific target disease itself in the presence of competing mortality. It is measured by the relative survival ratio, in which the observed survival proportion of the patients is divided by the expected survival proportion derived from a comparable reference population. Three commonly known methods of estimating the relative survival ratio have been proposed (Ederer & Heise 1959, Ederer *et al.* 1961, Hakulinen 1982). The method of Hakulinen has been preferred and widely adopted (Ries *et al.* 2007, Coleman *et al.* 2008, Sant *et al.* 2009), because it takes informative censoring due to heterogeneity in potential follow-up times among patients into account, thus producing an unbiased estimator of the relative survival ratio. The relative survival ratio is often interpreted as describing the net survival probability of the patients, i.e., the hypothetical survival probability, if the cancer of the patients were the only cause of death (Pokhrel & Hakulinen 2008). However, if the excess hazard of death due to cancer is not constant across the ages, the relative survival ratio gives a biased estimator of the net survival (Hakulinen *et al.* 2011). A new method that does not require modeling has been proposed for estimating the net survival in the context of relative survival (Pohar Perme *et al.* 2012).

Regression modelling has been applied in relative survival analysis in the framework of generalised linear models (Hakulinen & Tenkanen 1987, Dickman *et al.* 2004) based on assuming piecewise constant excess hazard rates. Nelson *et al.* (2007) proposed to model the baseline excess hazard on the log cumulative hazard scale using restricted cubic splines. Cure fraction models assume that the patient population can be divided into cured and non-cured patients (Verdecchia *et al.* 1998). A mixture cure fraction model was introduced to relative survival already by Berkson & Gage (1952). In mixture cure fraction models, the cause-specific or relative survival of the patients is often modelled by the Weibull distribution (Farewell 1982, Verdecchia *et al.* 1998, De Angelis *et al.* 1999, Francisci *et al.* 2009), but more flexible distributions have also been proposed (Yu *et al.* 2004, Lambert *et al.* 2010b, Andersson *et al.* 2011). Lambert *et al.* (2007) extended an alternative cure model presented by Chen *et al.* (1999), not

based on the mixture of cured and non-cured patients for the estimation of the cure fraction in relative survival.

Cure fraction models have been used in international comparisons of relative survival (Verdecchia *et al.* 1998, Francisci *et al.* 2009) where the expected survival can be estimated precisely by country. Within a country, cure fraction models with spatially correlated random effects have been applied to geographically clustered cause-specific survival data in a Bayesian modelling framework (Banerjee & Carlin 2004, Cooner *et al.* 2006). Peng & Taylor (2011) proposed estimation methods to obtain maximum likelihood estimates of a cure fraction model with random effects for clustered cause-specific survival data. The Poisson regression model of relative survival (Dickman *et al.* 2004) was extended by Kuss *et al.* (2008) to account for clustered responses using regional random effects. This model was fitted within the class of generalised linear mixed models. Recently, Saez *et al.* (2012) also included spatially correlated random effects in the similar type of model in a Bayesian framework. In both of these models for regional relative survival, the mortality rates of the reference population were considered to be fixed.

2.2 Survival concepts with competing risks

Let T be a random variable representing the survival time of a patient from diagnosis to death from any cause (Putter *et al.* 2007). In addition, let C be a random variable of the cause of death of the patient such that $C = 1$, if the patient dies due to the cancer, and $C = 2$, if the patient dies due to other causes. If the exact survival time T from diagnosis to death is unknown, the survival time of the patient is said to be censored. Let T_0 be the time from diagnosis to censoring and $U = \min\{T, T_0\}$ is the observed follow-up time of the patient. If $U = T_0$, the follow-up is called incomplete. This can happen, for example, if the study is closed before everyone has died or if a patient has moved abroad, after which the survival status of the patient is missing.

The overall survival function $S(t)$ describes a patient's probability of being alive at time t after diagnosis:

$$S(t) = P(T > t) = \exp\left\{-\int_0^t h(u)\,du\right\}$$

where the overall hazard function $h(t)$ is the instantaneous rate at which death occurs for a patient alive at t:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t \mid T > t)}{\Delta t}.$$

The mean of the survival time T is given by the integral of the survival function over the time axis:

$$\tau = E(T) = \int_0^\infty S(t)\, dt.$$

The cause-specific hazard $h_c(t)$ of death from cause c is the instantaneous rate at which death due to cause c occurs for a patient alive at t:

$$h_c(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t \cap C = c \mid T > t)}{\Delta t}. \tag{1}$$

When the causes of death are mutually exclusive, the sum of the cause-specific hazards is equal to the overall hazard, i.e. $h(t) = h_1(t) + h_2(t)$. However, the additivity of the hazards does not imply that the different causes of death are necessarily independent (Hakulinen & Rahiala 1977).

The probability of dying from cause c by time t in the presence of competing risk of dying from other causes is obtained as

$$F_c(t) = P(T \leq t \cap C = c) = \int_0^t h_c(u) S(u)\, du.$$

This function of t has also been called as the crude probability of death from cause c (Chiang 1968, p. 242), or as the cumulative incidence function of outcome event c in the more general framework of multi-state models (Putter *et al.* 2007). If the causes of deaths are mutually exclusive, the sum of the crude probabilities equals the overall probability of dying. Under the assumption of a constant relative hazard, i.e., when the ratio $h_c(t)/h(t)$ is a constant within time interval $[t_{j-1}, t_j)$, the probability of dying due to cause c during the interval j can be written as (Chiang 1968, p. 244–245):

$$q_{cj} = P(t_{j-1} < T \leq t_j \cap C = c \mid T > t_{j-1}) = \frac{h_c(t_{j-1})}{h(t_{j-1})} \left\{1 - S(t_j)/S(t_{j-1})\right\}. \tag{2}$$

Suppose T is the minimum of two random variables $T = \min\{T_1, T_2\}$ where T_1 is the time to death from a specific cancer and T_2 is the time to death from other causes than the cancer in question. We cannot observe both T_1 and T_2 but only T at best. If survival times T_1 and T_2 were independent, we could define the net (marginal) survival function, that is the survival probability of a patient in absence of competing hazards of death, i.e.

in a hypothetical world where the other causes of death than cause c are eliminated:

$$S_c^*(t) = P(T_c > t) = \exp\left\{ -\int_0^t h_c^*(u)\, du \right\}$$

where h_c^* is the net (marginal) hazard function:

$$h_c^*(t) = \lim_{\Delta t \to 0} \frac{P(t < T_c \leq t + \Delta t \mid T_c > t)}{\Delta t}.$$

Furthermore, if T_1 and T_2 are independent, the net hazard is the same as the cause-specific hazard, i.e. $h_c^* = h_c$. Hence, $S(t) = S_1^*(t)S_2^*(t)$, and the net survival associated with the target disease $S_1^*(t)$ could be estimated based on the overall survival $S(t)$ and the net survival $S_2^*(t)$, the latter referring to the hypothetical probability for a patient to be alive at time t, if deaths due to cancer of interest were eliminated.

In general, we are interested in the distributions of T, T_1 and T_2 in the absence of censoring. Marginal distribution of T exists, if censoring is noninformative, i.e., T_0 and T are independent. The marginal distributions of T_1 and T_2 exist if T_0, T_1 and T_2 are mutually independent. One may view the assumption of independence more realistic, if it is made conditional on important determinants of survival, such as age, gender, calendar time and region. Even in spite of this, it is almost always essential to stratify the cancer patients at least by age, because the hazards of death due to cancer and due to other causes and even the hazard of censoring often depend on age.

2.3 Relative survival

In the context of relative survival, $S_2^*(t)$ is estimated by the expected survival function $S_P(t)$ that refers to the survival of a group in a relevant reference population that is sufficiently similar to the patient group with respect to characteristics affecting the mortality, but free of the cancer under study.

When the expected survival is estimated from a large national population stratified by sex, age and calendar year, it can be considered as fixed, i.e., essentially free from random error. Although the reference population also includes cancer patients under study, this can usually be considered to have a negligible effect on the estimated relative survival (Ederer *et al.* 1961). However, in prostate cancer and in all cancer sites combined, it is recommended to correct for the mortality due to the target cancer in the reference population (Talbäck & Dickman 2011, Hinchliffe *et al.* 2012). Even if no need exists to correct for regional differences in the expected survival, when estimating

relative survival in all regions combined, it is recommended to use region-specific expected survival in the estimation of relative survival by region (Dickman & Hakulinen 1996).

If the assumption of the independence of the survival times T_1 and T_2 is not valid, the net survival and related concepts have no meaningful probability interpretation as such (Andersen & Keiding 2012). However, the relative survival

$$S_R(t) = \frac{S(t)}{S_P(t)}$$

can always be interpreted as the ratio between the probability of a patient of being alive at time t and the corresponding probability of a comparable healthy person in the reference population. In addition, the excess hazard

$$h_R(t) = h(t) - h_P(t)$$

that is the excess rate of death that the patient has as compared to the rate of death $h_P(t)$ in a comparable healthy person, can be estimated as a surrogate for the net hazard.

The relative survival ratio in a group of n patients is the ratio between the averages of the patient-specific observed and expected survival probabilities $S_i(t)$ and $S_{Pi}(t)$, respectively:

$$S_R(t) = \frac{\sum_{i=1}^{n} S_i(t)}{\sum_{i=1}^{n} S_{Pi}(t)}. \tag{3}$$

On the other hand, the net survival in a group of n patients is the average of the patient-specific net survival probabilities $S_{1i}^*(t)$:

$$S_1^*(t) = \frac{1}{n}\sum_{i=1}^{n} S_{1i}^*(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{S_i(t)}{S_{Pi}(t)} \tag{4}$$

where the net survival $S_{1i}^*(t)$ of patient i is written as the ratio between the observed survival $S_i(t)$ and the expected survival $S_{Pi}(t)$ of patient i. This equals the traditional method of internal age standardisation of relative survival ratios (Pokhrel & Hakulinen 2008), if the excess hazard of death is the same for all ages within each age group.

2.4 Life table method and expected survival

The overall survival function $S(t)$ can be estimated by the life table (actuarial) method (Cutler & Ederer 1958) that does not require specification of any parametric model for

$S(t)$, and is especially useful if exact survival times are not known. This is the case, for example, with cancer registry data where survival times of patients are known to the nearest month as best, as the time of cancer diagnosis is usually recorded with the accuracy of the month of the diagnosis.

In the life table method, survival data are grouped by dividing the follow-up time axis into J disjoint time intervals: $[t_{j-1}, t_j),\ j = 1, \ldots, J$. Let l_j be the number of patients alive and under follow-up at the beginning of interval j, d_j be the number of deaths during interval j and w_j be the number of patients whose survival time was censored during interval j. The probability $S(t_k)$ of being alive at the end of interval k, at time t_k, is estimated by

$$\hat{S}(t_k) = \prod_{j=1}^{k} \left(1 - \frac{d_j}{l_j - w_j/2}\right).$$

Here the actuarial assumption is made, i.e., patients whose survival times were censored during interval j were assumed to be at risk of dying, on average, for half of the interval. As the cumulative survival $\hat{S}(t_k)$ is calculated as the product of the interval-specific conditional survival probabilities, which are based on only those patients who were alive and under follow-up at the beginning of the corresponding follow-up intervals, the censoring mechanism is assumed to be noninformative. Otherwise, the l_{j+1} patients remaining alive and under follow-up at time t_j may not be representative for the group of patients that would be alive at t_j under the complete follow-up.

To estimate the mean survival time, estimates of the overall survival function $S(t)$ are required for all time points t until $S(t)$ reaches zero. The life table method can be utilised in the estimation. If the patients with the longest follow-up are not dead by the end of the last follow-up interval J, the estimate $\hat{S}(t_J)$ of the overall survival probability at the end of interval J is larger than zero. The survival of the patients remaining alive after a given time t_k from diagnosis can be extrapolated into the future using their expected survival probabilities. The mean survival time τ can be estimated as the sum over cumulative survival proportions (Hakama & Hakulinen 1977):

$$\hat{\tau} = \frac{1}{2} + \sum_{j=1}^{k} \hat{S}(t_j) + \hat{S}(t_k) \sum_{j=1}^{\infty} S_P(t_{k+j} \mid t_k) \tag{5}$$

where intervals of the same length are used, and the conditional expected survival proportion at t_{k+j} for the l_{k+1} patients remaining alive and under follow-up after time t_k from diagnosis is denoted by $S_P(t_{k+j} \mid t_k)$.

The expected survival proportion $S_P(t)$ in a group of n cancer patients can be estimated simply as the average of the patient-specific expected survival probabilities $S_{Pi}(t)$:

$$S_P^{\mathrm{EI}}(t) = \frac{1}{n}\sum_{i=1}^{n} S_{Pi}(t).$$

This is called the Ederer I method (Ederer *et al.* 1961), and it gives an unbiased estimator of the expected survival proportion. The expected hazard $h_P(t)$ of death in the patient group, associated with the expected survival $S_P^{\mathrm{EI}}(t)$, is then

$$h_P^{\mathrm{EI}}(t) = \sum_{i=1}^{n} \frac{W_i^{\mathrm{EI}}(t) h_{Pi}(t)}{\sum_{i=1}^{n} W_i^{\mathrm{EI}}(t)} \tag{6}$$

where the expected hazard $h_{Pi}(t)$ of patient i is weighted by the expected survival probability of being alive at time t after diagnosis, i.e., $W_i^{\mathrm{EI}}(t) = S_{Pi}(t)$. However, a biased estimator of the relative survival ratio may be obtained using the Ederer I method, because informative censoring due to heterogeneity in the potential follow-up times may lead to a biased estimator $\hat{S}(t)$ of the overall survival.

Hakulinen proposed an alternative method for deriving the expected survival in the estimation of the relative survival ratio (Hakulinen 1982). In this method, the estimate of the expected hazard $h_P^{\mathrm{H}}(t)$ of death at time t is obtained as the weighted average of expected hazards of the patients whose potential follow-up times are greater than t, by replacing the weight function $W_i^{\mathrm{EI}}(t)$ in (6) with function

$$W_i^{\mathrm{H}}(t) = \begin{cases} W_i^{\mathrm{EI}}(t) = S_{Pi}(t) & \text{if } t < V_i \\ 0 & \text{otherwise} \end{cases}$$

where the potential follow-up time of patient i is denoted by V_i. The method of Hakulinen introduces a biased estimator of the expected survival, the bias being similar to that in the observed survival, producing an unbiased estimator of the relative survival ratio (3). The method gives an unbiased estimator for the net survival, if the net survival is constant across patients, which usually is not true (Hakulinen *et al.* 2011). Otherwise, the estimator of the relative survival ratio converges towards the net survival of patients who have the highest expected survival (Hakulinen 1977).

The third method for estimating the expected survival (Ederer & Heise 1959) is called the Ederer II method. In this method, the estimate of the expected hazard $h_P^{\mathrm{EII}}(t)$ of death at time t is obtained as the weighted average of expected hazards of the patients who are alive and under follow-up at time t, by replacing the weight function $W_i^{\mathrm{EI}}(t)$ in

(6) with function

$$W_i^{\mathrm{EII}}(t) = \begin{cases} 1 & \text{if } t < U_i \\ 0 & \text{otherwise} \end{cases}$$

where U_i is the observed follow-up time of patient i. This method allows for heterogeneity in potential follow-up times, but the estimator of the expected survival depends on the observed survival of the patients, leading to a biased estimator of the relative survival ratio (Hakulinen 1982). However, quite recently Hakulinen *et al.* (2011) recommended the Ederer II method rather than the Hakulinen method for estimating the net survival, as the estimates of the Ederer II method were closest to those of the traditional method of internal age standardisation in practical applications. The Ederer II method estimates the observable net survival associated with the cause-specific hazard $h_1(t)$ of death due to cancer (1), but this method also gives a biased estimator of the net survival (4), because the hazards of death due to cancer and due to other causes share the influence of the same demographic covariates such as age, gender and calendar time (Pohar Perme *et al.* 2012).

2.5 Cure fraction model

Suppose that patients can be latently divided into two distinct subgroups: statistically cured patients D_0 and non-cured patients D_1 (Verdecchia *et al.* 1998, De Angelis *et al.* 1999). In this model, a proportion $\pi = P(D_0)$ of cured patients shares the same hazard $h_P(t)$ of death as that in a comparable group in the reference population, i.e., their excess hazard of death attributable to diagnosis of the cancer is zero. The proportion $1-\pi = P(D_1)$ of non-cured patients has the hazard of death that can be expressed as the sum of the expected hazard $h_P(t)$ and the excess hazard $h_D(t)$. The survival function for the entire patient population can be written as a mixture of these two components

$$\begin{aligned} S(t) &= P(D_0)P(T > t \mid D_0) + P(D_1)P(T > t \mid D_1) \\ &= \pi S_P(t) + (1-\pi)S_P(t)S_D(t) \end{aligned} \tag{7}$$

where $S_P(t) = \exp(-\int_0^t h_P(u)\,du)$ is the expected survival function and $S_D(t) = \exp(-\int_0^t h_D(u)\,du)$ is the relative survival function of the non-cured patients. The relative survival function $S_R(t)$ for the entire patient population can be written as

$$S_R(t) = \frac{S(t)}{S_P(t)} = \pi + (1-\pi)S_D(t).$$

The two relative survival functions $S_R(t)$ and $S_D(t)$ can be interpreted as net survival functions $S_1^*(t)$ and $S_D^*(t)$ of the entire patient population and of the non-cured patients, respectively, in a hypothetical situation in which other causes of deaths than the cancer would be eliminated, if T_1 and T_2 are independent and the hazard $h_P(t)$ of death in the reference population equals the net hazard $h_2^*(t)$ of death due to the other causes than the cancer.

In practice, a parametric form of cause-specific (net) or relative survival function of the non-cured patients may strongly affect the estimate of the cure fraction that is negatively correlated with the estimate of the mean survival time of the non-cured patients. This correlation may lead to poor identifiability between the two components (Li *et al.* 2001). This problem could be alleviated by increasing the number of patients, by extending the follow-up time or by decreasing the number of censored survival times (Farewell 1986, Yu *et al.* 2004).

3 Aims of the thesis

The overall objective of this series of studies is to provide new methods for assessing regional variability in the cause-specific and the relative survival of cancer patients in a country and to develop comprehensible summary measures for quantifying the public health importance of the regional differences in the net survival in the presence of competing hazards of death.

The specific aims are:

1. To provide methods for estimating unbiasedly the mean and median survival times of cancer patients by taking into account informative censoring due to the selection of patients with respect to their age during the period of diagnosis, and to demonstrate the effect of outdated mortality rates of the reference population on the predictions of the mean and the median survival times (I).
2. To develop methods for obtaining valid confidence intervals for the numbers of avoidable deaths (II).
3. To implement the cure fraction model with random effects in a Bayesian framework using MCMC simulation and to obtain valid estimates of the proportion of cured patients and the net survival of the non-cured patients also for cancer patients in small or sparsely populated regions (III, IV).
4. To extend a cure fraction model with regional random effects for the estimation of relative survival where the potential heterogeneity in the expected survival is taken into account and the expected survival is considered as a random quantity (IV).

4 Methodological developments

This chapter summarises the methodological contributions of the thesis, and is organised into five sections. Analysis of the cause-specific and the relative survival requires information on either the causes of deaths or the expected survival of the patients. Estimation of the expected and the relative survival by assuming piecewise constant hazard functions is described in section 4.1. A mixture cure fraction model with random effects for estimation of the regional variation in the relative and the cause-specific survival is developed in section 4.2. Methods for summarising survival experience in a group of patients in terms of the mean and the median survival times and the number of deaths due to cancer and other causes within a given period after diagnosis are presented in section 4.3. Methods for quantifying regional variation in relative survival in the presence of competing mortality are developed in section 4.4 using the summary estimates in section 4.3. Implementations of the developed methods are described in section 4.5.

4.1 Expected survival (I, II, IV)

In II and IV, the expected hazard of death due to competing causes of death $h_{Pr}(t)$ was assumed to be constant within any combination of year of age a and calendar year v for gender s in region r, i.e., $h_{Pr}(t) = \lambda_{rsva}$, and it was estimated from the region-specific population life tables (stratified by sex, age and calendar year). Because the number of deaths d_{rsva} in each combination of sex, age, calendar year and region was available only in 5-year age groups, the regional hazard λ_{rsva} of death was assumed to be proportional to the hazard λ^*_{sva} of death in the whole country within the five-year age groups in order to estimate the region-specific hazards in 1-year age groups, i.e.:

$$\lambda_{rsv,a+j} = \frac{\lambda^*_{sv,a+j}}{\lambda^*_{sva}} \lambda_{rsva}$$

where $a = 0, 5, \ldots, 95$ and $j = 0, 1, \ldots, 4$.

For modelling the relative survival in paper II, the excess hazard $h_R(t)$ of death due to cancer was also assumed to be constant within J pre-specified follow-up intervals, i.e., $h_P(t) = \lambda_j = \lambda_{rsv_j a_j}$ and $h_R(t) = \nu_j$, when $t_{j-1} \leq t < t_j$ for $j = 1, 2, \ldots, J$. The excess hazard ν_{ji} of patient i is modelled as a multiplicative function of covariates

$\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ib})'$, such that $\nu_{ji} = \exp(\boldsymbol{\beta}'\boldsymbol{x}_i)$ where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_b)'$ is the vector of regression coefficients, in which each β_l, $l = 1, \ldots, b$, is interpreted as the additive effect of the lth covariate on the logarithm of the excess hazard. If the expected hazard λ_{ji} of patient i can be considered essentially free of random variation, the model can be fitted within the framework of the generalised linear models by assuming a Poisson error for the observed number of deaths and specifying link function $\ln(\mu_{ji} - \lambda_{ji} y_{ji})$, where y_{ji} is time at risk in interval j for patient i and $\mu_{ji} = (\lambda_{ji} + \nu_{ji}) y_{ji}$ is the expectation of the Poisson distribution (Dickman *et al.* 2004).

In order to take into account random variation in the region-specific expected survival figures (IV), the survival of the reference population was described by a piecewise exponential model. For estimating the survival in a cancer-free population the survival times of individuals who were diagnosed with cancer were censored at the time of diagnosis. Hence, the log-likelihood of a cancer-free population can be written as a Poisson type of log-likelihood:

$$\begin{aligned} \ell_1 &= \sum_{r,s,v,a} \{d_{rsva} \log(\lambda_{rsva}) - \lambda_{rsva} y_{rsva}\} + B \\ &= \sum_{r,s,v} \cdots \sum_{a=0,5,\ldots,95} \left\{ \log(\lambda_{rsva}) \sum_{j=0}^{4} d_{rsv,a+j} - \sum_{j=0}^{4} \frac{\lambda^*_{sv,a+j}}{\lambda^*_{sva}} \lambda_{rsva} y_{rsv,a+j} \right\} + B \end{aligned} \qquad (8)$$

where y_{rsva} is the number of person years and d_{rsva} is the number of deaths in the cancer-free population calculated by subtracting the numbers of person years and deaths pertaining to the patients (after diagnosis of the cancer) from those of the general population, respectively, and B is a constant not depending on parameters λ_{rsva}.

4.2 Mixture cure fraction model with random effects for regional variation in cause-specific and relative survival (III–IV)

A relative survival model with random effects that takes into account random variation in the expected survival was developed for estimating survival in small or sparsely populated regions within a country (IV). In the model, the excess hazard of death due to cancer was described by the parametric mixture cure fraction model presented in section 2.5. With small modifications (see the last paragraph of this section), the model can be applied in the analysis of the cause-specific survival (III).

In III, the cause-specific survival of the non-cured patients was modelled by the generalised gamma distribution with location (μ), scale (σ) and shape (κ) parameters (Cox *et al.* 2007):

$$S_D^*(t) = \begin{cases} 1 - G[\kappa^{-2}(t\mathrm{e}^{-\mu})^{\kappa/\sigma}; \kappa^{-2}] & \text{if } \kappa > 0 \\ G[\kappa^{-2}(t\mathrm{e}^{-\mu})^{\kappa/\sigma}; \kappa^{-2}] & \text{if } \kappa < 0 \\ 1 - \Phi[(\log(t) - \mu)/\sigma] & \text{if } \kappa = 0 \end{cases}$$

where $G(t;a) = \int_0^t x^{a-1}\mathrm{e}^{-x}\,dx/\Gamma(a)$ is the cumulative distribution function for the particular case of the gamma distribution with mean and variance equal to a, and Φ is the cumulative distribution function of the standard normal distribution.

In IV, the relative survival of the non-cured patients was modelled by the Weibull distribution, that is a special case of the generalised gamma distribution in which $\kappa = 1$ (Cox *et al.* 2007). This distribution was parameterised by the scale and shape parameters $\tilde{\mu} = \mathrm{e}^{-\mu/\sigma}$ and $\tilde{\sigma} = 1/\sigma$, respectively: $S_D(t) = \exp(-\tilde{\mu}t^{\tilde{\sigma}})$.

All the parameters in the model were allowed to vary by age group. In addition, the cure fraction π, the location parameter μ in the generalised gamma distribution and the scale parameter $\tilde{\mu}$ in the Weibull distribution were allowed to vary by region ($r = 1, \ldots, R$). The proportion of cured π was modelled using the logit link function (III, IV):

$$\pi_r(\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{\alpha}'\boldsymbol{x}_i + \omega_{\pi r})}{1 + \exp(\boldsymbol{\alpha}'\boldsymbol{x}_i + \omega_{\pi r})}$$

where $\boldsymbol{x}_i$ is the vector of covariates for patient i.

In the generalised gamma distribution (III), the location μ, scale σ and shape κ parameters were modelled as follows:

$$\mu_r(\boldsymbol{x}_i) = \boldsymbol{\beta}'\boldsymbol{x}_i + \omega_{\mu r}, \;\; \sigma(\boldsymbol{x}_i) = \exp(\boldsymbol{\gamma}'\boldsymbol{x}_i) \;\text{ and }\; \kappa(\boldsymbol{x}_i) = \boldsymbol{\delta}'\boldsymbol{x}_i.$$

In the Weibull distribution (IV), the scale $\tilde{\mu}$ and shape $\tilde{\sigma}$ parameters were modelled as follows:

$$\tilde{\mu}_r(\boldsymbol{x}_i) = \exp(\tilde{\boldsymbol{\beta}}'\boldsymbol{x}_i - \omega_{\tilde{\mu} r}) \;\text{ and }\; \tilde{\sigma}(\boldsymbol{x}_i) = \exp(\tilde{\boldsymbol{\gamma}}'\boldsymbol{x}_i)$$

The regional random effects $\omega_{\pi r}$, $\omega_{\mu r}$ and $\omega_{\tilde{\mu} r}$ were assumed to be drawn independently, each from a pertinent normal distribution:

$$\omega_{\pi r} \sim \mathrm{N}(0, \sigma_\pi^2), \quad \omega_{\mu r} \sim \mathrm{N}(0, \sigma_\mu^2) \quad \text{and} \quad \omega_{\tilde{\mu} r} \sim \mathrm{N}(0, \sigma_{\tilde{\mu}}^2)$$

with unknown variance parameters σ_π^2, σ_μ^2 and $\sigma_{\tilde{\mu}}^2$. This implies that the regions are assumed to be exchangeable i.e. the joint distribution of the regional effects is

invariant to permutations of the indexes $(1,\ldots,R)$ (Gelman *et al.* 2004, p. 121–124). If the hierarchical variance parameters are zero, no variation exists across the regions. This hierarchical model also practically includes a fixed effect formulation in which $\sigma_\pi^2 = \sigma_\mu^2 = \sigma_{\tilde{\mu}}^2 = \infty$, i.e., separate parameters are assigned for each region without reference to parameters in other regions. In III, an alternative specification, in which the random effect pairs $(\omega_{\pi r}, \omega_{\mu r})'$ were assumed to be drawn independently from a bivariate normal distribution with a correlation coefficient ρ_ω was also considered.

Based on model (7) the log-likelihood function on all cancer patients $i = 1,\ldots,n$ can be written as

$$\ell_2 = \sum_{i=1}^{n} \Big[\log\{S_{Pr}(t_i)\} + (1-c_i)\log\{\pi_r + (1-\pi_r)S_D(t_i \mid \boldsymbol{\theta}_r)\} + c_i \log\{\pi_r h_{Pr}(t_i) + (1-\pi_r)S_D(t_i \mid \boldsymbol{\theta}_r)[h_{Pr}(t_i) + h_D(t_i \mid \boldsymbol{\theta}_r)]\}\Big] \qquad (9)$$

where $\boldsymbol{\theta}_r$ is a vector of the parameters of the relative survival distribution of the non-cured patients; t_i is the observed follow-up time; $c_i = 1$ for a patient who was observed to die at t_i and $c_i = 0$ for a patient whose survival time was censored at t_i; $S_{Pr}(t_i)$ is the expected cumulative survival proportion and $h_{Pr}(t_i)$ is the expected hazard of death at time t_i for patient i diagnosed in region r.

For taking into account the random variation in the expected hazard $h_{Pr}(t) = \lambda_{rsva}$ (see section 4.1), the cancer-free population was combined with the patient population (IV). Given the population to which each individual belongs, the common log-likelihood ℓ can be written as the sum of the log-likelihood functions (8) and (9) of the two populations: $\ell = \ell_1 + \ell_2$.

In the estimation of the cause-specific survival (III), the functions related to the expected and relative survival S_{Pr}, h_{Pr}, S_D and h_D are replaced with those related to net survival S_2^*, h_2^*, S_D^* and h_D^*, respectively, in the log-likelihood (9). Because the net hazard due to the other causes than the cancer was not of interest, the deaths due to other causes were regarded as censored events, i.e., $c_i = 1$ for patients who died due to the cancer and $c_i = 0$ otherwise. The likelihood was further simplified by specifying $S_2^* = 1$ and $h_2^* = 0$.

4.3 Mean and median survival times and mean number of deaths (I, II, IV)

The mean and median survival times and the mean number of deaths within a given time period after diagnosis were estimated for summarising survival experience in a group of

patients in the presence of competing mortality due to other causes of death than the cancer.

When the estimation of the mean and the median survival time was based on the life table method (I), it was essential to stratify the estimation by age when age range of the patients was wide, because the observed survival times are subject to the risk of informative censoring due to a common closing date of the follow-up. The observed survival proportions in formula (5) are estimated using the life table method where the censoring mechanism is assumed to be noninformative. In addition, although the Ederer I method provides an unbiased estimator of the conditional expected survival proportion $S(t_{k+j} \mid t_k)$ for the l_{k+1} patients remaining alive and under follow-up after time t_k from diagnosis, these patients may not be representative for patients who would be alive at t_k under complete follow-up. For example, if the potential follow-up times of old patients are on average shorter than those of the younger patients, the conditional expected survival proportion $S(t_{k+j} \mid t_k)$ is estimated on the basis of patients who are on average too young.

The mean survival time τ for the whole group of n patients was estimated as the weighted average of the age-specific estimates τ_a of A age groups:

$$\tau = \frac{1}{n} \sum_{a=1}^{A} n_a \tau_a$$

where the number of patients n_a alive in age group a at the beginning of the follow-up was used as the weight for the age group and the mean τ_a of the survival time in age group a was estimated using formula (5). The estimate of the median was observed at a time point in which the weighted average of the age-specific cumulative survival proportions $S_a(t)$

$$S(t) = \frac{1}{n} \sum_{a=1}^{A} n_a S_a(t)$$

reaches the value of 50%.

In IV, the overall survival function was modelled by the mixture cure fraction model where a proportion of patients was assumed to share the same hazard of death as that in a comparable group in the reference population (IV). In this model, the mean survival time τ_r in region r can be written as

$$\tau_r = \int_0^\infty S_{Pr}(t) S_R(t \mid \pi_r, \tilde{\mu}_r, \tilde{\sigma})\, dt$$

where the expected survival function $S_{Pr}(t)$ and the two parameters π_r and $\tilde{\mu}_r$ in the relative survival function depend on region r.

The mean number of deaths in region r accumulated until time t from diagnosis is given as the sum of probabilities of death of n_r patients: $N_r(t) = \sum_{i=1}^{n_r}\{1 - S_i(t)\}$. The mean number of deaths from cancer was estimated as the sum of the patient-specific crude probabilities $F_{1i}(t)$ of dying from cancer within time t from diagnosis, i.e., $M_r(t) = \sum_{i=1}^{n_r} F_{1i}(t)$. In the model with piecewise constant excess hazard rates (II), the mean number of deaths from cancer $M_r^{\mathrm{II}}(t_k)$ accumulating during time t_k from diagnosis can be written using the simplified formula (2) for the probability of dying due to cancer in interval j:

$$M_r^{\mathrm{II}}(t_k) = \sum_{i=1}^{n_r}\sum_{j=1}^{k} S_i(t_{j-1})q_{1ji} = \sum_{i=1}^{n_r}\sum_{j=1}^{k} \frac{\nu_{ji}}{\lambda_{ji}+\nu_{ji}}\{S_i(t_{j-1}) - S_i(t_j)\}$$

where $S_i(t_j)$ is the probability of being alive at the end of follow-up interval j and q_{1ji} is the probability of dying from cancer in interval j for patient i. In the mixture cure fraction model (IV), the mean number of deaths $M_r^{\mathrm{IV}}(t_k)$ from cancer accumulating during time t from diagnosis is

$$M_r^{\mathrm{IV}}(t) = \sum_{i=1}^{n_r}(1-\pi_r)\int_0^t S_{Pr}(u) f_D(u \mid \tilde{\mu}_r, \tilde{\sigma})\, du$$

where $f_D(t) = \tilde{\mu}_r\tilde{\sigma}t^{\tilde{\sigma}-1}\exp(-\tilde{\mu}_r t^{\tilde{\sigma}})$ is the probability density function related to the relative survival of the non-cured patients.

4.4 Quantifying regional variation in the presence of competing mortality (II, IV)

For quantifying the impact of regional variation in relative survival, the region-specific mean survival times and the numbers of deaths were compared with those estimated in a hypothetical situation in which all regions compared would achieve the same level of relative survival (II, IV).

In an extreme situation, there would be no excess hazard attributable to diagnosis of the cancer, i.e., the survival experience of the patients is the same as that of a comparable group in the reference population (I). In an alternative hypothetical situation, all the regions would achieve the relative survival of the "best" region where the relative survival is estimated to be the largest, on average (II). This would be a reasonable reference level, if the regions were relatively large and there were not too much random variation in the estimators, so that the reference level would not be extremely high by chance.

A more stable estimator of the most favourable survival was defined by the standard deviations of the distributions of the regional random effects (IV). As the regional effects were assumed to follow the N(0, σ^2) distribution where σ is the standard deviation of the distribution, a random effect ω is larger than $z_{1-\delta}\sigma$ with probability of δ, where $z_{1-\delta} = \Phi^{-1}(1-\delta)$ and Φ^{-1} is the quantile function of the standard normal distribution. Because the two sets of the regional effects are assumed to be independent, the proportion δ of the region-specific parameter pairs $(\omega_\pi, \omega_{\tilde{\mu}})$ should lie in the set

$$\{(\omega_\pi, \omega_{\tilde{\mu}}) \in R^2 \mid (\omega_\pi, \omega_{\tilde{\mu}}) > z_{1-\sqrt{\delta}}(\sigma_\pi, \sigma_{\tilde{\mu}})\}.$$

Level $(\omega_\pi^*, \omega_{\tilde{\mu}}^*) = z_{1-\sqrt{0.025}}(\sigma_\pi, \sigma_{\tilde{\mu}})$ was used, i.e. $\delta = 2.5\%$, to define the hypothetical values for π and $\tilde{\mu}$, and the most favourable level of relative survival.

By replacing the regional parameters π_r and $\tilde{\mu}_r$ in the relative survival function S_R with their hypothetical values π^* and $\tilde{\mu}^*$, respectively, we get the hypothetical mean survival time τ_r^*, and the hypothetical number of deaths due to any cause and to cancer in region r, $N_r^*(t)$ and $M_r^*(t)$, respectively.

The impact of the regional variation in relative survival was quantified in the presence of competing mortality by the extra survival time (increase in the mean survival time), the avoidable deaths (decrease in the number of deaths) due to any cause and due to cancer, and the proportions of those:

$$\psi_r = \frac{\tau_r^* - \tau_r}{\tau_r}, \quad \phi_r(t) = \frac{N_r(t) - N_r^*(t)}{N_r(t)} \quad \text{and} \quad \phi_r^{\mathrm{c}}(t) = \frac{M_r(t) - M_r^*(t)}{M_r(t)},$$

respectively, that could be hypothetically gained in region r, if the patients achieved the most favourable level of relative survival. The proportion of avoidable deaths due to any cause can also be interpreted as the potential impact fraction (Barendregt & Veerman 2010) measuring the fractional change in the average probability of death after a change in the hazard of death due to the target cancer.

The variances for the estimated numbers and proportions of avoidable deaths were approximated by the delta method (Casella & Berger 2001, p. 240–244) in paper II. If α_{rsva} is the natural logarithm of the expected hazard, i.e., $\lambda_{rsva} = \exp(\alpha_{rsva})$, and β_l is the regression coefficient of the lth covariate in the excess hazard, the variance of the estimated proportion of avoidable deaths due to cancer can be approximated using the first-order partial derivatives of the proportion of avoidable death with respect to

parameters α_{rsva} and β_l:

$$\mathrm{Var}(\hat{\phi}_r^c(t_k)) \approx \left\{ \sum_{l,m} D(M_r(t_k), M_r^*(t_k), \beta_l) D(M_r(t_k), M_r^*(t_k), \beta_m) \mathrm{Cov}(\hat{\beta}_l, \hat{\beta}_m) + \sum_{s,v,a} D(M_r(t_k), M_r^*(t_k), \alpha_{rsva})^2 \mathrm{Var}(\hat{\alpha}_{rsva}) \right\} \{M_r(t_k)\}^{-4}$$

where $D(M, M^*, \beta)$ is a shorthand for $\frac{\partial M}{\partial \beta} M^* - \frac{\partial M^*}{\partial \beta} M$. The variance of the estimator of α_{rsva} was estimated by the inverse of the number of deaths in the national population stratified by region, sex, calendar year and age group. The partial derivative of the number of deaths from cancer with respect to β_l parameter is given by

$$\frac{\partial M_r(t_k)}{\partial \beta_l} = \sum_{i=1}^{n_r} \sum_{j=1}^{k} S_i(t_{j-1}) q_{1ji} \left\{ I_{\beta_l}(\nu_{ji}) q_{\cdot ji}^{-1} (\Delta_j \nu_{ji} p_{ji} + q_{2ji}) - \sum_{m=1}^{j-1} I_{\beta_l}(\nu_{mi}) \Delta_m \nu_{mi} \right\}$$

where $S_i(t_j)$ is the probability of remaining alive at least until the end of interval j for a patient i where in particular $S_i(t_0) = 1$; $q_{\cdot ji} = 1 - p_{ji} = 1 - S_i(t_j)/S_i(t_{j-1})$ and q_{cji} are the conditional probabilities of dying from any cause and from cause c, respectively, for patient i in interval j; $\Delta_j = t_j - t_{j-1}$ is the length of interval j; and $I_{\beta_l}(\nu_{ji})$ is an indicator which equals 1, if the regression coefficient β_l is included in the predicted excess hazard ν_{ji} of patient i in follow-up interval j, and $I_{\beta_l}(\nu_{ji}) = 0$ otherwise.

The partial derivative $\partial M_r(t_k)/\partial \alpha_{rsva}$ is otherwise similar to $\partial M_r(t_k)/\partial \beta_l$ but q_{2ji}, ν_{ji} and I_{β_l} are replaced with $-q_{2ji}$, λ_{ji} and $I_{\alpha_{rsva}}$, respectively, where $I_{\alpha_{rsva}}(\lambda_{ji}) = 1$, if parameter α_{rsva} is included in the expected hazard λ_{ji} of patient i in follow-up interval j, and $I_{\alpha_{rsva}}(\lambda_{ji}) = 0$ otherwise.

4.5 Implementation of the methods

In papers I and II, R environment for statistical computing and graphics (R Development Core Team 2012) was used in all the analyses. In II, the piecewise exponential model presented in section 4.1 was estimated in the framework of generalised linear models (Dickman *et al.* 2004) using the `glm` function to fit the model. This function provided the maximum likelihood estimates of the parameters of the excess hazard and the estimated covariance matrix of the estimators of these parameters using the iterative weighted least squares algorithm (McCullagh & Nelder 1989, p. 40–43). Based on these estimates, the numbers of deaths, the numbers and proportions of avoidable deaths and their confidence intervals were calculated.

In papers III and IV, the statistical inference was based on the MCMC simulation of the posterior distributions. The random walk Metropolis algorithm was used within the Gibbs sampler to sample from the full conditional distributions of the joint posterior distribution. Because the hierarchical standard deviation depends on the random effects in the posterior distribution, the Markov chain got easily stuck, if the standard deviation was close to zero. To improve the mixing of the chain, the samples were drawn from a parameter-expanded model (Gelman & Hill 2007, p. 424–427; Gelman *et al.* 2008), where each set of random effects is multiplied by an additional parameter. In the Metropolis algorithm, it was necessary to let the variances of the proposal distributions of the hierarchical standard deviations, the random effects and the additional redundant parameters to depend on the current state of the chain. The simulation algorithms were written with C++ programming language using the GNU Scientific Library (Galassi *et al.* 2011). R with the `coda` package was used in assessing the convergence of the Markov chain and in further calculations with simulated samples.

5 Empirical applications

5.1 Cancer patients and their follow-up

The methods proposed in the papers were applied to population-based data on cancer patients registered by the Finnish Cancer Registry. The Finnish Cancer Registry was founded in 1952. It receives notifications on cancer patients independently from hospitals, pathological and haematological laboratories, physicians, dentists, forensic autopsies and death certificates. Multiple sources of notifications at different phases of the disease improve the coverage of registration (Teppo *et al.* 1994). There are, on average, five notifications per cancer case. The cancer registration is compulsory and covers the whole of Finland. The official causes of death are obtained from the Cause of Death Register located at Statistics Finland. The Finnish Cancer Registry compares the official causes of death of each cancer patient to all data available for that cancer and evaluates whether the patient died from that cancer or something else (Pukkala *et al.* 2001, p. 50–51). The following data sets were used in the empirical applications of the papers I–IV:

I Patients diagnosed with localised colon cancer in 1970–1979 and patients diagnosed with localised thyroid cancer in 1978–1987 being followed up until the end of 2005.

II Patients diagnosed with colon cancer in 2000–2007 at 0–89 years of age being followed up until the end of 2007.

III Female patients diagnosed with breast cancer in 1953–2000 at 40–69 years of age being followed up for deaths due to breast cancer until the end of 2007.

IV Patients diagnosed with colon cancer in 1975–2004 at 0–79 years of age being followed up until the end of 2009.

5.2 Mortality in the general population of Finland

Mortality figures in the population of Finland were obtained from Statistics Finland. In paper I, life tables of the Finnish population stratified by age, sex and calendar year were utilised to extrapolate the survival of the patients into the future. In papers II and IV, municipality-specific population counts and numbers of deaths for males and females in

each calendar year were utilised, as the relative survival was estimated by region. The population counts and the numbers of deaths were available in 1 and 5-year age groups, respectively. In papers I and II, the expected survival was considered as fixed, and the reference population included the cancer patients, too. In paper IV, the expected survival was considered as random, and it was corrected for the mortality due to the cancer by estimating the mortality rates of a cancer-free population.

5.3 Main findings

5.3.1 Mean and median survival times (I)

The mean and the median survival times were estimated for female patients diagnosed with localized colon cancer in 1970–1979 and localized thyroid cancer in 1978–1987.

After the first 9 years of follow-up, the mortality of the patients remaining alive was extrapolated by the mortality in a general population group comparable with the patients with respect to age, sex and calendar year. The effects of these cancer diseases on survival were quantified by comparing the estimates of the patients with those in a comparable group in the general population. Both data sets were subject to informative censoring, because patients' follow-up times were censored at the end of each diagnosis period. The overall survival proportions in colon and thyroid cancer patients were over- and underestimated, respectively, if the analyses were not stratified by age. The mean and the median survival times were estimated in a more accurate way by combining the results of age-specific analyses and utilising the life tables of the general population up to the year 2005.

In colon cancer patients, for example, the median age at diagnosis increased over the diagnosis period, as relatively more elderly patients were diagnosed in the end of the period. Hence, the potential follow-up times of elderly patients were on average shorter than those of the younger ones. If the analysis was not stratified by age, the estimate of the mean survival time was 15.6 years (CI 14.5–16.7) which was 18% (CI 13–24) smaller than the estimate in a general population group comparable with the patient group with respect to age, sex and calendar year. By estimating the mean survival time as the weighted average of the age-specific means, the estimate of the mean survival time was 14.0 years (CI 13.2–14.8) which was 27% (CI 23–31) smaller than the estimate in the comparable group being very close to the estimates in a more complete data set where the follow-up was extended until the end of 2005: 13.9 years (CI 13.2–14.6) and

27% (CI 24–31). In the thyroid cancer patients, bias due to informative censoring went in the other direction, because the potential follow-up times of elderly patients were on average longer than those of the younger patients.

5.3.2 Numbers of avoidable deaths (II)

Numbers of deaths from cancer and from other causes within the first 5-year period after diagnosis were estimated for patients diagnosed with colon cancer in the five cancer control regions in 2000–2007. The public health impact of the regional differences in the relative and the expected survival, respectively, were quantified by the numbers of avoidable deaths.

The estimated numbers of deaths from cancer and from other causes than cancer were 4139 and 1335 deaths, respectively. These estimates were compared with those in a hypothetical scenario in which all patients would have achieved the same level of relative survival as that in the largest cancer control region where the capital is located. Under this scenario, 176 deaths from cancer itself were estimated to be avoidable. However, patients were also at risk of dying from other causes of death. An additional 30 deaths were estimated to occur due to other causes, and the vast majority of the additional deaths, 28 deaths, were estimated to occur in patients 65–89 years at diagnosis. Hence, the total number of avoidable deaths was estimated with a wide error margin as 146 deaths (CI 3–290), that is 3% (CI 0–5) of all deaths. In the number of avoidable deaths from cancer itself, the confidence interval was even wider ranging from 3 to 349 deaths.

If all the patients had also shared the same expected survival as that in the region where the background mortality was the lowest, the estimated number of avoidable deaths due to any cause would have been 172 (CI 25–319).

5.3.3 Regional variation in cause-specific survival (III)

The cause-specific survival of female breast cancer patients diagnosed in 1953–1969, 1970–1985 and 1986–2000, respectively, was estimated in each of the 21 hospital districts (the autonomous Province of Åland was excluded).

The posterior means of the shape parameter κ in the generalised gamma distribution ranged from 0.24 to 0.97 being closest to unity (the Weibull model) in the oldest age group: 60–69 years at diagnosis. The proportion of cured patients and the mean survival time until death from breast cancer in the non-cured patients were negatively correlated

in the posterior distribution. The correlation between the two components of survival was inversely related to the follow-up time and was the largest in the last period.

The posterior means of the 10-year cause-specific survival proportions for patients in the whole country were 44, 61 and 77% in the three periods of diagnosis, respectively. In 1953–1969 and 1970–1985, the region-specific estimates in Helsinki were clearly larger than the estimates in the whole county: by 7.4 percentage points (PI 5.2–9.8) in the first period, and by 4.1 percentage points (PI 2.1–6.0) in the second period. The posterior medians of the standard deviation of the population distribution of the 10-year cause-specific survival proportions were 3.3 (PI 2.2–5.2), 2.9 (PI 1.9–4.2) and 2.0 percentage points (PI 1.3–3.1) in the three periods, respectively. Allowing for correlation between the random effects $\omega_{\pi r}$ and $\omega_{\mu r}$ did not affect these results.

5.3.4 Quantifying regional variation in relative survival (IV)

The relative survival of colon cancer patients diagnosed in 1975–1984, 1985–1994 and 1995–2004, respectively, was estimated by hospital district. Public health impacts of the regional differences between the hospital districts were quantified by the extra survival times and the numbers of avoidable deaths.

Whether the fixed expected survival was calculated by district or not, the largest difference in the posterior means of the cure fraction and the mean survival time of the non-cured patients (in the absence of competing mortality) were 0.5 percentage points and 1.1 months, respectively. Whether the district-specific expected survival was considered as random or not, the largest difference in the cure fraction and the mean survival time of the non-cured patients were 0.2 percentage points and 0.3 months, respectively. The following results are based on the model where the district-specific expected survival is treated as a random quantity.

According to the posterior medians of the hierarchical standard deviations σ_π and $\sigma_{\tilde{\mu}}$, regional variation in the cure fraction and in the relative survival function of the non-cured patients was the largest in the first and the second period, respectively, both in males and females. In the last period, the posterior means of the cure fraction ranged from 50 to 54% in males and from 57 to 59% in females, whereas the estimates were 54% in males and 59% in females in the most favourable level of relative survival. The mean survival time of the non-cured patients ranged from 2.4 to 2.6 years in males and 2.1 to 2.4 years in females, whereas the estimates were 2.7 years in males and 2.5 years in females in the most favourable level of relative survival.

In the presence of competing mortality, the mean survival time of patients diagnosed in 1995–2004 would have increased by 4% (PI 1–10) in males and by 2% (PI 0–6) in females and the number of deaths due to any cause within the first 5-year period would have decreased by 5% (PI 1–10) in males and by 4% in females (PI 1–10), if the patients had shared the most favourable level of relative survival. In the numbers of deaths due to cancer, 7% (PI 1–15) in males and 5% (PI 1–13) in females were estimated to be avoidable.

6 Discussion

In this thesis, methods were developed for incorporating regional variability in the cause-specific and the relative survival of cancer patients in a country (III, IV), for summarising survival of patients in terms of the mean and the median survival times (I) and the numbers of deaths due to cancer and other causes, respectively (II), and for quantifying regional variation in terms of the extra mean survival time (IV) and the number of avoidable deaths (II, IV).

6.1 Estimation of the expected survival

When the relative survival is estimated by region, the expected survival also needs to be estimated by region (Dickman & Hakulinen 1996). In the cure fraction model of paper IV, the expected survival was treated as a random quantity, because it was estimated separately for the 22 hospital districts within which the expected survival cannot be estimated as precisely as for the five cancer control regions with larger populations (II), or for the whole country (I). However, the district-specific expected survival could also have been considered fixed in paper IV without any substantial impact on the point estimates or on the random errors of the target parameters. It might be more important to take into account the random variation in the expected survival, if the size of the reference population were closer to that of the patient population. This would be the case, for example, if the relative survival of all cancers combined were estimated, or if the expected survival were based on a clearly smaller reference population than an entire national population.

6.2 Modelling the cause-specific or the relative survival

In the mixture cure fraction model, patients are assumed to be latently divided into the cured and the non-cured patients. In breast cancer patients, however, deaths from the cancer are still observed even after decades since diagnosis, and the point of statistical cure may not be well achieved during the realistic lifespan of the patients (Brenner & Hakulinen 2004, Woods *et al.* 2009). Hence, the two components, the proportion of cured patients and the mean survival time until cancer death in non-cured patients,

must be interpreted with great caution. Correlations between these components can be assessed by estimating 95% posterior regions as was done in paper III. The 10-year cause-specific survival provides a more appropriate summary measure for survival of the breast cancer patients (III). In colon cancer patients, the statistical cure was seemingly achieved within the first 10 years after diagnosis, and the two components of relative survival had meaningful interpretations (IV).

The Weibull distribution was well suited for estimating the relative survival of the non-cured colon cancer patients diagnosed below 80 years of age. For older colon cancer patients, a mixture of the Weibull distributions should perhaps have been considered (Lambert *et al.* 2010b). In breast cancer, the generalised gamma distribution was used instead of the Weibull distribution, which provided an insufficient fit especially in the youngest age group: 40–49 years old at diagnosis.

If the proportion of cured patients is irrelevant or not of interest, the cause-specific and the relative survival of the patients can be modelled by assuming a piecewise constant function for the hazard of death due to cancer (II). In general, this is a more simple and flexible model than the mixture cure fraction model where the survival of non-cured patients was modelled within the generalised gamma family. However, intervals of follow-up time within which a constant excess hazard is assumed should not be too long especially at the beginning of the follow-up, when the hazard due to cancer decreases fast. Moreover, this model may not be well applicable for sparse data, because the piecewise constant hazards become very unstable in short follow-up intervals. Instead, the cure fraction model or flexible parametric models that use restricted cubic splines (Nelson *et al.* 2007) can be preferred to obtain smoother estimates for the excess hazard of death.

6.3 Modelling regional variation in survival

Regional variation between the five cancer control regions was described by including in the model a separate fixed effect parameter for each region (II). This is an appropriate way of modelling the regional effects, when there are not too many regions and they are sufficiently large such that the random errors of the regional effects remain moderate. However, it has been strongly argued that a random-effects model will generally be appropriate here, too, as it reflects a judgment of exchangeability between providers (here regions), will shrink in estimates and hence automatically adjust for ‘regression to the mean’, and will provide improved precision when estimating each provider effect

(Ohlssen *et al.* 2007a). When cancer survival was estimated by hospital district (III, IV), two sets of random effects modelled by two independent normal distributions were used to describe regional variation in the cure fraction and in the survival of the non-cured patients. In the estimation of the cause-specific survival of the breast cancer patients (III), a model with a bivariate normal distribution for the random effect pairs was also fitted, but the correlation structure did not turn out to be necessary.

As the random effects were drawn from a common distribution, the hospital districts were assumed to be exchangeable. Because the regional differences in background mortalities and age distributions were known to create variation in overall survival of the patients between regions, the differences were taken into account by focusing on either cause-specific or relative survival and including age in the models. In addition, two alternative models, in which the cause-specific survival of the patients were allowed to vary systematically across the five cancer control regions or within them, were fitted for assessing the exchangeability assumption in paper III, but these more complicated models had only minor effects on the results. The assumption of the normal distribution for the random effects can be justified, to some extent at least, by the central limit theorem, which tells us that the normal distribution is an appropriate model if the regional effects are formed as the sum of a large number of independent components. However, if there are outlying regions, the normal distribution assumption can lead both to undue influence of larger outliers, and undue impact on smaller outlying regions as they are shrunk towards the overall mean (Ohlssen *et al.* 2007a). Because the survival from breast cancer in Helsinki essentially differed from the other hospital districts, especially in the first period (III), specific fixed intercept parameters were tried for the capital district, but the estimates of the target parameters remained very similar. If some discrepant regions do not accommodate within a normal distribution, a heavier-tailed t-distribution with a low number of degrees of freedom or other more flexible distributions should be considered (Ohlssen *et al.* 2007a,b).

Even though survival was analysed across regions, information on geographical proximity was not incorporated in the models. The spatial correlation between the hospital districts was considered irrelevant, because so far in the Finnish system patients have not crossed over to other districts for cancer treatment, as the patients in a given municipality are treated in a designated central hospital and further referred to the pertinent university hospital, if more advanced care is needed. If the random variation within each hospital district were to be estimated, conditionally autoregressive models

(Cooner *et al.* 2006) could be employed for describing spatial correlation, for example, across municipalities.

6.4 Bayesian approach as a computational framework

Bayesian methods were employed for fitting the hierarchical cure fraction model, where the parameters of the effects of the hospital districts were modelled by the normal priors whose standard deviations were further assigned noninformative hyperprior distributions (III, IV). Although the possibility to incorporate prior information for the parameters was not utilised, the Bayesian approach was computationally more appealing than the non-Bayesian approach based on maximum likelihood, because by using MCMC methods it was in principle straightforward to fit the hierarchical models. In addition, estimates of the posterior distribution of the parameters and their functions can be obtained easily, and probability statements concerning the parameters can be estimated using simulated samples from the posterior distribution. These estimates are appropriate, even if the number of patients is small (Gelman *et al.* 2004). Spatial correlation and missing data could also be handled conveniently in a Bayesian framework (Saez *et al.* 2012).

It is essential to carefully examine the convergence of the Markov chain by various diagnostic tools whenever using MCMC methods in the Bayesian inference. When the hierarchical standard deviation of either set of the random effects was small, the Markov chain easily got stuck close to zero for many simulation rounds, and the chain mixed slowly. This kind of slow convergence is typical for the simulation algorithm, in which random effects and their hierarchical standard deviation are updated one at time, and it emerged with both empirical data sets (III, IV). The simulation algorithm could seemingly traverse the entire area of the posterior distribution associated with non-negligible probability mass, but to achieve this requires a considerable amount of simulation time. Fortunately, the mixing could be improved by the parameter expansion in which each set of random effects is rescaled by multiplying the random effects by an additional parameter (Gelman & Hill 2007, p. 424–427).

As an alternative to the Bayesian approach, the maximum likelihood estimation of the cure fraction model with random effects is possible using the Gaussian quadrature method or the Monte Carlo EM algorithm (Peng & Taylor 2011). Methods for obtaining residual maximum likelihood estimates have also been proposed (Xiang *et al.* 2011).

6.5 Summarising survival experience in the presence of competing mortality

In addition to the hypothetical measures of the cause-specific and the relative survival, the survival of cancer patients should also be described in the presence of competing causes of death that cannot be ignored in elderly patient populations. The mean and the median survival times are unique measures that do not depend on follow-up time (I). Yet, for estimating the mean and the median, the survival function of the patients has to be estimable until it goes to 0 and reaches level 0.5, respectively. In paper I, the survival function of the patients was extrapolated into the future by assuming that after 9 years from diagnosis, patients remaining alive would share the same hazard of death as that in a comparable group in the general population. In the cure fraction model (IV), the estimation of the mean survival time relies on the estimated level of the cure fraction that the survival function asymptotically approaches. The model gave an estimate for the relative survival at every time point, but estimates of the expected survival were needed for the future. If life tables of the general population are not available or cannot be predicted for calendar years after the end of the follow-up period, the mean survival time will be underestimated. In paper IV, the expected survival in 2009 was used for calendar years in the future. Hence, the underestimation of the mean survival time was likely to be the largest in patients diagnosed with colon cancer in 1995–2004 but smaller in patients diagnosed earlier.

Numbers of deaths were estimated within a given period of follow-up time. The number of deaths in the first 5-year period after diagnosis was a reasonable summary measure for colon cancer patients, because the patients experienced most of the excess hazard due to cancer within the first 5 years (II, IV). Within this period, the underestimation of the expected survival was not a major problem, because the mortality of the reference population was needed for only a few calendar years in the future. The methods of relative survival allowed to divide the total number of deaths to the numbers of deaths from cancer and from other causes, respectively, without using information on causes of death. However, this division is not valid if the excess hazard of death due to cancer is allowed to be negative. If negative excess hazards are forced to zero, this may lead to a biased estimate of the cumulative crude probability of dying from cancer (Cronin & Feuer 2000). Negative estimates of the excess hazard may easily emerge in some follow-up time intervals by chance, but statistical modelling can be used to control such a random variation. When the relative survival of colon cancer patients was

modelled by the piecewise constant excess hazard (II), not all interactions between the age, sex, region and follow-up time interval were required. In the cure fraction model (IV), this problem did not exist, because proper parametric survival distributions were assigned to the excess hazard of the non-cured patients. As an alternative to these models, flexible parametric models could also be used to smooth the excess hazard in the estimation of the crude probability (Lambert *et al.* 2010a).

6.6 Quantifying the public health impact of regional variation

If there is systematic variation in the cause-specific or relative survival between regions, it would be important to quantify the effects of the regional variation in the presence of competing mortality. The public health impact was quantified by estimating the extra survival time per patient (increase in the mean survival time, IV) and the number of avoidable deaths (decrease in the number of deaths) from cancer and other causes, respectively (II, IV), under different scenarios in which an optimal level of the relative or the expected survival was achieved by all patients. In earlier studies, point estimates for the numbers of avoidable deaths have been reported without considering their random errors (Dickman *et al.* 1997, Abdel-Rahman *et al.* 2009, Pokhrel *et al.* 2010, Holmberg *et al.* 2012). However, substantial error margins in the number of avoidable deaths were estimated even when the whole population of Finland and the relatively common cancer site (colon) was considered (II). When regional variation across the five cancer control regions was quantified, the level of relative survival in the largest region where the capital of the country is located was chosen as the optimal, because the estimated excess hazard in the capital region was on average lower than that in the other regions. In addition, it would be desirable that patients in the more remote areas of Finland could achieve the same level as that in the capital region. This approach does not necessarily apply when choosing an optimal level of relative survival among smaller or sparsely populated regions, because then the region-specific estimates are more prone to random variation. Hence, when the impact of the regional variation across the hospital districts was quantified, the optimal level of relative survival was defined by the population distributions of the regional random effects (IV).

6.7 Implications for further research

A new method for estimating the net survival in the context of relative survival has recently been proposed by Pohar Perme *et al.* (2012). The method does not require parametric modelling of the survival function, and provides (under noninformative censoring) a consistent estimator as opposed to the estimators of the Ederer I, Ederer II and Hakulinen methods, which usually overestimate the net survival (Pohar Perme *et al.* 2012). Therefore, either the estimator by Pohar Perme *et al.* or an estimator based on an adequate regression model including important covariates have been recommended to be used (Pohar Perme *et al.* 2012, Danieli *et al.* 2012). However, informative censoring due to heterogeneity in the potential follow-up times leads to bias in all these estimators, except when a well-specified regression model is used.

From the modelling point of view, it is important to specify a model that enables estimation of the excess hazard in homogeneous groups for obtaining a consistent estimator of the net survival (Pohar Perme *et al.* 2012). In paper II, the excess hazard was modelled by the piecewise exponential model where the effects of age group, sex and cancer control region were included. In paper IV, the effects of age group and hospital district were included in the mixture cure fraction model that was separately fitted in males and females and in each 10-year calendar period of diagnosis. Instead of using categorised covariates, incorporating flexible parametric functions for the effects of age and calendar time of diagnosis is worth being considered in the future studies of the developed models.

Instead of the new method by Pohar Perme *et al.*, the Ederer II method was used in this thesis for non-parametric estimation of the net survival in the whole country in each stratum of combination of age group, sex and calendar period (II, IV). The traditional method of internal age standardisation (Pokhrel & Hakulinen 2008) was used to obtain the estimates for all ages combined (II). One could expect that the estimate of the new method would be very close to that of the Ederer II method, especially within each stratum, because the Ederer II method has been shown to work quite well even without stratification by age, being close to the estimates of the traditional age standardisation (Hakulinen *et al.* 2011). Yet, the new estimator by Pohar Perme *et al.* is consistent (under noninformative censoring), even if the excess hazard cannot be assumed to be constant within each stratum. A detailed study concerning the choice of the method for practical applications, in which potential follow-up times are often heterogeneous, is an issue for further research.

7 Conclusions

In the estimation of overall survival by the life table method, stratifying by age is essential, because administrative censoring caused by a common closing date at the end of the study can be informative. For reducing the bias due to the informative censoring, the mean and the median survival times of patients with a wide age range should be estimated based on the weighted averages of the age-specific results (I).

When the numbers of avoidable deaths are estimated on the basis of relatively large regional units, the simple relative survival model that can be fitted within the framework of the generalised linear models can be applied. Random error in the numbers of avoidable deaths which may be substantial can be assessed within this model using the delta method (II).

When small or sparsely populated regions within a country are considered, the mixture cure fraction model with two sets of random effects allows the estimation of cause-specific and relative survival by region with a parsimonious number of parameters yielding reasonable estimates also for the smallest regions (III, IV). This model can be used to quantify the public health impact of the regional variation in cancer survival by comparing regional survival figures with the most favourable survival defined by the distributions of the random effects (IV). The model can be fitted in a Bayesian framework using MCMC simulation that provides realistic posterior intervals for all target parameters (III), also taking the random variation in the region-specific expected survival into account (IV). Mixing of the posterior simulation can be improved by the method of parameter-expansion (III).

References

Abdel-Rahman M, Stockton D, Rachet B, Hakulinen T & Coleman MP (2009) What if cancer survival in Britain were the same as in Europe: how many deaths are avoidable? British Journal of Cancer 101: s115–s124.

Andersen PK & Keiding N (2012) Interpretability and importance of functionals in competing risks and multistate models. Statistics in Medicine 31: 1074–1088.

Andersson TM, Dickman PW, Eloranta S & Lambert PC (2011) Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models. BMC Medical Research Methodology 11: 96.

Banerjee S & Carlin BP (2004) Parametric spatial cure rate models for interval-censored time-to-relapse data. Biometrics 60: 268–275.

Barendregt JJ & Veerman JL (2010) Categorical versus continuous risk factors and the calculation of potential impact fractions. Journal of Epidemiology and Community Health 64: 209–212.

Berkson J & Gage RP (1950) Calculation of survival rates for cancer. Proceedings of the Staff Meetings of the Mayo Clinic 25: 270–286.

Berkson J & Gage RP (1952) Survival curve for cancer patients following treatment. Journal of the American Statistical Association 47: 501–515.

Brenner H & Hakulinen T (2004) Are patients diagnosed with breast cancer before age 50 years ever cured? Journal of Clinical Oncology 22: 432–438.

Casella G & Berger RL (2001) Statistical Inference. Duxbury Press, Pacific Grove, CA, 2nd edition.

Chen MH, Ibrahim JG & Sinha D (1999) A new Bayesian model for survival data with a surviving fraction. Journal of the American Statistical Association 94: 909–919.

Chiang CL (1968) Introduction to Stochastic Processes in Biostatistics. Wiley, New York.

Coleman MP, Quaresma M, Berrino F, Lutz JM, De Angelis R, Capocaccia R, Baili P, Rachet B, Gatta G, Hakulinen T, Micheli A, Sant M, Weir HK, Elwood JM, Tsukuma H, Koifman S, E Silva GA, Francisci S, Santaquilani M, Verdecchia A, Storm HH, Young JL & CONCORD Working Group (2008) Cancer survival in five continents: a worldwide population-based study (CONCORD). The Lancet Oncology 9: 730–756.

Cooner F, Banerjee S & McBean AM (2006) Modelling geographically referenced survival data with a cure fraction. Statistical Methods in Medical Research 15: 307–324.

Cox C, Chu H, Schneider MF & Muñoz A (2007) Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. Statistics in Medicine 26: 4352–4374.

Cronin KA & Feuer EJ (2000) Cumulative cause-specific mortality for cancer patients in the presence of other causes: a crude analogue of relative survival. Statistics in Medicine 19: 1729–1740.

Cutler SJ & Ederer F (1958) Maximum utilization of the life table method in analyzing survival. Journal of Chronic Diseases 8: 699–712.

Danieli C, Remontet L, Bossard N, Roche L & Belot A (2012) Estimating net survival: the importance of allowing for informative censoring. Statistics in Medicine 31: 775–786.

De Angelis R, Capocaccia R, Hakulinen T, Söderman B & Verdecchia A (1999) Mixture models for cancer survival analysis: application to population-based data with covariates. Statistics in

Medicine 18: 441–454.
Dickman PW, Gibberd RW & Hakulinen T (1997) Estimating potential savings in cancer deaths by eliminating regional and social class variation in cancer survival in the Nordic countries. Journal of Epidemiology and Community Health 51: 289–298.
Dickman PW & Hakulinen T (1996) Adjusting for region of residence in relative survival analysis. Journal of Epidemiology and Biostatistics 1: 213–218.
Dickman PW, Sloggett A, Hills M & Hakulinen T (2004) Regression models for relative survival. Statistics in Medicine 23: 51–64.
Ederer F, Axtell LM & Cutler SJ (1961) The relative survival rate: A statistical methodology. National Cancer Institute Monograph 6: 101–121.
Ederer F & Heise H (1959) Instructions to IBM 650 Programmers in Processing Survival Computations. Methodological note no. 10. End Results Evaluation Section, National Cancer Institute, Bethesda, MD.
Estève J, Benhamou E, Croasdale M & Raymond L (1990) Relative survival and the estimation of net survival: elements for further discussion. Statistics in Medicine 9: 529–538.
Farewell VT (1982) The use of mixture models for the analysis of survival data with long-term survivors. Biometrics 38: 1041–1046.
Farewell VT (1986) Mixture models in survival analysis: Are they worth the risk? The Canadian Journal of Statistics 14: 257–262.
Francisci S, Capocaccia R, Grande E, Santaquilani M, Simonetti A, Allemani C, Gatta G, Sant M, Zigon G, Bray F, Janssen-Heijnen M & the EUROCARE Working Group (2009) The cure of cancer: a European perspective. European Journal of Cancer 45: 1067–1079.
Galassi M, Davies J, Theiler J, Gough B, Jungman G, Alken P, Booth M & Rossi F (2011) GNU Scientific Library Reference Manual, Edition 1.15, for GSL Version 1.15. URI: http://www.gnu.org/software/gsl/, accessed October 24, 2012.
Gelman A, Carlin JB, Stern HS & Rubin DB (2004) Bayesian Data Analysis. Chapman and Hall/CRC, Boca Raton, FL, 2nd edition.
Gelman A & Hill J (2007) Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, New York.
Gelman A, van Dyk DA, Huang Z & Boscardin JW (2008) Using redundant parameterizations to fit hierarchical models. Journal of Computational and Graphical Statistics 17: 95–122.
Hakama M & Hakulinen T (1977) Estimating the expectation of life in cancer survival studies with incomplete follow-up information. Journal of Chronic Diseases 30: 585–597.
Hakulinen T (1977) On long-term relative survival rates. Journal of Chronic Diseases 30: 431–443.
Hakulinen T (1982) Cancer survival corrected for heterogeneity in patient withdrawal. Biometrics 38: 933–942.
Hakulinen T & Rahiala M (1977) An example on the risk dependence and additivity of intensities in the theory of competing risks. Biometrics 33: 557–559.
Hakulinen T, Seppä K & Lambert PC (2011) Choosing the relative survival method for cancer survival estimation. European Journal of Cancer 47: 2202–2210.
Hakulinen T & Tenkanen L (1987) Regression analysis of relative survival rates. Journal of the Royal Statistical Society: Series C (Applied Statistics) 36: 309–317.
Hinchliffe SR, Dickman PW & Lambert PC (2012) Adjusting for the proportion of cancer deaths in the general population when using relative survival: a sensitivity analysis. Cancer Epidemiology 36: 148–152.

Holmberg L, Robinson D, Sandin F, Bray F, Linklater KM, Klint Å, Lambert PC, Adolfsson J, Hamdy FC, Catto J & Møller H (2012) A comparison of prostate cancer survival in England, Norway and Sweden: a population-based study. Cancer Epidemiology 36: e7–e12.

Karjalainen S (1990) Geographical variation in cancer patient survival in Finland: chance, confounding, or effect of treatment? Journal of Epidemiology and Community Health 44: 210–214.

Kuss O, Blankenburg T & Haerting J (2008) A relative survival model for clustered responses. Biometrical Journal 50: 408–418.

Lambert PC, Dickman PW, Nelson CP & Royston P (2010a) Estimating the crude probability of death due to cancer and other causes using relative survival models. Statistics in Medicine 29: 885–895.

Lambert PC, Dickman PW, Weston CL & Thompson JR (2010b) Estimating the cure fraction in population-based cancer studies by using finite mixture models. Journal of the Royal Statistical Society: Series C (Applied Statistics) 59: 35–55.

Lambert PC, Thompson JR, Weston CL & Dickman PW (2007) Estimating and modeling the cure fraction in population-based cancer survival analysis. Biostatistics 8: 576–594.

Li CS, Taylor JMG & Sy JP (2001) Identifiability of cure models. Statistics & Probability Letters 54: 389–395.

McCullagh P & Nelder JA (1989) Generalized Linear Models. Chapman and Hall, London, 2nd edition.

Nelson CP, Lambert PC, Squire IB & Jones DR (2007) Flexible parametric models for relative survival, with application in coronary heart disease. Statistics in Medicine 26: 5486–5498.

Ohlssen DI, Sharples LD & Spiegelhalter DJ (2007a) A hierarchical modelling framework for identifying unusual performance in health care providers. Journal of the Royal Statistical Society: Series A (Statistics in Society) 170: 865–890.

Ohlssen DI, Sharples LD & Spiegelhalter DJ (2007b) Flexible random-effects models using Bayesian semi-parametric models: applications to institutional comparisons. Statistics in Medicine 26: 2088–2112.

Peng Y & Taylor JM (2011) Mixture cure model with random effects for the analysis of a multi-center tonsil cancer study. Statistics in Medicine 30: 211–223.

Pohar Perme M, Stare J & Estève J (2012) On estimation in relative survival. Biometrics 68: 113–120.

Pokhrel A & Hakulinen T (2008) How to interpret the relative survival ratios of cancer patients. European Journal of Cancer 44: 2661–2667.

Pokhrel A, Martikainen P, Pukkala E, Rautalahti M, Seppä K & Hakulinen T (2010) Education, survival and avoidable deaths in cancer patients in Finland. British Journal of Cancer 103: 1109–1114.

Pukkala E, Söderman B, Okeanov A, Storm H, Rahu M, Hakulinen T, Becker N, Stabenow R, Bjarnadottir K, Stengrevics A, Gurevicius R, Glattre E, Zatonski W, Men T & Barlow L (2001) Cancer Atlas of Northern Europe. Cancer Society of Finland Publication No. 62, Helsinki.

Putter H, Fiocco M & Geskus RB (2007) Tutorial in biostatistics: competing risks and multi-state models. Statistics in Medicine 26: 2389–2430.

R Development Core Team (2012) R: A Language and Environment for Statistical Computing, version 2.15.1. R Foundation for Statistical Computing, Vienna. URI: http://www.R-project.org/, accessed October 24, 2012.

Ries LAG, Melbert D, Krapcho M, Mariotto A, Miller BA, Feuer EJ, Clegg L, Horner MJ, Howlader N, Eisner MP, Reichman M & Edwards BK (editors) (2007) SEER Cancer Statistics Review, 1975–2004. National Cancer Institute, Bethesda, MD.

Saez M, Barceló MA, Martos C, Saurina C, Marcos-Gragera R, Renart G, Ocaña-Riola R, Feja C & Alcalá T (2012) Spatial variability in relative survival from female breast cancer. Journal of the Royal Statistical Society: Series A (Statistics in Society) 175: 107–134.

Sant M, Allemani C, Santaquilani M, Knijn A, Marchesi F, Capocaccia R & the EUROCARE Working Group (2009) EUROCARE-4. Survival of cancer patients diagnosed in 1995–1999. Results and commentary. European Journal of Cancer 45: 931–991.

Talbäck M & Dickman PW (2011) Estimating expected survival probabilities for relative survival analysis – exploring the impact of including cancer patient mortality in the calculations. European Journal of Cancer 47: 2626–2632.

Teppo L, Pukkala E & Lehtonen M (1994) Data quality and quality control of a population-based cancer registry. Experience in Finland. Acta Oncologica 33: 365–369.

Verdecchia A, De Angelis R, Capocaccia R, Sant M, Micheli A, Gatta G & Berrino F (1998) The cure for colon cancer: results from the EUROCARE study. International Journal of Cancer 77: 322–329.

Woods LM, Rachet B, Lambert PC & Coleman MP (2009) 'Cure' from breast cancer among two populations of women followed for 23 years after diagnosis. Annals of Oncology 20: 1331–1336.

Xiang L, Ma X & Yau KK (2011) Mixture cure model with random effects for clustered interval-censored survival data. Statistics in Medicine 30: 995–1006.

Yu B, Tiwari RC, Cronin KA & Feuer EJ (2004) Cure fraction estimation from the mixture cure models for grouped survival data. Statistics in Medicine 23: 1733–1747.

Original articles

I Seppä K & Hakulinen T (2009) Mean and median survival times of cancer patients should be corrected for informative censoring. Journal of Clinical Epidemiology 62: 1095–1102.

II Seppä K, Hakulinen T & Läärä E (2012) Avoidable deaths and random variation in patients' survival. British Journal of Cancer 106: 1846–1849.

III Seppä K, Hakulinen T, Kim H-J & Läärä E (2010) Cure fraction model with random effects for regional variation in cancer survival. Statistics in Medicine 29: 2781–2793.

IV Seppä K, Hakulinen T & Läärä E (2012) Regional variation in relative survival — Quantifying the effects of the competing risks of death using cure fraction model with random effects. Manuscript.

Original publications are not included in the electronic version of the dissertation.

Book orders:
Granum: Virtual book store
http://granum.uta.fi/granum/

UNIVERSITY OF OULU P.O.B. 7500 FI-90014 UNIVERSITY OF OULU FINLAND

ACTA UNIVERSITATIS OULUENSIS

ISBN 978-952-62-0010-1 (Paperback)
ISBN 978-952-62-0011-8 (PDF)
ISSN 0355-3191 (Print)
ISSN 1796-220X (Online)

UNIVERSITY of OULU
OULUN YLIOPISTO